Sistema automàtic d'aprenentatge de preferències personals sobre notícies classificades en seccions i definides per paraules clau

Jordi Pueyo Busquets
Enginyeria en Informàtica

Tutor:
David Isern Alarcón

11 de juny del 2014

Resum

Entrar a una pàgina web de continguts i trobar a l'instant allò que et ve de gust llegir, escoltar o veure en aquell precís instant és un dels somnis de molts internautes. Aquest projecte ha tingut com a objectiu dissenyar un prototip de sistema d'aprenentatge automàtic que va en aquesta línia, dins les possibilitats d'un projecte de final de carrera d'Enginyeria en Informàtica.

El programa desenvolupat treballa sobre una mostra de documents classificats en una secció principal i definits per paraules clau. Està pensat per ser vàlid per a qualsevol repositori de dades però s'ha aplicat a un cas concret, l'aprenentatge de gustos sobre notícies del diari britànic 'The Guardian'. L'algoritme té un perfil ideal que modela el cervell d'un lector i un perfil evolutiu, que comença de zero i va aprenent a mesura que l'usuari va consumint notícies.

L'aprenentatge s'ha aconseguit fent una simulació d'aquestes tries en dos corpus de notícies d'uns 6.000 articles cadascun. En cada iteració, l'algorisme té en compte un petit grup notícies, a les quals s'assignen dues valoracions, una d'acord amb el perfil ideal i una altra segons el perfil evolutiu. La diferència entre les dues seleccions és la que ens dóna informació per a l'aprenentatge, que s'ha abordat amb dues estratègies. L'aprenentatge online fa una petita variació al perfil evolutiu després de cada tria, mentre que l'offline s'espera a tenir més dades per trobar patrons de conducta i poder fer modificacions més de més magnitud.

Després de diverses proves, s'ha comprovat que s'aconsegueix reduir la distància entre valoracions de les seccions principals i *tags*, comparant el perfil ideal i l'evolutiu abans i després de l'execució de l'algorisme d'aprenentatge.

Índex

1. INTRODUCCIÓ 5

1.1 PROBLEMA A RESOLDRE 7
1.2 OBJECTIUS 8
1.3 ENFOCAMENT I MÈTODE SEGUIT 9
1.4 PLANIFICACIÓ INICIAL 10
1.5 ESTRUCTURA DE LES FASES DEL PROJECTE 15
1.6 DESVIACIONS RESPECTE LA PLANIFICACIÓ INICIAL 15

2. ANÀLISI I DISSENY 16

2.1 DEFINICIÓ DE REQUISITS 16
2.1.1 Requisits funcionals 16
2.1.2 Requisits no funcionals 17
2.2 IDENTIFICACIÓ I DESCRIPCIÓ DELS CASOS D'ÚS 18
2.2.1 Gestió de dades 18
2.2.2 NewsRecommender 19
2.3 DIAGRAMA DE CLASSES 20
2.3.1 Tag 20
2.3.2 Section 21
2.3.3 Article 22
2.3.4 Corpus 23
2.3.5 Profile 23
2.3.6 Learner 25
2.3.7 Rating 26
3.3.8 Counter 26
2.4 ALGORISME D'APRENENTATGE 27
2.4.1 Valoració d'una notícia 29
2.4.2 Adaptació online 31
2.4.3 Adaptació offline 32
2.4.4 Valorar eficàcia 33
2.4.5 Definició del perfil ideal 34
2.4.6 Seccions a ignorar 35
2.5 REPOSITORIS DE DADES 36
2.5.1 API The Guardian 36

3. IMPLEMENTACIÓ 38

3.1 ENTORN DE DESENVOLUPAMENT 38
3.2 CLASSES JAVA 39
3.2.1 Tag 39
3.2.2 Section 40
3.2.3 Article 40
3.2.4 Corpus 41
3.2.5 Profile 43
3.2.6 Learner 49
3.2.7 NewsRecommender 55
3.3 COMENTARIS 56

4 ESTUDI DE DIVERSOS SUPÒSITS 57

4.1 CÀRREGA DE LES DADES 57
4.2 CREAR ESTRUCTURA DE PERFIL 59

4.3 INICIALITZACIÓ D'UN PERFIL IDEAL 64
4.4 ESTUDI DELS VALORS DELS PARÀMETRES 64
4.4.1 ALFA i BETA 64
4.4.2 OVER_RANKED_SIGNIFICATIU / LOVED_SIGNIFICATIU 68
4.4.3 PERCENTATGE_SUPERAR 72
4.4.4 NUM_TAGS_OFFLINE 76
4.4.5 UP_LOVED / DOWN_OVER_RANKED 79
4.5 SUPÒSITS D'APRENENTATGE 84
4.5.1 Persona de negocis 84
4.5.2 Persona aficionada al cricket i als cavalls 110
4.5.3 Hipster 121
4.6 CAP AL PERFIL CANVIANT 134

5 CONCLUSIONS I FUTUR 141

5.1 APRENENTATGE DE SECCIONS 141
5.1.1 Reducció d'un 47% de la distància euclidiana, patró constant 141
5.1.2 Seccions destacades, apreses 143
5.2 APRENENTATGE DE *TAGS* 144
5.2.1 Reducció mitjana d'un 14% 144
5.2.2 Lluny del rating ideal i proporcional a nombre d'aparicions 145
5.3 ALTRES CONSIDERACIONS SOBRE L'APRENENTATGE 148
5.3.1 Evolució de l'overranked 148
5.3.2 Paràmetres aprenentatge offline 148
5.3.3 Seccions a ignorar 149
5.3.4 Sobre el perfil canviant 150
5.4 POSSIBLES MILLORES 150
5.5 ESCALABILITAT I REUTILITZACIÓ 151
5.6 APLICACIONS 153
5.7 SOBRE L'EXPERIÈNCIA 153

6 BIBLIOGRAFIA 154

7 ANNEXOS 155

7.1 MANUAL D'INSTRUCCIONS 155
7.2 ESTADÍSTIQUES DELS CORPUS UTILITZATS 158
7.2.1 Corpus 1 158
7.2.2 Corpus 2 161

1. Introducció

Els comunicòlegs Maxwel McCombs i Donal Shaw van introduir l'any 1972 el concepte d'*agenda setting*. Aquest terme ha servit des de llavors per referir-se al poder que tenen els mitjans de comunicació de decidir quines notícies són més importants que les altres i, en conseqüència, d'influir en l'opinió pública i el pensament col·lectiu. És a dir, si una notícia apareix molt destacada a la portada d'un diari i amb una foto molt gran, els lectors tendiran a considerar-la una informació primordial i li prestaran molta més atenció que a una altra que quedi més camuflada entre les pàgines del rotatiu.

Aquest concepte, il·lustrat a la *Fig. 1a*, encara acompanya avui dia els mitjans de comunicació moderns i és clau en les teories de la comunicació. En funció de com ens presenten les notícies, els editors dels mitjans poden influir en el nostre perfil com a lector. Si ens van venent una informació com a molt rellevant, tendirem a pensar que realment ho és. Li prestarem més atenció i és probable que, si ens desperta interès, la temàtica que l'envolta passi a formar part de les nostres prioritats com a lectors.

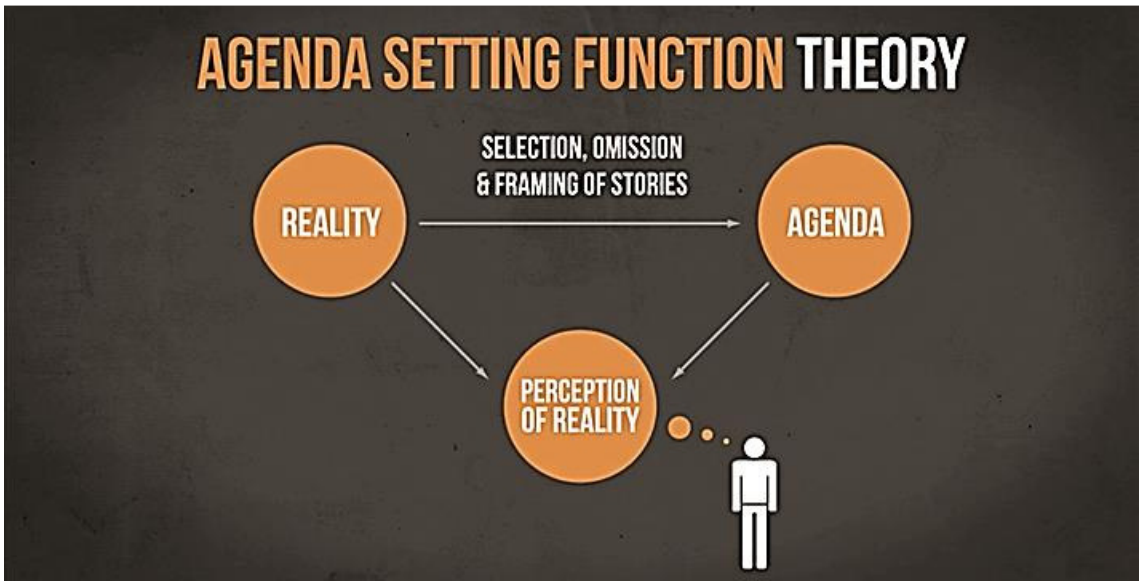


Fig. 1a / Font: lessonbucket.com

La limitació de la televisió i la ràdio convencionals ha estat, des dels seus orígens, el temps. No poden emetre més de 24 hores al dia. D'altra banda, els mitjans de comunicació escrits s'han vist encotillats per un límit de pàgines a imprimir. És cert que, en aquest sentit, res és impossible però seria molt difícil d'imaginar, per exemple, un diari de mil pàgines. També molt car d'imprimir i poc rendible, per suposat.

Amb l'arribada d'Internet, tots aquests límits s'han anat trencant cada vegada més. És possible trobar hores i hores de vídeos a la xarxa, topar amb àudios de tot tipus i llegir notícies sobre els temes més especialitzats que ens podem imaginar en quantitats industrials. Els mitjans de comunicació tradicionals han conquerit la xarxa i han pogut saltar-se les històriques fronteres del temps i l'espai. Continuen tenint el poder de *l'agenda-setting* però ja no és l'única manera a través de la qual es poden servir continguts a l'audiència. A la gran xarxa, el consum de notícies pot esdevenir molt més personalitzat.

Els mitjans de comunicació tenen la possibilitat de proporcionar peces informatives a la carta, adaptades a cada perfil de lector i, en aquest sentit, la informàtica i els mètodes d'intel·ligència artificial hi estan tenint un paper molt important. Aconseguir modelar els gustos i interessos del lector, però, no és una feina gens trivial.

A Internet el temps és or i és complicat que un usuari estigui disposat a dedicar deu minuts a definir-se ell mateix com a lector per tal que li puguem mostrar en un diari digital les notícies que més li interessen. No obstant això, en cas de fer-ho, serà molt difícil que es pugui definir a si mateix de manera correcta perquè les persones i els seus interessos van canviant per diverses circumstàncies de la vida i difícilment es pot saber amb precisió què ens agradarà llegir d'aquí sis mesos o un any. Més que res perquè tampoc sabem al 100% els continguts que se'ns oferiran. L'actualitat té un efecte sorpresa.

Per aquest motiu, aquest projecte de final de carrera es vol centrar en fer un prototipus de recomanador personal de notícies que, a partir de mètodes d'aprenentatge d'intel·ligència artificial, sigui capaç d'anar construint de manera automàtica el perfil d'un usuari en concret a partir del seu comportament. Es pretén també que aquest perfil pugui anar evolucionant en funció dels interessos canviants de l'usuari.

L'*agenda setting* dels mitjans de comunicació no desapareixerà mai perquè la funció de filtre dels *mass media* és més vital que mai en l'era de la sobreinformació. Els mitjans tenen un paper bàsic a la societat, el compromís d'explicar la realitat i crear opinió pública. No obstant això, el que es pretén amb aquest recomanador és introduir una altra manera de prioritzar les notícies, una *agenda setting* pensada per un usuari individual que pot ser molt útil per trobar les notícies més escaients per una persona en concret més enllà de les portades dels diaris i els titulars dels informatius de televisió. Aquesta forma de consum es pot considerar més lliure en contraposició amb l'habitual, en què els diaris juguen amb el poder de fer la priorització de notícies en funció de la seva línia editorial.

Tot i que aquest treball se centrarà en l'aprenentatge de preferències personals sobre continguts periodístics, l'objectiu és que l'algorisme desenvolupat es pugui aplicar a tot tipus de continguts de qualsevol caire classificats en grups o seccions i cadascun d'ells definit per un conjunt de paraules clau.

1.1 Problema a resoldre

La voluntat d'aquest treball és crear un prototipus que simuli i aprengui les accions d'un lector de manera intel·ligent. El sistema a desenvolupar, que durà a terme un aprenentatge no supervisat, ha de ser capaç de fer un seguiment dels tipus de continguts que llegeix un lector. D'aquesta manera, el sistema anirà elaborant un patró de conducta que li permetrà suggerir continguts que siguin d'interès per cada usuari en concret d'aquesta plataforma.

Si bé aquest treball es concretarà en un sistema de recomanació de notícies, la vocació de l'algorisme desenvolupat està pensat per tot tipus de continguts, sempre que estiguin ordenats en grups o seccions i cadascun d'ells definit per un conjunt de paraules clau associades, cada una d'elles, a un grup en concret. Podríem parlar, per exemple, d'imatges mèdiques agrupades en diferents tipus de mètodes de diagnòstic i cadascuna d'elles definida per un conjunt de paraules clau. També ens podríem imaginar un conjunt de documents històrics de diferents èpoques (criteri d'agrupament) i cadascun d'ell amb uns *tags* associats, etcètera.

El prototip anirà assignant pesos als diversos *tags* que vagin apareixent i, aquells que es vagin repetint en les successives seleccions de l'usuari, aniran esdevenint cada vegada més rellevants. D'altra banda, quan l'algorisme consideri que ha suggerit continguts no interessants a l'usuari, podrà abaixar el pes de les paraules clau que tinguin associades i/o a les seccions que les agrupin.

L'algorisme desenvolupat no treballarà tan sols amb paraules clau, sinó que també donarà rellevància a seccions de manera global. Serà capaç de descobrir, per exemple, que quan a un usuari li agrada molt la música, serà bo que a banda d'oferir-li notícies de música també n'hi ofereixi d'art o de literatura perquè és probable que li puguin agradar. En tots dos casos estarem parlant de cultura.

1.2 Objectius

Els objectius d'aquest treball són els següents:

- Crear un sistema que permeti extreure dades de l'API del diari 'The Guardian' entre unes dates en concret i doni la possibilitat de guardar un corpus de notícies en un fitxer separat per comes (CSV).
- Saber classificar cadascuna de les peces informatives dins d'un conjunt de seccions que es definiran en els pròxims apartats per intentar representar amb el màxim de precisió possible l'essència de cada notícia.
- Crear un perfil ideal d'un usuari. Serien els gustos d'una persona en un moment determinat de la seva vida. Vindria a ser la representació virtual de la ment d'un lector que en un futur podria fer servir una aplicació que fes ús de l'algorisme que es desenvoluparà. En una aplicació real d'aquest programari, aquest paper el faria una persona real però en un prototipus és necessari l'ús d'un perfil ideal per poder mesurar el bon funcionament de l'algorisme. Per a cada categoria aquest usuari tindrà un interès variant que podrà prendre uns pesos determinats.
- Es vol que l'algorisme vagi modelant un perfil en cadascuna de les seves iteracions que ha de tendir cap a aquest perfil ideal una vegada recorregudes les aproximadament 6.000 notícies del corpus.
- El programa permetrà obtenir informació per veure com de bé o de malament es fa l'aprenentatge.
- Es provarà si els resultats milloren o empitjoren en modificar dinàmicament el perfil ideal de l'usuari durant l'execució del programa. Aquest punt servirà per modelar el canvi de preferències que podria tenir un usuari real al llarg del temps.
- El mètode se centrarà en 'The Guardian' i la seva estructura de representació e les notícies si bé es vol que el nucli de l'algorisme i els mètodes de càlcul estiguin oberts a ser utilitzats amb altres repositoris de dades.
- Intentar aportar el punt de vista singular d'un estudiant d'Enginyeria Informàtica que compta amb una trajectòria professional com a periodista.

1.3 Enfocament i mètode seguit

Per tal de construir el prototip de recomanador personal s'agafarà un corpus d'aproximadament 6.000 notícies a partir de l'API que ofereix el diari britànic 'The Guardian' acotat entre dues dates. Cadascuna d'aquestes notícies es podrà identificar per un conjunt determinat de paraules clau cadascuna de les quals es podrà englobar en una secció –o grup- en concret. El criteri principal per descriure les notícies serà la secció del diari a la qual pertanyen (Politics, Business, Health, Society, Culture, entre d'altres).

L'algorisme que s'implementarà anirà fent iteracions en els quals cada vegada agafarà un grup de 15 notícies del corpus. Automàticament serà capaç de donar una valoració a cadascuna d'aquestes notícies en funció del perfil d'usuari que s'hagi anat definint fins aquell moment. Quan hagi fet aquesta valoració, ordenarà les notícies en funció de l'interès. De més a menys. Llavors, se simularà la tria d'un usuari a partir d'un perfil ideal que s'haurà definit prèviament. Triarà una notícia que no té perquè ser la de dalt de tot. En funció de l'elecció, s'actualitzarà el perfil i l'objectiu és que al final de totes aquestes iteracions el perfil viu i canviant amb què treballem s'hagi acostat al màxim possible al perfil ideal. La diferència entre un i l'altre indicarà el nivell d'aprenentatge assolit.

A banda de les dades de la notícia escollida per anar perfilant el perfil d'usuari, també s'utilitzaran les d'aquelles que estiguin per sobre de la notícia seleccionada en el rànquing de classificació per interès. En veure's que l'usuari no els ha prestat atenció, serviran per actualitzar el perfil amb pesos negatius per a les paraules que les representen. A més a més, s'anirà guardant l'històric de les notícies seleccionades, que serviran per buscar patrons de gustos.

El resultat d'aquest projecte serà un prototip d'algorisme de classificació que es basarà en un aprenentatge no supervisat. S'obtindrà un codi font que modelarà el que s'ha explicat fins ara i que permetrà valorar l'eficàcia de l'algorisme a partir de diversos gràfics, fruit de les dades de sortida de cadascuna de les iteracions de les diferents execucions que es vagin fent. L'algorisme modelat per aquesta eina s'haurà de poder adaptar per a poder ser utiltizat en qualsevol repositori de continguts.

Com a punt de partida d'aquest projecte es faran servir dos articles, [1] i [2], de Marin L, Isern D i Moreno A de la Universitat Rovira i Virgili (URV).

1.4 Planificació inicial

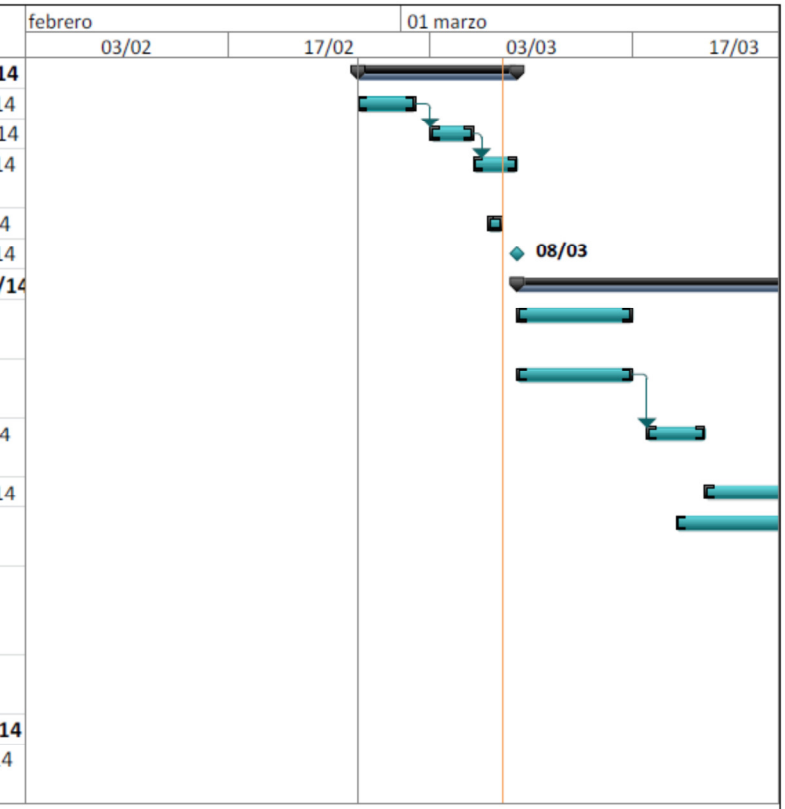
Amb l'objectiu de facilitar el desenvolupament del projecte i de cenyir-se al calendari previst per a la seva execució, es consideraran les següents fases:

- Planificació. L'objectiu d'aquesta fase és definir l'enunciat del projecte de final de carrera després d'haver-se documentat i haver parlat amb el tutor sobre el tema d'interès. El resultat final és un document amb una introducció, una descripció del problema a resoldre, els objectius del projecte i la seva planificació inicial.
- Anàlisi i disseny. En aquesta fase es conceptualitzaran les estructures de dades i es definirà l'algorisme que es farà servir per aconseguir els objectius del projecte. També s'analitzarà com es farà l'extracció de dades de la font mencionada en els punts anteriors.
- Implementació. En aquest punt es materialitzarà l'eina dissenyada. És una fase molt important i a la qual s'ha de donar temps de marge per als possibles imprevistos que puguin anar sorgint. No obstant això, és important que es faci aviat per poder treure bones conclusions de l'algorisme i poder fer diverses proves una vegada implementat.
- Estudi de diversos supòsits i conclusions. Fruit de la implementació del projecte obtindrem una sèrie de dades que permetran valorar els resultats obtinguts. En aquesta fase s'analitzaran tots els *outputs* obtinguts de l'algorisme i es podrà veure quina és la seva eficàcia tenint en compte diversos supòsits.
- Lliurament final. En aquesta fase s'acabarà de tancar i unificar la documentació que s'haurà anat obtenint en cadascuna de les fases del projecte.

Tasca	Inici	Fi
Planificació	26/02/14	08/03/14
Entrega PAC1		08/03/14
Anàlisi i disseny	09/03/14	30/03/14
Implementació	31/03/14	30/04/14
Entrega PAC2		12/04/14
Conclusions	01/05/14	15/05/14
Entrega PAC3		17/05/14
Lliurament final	19/05/14	11/06/14

A les pàgines següents *[Fig. 1b]* es pot veure una planificació més detallada feta amb l'eina Microsoft Project.

Id	Modo de tarea	Nombre de tarea	Duración	Comienzo	Fin
1		**Planificació**	**10 días**	**mié 26/02/14**	**sáb 08/03/14**
2		Recerca i lectures	4 días	mié 26/02/14	sáb 01/03/14
3		Participació al fòrum	2 días	lun 03/03/14	mié 05/03/14
4		Redacció document inicial	3 días	jue 06/03/14	sáb 08/03/14
5		Planificació inicial	1 día	vie 07/03/14	vie 07/03/14
6		Entrega PAC1	0 días	sáb 08/03/14	sáb 08/03/14
7		**Anàlisi i disseny**	**19 días**	**dom 09/03/14**	**dom 30/03/14**
8		Estudi API 'The Guardian'	7 días	dom 09/03/14	dom 16/03/14
9		Definició d'atributs categòrics	7 días	dom 09/03/14	dom 16/03/14
10		Definició del perfil d'usuari	4 días	mar 18/03/14	vie 21/03/14
11		Extracció de les dades	7 días	sáb 22/03/14	sáb 29/03/14
12		Disseny d'algorisme en pseudo-codi	10 días	jue 20/03/14	dom 30/03/14
13		Disseny dels diferents supòsits a provar	1 día	dom 30/03/14	dom 30/03/14
14		Redacció de la memòria	1 día	dom 30/03/14	dom 30/03/14
15		**Implementació**	**27 días**	**lun 31/03/14**	**mié 30/04/14**
16		Instal·lació entorn de treball Java	1 día	lun 31/03/14	lun 31/03/14



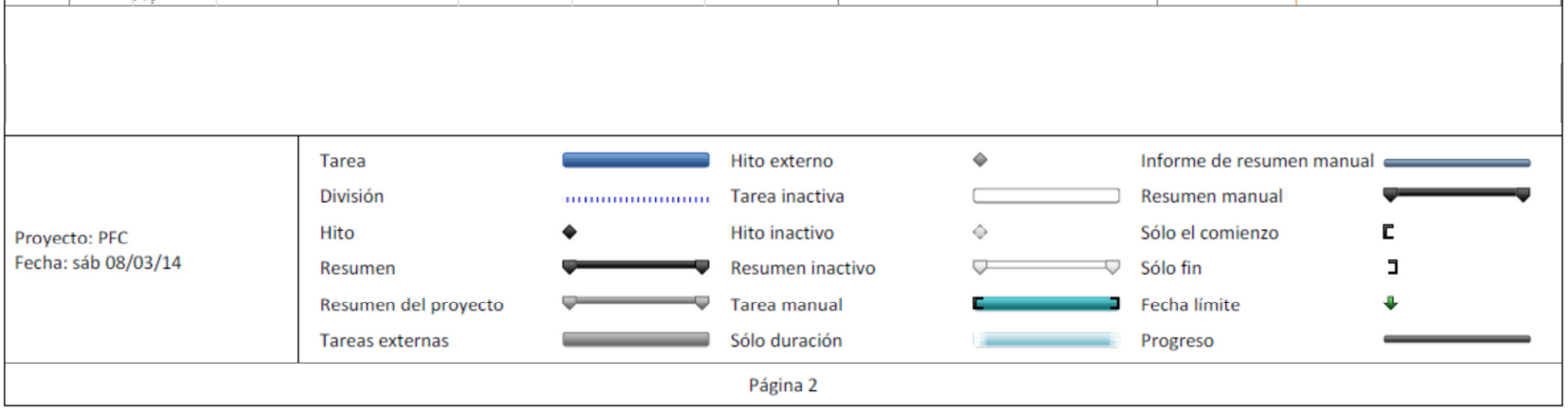
Proyecto: PFC
Fecha: sáb 08/03/14

Tarea	Hito externo	Informe de resumen manual
División	Tarea inactiva	Resumen manual
Hito	Hito inactivo	Sólo el comienzo
Resumen	Resumen inactivo	Sólo fin
Resumen del proyecto	Tarea manual	Fecha límite
Tareas externas	Sólo duración	Progreso

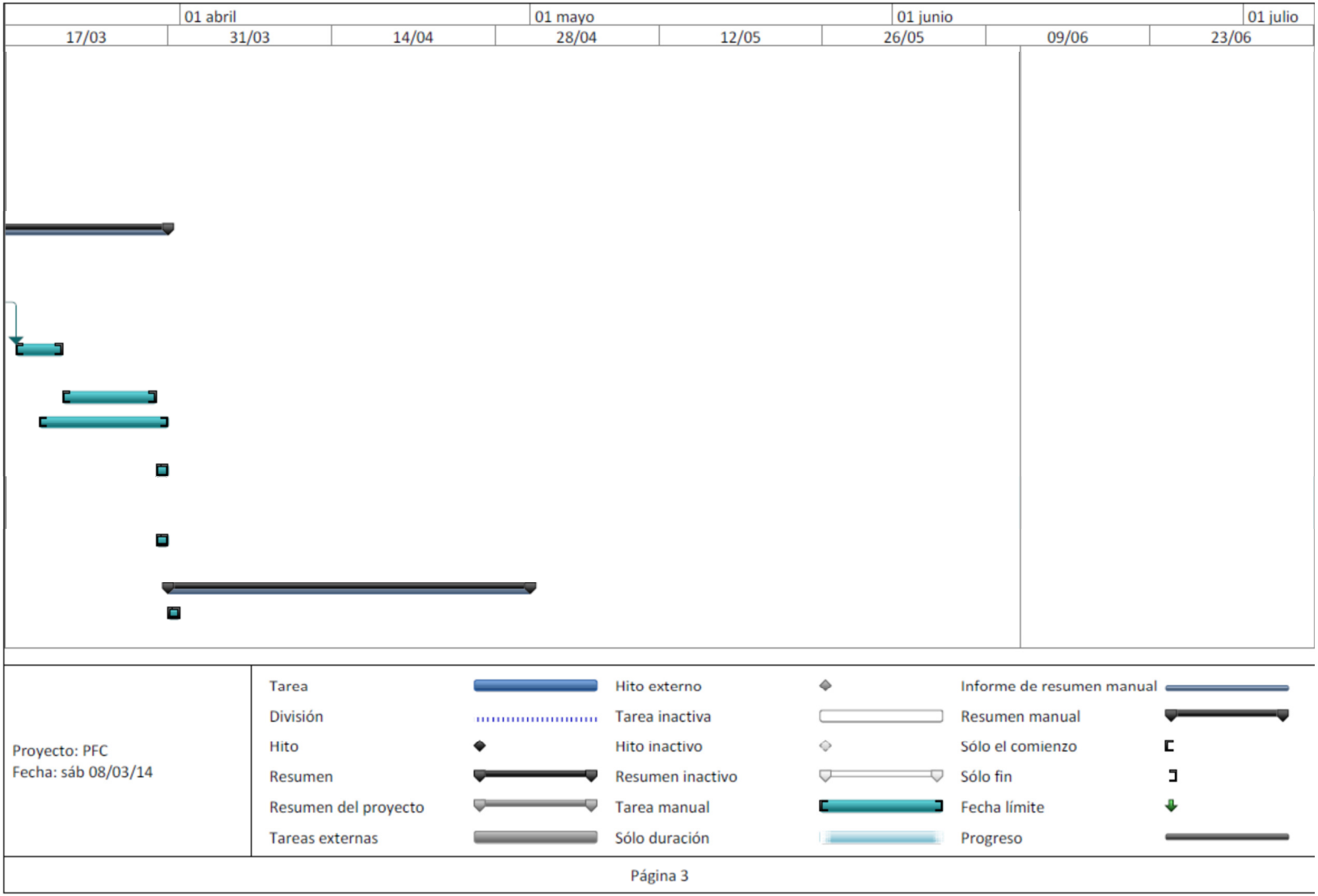
Página 1

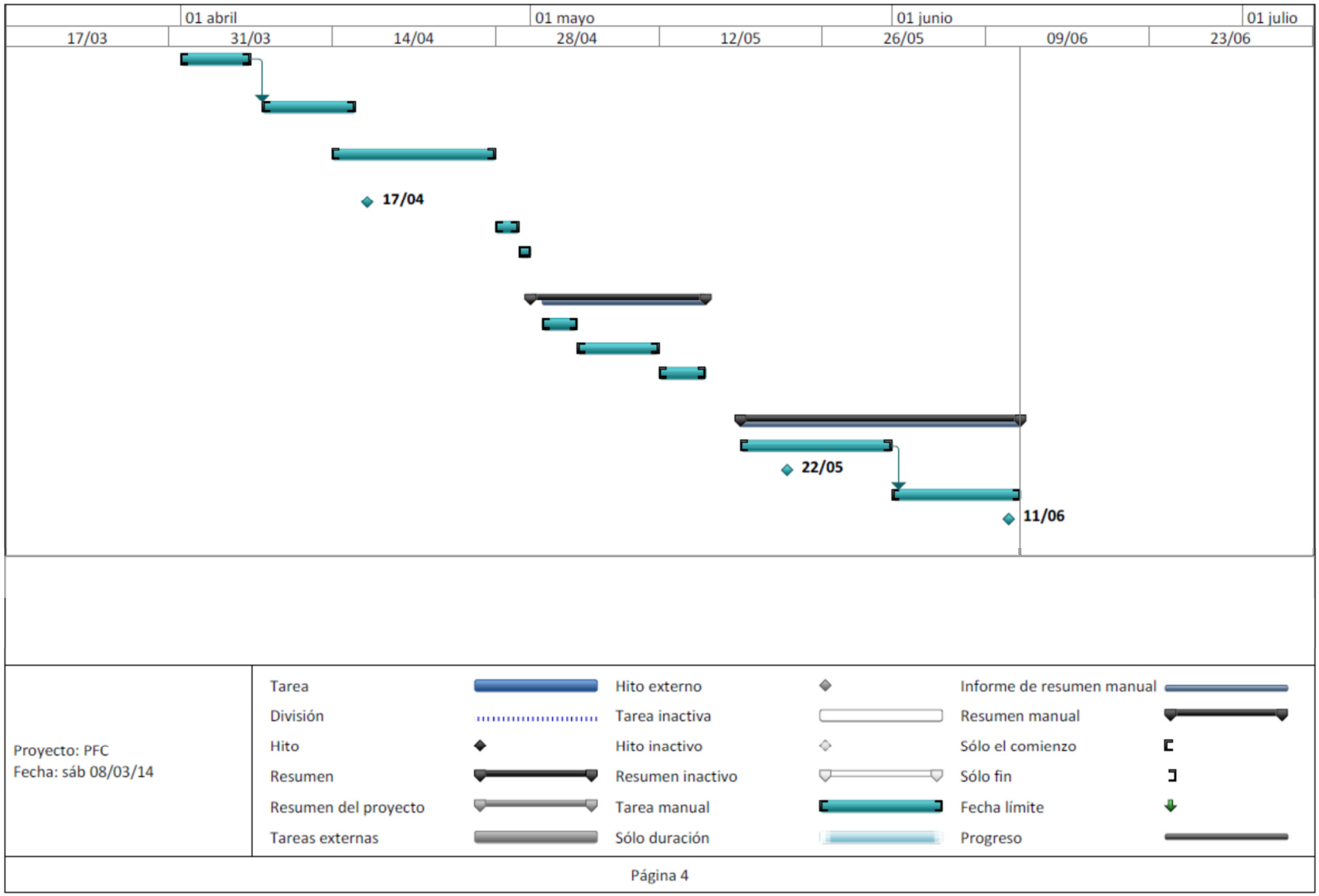
Fig. 1b

Id	Modo de tarea	Nombre de tarea	Duración	Comienzo	Fin
17		Disseny classes per a estructures de dades	6 días	mar 01/04/14	dom 06/04/14
18		Lectura fitxer CSV i càrrega de dades	7 días	mar 08/04/14	mar 15/04/14
19		Programació de l'algorisme	13 días	lun 14/04/14	dom 27/04/14
20		Entrega PAC2	0 días	jue 17/04/14	jue 17/04/14
21		Proves	2 días	lun 28/04/14	mar 29/04/14
22		Redacció de la memòria	1 día	mié 30/04/14	mié 30/04/14
23		**Conclusions**	**13 días**	**jue 01/05/14**	**jue 15/05/14**
24		Recol·lecció de dades	3 días	vie 02/05/14	dom 04/05/14
25		Elaboració de gràfics	7 días	lun 05/05/14	dom 11/05/14
26		Redacció de la memòria	4 días	lun 12/05/14	jue 15/05/14
27		**Lliurament final**	**21 días**	**lun 19/05/14**	**mié 11/06/14**
28		Integració documents	12 días	lun 19/05/14	sáb 31/05/14
29		Presentació PAC3	1 día	jue 22/05/14	jue 22/05/14
30		Presentació amb àudio	9 días	dom 01/06/14	mié 11/06/14
31		Entrega final	0 días	mié 11/06/14	mié 11/06/14
32					



Página 2


01 abril
01 mayo
01 junio
01 julio
17/03
31/03
14/04
28/04
12/05
26/05
09/06
23/06
Proyecto: PFC
Fecha: sáb 08/03/14
Tarea
División
Hito
Resumen
Resumen del proyecto
Tareas externas
Hito externo
Tarea inactiva
Hito inactivo
Resumen inactivo
Tarea manual
Sólo duración
Informe de resumen manual
Resumen manual
Sólo el comienzo
Sólo fin
Fecha límite
Progreso
Página 3


01 abril
01 mayo
01 junio
01 julio
17/03
31/03
14/04
28/04
12/05
26/05
09/06
23/06
17/04
22/05
11/06
Proyecto: PFC
Fecha: sáb 08/03/14
Tarea
División
Hito
Resumen
Resumen del proyecto
Tareas externas
Hito externo
Tarea inactiva
Hito inactivo
Resumen inactivo
Tarea manual
Sólo duración
Informe de resumen manual
Resumen manual
Sólo el comienzo
Sólo fin
Fecha límite
Progreso
Página 4

1.5 Estructura de les fases del projecte

No podem donar per conclosa la introducció del projecte sense fer una breu introducció a les pròximes fases. A l'apartat d'anàlisi i disseny es definiran els requisits del prototipus a desenvolupar, així com els diversos casos d'ús amb què treballarem. També es presentarà el disseny de classes que es farà servir a la fase d'implementació, amb una programació orientada a objectes. En aquesta fase també es definiria el nucli principal del prototipus: l'algorisme d'aprenentatge, que girarà en torn a tres grans eixos: la valoració d'una notícia i dos tipus d'adaptació del perfil evolutiu, l'adaptació online i l'adaptació offline. Finalment, es presentarà l'API del diari 'The Guardian', que es farà servir per obtenir els corpus de notícies a analitzar.

Després de l'anàlisi i disseny, entrarem en la fase d'implementació del prototipus. Primer de tot es farà una breu introducció de la tecnologia utilitzada per materialitzar el projecte. Seguidament, s'entrarà en una explicació de cadascuna de les classes implementades en Java. No hi faltarà tampoc el tractament de les dades d'entrada i sortida, fent servir JSON (J*avaScript Object Notation)* i fitxers separats per comes CSV.

Quan tinguem el programa funcionant d'acord amb les especificacions inicials, entrarem en l'estadi d'estudi de diversos supòsits. En aquesta fase, l'objectiu serà analitzar el procés d'aprenentatge a partir de diversos escenaris: tipus de repositori de dades, segons perfil ideal i segons diversos valors de les variables de l'algorisme que es podran variar. En aquest apartat també es crearan els perfils ideals a tenir en compte i s'explicarà el procés que s'haurà fet servir.

Tot plegat servirà per treure conclusions sobre el funcionament del prototipus presentat, idees que es veuran en l'última fase. En aquest punt es presentaran millores que es puguin aplicar a l'algorisme i es posaran diversos exemples d'aplicacions que es podrien nodrir del prototipus desenvolupat.

Finalment, disposarem d'una secció on es podrà consultar la bibliografia utilitzada i un apartat dedicat als annexos.

1.6 Desviacions respecte la planificació inicial

En general s'ha pogut complir la planificació inicial tal com s'havia establert. Només hi ha hagut un retard d'una setmana en l'inici de l'apartat de conclusions perquè la implementació es va allargar una mica més del que estava previst.

2. Anàlisi i disseny

En aquest apartat s'analitzarà a fons el problema plantejat a la introducció. Es començarà per definir els requisits de l'aplicació a desenvolupar i llavors s'entrarà a la fase de disseny. Amb tot, la voluntat és obtenir una especificació a gran nivell de detall de l'aplicació que es desenvoluparà a la fase d'implementació del projecte.

2.1 Definició de requisits

Els objectius definits a la secció anterior serviran per definir els requisits funcionals i no funcionals de l'aplicació que es desenvoluparà. Els primers, relacionats amb les opcions que el prototip oferirà a l'usuari i els segons, relacionats amb tota la resta de necessitats que tindrà l'aplicació per sostenir les funcionalitats especificades.

2.1.1 Requisits funcionals

1. El prototip s'haurà de connectar via Internet amb l'API de 'The Guardian' i permetre descarregar-se un corpus de notícies comprès entre dues dates determinades. Els atributs mínims que caldrà obtenir d'una notícia en concret seran un identificador, la secció principal a la qual pertany i un conjunt de paraules clau que en defineixin el contingut.
2. Si bé aquesta informació es podrà descarregar en cada execució del programa, és necessari també poder emmagatzemar un corpus de notícies en local, per tal d'agilitzar l'execució de l'algorisme.
3. Per tant, també caldrà que el prototipus sigui capaç de llegir un corpus de notícies guardat en una màquina local sense haver-se de connectar cada vegada a l'API de 'The Guardian'.
4. Una vegada carregat en memòria dinàmica un corpus de notícies en concret, el prototipus haurà de ser capaç de crear una estructura de perfil d'usuari buida. Per aconseguir-ho, caldrà que llegeixi totes les seccions i paraules clau de les notícies.
5. El programa haurà de donar l'opció d'obtenir estadístiques sobre un corpus de notícies en concret: nombre de notícies de cada secció, aparicions de cada *tag* i també nombre total de seccions i paraules clau. També oferirà un rànquing de les seccions i *tags* amb més articles.
6. El prototip permetrà crear un perfil ideal d'usuari. Es podrà fer de manera aleatòria o bé donant pes a unes seccions determinades.

7. La informació de tots els perfils també s'ha de poder emmagatzemar per tal de poder-la tornar a consultar sempre que es vulgui.
8. Donat un corpus de notícies, un perfil en blanc i un perfil ideal, el programa oferirà la possibilitat d'executar l'algorisme d'aprenentatge –nucli central d'aquest projecte- amb la hipòtesi inicial que, en cadascuna de les iteracions de l'algorisme, el perfil que començarà en blanc anirà tendint cada vegada més al perfil ideal.
9. El prototip donarà l'opció d'alterar el perfil ideal a mitja execució per simular un canvi de gustos del lector.
10. El programa oferirà la possibilitat de mirar quant de bé està aprenent l'algorisme, fent un càlcul de la distància euclidiana entre el perfil après i el perfil ideal, a nivell de seccions i a nivell de *tags*.

2.1.2 Requisits no funcional

1. Un entorn via terminal que, a través de menús, permeti a l'usuari executar la majoria de les opcions que oferirà el prototipus (algunes requeriran suport manual d'edició de fitxers).
2. El programa comptarà amb el conjunt d'executables, arxius de codi font i fitxers d'informació necessaris per tal de poder-lo executar de manera correcta.
3. El prototipus anirà acompanyat d'una documentació clara i ben estructurada entre la qual hi haurà inclòs un manual d'usuari.
4. El desenvolupament del programa es farà de manera modular i facilitant el màxim la seva aplicació a tipus de repositoris de dades diferents al que es farà servir en aquest projecte de final de carrera.
5. A banda de poder treballar amb diversos repositoris de dades, l'estructura també ha de ser capaç d'adaptar-se a diverses tecnologies que en un futur es puguin haver de nodrir de l'algorisme desenvolupat (aplicació web, *app* per a mòbil, etc.).
6. El programa haurà de facilitar l'obtenció dels diversos fitxers de dades que portarà associats per poder-los utilitzar en altres plataformes per tal de fer l'anàlisi de resultats i d'elaborar estadístiques.

2.2 Identificació i descripció dels casos d'ús

D'acord amb el llistat de requeriments del punt anterior, s'han establert tota una sèrie de casos d'ús *[Fig. 2a].* En la major part de les execucions del programa es cridaran tots ells de manera seqüencial ja que, a banda de la part de la gestió de dades, per poder provar el prototipus caldrà fer ús de totes les funcionalitats, des de la càrrega del perfil fins a l'obtenció de resultats.

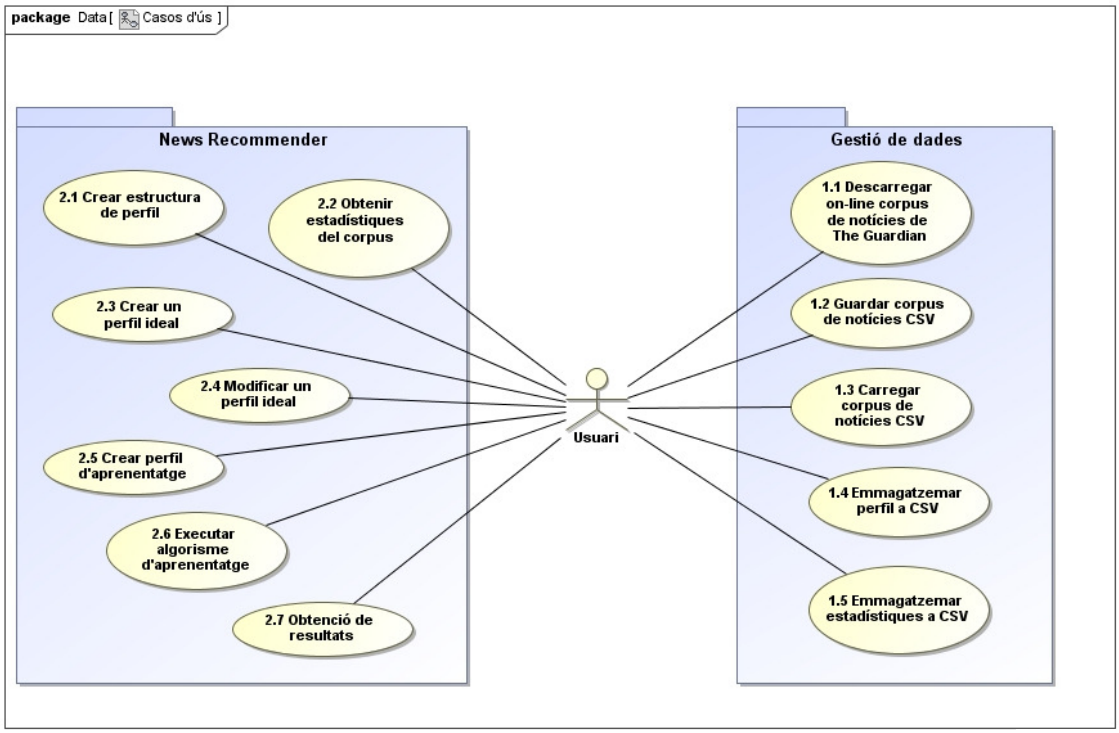


Fig. 2a

2.2.1 Gestió de dades

1.1 **Descarregar online el corpus de notícies de 'The Guardian'.** Aquest cas d'ús serà la primera fase de contacte que es tindrà amb el prototipus. Permetrà que l'usuari especifiqui un rang de temps comprès entre dues dates.

1.2 **Guardar el corpus de notícies a un CSV.** Permetrà guardar en local el corpus de notícies que hi hagi carregat a la memòria del programa en un moment determinat. Aquest fet evitarà haver-se de connectar cada vegada a l'API de 'The Guardian' i farà l'aplicació més ràpida.

1.3 **Carregar el corpus de notícies d'un CSV.** Donarà la possibilitat llegir un corpus de notícies prèviament emmagatzemat en un fitxer CSV.

1.4 **Emmagatzemar un perfil a un CSV.** Permetrà emmagatzemar un perfil en un fitxer CSV.

1.5 **Emmagatzemar estadístiques a un CSV**. Servirà per guardar dades estadístiques d'un perfil en un fitxer separat per comes.

2.2.2 NewsRecommender

2.1 **Crear l'estructura del perfil.** Servirà per crear una estructura de perfil donat un corpus de dades determinat. Farà un escaneig de totes les seccions i els *tags*.

2.2 **Obtenir estadístiques del corpus.** Aquest cas d'ús donarà la possibilitat de tenir un recompte del nombre de seccions i *tags*, això com també l'aparició de cadascun d'ells. Permetrà guardar les dades en un CSV per poder fer tractaments de les estadístiques més avançats amb un full de càlcul extern.

2.3 **Crear un perfil ideal.** Servirà per crear un perfil ideal, que modelarà els gustos d'una persona en concret.

2.4 **Modificar un perfil ideal.** Permetrà modificar paràmetres d'un perfil ideal per simular l'evolució de gustos d'un usuari. També per configurar inicialment un perfil ideal en concret.

2.5 **Crear un perfil d'aprenentatge.** Inicialitzar el perfil evolutiu abans d'executar l'algorisme d'aprenentatge.

2.6 **Executar algorisme d'aprenentatge.** Cas d'ús en què s'executarà l'algorisme que tindrà com objectiu que el perfil evolutiu tendeixi al perfil ideal.

2.7 **Obtenció de resultats.** Quan s'hagi executat l'algorisme, es mostraran per pantalla els resultats de l'aprenentatge.

2.3 Diagrama de classes

Veiem el diagrama de classes a la *Fig. 2b*:

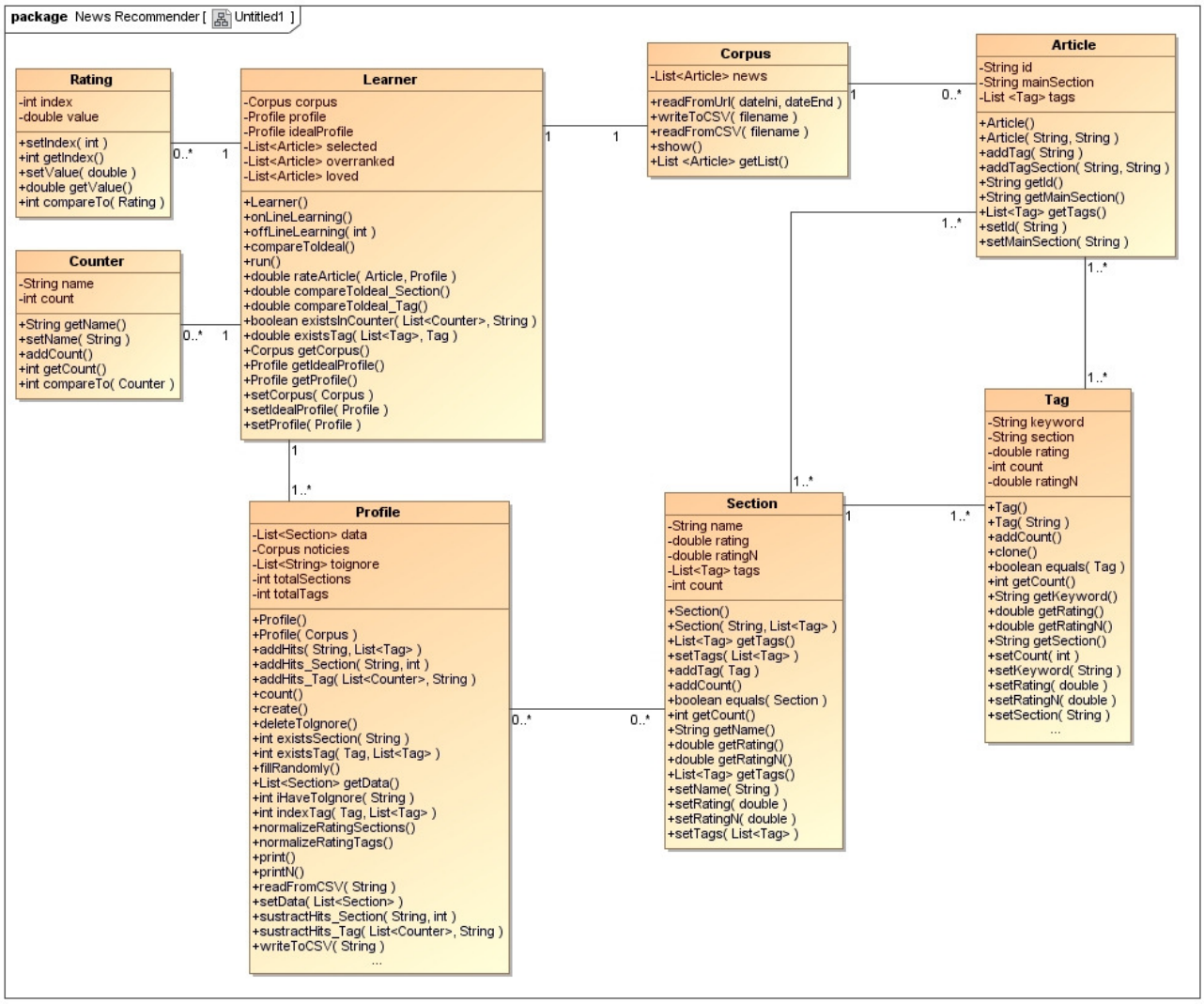


Fig. 2b

2.3.1 Tag

Aquesta classe tindrà els següents atributs:

- **keyword:** Paraula clau que representa el *tag*
- **section:** Secció a la qual pertany el *tag*
- **rating:** Valoració d'aquest *tag* en el perfil en què estigui instanciat
- **ratingN:** Valor del *rating* normalitzat entre 0 i 1
- **count:** Nombre d'aparicions de l'etiqueta en tot el corpus de notícies

I els següents mètodes:

- **Tag():** Constructor de classe
- **Tag(String):** Crea un Tag en blanc només amb el nom
- **clone():** Crea una còpia del Tag
- **boolean equals(Tag):** Retorna 'cert' si dos *tags* són iguals i 'fals' si són diferents
- **addCount**: Incrementarà en una unitat el nombre d'aparicions de l'etiqueta.
- **int getCount():** Retorna el nombre d'aparicions de l'etiqueta
- **String getKeyword():** Retorna la paraula clau que defineix el Tag
- **double getRating():** Retorna la valoració del Tag en un moment determinat
- **double getRatingN():** Retorna la valoració normalitzada del Tag en un moment determinat
- **String getSection():** Retorna la secció a la qual pertany el Tag
- **setCount(int):** Permet establir el paràmetre que defineix el nombre d'aparicions del Tag
- **setKeyword(String):** Defineix la paraula clau que defineix el Tag
- **setRating(double):** Estableix la valoració d'un Tag
- **setRatingN(double):** Estableix la valoració normalitzada d'un Tag
- **setSection(String):** Defineix a quina secció pertany un Tag

2.3.2 Section

Aquesta classe tindrà els següents atributs:

- **name:** El nom de la secció
- **rating:** Valoració de la secció en el perfil en què estigui instanciada
- **ratingN:** Valoració de la secció normalitzada entre 0 i 1
- **tags:** Llistat de *tags* que té aquesta secció en tota la mostra
- **count:** Nombre d'aparicions de la secció en el corpus de notícies

I els següents mètodes:

- **Section():** Constructor de classe
- **Section(String, List<Tag>):** Crea una secció nova amb el nom i la llista de *tags* que se li passen com a paràmetres
- **getTags:** Retorna el llistat de *tags* de la secció
- **setTags:** Assigna una llista de *tags* a la secció
- **addTag:** Afegeix un *tag* a la secció

- **addCount:** Incrementa en una unitat el nombre d'aparicions de la secció en el corpus de notícies.
- **boolean equals(Section):** Retorna 'cert' si les dues seccions són la mateixa i 'fals' en cas contrari
- **int getCount():** Retorna el nombre d'aparicions de la secció en el corpus de notícies
- **String getName():** Informa sobre el nom d'una secció
- **double getRating():** Retorna la valoració d'una secció
- **double getRatingN():** Retorna la valoració d'una secció normalitzat entre 0 i 1
- **List<Tag> getTags():** Retorna el llistat de *tags* associat a una secció
- **setName(String):** Defineix el nom d'una secció
- **setRating(double):** Estableix la valoració d'una secció en concret
- **setRatingN(double):** Estableix la valoració d'una secció en concret, normalitzada entre 0 i 1
- **setTags(List<Tag>):** Associa un llistat de *tags* a una secció

2.3.3 Article

Aquesta classe tindrà els següents atributs:

- **id:** Identificador de la notícia
- **mainSection:** Secció principal de la notícia
- **tags**: Etiquetes que defineixen la notícia. Convé notar que no tots els *tags* pertanyeran a la secció principal. Una notícia pot tenir *tags* classificats en diverses seccions

I els següents mètodes:

- **Article()**: Constructor de classe
- **Article(String, String):** Crea un article amb l'identificador i el títol que li passem com a paràmetre
- **addTag(String):** Afegeix un *tag* a l'article passant-li el seu nom com a paràmetre
- **addTagSection(String, String):** Afegeix un *tag* a l'article informant també de la secció a la qual pertany
- **String getId()**: Retorna l'identificador de l'article
- **String getMainSection():** Retorna la secció principal d'un article

- **List<Tag> getTags():** Retorna els *tags* associats a un article
- **setId(String):** Assigna un identificador a l'article
- **setMainSection(String):** Assigna a un article la seva secció principal
- **List<Tag> getTags():** Retorna el llistat de *tags* de l'article

2.3.4 Corpus

Aquesta classe tindrà els següents atributs:

- **news:** Llistat d'articles

I els següents mètodes:

- **readFromUrl (Date, Date):** Mètode que es connecta a l'API de The Guardian per descarregar les notícies entre dues dates determinades
- **writeToCSV (String):** Emmagatzema el corpus de notícies en un fitxer CSV amb el nom que se li passa com a paràmetre
- **readFromCSV (String):** Recupera el corpus de notícies d'un fitxer CSV, donat el seu nom passat com a paràmetre
- **show():** Mostra el corpus de notícies per pantalla
- **getList():** Retorna el llistat d'articles associats al corpus

2.3.5 Profile

Aquesta classe tindrà els següents atributs:

- **data:** Llistat de seccions del perfil.
- **corpus:** Corpus d'articles sobre el qual s'elabora el perfil
- **List<String> toignore:** Seccions que s'ignoraran per evitar biaixos en l'aprenentatge
- **int totalSections:** Recompte del total de seccions que es tenen en compte en el perfil
- **int totalTags:** Recompte del total de *tags* que es tenen en compte en el perfil

I els següents mètodes:

- **Profile():** Constructor de classe
- **Profile(Corpus):** Construeix un perfil i li assigna el corpus que se li passa com a paràmetre

- **addHits(String, List<Tag>):** Mètode per afegir valoració a una secció i a un llistat de *tags*. És el mètode que es crida des de l'aprenentatge online, en què s'afegeix valor a les paraules que defineixen un article que s'ha considerat interessant per a l'usuari
- **addHits_Section(String, int):** Mètode per afegir una valoració en concret –que es passa com a paràmetre- a una secció en concret. Es crida des de l'aprenentatge offline.
- **addHits_Tag(List<Counter>, String):** Afegeix valoració a un llistat de *tags* indexats en una llista de comptadors i que pertanyen a una secció en concret el nom de la qual es passa també com a paràmetre
- **count():** Fa un recompte del nombre de seccions que hi ha al perfil i de les vegades que apareix cadascuna en el corpus. També en el cas dels *tags*. Ho emmagatzema tot en un fitxer amb finalitats estadístiques
- **deleteToIgnore():** Esborra del perfil aquelles seccions que s'ha decidit ignorar per evitar biaixos en l'aprenen*tag*e
- **int existsSection(String):** Mètode que esbrina si una secció –se li passa el seu nom com a paràmetre- existeix o no en un perfil.
- **int existsTag(Tag, List<Tag):** Mira si existeix un *tag* en una llista de *tags*
- **fillRandomly():** Omple les valoracions d'un perfil (normalitzades) de manera aleatòria
- **List<Section> getData():** Retorna el llistat de seccions associat al perfil
- **int iHaveToIgnore(String):** Mira si una secció en concret pertany a aquelles que s'han d'ignorar.
- **int indexTag(Tag, List<Tag>):** Retorna l'índex de la llista que identifica el *tag* que es passa com a paràmetre
- **normalizeRatingSections():** Normalitza els *ratings* de les seccions del perfil entre 0 i 1 utilitzant ranging
- **normalizeRatingTags():** Normalitza els *ratings* dels *tags* del perfil entre 0 i 1 utilitzant ranging
- **create():** Crea una estructura buida de perfil d'acord amb el corpus de notícies que hi hagi carregat en memòria
- **print():** Ensenya les dades de les seccions i els *tags* del perfil, així com les valoracions –sense normalitzar- de cadascun dels components
- **printN()** Ensenya les dades de les seccions i els *tags* del perfil, així com les valoracions –normalitzades- de cadascun dels components

- **readFromCSV(String):** Llegeix un perfil d'un fitxer –rep el seu nom com a paràmetre- i el carrega en memòria
- **setData(List<Section>):** Assigna una llista de seccions a un perfil
- **sustractHits_Section(String, int):** Treu un nombre determinat d'unitats a la secció de la qual rep el nom com a paràmetre
- **sustractHits_*tag*(List<Counter>, String):** Resta valoració a un llistat de *tags* d'una secció en concret
- **writeToCSV(String):** Escriu el perfil en un fitxer CSV
- **count():** Ensenya el recompte de seccions i *tags* del perfil.

2.3.6 Learner

Aquesta classe tindrà els següents atributs:

- **corpus:** Conté el llistat de notícies en base al qual es durà a terme l'aprenentatge
- **profile:** Perfil evolutiu
- **idealProfile:** Perfil ideal
- **selected:** Nombre d'articles seleccionats en una iteració en concret
- **overranked:** Llista d'articles que han estat sobrevalorats en un moment determinat de l'algorisme d'aprenentatge
- **loved:** Llistat d'articles que han estat seleccionats d'acord amb el perfil ideal en un moment determinat de l'algorisme d'aprenentatge

I els següents mètodes:

- **Learner():** Constructor de classe
- **onLineLearning():** Aprenentatge online que s'executa després de cada elecció.
- **offLineLearning:** Aprenentatge online que s'executarà quan hi hagi una mostra significativa de notícies (loved o overranked)
- **run:** Executa l'algorisme d'aprenentatge
- **rateArticle(Article, Profile):** Valora un article en funció del perfil que se li passa com a paràmetre
- **double compareToIdeal_Section():** Compara –a nivell de seccions- el perfil evolutiu amb l'ideal per mesurar en quina mesura s'ha efectuat bé o malament l'aprenentatge

- **double compareToIdeal_Tag():** Compara –a nivell de *tags*- el perfil evolutiu amb l'ideal per mesurar en quina mesura s'ha efectuat bé o malament l'aprenentatge
- **boolean existsInCounter(List<Counter>, String):** Mira si un text en concret –que representarà una secció o un *tag*- existeix en una llista de contadors
- **double existsTag(List<Tag>, <Tag>):** Mira si existeix un *tag* en una llista de *tags*
- **Corpus getCorpus():** Retorna el corpus de notícies associat a l'objecte Learner
- **Profile getIdealProfile():** Retorna el perfil ideal associat a l'estructura Learner
- **Profile getProfile():** Retorna el perfil evolutiu associat a l'estructura Learner
- **setCorpus (Corpus):** Assigna un corpus de notícies a l'objecte Learner
- **setIdealprofile(Profile):** Assigna un perfil ideal a l'objecte Learner
- **setProfile(Profile):** Assigna un perfil evolutiu a l'objecte Learner

2.3.7 Rating

Aquesta classe tindrà els següents atributs:

- **index:** Índex que anirà del 0 al 14 per representar la quinzena de notícies seleccionades en cada iteració
- **value:** Emmagatzemarà el *rating* que s'ha donat a un article en concret

I els següents mètodes:

- **setIndex(int):** Assigna un índex a un objecte Rating
- **int getIndex():** Retorna el valor de l'índex
- **setValue(double):** Assigna un valor a l'objecte
- **double getValue():** Retorna el valor de l'objecte
- **int compareTo(Rating):** Compara dos objectes Rating (ens servirà per ordenar la quinzena d'articles segons el seu *rating*)

3.3.8 Counter

Aquesta classe tindrà els següents atributs:

- **name:** Representa el nom d'una secció o un *tag*
- **count:** Nombre d'aparicions de la secció o *tag*

I els següents mètodes:

- **String getName():** Retorna el nom de l'objecte Counter (representarà una secció o un *tag*)
- **setName(String):** Assigna un nom a un objecte Counter
- **addCount():** Afegeix una aparició a l'objecte
- **int getCount():** Retorna el nombre d'aparicions de l'objecte
- **int compareTo(Counter):** Compara dos objectes Counter segons el seu nombre d'aparicions (ens servirà per fer rankings d'aparicions a l'aprenentatge offline)

2.4 Algorisme d'aprenentatge

El nucli central del prototip a desenvolupar serà l'algorisme d'aprenentatge que s'encabirà a la classe Learner que s'ha pogut veure en el diagrama de classes del projecte. Hem vist fins ara que aquest algorisme partirà d'una estructura de perfil construïda d'acord amb un repositori de dades concret, fruit de les dades extretes d'un repositori de continguts que podria ser qualsevol que organitzi els seus documents per seccions i paraules clau.

L'estructura de perfil, una vegada carregades les dades al programa, tindrà la següent forma:

Secció1 (r1)	**Tag1a (r1a)**	**Tag1b (r1b)**	**Tag1c(r1c)**	**Tag1d(r1d)**	**...**	**Tag1M(r1M)**
Secció2 (r2)	**Tag2a (r2a)**	**Tag2b (r2b)**	**Tag2c(r2c)**	**Tag2d(r2d)**	**...**	**Tag2M(r2M)**
Secció3 (r3)	**...**					
Secció4 (r4)	**...**					
Secció5 (r5)	**...**					
...	**...**					
Secció N (rN)	**TagNa (rNa)**					**TagMN (rMN)**

Quan totes les dades estiguin carregades, se'ls farà un escaneig per determinar a quantes seccions pertanyen i quins *tags* té associats cadascuna d'aquestes seccions. Convé deixar clar que el llistat de *tags* serà una llista dinàmica, ja que cada secció pot tenir un nombre diferent d'etiquetes associades. És a dir, una en podria tenir tres i una altra cent, per posar un exemple.

Cada secció tindrà la seva valoració global associada, així com cada etiqueta. S'ha decidit que es farà una valoració en aquests dos nivells per tenir més dades sobre els gustos d'un usuari. Si li agrada molt una secció en general, li podran agradar altres continguts englobats dins la mateixa temàtica tot i tenir etiquetes diferents.

Per tal de poder exemplificar millor a partir d'ara el procés d'aprenentatge, ens traslladarem ja al cas particular al qual s'aplicarà l'algorisme en aquest projecte, a un repositori de notícies. Es tindran en compte les següents consideracions:

1. Es treballarà sobre corpus d'aproximadament 6.000 notícies agafades a través de l'API del diari 'The Guardian' d'un període de temps acotat entre dues dates.
2. De cada una de les notícies, s'agafarà un ID, la secció principal a la qual pertany i una llista amb totes les paraules clau que la defineixen. Les paraules clau van acompanyades de la secció a la qual pertanyen. Aquesta informació es farà servir per ordenar-les dins l'estructura del perfil, així també es podrà considerar la informació de les seccions secundàries a les quals pertany la notícia.
3. El sistema treballarà amb dos perfils. Un al que anomenarem perfil ideal (el que volem aprendre) i el perfil evolutiu, que serà el que va evolucionant durant l'execució de l'algorisme i que es pretén aconseguir que vagi tendint cap al perfil ideal.
4. L'algorisme d'aprenentatge no supervisat anirà recorrent el corpus d'aproximadament 6.000 notícies i anirà treballant amb elles de 15 en 15 (conjunts d'entrenament), en grups escollits de manera aleatòria. Una notícia només apareixerà en un sol grup. Per tant, tindrem un total d'unes 400 iteracions (6000/15).
5. En cada iteració, per a cadascuna de les 15 notícies s'obtindrà una valoració (*rating*) en funció de les seves característiques i la relació que tenen amb el perfil evolutiu. Aquestes notícies s'ordenaran de més a menys valoració.
6. Es simularà quina seria la notícia preferida del grup de 15 escollida pel perfil ideal que hem definit. Serà la que tingui el *rating* més alt d'acord amb el perfil ideal. D'alguna manera, aquest perfil simularà els gustos de l'usuari de l'aplicació, la seva manera de pensar o escollir, per dir-ho d'alguna manera.

Exemple de l'aparença que podrà tenir un perfil de l'algorisme adaptat a aquest cas particular (els números entre parèntesi són les valoracions dels *tags* i les seccions)

section (name/*rating*)	***tag*1**	***tag*2**	***tag*3**	***tag*4**	***tag*5**	***tag*6**	**...**	***tag*N**
Education/ 0.3	Primary-schools / 0.2	Schools / 0.1	Education /0.3					
Uk / 0.2	Uk/ 0.2							
Lifeandstyle /0.5	Health-and-wellbeing /0.8	Food-and-drink /0.8	Lifeandstyle /0.2	Parents-and-parenting /0.5	Family / 0.2	Pregnancy /0		
Society /0.2	Children /0.1	Society /0	Caesareans /0	Nhs /0.1	Mental-health /0	Health / 0.1		

2.4.1 Valoració d'una notícia

La classe Learner comptarà amb un mètode que servirà per valorar un article tal com s'ha introduït en el disseny dels punts anteriors. Anem a conèixer el seu funcionament en pseudocodi. Parlarem de hits com a unitats de rating. És a dir, si una secció té un *rating* de 5, tindrà 5 *hits*.

rateArticle (Article, Profile)

R*ate_tags*

Calcular el sumatori de pesos dels *hits* de les paraules clau coincidents en tot el perfil. És a dir, s'anirà a cercar cadascuna de les etiquetes que aparegui en l'article corresponent al perfil passat com a paràmetre. Convé notar que no totes les etiquetes associades a un article pertanyeran a la mateixa secció.

Els *ratings* amb els quals es treballarà estaran normalitzats entre 0 i 1 entre totes les etiquetes que hi hagi en un perfil concret i s'emmagatzemaran a la classe Article.

Rate_section

Agafar la valoració global que la secció principal de l'article a tractar té al perfil (també normalitzada). Es valorarà la secció en conjunt per tenir en compte que si a un usuari li agrada molt una secció, encara que tingui *ratings* a paraules clau diferents a la notícia que estem valorant,

segurament li poden agradar altres notícies de la secció definides amb altres *tags* diferents

Una vegada obtingudes les valoracions de les paraules clau de la notícia i de la seva secció principal, procedirem a obtenir el *rating* global, que es calcularà d'acord amb la següent fórmula:

rating = α*valoració_*tags* + β*valoració_secció

Caldrà veure amb quins valors d'alfa i beta es du a terme millor l'aprenentatge. Inicialment s'ha començat provant un 60 i un 40%, respectivament. No obstant això, en la fase de proves (Apartat 4.4) s'ha descobert que un 50% i 50% dóna millors resultats. S'ha considerat que valorar els *tags* per separat i el conjunt de la secció és una solució pels casos en què, per exemple, si a algú li agraden molt les notícies de *society/health*, potser també li agraden altres notícies de societat però d'altres branques i no només les de salut, que seria allà on arribaríem si només féssim cas dels *tags*.

Una vegada fet el *rating* de les 15 notícies d'una iteració, s'ordenaran de major a menor valoració. La de més amunt serà la preferida del perfil evolutiu.

Per dur a terme l'aprenentatge i anar actualitzant els *ratings* dels *tags* i les seccions del perfil es farà a cada iteració mitjançant l'aprenentatge online. Així mateix, s'anirà executant periòdicament un altre procés al qual anomenarem aprenentatge offline. Aquest últim se centrarà en l'històric d'articles seleccionats i rebutjats per part de l'usuari.

Una vegada ordenada una llista de 15 notícies segons els pesos del perfil evolutiu, si el perfil ideal selecciona, per exemple, la que queda a la posició 6, es considerarà que els cinc articles que li queden per sobre són articles rebutjats.

A tall d'introducció, l'aprenentatge online es repetirà a cada iteració i, per tant, no farà canvis substancials, ja que no es pot treure un gran patró de comportament a partir d'una notícia aïllada. D'altra banda, l'aprenentatge offline farà canvis més substancials, ja que buscarà similituds entre els articles seleccionats i entre els rebutjats amb l'objectiu de trobar patrons. En trobar repeticions en les notícies preferides i les rebutjades, podrà actualitzar els pesos positivament o negativament.

2.4.2 Adaptació online

L'aprenentatge online s'executarà a cada iteració, després de la selecció de l'usuari (perfil ideal) d'una notícia d'entre la quinzena de seleccionades.

Treballarem amb els següents paràmetres:

- **UNIT:** Unitat amb la que treballarem (serà igual a 1 però es deixarà parametritzable)
- **ONLINE_MAINSECTION_RATE:** Coeficient multiplicador de *hits* a la secció principal. Valdrà dos punts per donar-li més importància que els *tags*.
- **ONLINE_TAG_RATE:** Coeficient multiplicador de *hits* d'un *tag* que valdrà una unitat.

El seu comportament, en pseudocodi, serà el següent:

onLineLearning()

per a cada grup de MIDA_MAX_GRUP notícies del corpus **fer**

Calculem el *rating* de cadascuna de les notícies segons el perfil ideal i el perfil evolutiu

Ordenem les notícies segons el *rating*

si (noticia_preferida_ideal = noticia preferida_evolutiu) **llavors**

No podem aprendre res perquè ja ho ha fet bé. Afegim la notícia al grup de loved

altrament

Segons les dades (secció i *tags*) de la notícia preferida ideal, modifiquem el perfil evolutiu de la següent manera:

- Augmentem els hits de la secció principal de l'article en ONLINE_MAIN_SECTON_RATE (2) unitats
- Augmentem els hits de cada *tag* en ONLINE_TAG_RATE (1)unitats
- Recalculem les valoracions de les seccions. Hi afegim el sumatori dels hits afegits a les paraules clau.
- Posem "en quarantena" les notícies overranked i guardem en un històric la notícia seleccionada pel perfil ideal, dades per a l'adaptació offline.

- Mirem si és necessari entrar a l'adaptació offline (quan hi hagi nombre significatiu de notícies). Si s'escau, es crida el mètode en aquest punt.

fsi

fper

2.4.3 Adaptació offline

L'adaptació offline s'executarà quan tinguem un nombre significatiu de notícies overranked que anomenarem η (OVER_RANKED_SIGNIFICATIU). Inicialment el definirem a 100 però estarà parametritzat per tal de poder provar diferents valors. A l'apartat de la secció 4.4 es justificarà l'elecció d'aquest paràmetre i de tots els altres. Els canvis en els *hits* de les paraules clau i seccions seran de més magnitud que en l'adaptació online, ja que tindrem en compte més dades i el patró de conducta tindrà més pes que quan tinguem només un sol article. Si en l'adaptació online els canvis que farem seran d'una unitat, en l'offline els establirem en μ unitats (nombre d'aparicions). A nivell de pseudocodi, el procediment definit en aquest apartat actuarà de la següent manera:

offLineLearning()

S'executarà des de l'onLineLearning cada vegada que a la llista de loved o overranked hi hagi un nombre significatiu de notícies.

cas (tipus)

overranked:

1. Fer un recompte de les aparicions de seccions d'entre totes les notícies que hi hagi en aquest moment determinat a la llista overranked
2. Mirar quines seccions superen el PERCENTATGE_SUPERAR* (10%) i esborrar les que no el superin. S'ha considerat que un 10% demostra que hi ha una quantitat significativa de notícies rebutjades d'una mateixa secció.
3. Ordenar-les segons el número d'aparicions
4. Per a cada secció d'aquesta llista, restar al perfil evolutiu el nombre de hits que té associats. La magnitud de la resta la marcarà el nombre d'aparicions (superior al 10%). S'ha establert així perquè la resta sigui proporcional en cada cas. Una secció que tingui un 30% d'aparicions en aquesta llista haurà de ser més rebutjada que una que només en tingui un 11%.

5. De les seccions overranked que superen el PERCENTATGE_SUPERAR Ω %, n'agafem els *tags* més repetits. Per fer-ho, caldrà fer un recompte de tots els *tags* i les seves aparicions i ordenar aquesta llista.
6. Ens quedarem només amb els més repetits. NUM_TAGS_OFFLINE* tindrà un valor de deu.
7. A aquests deu *tags*, els restarem DOWN_OVER_RANKED* (5) unitats al perfil evolutiu. La resta s'ha configurat com a més gran que en l'aprenentatge online perquè aquí tenim un patró de conducta més elaborat.

loved:

1. Funciona exactament que overranked però en comptes de restar hits, en aquest cas en sumarem, ja que no estem davant d'un patró de rebuig, sinó d'atracció. Treballarem amb una llista de loved (històric de notícies seleccionades pel perfil ideal). Els paràmetres seran els mateixos, UP_LOVED* (5)

Fcas

* En l'apartat d'estudi de l'algorisme se seleccionaran els valors que van millor per a tots aquests paràmetres, en aquest punt s'han explicat amb els valors de la hipòtesi inicial.

2.4.4 Valorar eficàcia

Una vegada acabat el procés d'aprenentatge, serà necessari calcular en quina mesura el perfil evolutiu s'assembla al perfil ideal. La voluntat de tot el procés és que aquests dos usuaris modelats s'assemblin al màxim possible. Com més petita sigui la diferència entre tots dos, més bona es podrà dir que és l'eficàcia de l'algorisme d'aprenentatge.

Per determinar la similitud entre els dos perfils calcularem la seva distància euclidiana:

$$d(x_j, x_k) = (\Sigma_{i=1}^{M} (A_i(x_j) - A_i(x_k))^2)^{1/2}$$

Per a cadascuna de les seccions tindrem un vector de *ratings* de *tags*. A la fórmula anterior, el vector x_j seria el perfil ideal i x_k el perfil evolutiu.

Per exemple,

Society (*rating tag*1, *rating tag*2, *rating tag*3.... *rating tag*N)

Podríem tenir que al perfil ideal aquest vector fos X_j (Societat) = (0.2, 0.9, 0.5,0.3) i que al perfil evolutiu fos X_k (Societat) = (0.1, 1, 0.1,0.3). Per a cadascuna de les seccions caldrà calcular la seva distància euclidiana amb la fórmula exposada. Finalment, se sumaran totes les distàncies i s'obtindrà el sumatori de distàncies de *tags*.

D'altra banda, es calcularà una altra distància euclidiana entre els *ratings* de seccions de manera que, per exemple, recuperant un exemple ja esmentat anteriorment:

section (name/*rating*)	***tag*1**	***tag*2**	***tag*3**	***tag*4**	***tag*5**	***tag*6**	**...**	***tag*N**
Education/ 0.3	Primary-schools / 0.2	Schools / 0.1	Education /0.3					
Uk / 0.2	Uk/ 0.2							
Lifeandstyle /0.5	Health-and-wellbeing /0.8	Food-and-drink /0.8	Lifeandstyle /0.2	Parents-and-parenting /0.5	Family / 0.2	Pregnancy /0		
Society /0.2	Children /0.1	Society /0	Caesareans /0	Nhs /0.1	Mental-health /0	Health / 0.1		

Tindríem el següent vector de seccions. X_m=(0.3, 0.2, 0.5, 0.2) corresponents a Education, Uk, Lifeandstyle i Society.

Així doncs, tindrem dos indicadors sobre com de bé o de malament està aprenent l'algorisme: la distància total entre *tags* d'una secció determinada i la distància total entre seccions. Aquests càlculs es faran a compareToIdeal() de la classe Learner. Es treballarà amb dades normalitzades.

2.4.5 Definició del perfil ideal

L'objectiu de l'estructura del perfil ideal és simular el perfil d'un usuari que, en un futur, podria utilitzar una aplicació de recomanació de continguts que es nodrís de l'algorisme implementat en aquest projecte, com ja s'ha exposat en algun dels punts anteriors. La idea del projecte és aconseguir que un perfil evolutiu (inicialment amb tots els seus valors a zero) s'acabi assemblant el màxim possible al perfil ideal a aprendre.

Davant d'un contingut concret, tota persona té una determinada atracció o rebuig envers al seu contingut. Per omplir el perfil ideal, es crearà un repositori de perfils d'usuari ficticis i es donaran valors especialment alts a unes seccions i *tags*

determinats, per simular els gustos més destacats del tipus de persona en concret que ens interessi modelar. Aquestes valoracions marcaran l'eix general d'un perfil ideal. No obstant això, com que la magnitud de seccions i *tags* és difícilment tractable manualment, la resta d'informació del perfil s'omplirà amb valors aleatoris d'entre 0 i 1.

Per no provocar un biaix, a les seccions que tinguin molts *tags* només s'assignaran valoracions a uns quants i els altres es deixaran a 0, notant que no hi ha una valoració coneguda sobre determinats temes. Si no fos així, una secció amb molts *tags* jugaria amb molt d'avantatge davant d'una que en tingui pocs pel que fa a la valoració global de la secció. El nombre màxim de *tags* a assignar valor serà la mitjana de nombre de *tags* de cada secció, que se situa als 38, segons s'ha calculat fent la mitjana del corpus1 i el corpus2.

2.4.6 Seccions a ignorar

Arran de les diverses idiosincràsies amb les quals l'algorisme pugui treballar, en alguns casos tindrem seccions que ens podran esbiaixar les dades. Això passa en el cas de 'The Guardian'. Fixem-nos amb aquesta estructura de perfil:

section	***tag*1**	***tag*2**	***tag*3**	***tag*4**	***tag*5**	***tag*6**	**...**	***tag*N**
Education	Primary-schools	Schools	Education					
Uk	Uk							
Lifeandstyle	Health-and-wellbeing	Food-and-drink	lifeandstyle	Parents-and-parenting	Family	Pregnancy		
Society	Children	Society	caesarens	nhs	Mental-health	Health		
Profile	Benquinn	Editorial	Barryglendenining	Jacob-steinberg	Scotmurray			
Tone	News	Editorials	Comment	Minutebyminute	Blog	Interview		Features
Type	Article	Profile						
Sport	Daviscup	Australia-sport	Sport	Lleyton-hewitt	Jo-wilfried-tsonga			

Veiem que les seccions 'Tone' i 'Type' serveixen per definir el gènere de l'article. Com que tots els articles tenen un gènere, serien seccions que agradarien molt als usuaris perquè apareixen en cadascun dels articles. Per tant, no es tindran en compte en el moment de l'aprenentatge i es col·locaran en una llista de seccions a ignorar.

Val a dir que les dades que proporcionen les seccions esmentades, i altres que poden tenir característiques similars que es cercaran en la fase d'estudi, poden aportar altres

valors que serien interessants en futures ampliacions/millores de l'algorisme d'aprenentatge que estem desenvolupant.

2.5 Repositoris de dades

Per tal de no esbiaixar els resultats de l'aprenentatge depenent de les dades de les quals es parteixi i poder validar millor el comportament de l'algorisme s'ha optat per treballar amb dos repositoris de dades diferents, que seran, per una banda, totes les notícies de la primera quinzena de gener del 2014 i, per l'altra, els articles de la primera quinzena de febrer del mateix any.

La primera mostra serà de 5.836 notícies, mentre que, la segona, de 6.431. També es faran servir repositoris més petits per a fer proves durant la implementació de l'algorisme.

2.5.1 API The Guardian

Per tal de descarregar les notícies, caldrà connectar el nostre programa amb l'API del diari 'The Guardian', l'eina que ofereix als desenvolupadors per incloure a les seves aplicacions dades sobre les notícies que publica el rotatiu britànic. Per poder-la utilitzar amb llicència, hem registrat la nostra aplicació, de finalitat acadèmica.

L'aspecte de la interfície d'usuari de l'API és la següent *[Fig. 2c]*:

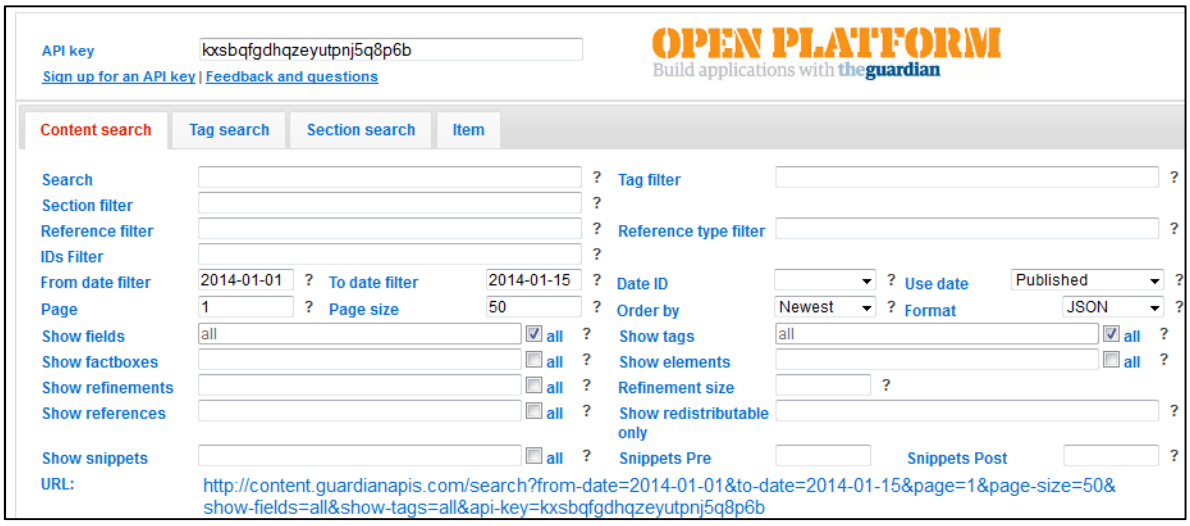


Fig. 2c

L'apartat que ens interessa és el 'Content search'. Com que volem obtenir totes les notícies entre dues dates, li hem d'especificar la data d'inici i de fi de la mostra que ens interessa. Llavors, li especifiquem com a 50 el nombre de notícies que ens presentarà en una pàgina (és el màxim) i li diem que ens ensenyi la pàgina 1. Seleccionem que

ens mostri tots els camps d'una notícia amb 'Show fields' i que ens mostri totes les paraules clau amb 'Show *tags*'. Deixem els altres paràmetres tal com estan.

El resultat és una URL amb el llistat de notícies en format JSON *[Fig. 2d]*:

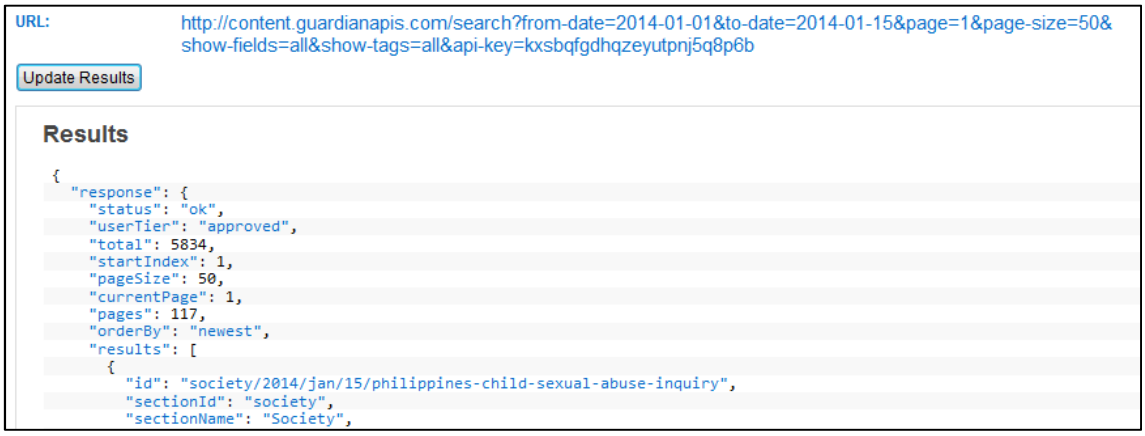


Fig. 2d

Veiem que la mateixa interfície de l'API ens mostra la primera pàgina dels resultats. No obstant això, el que ens interessarà amb vista a la implementació és la URL, amb la qual podrem navegar per les 117 pàgines de resultats –en el cas de la primera mostra- i obtenir els camps que ens interessen de cadascuna de les notícies.

És important destacar que l'API ens proporciona també el text de cada article, no es limita a proporcionar-nos només les seves metadades. També ens dóna l'enllaç a la pàgina web i aquesta informació ens seria molt útil en el cas de voler donar una aplicació real a l'algorisme, com es comenta a la secció de conclusions 5.6.

3. Implementació

En aquesta fase es materialitzarà el prototipus que s'ha definit a l'anàlisi i el disseny del projecte. En primer lloc es definirà l'entorn de desenvolupament utilitzat i llavors s'entrarà en un anàlisi a fons de cadascuna de les classes. Finalment, a tall de conclusió, es comentaran els punts del desenvolupament que hagin estat més especials i, en definitiva, una valoració global d'aquest procés. Comencem per un pla general de l'entorn de desenvolupament *[Fig. 3a]*:

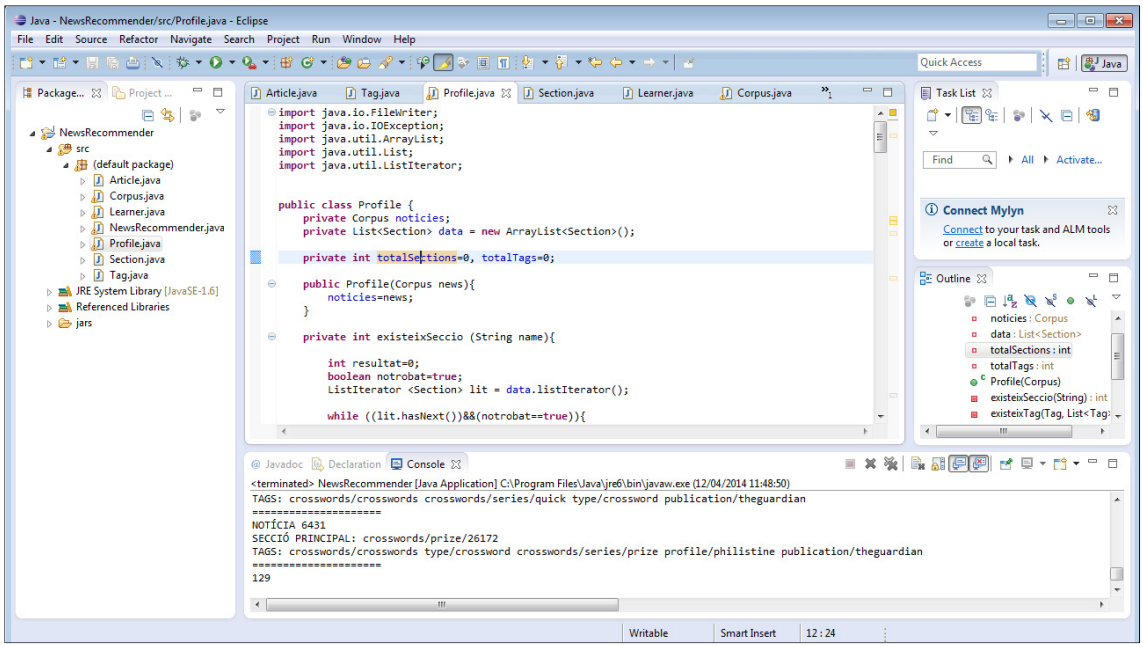


Fig. 3a

3.1 Entorn de desenvolupament

Per dur a terme el desenvolupament del projecte s'utilitzarà el següent entorn:

- **Eclipse IDE for Java Developers**. Version: Kepler Service Release 2. Build ID: 20140224-0627
- Compilador **Java SE Runtime Environment** 1.6.0_45
- **Llibreries JSON.simple** per a la lectura dels fitxers en aquest format proporcionats per l'API de The Guardian
- Ordinador amb **Windows 7** Home Premium. Service Pack 1 (64-bit)
- **Notepad++** v6.5.1 per a la lectura i l'edició d'arxius de text planer.
- **Microsoft Excel 2010** per al tractament de fitxers separats per comes i per al tractament d'estadístiques extretes del programa.

- Ordinador Acer Aspire 57733Z amb processador I**ntel Pentium P6200 @ 2.13GHz** i **4GB de RAM.**

3.2 Classes Java

En aquest apartat s'explicarà com l'algorisme s'ha implementat en Java fent un repàs de cadascuna de les classes que necessita per funcionar.,

3.2.1 Tag

És una de les classes més simples del sistema. Modela l'estructura d'una paraula clau i els seus constructors i mètodes per fixat i recuperar paràmetres. Com que l'API de 'The Guardian' dóna la paraula clau en aquest format: *section/keyword*, es pot veure *[Fig. 3b]* com el constructor separa aquests dos mots amb el mètode *split*:

```
public class Tag {
    private String section;
    private String keyword;
    private double rating;
    private double ratingN;
    private int count;

public Tag (String entrada){
    String results[] = entrada.split("/",2);
    section = results[0];
    if (results.length>1) keyword = results[1];
    rating=0;
    ratingN=0;
}
```

Fig. 3b

Aquesta classe té un mètode clone() *[Fig. 3c]* per tal de separar les referències dels *tags* relacionats amb un article i aquells lligats a un perfil. S'aconsegueix així que no interfereixin els *ratings* entre dos perfils diferents:

```
@Override
public Tag clone() {
    Tag t = new Tag();
    t.setKeyword(this.getKeyword());
    t.setSection(this.getSection());
    t.setCount(this.getCount());
    t.setRating(this.getRating());
    t.setRatingN(this.getRatingN());
    return t;
}
```

Fig. 3c

3.2.2 Section

Es tracta també d'una classe poc complexa, que serveix per representar una secció i compta amb una sèrie de constructors i mètodes per recuperar i fixar paràmetres *[Fig. 3d]*:

```
public class Section {
    private String name;
    private double rating;
    private double ratingN;
    private List<Tag> tags = new ArrayList<Tag>();
    private int count;

    public Section(){
        tags.clear();
    }

    public Section (String nom){
        name=nom;
        rating=0;
        ratingN=0;
        count=1;
        tags.clear();
    }
```

Fig. 3d

3.2.3 Article

Aquesta classe *[Fig. 3e]* modela l'estructura d'un article, amb la seva llista de *tags* corresponents:

```
import java.util.ArrayList;


public class Article {
    private String id;
    private String mainSection;
    private List<Tag> tags = new ArrayList<Tag>();

public Article (String idArticle, String mainSectionArticle) {
    id=idArticle;
    mainSection=mainSectionArticle;
}

public Article(){

}

public void setId(String idArticle){
    id=idArticle;
}
```

Fig. 3e

3.2.4 Corpus

És la classe que s'encarrega de descarregar-se notícies des de l'API de 'The Guardian' amb el mètode *readFromURL*. Per aconseguir-ho, s'utilitza un objecte JSONParser, que s'encarrega de llegir el fitxer JSON i carregar-lo dins de Java. Una vegada aconseguit, comptem amb un objecte JSONObject que ens permet recórrer el primer nivell de profunditat del document, que es correspon a aquestes línies:


```
{
  "response": {
    "status": "ok",
    "userTier": "approved",
    "total": 5834,
    "startIndex": 1,
    "pageSize": 10,
    "currentPage": 1,
    "pages": 584,
    "orderBy": "newest",
    "results": [
      {
        "id": "society/2014/jan/15/philippines-child-sexual-abuse-inquiry",
        "sectionId": "society",
        "sectionName": "Society",
    (...)
```


Amb això, podem saber quantes pàgines té la mostra agafada i, al mateix temps, podem recuperar l'atribut *results* que és on hi ha totes les notícies de la primera pàgina *[Fig. 3f]*:

```
public void readFromUrl(String dateIni, String dateEnd) throws Exception, IOException{

      JSONParser parser = new JSONParser();
      news.clear();
      try {

       int page = 1; int noticia=0;
       //Llegim la informació de l'API de The Guardian
       Object obj = parser.parse(readUrl("http://content.guardianapis.com/search?from-date="+dateIni+
       JSONObject jsonObject = (JSONObject) obj;
       JSONObject response = (JSONObject) jsonObject.get("response");
       //Mirem quantes pàgines ens retorna la resposta de l'URL
       Long numPaginesTemp = (Long) response.get("pages");
       int numPagines = numPaginesTemp.intValue();

       for (int j=0;j<numPagines;j++){ //Llegim pàgina a pàgina una mostra de notícies compresa entre

           JSONArray results = (JSONArray) response.get("results");
           Iterator<JSONObject> iterator = results.iterator();

           while (iterator.hasNext()) { //Anem llegint totes les notícies d'una pàgina

                System.out.println("N"+noticia);
                JSONObject noticiaActual = (JSONObject) iterator.next();
                String idNoticia = (String) noticiaActual.get("id");
                System.out.println("ID:"+ idNoticia);
```

Fig. 3f

Veiem que emmagatzemem results dins d'un objecte JSONArray i que després podem anar recorrent les notícies de la pàgina. Aquest mètode anirà iterant per totes les pàgines per tal d'obtenir a través d'Internet el corpus de notícies amb el qual treballarem.

El llistat de *tags* també es llegeix utilitzant una JSONArray. En aquest cas *[Fig. 3g]* es tracta en un tercer nivell de bucle, ja que aquesta llista és un atribut de cada notícia:

```
            System.out.print("TAGS: ");
            Iterator <JSONObject> iterator2 = tags.iterator();
            while (iterator2.hasNext()){
                JSONObject tagActual = (JSONObject) iterator2.next();
                String nomTag=(String)tagActual.get("id");
                System.out.print(nomTag+" ");
                articleTemp.addTag(nomTag);
            }
            news.add(articleTemp); //Carreguem l'article en memòria
            System.out.println();
            System.out.println("===========");
            noticia++;
        }
    //Canviem de pàgina
        if (page<numPagines){
         page++;
          obj = parser.parse(readUrl("http://content.guardianapis.com/search?from-date="
         jsonObject = (JSONObject) obj;
```

Fig. 3g

La classe Corpus té també mètodes per escriure el corpus i recuperar-lo d'un fitxer CSV. S'aconsegueix fent servir les classes FileWriter i BufferReader. Anem a veure un dels dos mètodes per fer-nos-en una idea *[Fig. 3h]*:

```
public void readFromCSV(String filename) throws IOException{

BufferedReader br = new BufferedReader(new FileReader(filename));
String text;
news.clear();

while ((text = br.readLine()) != null) {

    Article articleTemp = new Article();
    String resultat[] = text.split(",",3); //Separem camps per comes
    articleTemp.setId(resultat[0]);
    articleTemp.setMainSection(resultat[1]);
    String keywords[] = null;
    keywords = resultat[2].split(" "); //Separem els tags per espais

    int i=0;
    while (i<keywords.length) {
         //System.out.println(keywords[i]);
         articleTemp.addTag(keywords[i]);
        i++;
    }
    news.add(articleTemp);
}
```

Fig. 3h

Una vegada carregat un corpus d'articles, sigui via web o via CSV, se'ns mostra per pantalla el resultat en el següent format:

```
====================
NOTÍCIA 6194
SECCIÓ PRINCIPAL: crosswords/speedy/958
TAGS: crosswords/crosswords type/crossword crosswords/series/speedy
publication/theobserver
====================
NOTÍCIA 6195
SECCIÓ PRINCIPAL: crosswords/everyman/3513
TAGS: crosswords/crosswords crosswords/series/everyman type/crossword
publication/theobserver
====================
NOTÍCIA 6196
SECCIÓ PRINCIPAL: sport/2014/feb/01/france-women-england-women-six-nations
TAGS: sport/england-womens-rugby-union-team sport/rugby-union sport/sport
```

3.2.5 Profile

L'estructura de perfil, tal com hem insistit en els punts anteriors, es crea cada vegada en funció del corpus de notícies utilitzat amb un mètode que s'ha batejat amb el nom de *create*. Consisteix en fer un escaneig de totes les seccions i paraules clau que apareixen a la mostra.

El primer que es fa a l'hora de crear el perfil és recuperar el corpus de notícies llegit i es comença a navegar pel seu contingut a través del mètode *listIterator* que porten associat les llistes implementades en Java. El primer que es fa és navegar per totes les seccions principals i crear-les dins l'estructura de perfil. Llavors, cal navegar pels *tags*, que també porten seccions associades algunes de les quals poden no trobar-se a la llista de seccions principals.

Per cadascun dels *tags* de la notícia, cal mirar si la seva secció associada ja existeix dins l'estructura de perfil. En cas positiu, caldrà afegir el *tag* a la secció corresponent. Si no existeix, caldrà crear la secció amb el seu primer *tag* associat, aquell que estem tractant. Mentre es crea el perfil, també es va fent el recompte de aparicions que té cada secció i cada *tag*, amb la finalitat d'obtenir estadístiques.

Anem a veure *[Fig. 3i]* un fragment d'aquest mètode a la següent captura de pantalla:

```
while (lit.hasNext()){
    Article noticiaActual = (Article) lit.next();
    List<Tag> tagsActual = noticiaActual.getTags(); //Agafem els tags relacionats amb la notícia
    ListIterator <Tag> lit2 = tagsActual.listIterator();

    while (lit2.hasNext())  { //Recorrem tots els tags de la notícia
        Tag tagTemp = (Tag) lit2.next();
        //Anem a recuperar la secció a tractar de la llista 'data'
        ListIterator <Section> lit3 = data.listIterator();
        boolean notrobat=true;

        while ((lit3.hasNext())&&(notrobat==true)){
            seccioATractar= (Section) lit3.next();
            if (seccioATractar.getName().equals(tagTemp.getSection())){ //Recuperem la secció i deixem l'índex col·lo
                notrobat=false;
            }
        }
        if (notrobat==true) { //NO EXISTEIX SECCIÓ
            //Hem d'afegir la secció perquè no era una de les principals amb el seu primer tag
            Section novaSeccio2 = new Section();
            novaSeccio2.setName(tagTemp.getSection());
            tagTemp.addCount();
            novaSeccio2.addTag(tagTemp);
            novaSeccio2.addCount(); novaSeccio2.addCount(); //Un per la secció i un pel tag
            totalTags++;
            data.add(novaSeccio2);
            totalSections++;
            //System.out.println("NO EXISTEIX");

        }
        else { //JA EXISTEIX SECCIÓ
            //Ara hem de mirar si tagTemp existeix dins la llista de tags de seccioATractar
            seccioATractar.addCount();
            if (existeixTag(tagTemp, seccioATractar.getTags())==0){ //Si no existeix, l'afegim i actualitzem la secci
                tagTemp.addCount();
                seccioATractar.addTag(tagTemp);
```

Fig. 3i

Una vegada executat aquest mètode, podem veure per pantalla *[Fig. 3j]* l'estructura de perfil en blanc amb el mètode *print*. També existeix el mètode *printN* que ensenya les valoracions normalitzades.

```
18.- Nom secció: artanddesign/ Rating global: 0.0
Tags: photography (0.0) artanddesign (0.0) david-bailey (0.0) exhibition (0.0) art (0.0) pierre-auguste-renoir (0.0)
19.- Nom secció: film/ Rating global: 0.0
Tags: baftas (0.0) film (0.0) baftas-2014 (0.0) danny-boyle (0.0) berlin-film-festival-2014 (0.0) shia-labeouf (0.0)
20.- Nom secció: info/ Rating global: 0.0
Tags: developer-blog (0.0) series/guardian-and-observer-style-guide (0.0) digital-journalism-scheme-blog (0.0) info (
21.- Nom secció: public-leaders-network/ Rating global: 0.0
Tags: public-leaders-network (0.0) global-public-leaders-series (0.0) local-leadership (0.0) local-government (0.0) t
22.- Nom secció: local-government-network/ Rating global: 0.0
Tags: local-government-network (0.0) local-government-network-blog (0.0) local-government-careers (0.0) insight-engag
23.- Nom secció: higher-education-network/ Rating global: 0.0
Tags: higher-education-network (0.0) management-and-administration (0.0) series/academics-anonymous (0.0) blog (0.0)
```

Fig. 3j

El mètode *count* ens fa un recompte de les aparicions de cada secció i cada *tag* amb l'objectiu de tenir un coneixement més profund de la mostra de la qual estem parlant *[Fig 3k]* :

```
17.- Nom secció: theobserver/ Aparicions: 1043
Tags: sport/news (122) sport (122) news/uknews (60) news (147) new-review/discover (10) new-review (149) ne
18.- Nom secció: artanddesign/ Aparicions: 628
Tags: photography (106) artanddesign (165) david-bailey (6) exhibition (24) art (73) pierre-auguste-renoir
19.- Nom secció: film/ Aparicions: 897
Tags: baftas (8) film (263) baftas-2014 (7) danny-boyle (5) berlin-film-festival-2014 (10) shia-labeouf (6)
20.- Nom secció: info/ Aparicions: 19
Tags: developer-blog (3) series/guardian-and-observer-style-guide (10) digital-journalism-scheme-blog (1) i
21.- Nom secció: public-leaders-network/ Aparicions: 77
Tags: public-leaders-network (21) global-public-leaders-series (4) local-leadership (3) local-government (3
22.- Nom secció: local-government-network/ Aparicions: 77
Tags: local-government-network (25) local-government-network-blog (3) local-government-careers (2) insight-
23.- Nom secció: higher-education-network/ Aparicions: 122
```

Fig. 3k

Tal com s'ha comentat en la fase de disseny, hi ha unes seccions que es decidiran ignorar per no provocar biaixos i s'emmagatzemaran en una llista associada al perfil. Anem a veure com s'ha determinat quines han de ser, tant en el corpus1 com en el corpus2.

Veiem en la següent llista que algunes seccions principals ens aportaran poca informació en el tipus d'aprenentatge que estem plantejant. Les seccions marcades en groc les ignorarem. Per exemple, tipus i to defineixen la forma en què es presenta cada notícia i la secció profile té infinitat d'aparicions –sumatori de secció principal i secció de *tag*- perquè fa referència a actors que apareixen a les notícies. Se n'acumulen molts perquè de pro*tag*onistes dels articles n'hi ha milers de diferents. Les altres, s'ignoren també per motius similars.

Seccions que compten amb més de 1.000 aparicions al corpus1:

world	6283
type	5847
tone	5612
profile	4647
theguardian	4037
sport	3412
football	2402
publication	2326
uk	2032
business	2002
politics	1748
society	1719
culture	1640
media	1497
lifeandstyle	1368
technology	1259
music	1235
environment	1020

Detall d'alguns dels *tags* de les seccions ignorades al corpus1:

type/article	5203
type/video	203
type/gallery	177
type/picture	48
type/cartoon	36
type/crossword	33
tone/news	2281
tone/blog	881

tone/features	768	
tone/comment	571	
tone/reviews	258	
tone/matchreports	160	
profile/monkey	23	
profile/mustafa-khalili	2	
profile/nabeelah-shabbir	1	
profile/nancybankssmith	1	
profile/naomi-gryn	1	
profile/nataliebennett	1	
theguardian/mainsection		1254
theguardian/mainsection/uknews		315
theguardian/g2		222
theguardian/sport		203
theguardian/sport/news		200
theguardian/mainsection/international		156
publication/besttreatments	1	
publication/guardianweekly	92	
publication/theguardian	1819	
publication/theobserver	413	

Al corpus2 les seccions a ignorar gairebé són les mateixes:

world	6653
type	6439
tone	6114
profile	5005
theguardian	4067
sport	3912
publication	2411
football	2267
uk	2187
politics	1929
business	1890
lifeandstyle	1713
culture	1711
media	1709
society	1689
environment	1442
technology	1341
music	1183
books	1158
theobserver	1044

Veiem que també hi ha 'theobserver'

theobserver/new-review	149
theobserver/news	147
theobserver/sport	122
theobserver/sport/news	122
theobserver/new-review/critics	65
theobserver/news/uknews	60

També ignorarem aquesta secció en el procés d'aprenentatge. Al corpus1 tenia 822 aparicions i no aporta informació a la nostra tasca.

Per tant, la llista de seccions a ignorar quedarà de la següent manera *[Fig. 3l]*:

```
public class Profile {
    private Corpus noticies;
    private List<Section> data = new ArrayList<Section>();
    private List<String> toignore = Arrays.asList("type","tone", "profile", "theguardian", "publication", "theobserver");
```

Fig. 3l

Els mètodes deleteToIgnore i IHaveToIgnore s'encarregaran d'eliminar aquestes seccions del perfil.

addHits és el mètode que s'encarrega d'afegir valoracions a les seccions i als *tags* de l'aprenentatge online. Rep com a paràmetres una secció i un llistat de *tags*. En la primera part de l'algorisme *[Fig. 3m]* s'encarrega d'afegir hits a la secció principal de l'article:

```
public void addHits(String section, List<Tag> tags){

    //Afegim hits a la secció principal de l'article
    ListIterator<Section> iteradorSeccions = data.listIterator();
    while (iteradorSeccions.hasNext()){
        Section s = iteradorSeccions.next();
        if (s.getName().equals(section)) {
        //Hem trobat la secció
        double oldrating=s.getRating();
        s.setRating(oldrating+UNIT*ONLINE_MAINSECTION_RATE);
        iteradorSeccions.set(s);
```

Fig. 3m

La segona part de l'algorisme *[Fig. 3n]* és més complexa perquè per cadascun dels *tags* a tractar s'ha de buscar la seva secció corresponent i, llavors, recorre els *tags* del perfil un a un per trobar els que li interessen:

```
//Afegim hits a cada tag que apareix a l'article
ListIterator<Tag> iteradorTags = tags.listIterator();
while (iteradorTags.hasNext()){
    Tag tagACercar=iteradorTags.next();
    String nomSeccioTag=tagACercar.getSection();
    //System.out.println("NOM SECCIO TAG"+nomSeccioTag);
    ListIterator<Section> iteradorSeccions2=this.data.listIterator();
    boolean tagtrobat=false;
    while ((iteradorSeccions2.hasNext())){
        Section seccioATractar=iteradorSeccions2.next();
        //System.out.println("SECCIO A TRACTAR"+seccioATractar.getName());
            if (seccioATractar.getName().equals(nomSeccioTag)) {
            //Afegim la valoració que afegirem el tag també a la secció
                double oldrating=seccioATractar.getRating();
                seccioATractar.setRating(oldrating+UNIT*ONLINE_TAG_RATE);
                iteradorSeccions2.set(seccioATractar);

            //Hem trobat la secció, ara hem de trobar el tag
            List <Tag> llistaModificar = seccioATractar.getTags();
            if (existsTag(tagACercar,llistaModificar)==1) {
                int tagIndex=indexTag(tagACercar,llistaModificar);
                Tag tagTemp=llistaModificar.get(tagIndex);
                //System.out.println("son el mismo objecto? " + (tagTemp == tagACercar));
                double oldrating2=tagTemp.getRating();
                tagTemp.setRating(oldrating2+UNIT*ONLINE_TAG_RATE);
                llistaModificar.set(tagIndex,tagTemp);
                seccioATractar.setTags(llistaModificar);
            }
            }
    }
}
```

Fig. 3n

Per a l'aprenentatge offline s'han creat mètodes similars per poder actualitzar els pesos de cada secció i cada *tag*. addHits_Section és molt similar a la primera part d'addHits, amb la diferència que rep per paràmetre la quantitat de valoració a afegir, ja que en l'aprenentatge offline no és fixa *[Fig. 3o]*:

```
public void addHits_Section(String section, int units) {
    // TODO Auto-generated method stub

    //Afegim hits a la secció principal de l'article
    ListIterator<Section> iteradorSeccions = data.listIterator();
    while (iteradorSeccions.hasNext()){
        Section s = iteradorSeccions.next();
        if (s.getName().equals(section)) {
        //Hem trobat la secció
        double oldrating=s.getRating();
        System.out.println("Rating antic"+s.getRating()+" "+s.getRatingN());
        s.setRating(oldrating+(units*ONLINE_MAINSECTION_RATE));
```

Fig.3o

addHits_Tag es crida en l'aprenentatge offline, quan hi ha prou articles acumulats a la llista loved. Quan passa això, es miren quins són els deu *tags* més repetits de les seccions que es tenen en compte en aquest procés, aquelles que superen el 10% d'aparicions. En aquest mètode se li passa la secció a la qual pertanyen els *tags* i els seus noms en una estructura Counter. Aquest mètode va a buscar aquests *tags* a la secció corresponent del perfil i els augmenta la valoració *[Fig. 3p]*:

```
public void addhits_Tag(List<Counter> countTag, String section) {
    // TODO Auto-generated method stub
    ListIterator<Section> iteradorSeccions=data.listIterator();
    boolean trobada=false;
    System.out.println("entro addHits_Tag");
    while ((iteradorSeccions.hasNext())&&(!trobada)){

        Section s = iteradorSeccions.next();

        if (s.getName().equals(section)){

            trobada=true;
            List<Tag> tagsSeccio=s.getTags();
            ListIterator<Counter> iteradorContador = countTag.listIterator();
            while (iteradorContador.hasNext()){
                ListIterator<Tag> iteradorTags = tagsSeccio.listIterator();
                Counter c = iteradorContador.next();
                boolean trobat=false;
                while ((iteradorTags.hasNext())&&(!trobat)){
                    Tag t = iteradorTags.next();
                    if ((t.getKeyword().equals(c.getName()))){
                        trobat=true;
                        double oldrating=t.getRating();
                        System.out.println("Tag a tractar"+t.getKeyword());
                        System.out.println("Rating antic"+t.getRating()+" "+t.getRatingN());
                        t.setRating(oldrating+UP_LOVED);
                        System.out.println("add hits tag");
                        this.normalizeRatingTags();
                        System.out.println("Rating nou"+t.getRating()+" "+t.getRatingN());
                        iteradorTags.set(t);

                    }
            }
            }
        }
}
}
}
```

Fig. 3p

sustractHits_Section i sustractHits_Tag treballen de la mateixa manera però, en comptes de sumar, resten.

Com a mètodes secundaris tenim existsSection, existsTag, setData, getData, indexTag, normalizeRatingSections, fillRandomly, normalizeRatingTags, readFromCSV i writeToCSV.

3.2.6 Learner

Aquesta classe *[Fig. 3q]* és el nucli de tot l'algorisme i en el mètode run() veiem molt sintetitzada la seva funció:

```
profile.writeToCSV("evolutiu");
idealProfile.writeToCSV("ideal");
System.out.println("DISTÀNCIA sec: "+compareToIdeal_Section());
System.out.println("DISTÀNCIA tag: "+compareToIdeal_Tag());
onLineLearning();
System.out.println("DISTÀNCIA sec: "+compareToIdeal_Section());
System.out.println("DISTÀNCIA tag: "+compareToIdeal_Tag());
profile.writeToCSV("evolutiu2");
idealProfile.writeToCSV("ideal2");
```

Fig. 3q

Veiem que es calcula la distància entre seccions i entre *tags* dels perfils ideal i evolutiu abans i després de cridar l'onLineLearning. D'aquesta manera es mesura l'eficàcia de l'algorisme. Aquests dos mètodes es basen en la distància euclidiana tal com s'ha detallat a la fase d'anàlisi i disseny. Aquesta classe compta amb els valors amb els quals es pot configurar l'aprenentatge de manera parametritzable *[Fig 3r]* :

```
private int MIDA_MAX_GRUP=15; //Nombre d'articles a tractar a cada iteració
private double ALFA=0.6; //Coeficient valoració tags
private double BETA=0.4; //Coeficient valoració secció

private int OVER_RANKED_SIGNIFICATIU=100;
private int LOVED_SIGNIFICATIU=100;
private int PERCENTATGE_SUPERAR=10;
private int NUM_TAGS_OFFLINE=10;

private int UP_LOVED=5;
private int DOWN_OVER_RANKED=5;
```

Fig. 3r

Anem-nos a fixar ara en l'onLineLearning, que comença agafant el llistat de notícies de cada iteració:

```
    for (int i=0; i<(corpus.getList().size()/MIDA_MAX_GRUP+1);i++)
{
        //Seleccionem el grup de notícies de la iteració
        System.out.println("ITERACIÓ"+(i+1));
        selected.clear(); ratingIdeal.clear(); ratingEvolutive.clear();
        int j =index;
        boolean nofinal=true;



        while ((j<index+MIDA_MAX_GRUP)&&(nofinal)){
            if (j<corpus.getList().size()) {
                selected.add(corpus.getList().get(j));
            }
            else {
                nofinal=false;
                //System.out.println("entro aquí");
            }

            j++;
        }
        //System.out.println(selected.size());
    index=index+MIDA_MAX_GRUP;
```

Fig. 3s

Una vegada seleccionades, calcula el *rating* de cada notícia en funció del perfil evolutiu i el perfil ideal *[Fig. 3t]* :

```
ListIterator<Article> iterator1 = selected.listIterator();
int counter=0;
while (iterator1.hasNext()){ //Calculem els ratings de tots els seleccionats
    double rating1, rating2;

    Article a = iterator1.next();
    rating1=rate(a,idealProfile);
    rating2=rate(a, profile);
    Rating r1 = new Rating();
    Rating r2=new Rating();
    r1.setIndex(counter);
    r1.setValue(rating1);
    r2.setIndex(counter);
    r2.setValue(rating2);
    ratingIdeal.add(r1);
    ratingEvolutive.add(r2);
    counter++;
}

Collections.sort(ratingIdeal);
Collections.sort(ratingEvolutive);

int escollidaEvolutive=ratingEvolutive.get(0).getIndex();
int escollidaIdeal=ratingIdeal.get(0).getIndex();
//int escollidaIdeal=3; // per proves

System.out.println("Escollida evolutive: "+escollidaEvolutive);
System.out.println("Escollida ideal: "+escollidaIdeal);
```

Fig. 3t

Veiem que en aquest punt *[Fig. 3u]* fa una crida al mètode rate, que implementa la fórmula presentada en la fase anterior del projecte:

```
public double rate (Article a, Profile p){
    double ratingSeccio=0;
    double ratingTags=0;
    boolean trobat=false;

    String mainSection=a.getMainSection(); //Farem servir per calcular la valoració que depèn de BETA
    List<Tag> tags=a.getTags(); //Farem servir per calcular la valoració que depèn d'ALFA

    //Recorrem tot el perfil
    ListIterator<Section> iterator1 = p.getData().listIterator();

    //Calculem valoració que depèn de BETA
    while ((iterator1.hasNext())&&(!trobat)) {
        Section s = iterator1.next();
        if (s.getName().equals(mainSection)) {
            trobat=true;
            ratingSeccio=s.getRatingN(); //Agafem el rating normalitzat
        }
    }
    //Calculem valoració que depèn d'ALFA
    ListIterator<Tag> iterator2=tags.listIterator();
    int numTagsATrobar=tags.size();
    int numTagsTrobats=0;

    boolean totstrobats=false;

    while ((iterator2.hasNext()&&(!totstrobats))){
        Tag t = iterator2.next();
        ListIterator<Section> iterator3=p.getData().listIterator();
        while (iterator3.hasNext()){
            Section s = iterator3.next();
            List<Tag> tagsATractar = s.getTags();
            if (existsTag(tagsATractar, t)!=0) {
                numTagsTrobats++;
                ratingTags=ratingTags+existsTag(tagsATractar,t);
            }
```

Fig. 3u

Tornant a l'onLineLearning, veiem *[Fig. 3v]* la resta del codi i com s'acaba cridant a addHits, mètode que ja s'ha explicat en l'apartat de la classe Profile:

```
if (escollidaEvolutive==escollidaIdeal){
    //System.out.println("entro");
    //No tenim res a aprendre
    String liniaAEscriure=(i+1)+"; "+"1; "+overranked.size()+"; "+loved.size()+"\n";
    writer.write(liniaAEscriure);

    Article articleSeleccionat = selected.get(escollidaIdeal);
    loved.add(articleSeleccionat);

}
else{
    //Augmentem hits a les paraules clau presents a la notícia en una unitat / main section
    Article articleSeleccionat = selected.get(escollidaIdeal);

    String mainSectionArticleSeleccionat = articleSeleccionat.getMainSection();
    List<Tag> tagsArticleSeleccionat = articleSeleccionat.getTags();

    String liniaAEscriure=(i+1)+"; "+"0; "+overranked.size()+"; "+loved.size()+"\n";
    writer.write(liniaAEscriure);

    profile.addHits(mainSectionArticleSeleccionat, tagsArticleSeleccionat);


    //Fet a addHits: Recalculem les valoracions de les seccions. Hi afegim el sumatori de hits paraules clau
    //Posem en quarentena les noticies over-ranked i guardem en un històric (loved) la notícia seleccionada,
    ListIterator<Rating> listtes2t = ratingEvolutive.listIterator();
    boolean trobat=false;
    while ((listtes2t.hasNext())&&(!trobat)){
        Rating r = listtes2t.next();
        int indexATractar=r.getIndex();
        if (indexATractar!=escollidaIdeal){
            overranked.add(selected.get(indexATractar));
        }
        else {

            trobat=true;
        }
```

Fig. 3v

Fins aquí acaba estrictament l'onLineLearning, però un cop arribat a aquest punt *[Fig. 3w]* ell mateix mira si cal entrar a l'offLineLearning:

```
//Mirar si nombre significatiu de over-rated

if (overranked.size()>=OVER_RANKED_SIGNIFICATIU){

    //Cridar offline learning (paràmetre overrated)
    System.out.println("Entro offline learning over ranked");
    offLineLearning(1);
}



//Mirar si hi ha nombre significatiu de loved

if (loved.size()>=LOVED_SIGNIFICATIU){
    //Cridar offline learning (paràmetre loved)
    System.out.println("Entro offline learning loved");
    offLineLearning(0);

}
```

Fig. 3w

Per acabar, anem a veure *[Fig. 3x]* la implementació de la classe offLineLearning. Ens fixarem en com es gestionen les notícies overranked (mode 1 com a paràmetre) i les

loved (mode 0 com a paràmetre). Com que funcionen de manera anàloga i una resta i l'altra suma, només ens centrem en el detall d'una d'elles:

```
case 0: //loved
    //System.out.println("Entro loved");
    ListIterator<Article> iteratorOR = loved.listIterator();
    while (iteratorOR.hasNext()){
        Article a = iteratorOR.next();
        if (existsInCounter(countSection,a.getMainSection())) {
            //El mètode existsInCounter ja ha afegit una aparició al contador
        }
        else{
            Counter c = new Counter();
            c.setName(a.getMainSection());
            c.addCount();
            countSection.add(c);
        }

    }
```

Fig. 3x

Primer de tot, com hem pogut veure, recorre tots els articles acumulats a la llista loved i va elaborant un recompte d'aparicions de cada secció. Llavors, es queda només amb aquelles que superen el PERCENTATGE_SUPERAR, que es recorren i s'invoca addHits_Section *[Fig. 3y]*

```
ListIterator<Counter> iteradorCounter = countSection.listIterator();
while (iteradorCounter.hasNext()){
    Counter c = iteradorCounter.next();
    if (c.getCount()<PERCENTATGE_SUPERAR) {
        iteradorCounter.remove();
    }
}

Collections.sort(countSection);
iteradorCounter = countSection.listIterator();
while (iteradorCounter.hasNext()){
    Counter c = iteradorCounter.next();
    System.out.println("CONTADOR SECCIÓ "+c.getName()+" "+c.getCount());
    String nomSeccio=c.getName();
    int aparicions=c.getCount();
    profile.addHits_Section(nomSeccio,aparicions);
    //profile.normalizeRatingSections();
}
```

Fig. 3y

Llavors, de les seccions que superen el llindar, n'agafem els 10 *tags* més repetits, fent servir un objecte Counter *[Fig. 3z]*

```
//De les seccions loved que superen el llindar PERCENTATGE_SUPERAR, n'agafem els tags més repetits
System.out.println("Nombre de seccions a tractar"+countSection.size());
iteradorCounter = countSection.listIterator();
while (iteradorCounter.hasNext()){ //D'una secció en concret

    iteratorOR = loved.listIterator();
    String seccioATractar=iteradorCounter.next().getName();
    countTag.clear();
    System.out.println("****Secció a tractar: "+seccioATractar);
    while (iteratorOR.hasNext()){

        Article a = iteratorOR.next();
        //Si pertany a la secció a tractar
        if (seccioATractar.equals(a.getMainSection())){ //Hem de recórrer els seus tags
            List<Tag> tagsArticle=a.getTags();
            //Eliminar aquells tags que no són d'aquesta secció
            ListIterator<Tag> tagsArticleIterator = tagsArticle.listIterator();
            while (tagsArticleIterator.hasNext()){
                Tag t = tagsArticleIterator.next();

                if (t.getSection().equals(seccioATractar)) {

                    if (existsInCounter(countTag,/*t.getSection()+"/"+*/t.getKeyword())){
                        //El mètode existsInCounter ja ha afegit una aparició al contador
                        //System.out.println("ja existeix tag");
                    }
                    else
                    {
                        Counter c = new Counter();
                        c.setName(/*t.getSection()+"/"+*/t.getKeyword());
                        //System.out.println("Afegeixo tag: "+t.getSection()+"/"+t.getKeyword());
                        c.addCount();
                        countTag.add(c);
                    }
                }
            }

    }
    }
```

Fig. 3z

Finalment, ens quedem amb els 10 més repetits i acabem cridant addHits_Tag *[Fig. 3a']*

```
//Tenim counttag amb tota info q ens interessa
    //Llista
    ListIterator<Counter> iteradorCounter2 = countTag.listIterator();
    Collections.sort(countTag);

    int llindar=0;

    //Deixem a countTag només esl 10 primers, que són als que volem disminuir la valoració
    while (iteradorCounter2.hasNext()){
        Counter c = iteradorCounter2.next();
        //System.out.println("Tag: "+c.getName()+" "+c.getCount());
        if (llindar>=NUM_TAGS_OFFLINE) iteradorCounter2.remove();
        llindar++;
    }

    iteradorCounter2 = countTag.listIterator();

    while (iteradorCounter2.hasNext()){
        Counter c = iteradorCounter2.next();
        System.out.println("Tag: "+c.getName()+" "+c.getCount());

    }

profile.addhits_Tag(countTag, seccioATractar);
```

Fig. 3a'

Aquesta classe també compte amb els mètodes auxiliars existsinCounter, existsTag, getCorpus, getIdealProfile, getProfile, setCorpus, setIdealProfile i setProfile.

3.2.7 NewsRecommender

A la classe News Recommender és on trobem el main() del programa i el menú que ens permet escollir què volem fer *[Fig 3b', 3c']*. Convé notar que és una interfície pensada per provar el funcionament de l'algorisme i no dirigida a un usuari final:

```
public class NewsRecommender {

    public static void showMenu(){
        System.out.println("========================================================================");
        System.out.println("                          NEWS RECOMMENDER                              ");
        System.out.println("========================================================================");
        System.out.println("1.  Descarregar on-line corpus de notícies de The Guardian");
        System.out.println("2.  Guardar corpus de notícies carregat en un fitxer CSV");
        System.out.println("3.  Carregar un corpus de notícies des de CSV");
        System.out.println("4. Emmagatzemar perfil a un CSV");
        System.out.println("5. Carregar perfil de CSV");
        System.out.println("6.  Crear estructura de perfil");
        System.out.println("7.  Obtenir estadístiques del corpus");
        System.out.println("8.  Crear un perfil ideal amb dades aleatòries");
        System.out.println("9. Donar pes a seccions i tags del perfil ideal - Boost");  //Excel
        System.out.println("9.  Crear perfil aprenentatge");
        System.out.println("10. Executar algorisme aprenentatge");
        System.out.println("11. Sortir");
        System.out.println("");

    }

    public static void main(String[] args) {
        // TODO Auto-generated method stub
        Learner learner = new Learner();
        Corpus corpus = new Corpus();
        Profile profile = new Profile();
        Profile idealProfile = new Profile();
```

Fig. 3b'

```
case 4:
    System.out.println("Escull una opció:");
    System.out.println("12. Emmagatzemar el perfil evolutiu");
    System.out.println("13. Emmagatzemar el perfil ideal");
    Scanner in2 = new Scanner(System.in);
       int i2 = in2.nextInt();
    switch (i2){
    case 12:
        System.out.println("Amb quin nom vols guardar aquest perfil evolutiu? Sense extensió");
        sc=new Scanner(System.in);
        name=sc.nextLine();
        profile.writeToCSV(name+".csv");

        break;
    case 13:
        System.out.println("Amb quin nom vols guardar aquest perfil ideal? Sense extensió");
        sc=new Scanner(System.in);
        name=sc.nextLine();
        idealProfile.writeToCSV(name+".csv");
        break;
    }
```

Fig. 3c'

3.3 Comentaris

La implementació de l'algorisme ha estat bastant complexa i, de fet, s'ha prolongat més en la planificació del que estava previst i, probablement, per la mancança d'un coneixement profund del llenguatge de programació utilitzat, fet que ha obligat a combinar la programació amb l'aprenentatge d'alguns aspectes tècnics.

No obstant això, la valoració d'aquesta fase és positiva ja que s'ha aconseguit materialitzar tot allò que es pretenia a la fase d'anàlisi i disseny, fins i tot l'aprenentatge offline a través de *tags*, que era la part més complicada de modelar.

El problema principal de la implementació va ser una interferència entre els *tags* de dos perfils, un bug que va fer perdre força hores de feina. Finalment, es va descobrir que als perfils s'agafaven les mateixes referències que tenien els *tags* dels articles i, per aquest motiu, en modificar la valoració d'un *tag* en un perfil, també es modificava en l'altre. Es tractava de la mateixa instància de l'objecte. Es va solucionar afegint un mètode clone() a la classe Tag.

A tall de conclusió, cal dir que s'ha aconseguit un producte prou robust com per passar amb èxit a la fase d'estudi de diferent supòsits i de conclusions generals i planificació de la continuïtat i aplicació del projecte.

4 Estudi de diversos supòsits

En aquest apartat s'analitzarà l'eficàcia, de l'algorisme. A través de diverses proves es podran arribar a conèixer les seves fortaleses i debilitats principals. Per tal de dur a terme els diversos tests s'ha creat un repositori de perfils que es faran servir per provar l'aprenentatge a partir de diversos supòsits.

4.1 Càrrega de les dades

Anem a veure com és la interacció amb el programa a l'hora de carregar un corpus de notícies. Hem de fer clic a l'opció 1 del menú del programa. Com ja hem comentat, ens descarregarem les notícies de The Guardian de la primera quinzena de gener i la primera quinzena de febrer per comptar amb dos corpus d'articles. Veiem que el programa ens va descarregant online totes les notícies una a una i les carrega en una estructura dinàmica Corpus:

```
================================================================
                    NEWS RECOMMENDER
================================================================
1.  Descarregar online corpus de notícies de The Guardian
2.  Guardar corpus de notícies carregat en un fitxer CSV
3.  Carregar un corpus de notícies des de CSV
4. Emmagatzemar perfila un CSV
5. Carregar perfil de CSV
6.  Crear estructura de perfil
7.  Obtenir estadístiques del corpus
8.  Crear un perfil ideal amb dades aleatòries
9. Donar pes a seccions i tags del perfil ideal - Boost
9.  Crear perfil aprenentatge
10. Executar algorisme aprenentatge

Escull una opció:
1
Data d'inici de la mostra AAAA/MM/DD:
2014-01-01
Data final de la mostra  AAAA/MM/DD:
2014-01-15
N0
ID:society/2014/jan/15/philippines-child-sexual-abuse-inquiry
SECTION:society
TAGS: society/childprotection society/children society/social-care society/society
world/philippines world/asia-pacific world/world uk/police uk/uk tone/news profile/conalurquhart
type/article
===========
N1
ID:football/blog/2014/jan/15/manuel-pellegrini-manchester-city-quadruple-blackburn
SECTION:football
```

```
TAGS: football/manuel-pellegrini football/manchestercity football/blackburn football/jose-
mourinho football/football sport/sport sport/blog profile/paulwilson tone/comment
theguardian/sport/news theguardian/sport type/article publication/theguardian
==========
N2
ID:culture/australia-culture-blog/2014/jan/16/oedipus-schmoedipus-review
SECTION:culture
TAGS: culture/sydney-festival-2014 culture/culture stage/theatre culture/australia-culture-blog
type/article tone/reviews profile/vickyfrost
==========
N3
ID:football/2014/jan/15/cristiano-ronaldo-real-madrid-copa-del-rey
SECTION:football
TAGS: football/ronaldo football/realmadrid football/copa-del-rey football/football sport/sport
tone/news type/article
==========
N4
ID:commentisfree/2014/jan/15/2004-republican-convention-protests-new-york-witness
SECTION:commentisfree
TAGS: commentisfree/commentisfree world/george-bush world/usa world/protest world/new-
york world/nypd law/us-constitution-and-civil-liberties profile/j-iddhis-bing tone/comment
type/article

(...)

5833 articles carregats al sistema
==========
```

Una vegada repetit aquest procés per obtenir el corpus1 i el corpus2, seleccionem l'opció 2 del menú per tal d'emmagatzemar-los en fitxers CSV i així ja els tenim en local i no caldrà anar-los descarregant cada vegada, que triga una estona:

```
Escull una opció:
2
Escriu el nom amb el qual vols emmagatzemar el corpus, sense extensió:
corpus1
El corpus ha estat guardat amb el nom de corpus1.csv al directori arrel del programa
```

Amb l'opció 3, podrem carregar el corpus des del fitxer CSV que acabem de guardar:

```
============================================================
                NEWS RECOMMENDER
============================================================
1.  Descarregar online corpus de notícies de The Guardian
2.  Guardar corpus de notícies carregat en un fitxer CSV
3.  Carregar un corpus de notícies des de CSV
4. Emmagatzemar perfila un CSV
5. Carregar perfil de CSV
6.  Crear estructura de perfil
7.  Obtenir estadístiques del corpus
8.  Crear un perfil ideal amb dades aleatòries
9. Donar pes a seccions i tags del perfil ideal - Boost
9.  Crear perfil aprenentatge
10. Executar algorisme aprenentatge
```

```
11. Sortir

Escull una opció:
3
Escriu el nom del fitxer del corpus que vols importar, sense extensió:
corpus1
Corpus carregat satisfactòriament. Prem una tecla per veure'l

(...)

====================
NOTÍCIA 5833 lifeandstyle/2014/jan/01/sudoku-2696-medium
SECCIÓ PRINCIPAL: lifeandstyle
TAGS: lifeandstyle/series/sudoku theguardian/g2/puzzles theguardian/g2
lifeandstyle/series/sudoku-medium type/sudoku lifeandstyle/lifeandstyle publication/theguardian
====================
NOTÍCIA 5834 politics/2014/jan/01/david-cameron-scottish-independence-new-years-message
SECCIÓ PRINCIPAL: politics
TAGS: politics/davidcameron politics/scottish-independence uk/uk politics/scotland uk/scotland
politics/politics politics/economy society/public-sector-cuts society/policy society/public-finance
society/society politics/welfare uk/immigration tone/news profile/rowena-mason
theguardian/mainsection/uknews theguardian/mainsection type/article publication/theguardian
====================
NOTÍCIA 5835 crosswords/cryptic/26145
SECCIÓ PRINCIPAL: crosswords
TAGS: crosswords/crosswords type/crossword profile/brummie crosswords/series/cryptic
publication/theguardian
====================
NOTÍCIA 5836 crosswords/quick/13618
SECCIÓ PRINCIPAL: crosswords
TAGS: crosswords/crosswords crosswords/series/quick type/crossword publication/theguardian
```

4.2 Crear estructura de perfil

Un cop vist com es carreguen els corpus de notícies, anem a crear l'estructura de perfil d'un corpus en concret amb l'opció 6 del programa:

```
================================================================
                    NEWS RECOMMENDER
================================================================
1.  Descarregar online corpus de notícies de The Guardian
2.  Guardar corpus de notícies carregat en un fitxer CSV
3.  Carregar un corpus de notícies des de CSV
4. Emmagatzemar perfila un CSV
5. Carregar perfil de CSV
6.  Crear estructura de perfil
7.  Obtenir estadístiques del corpus
8.  Crear un perfil ideal amb dades aleatòries
9. Donar pes a seccions i tags del perfil ideal - Boost
9.  Crear perfil aprenentatge
10. Executar algorisme aprenentatge
11. Sortir

Escull una opció:
6
```

```
Estructura de perfil creada satisfactòriament. Prem una tecla per veure-la

(...)

100.- Nom secció: partner-zone-path/ Rating global: 0.0
Tags: partner-zone-path (0.0)
101.- Nom secció: social-care-network-skills-for-care-partner-zone/ Rating global: 0.0
Tags: social-care-network-skills-for-care-partner-zone (0.0)
102.- Nom secció: teacher-network-hays-partner-zone/ Rating global: 0.0
Tags: teacher-network-hays-partner-zone (0.0)
103.- Nom secció: direct-line-for-business-partner-zone/ Rating global: 0.0
Tags: direct-line-for-business-partner-zone (0.0)
104.- Nom secció: teacher-network-advertisement-features/ Rating global: 0.0
Tags: teacher-network-advertisement-features (0.0)
105.- Nom secció: housing-network-partner-zone-pinnacle/ Rating global: 0.0
Tags: housing-network-partner-zone-pinnacle (0.0)
106.- Nom secció: sustainable-business-fairtrade-partner-zone/ Rating global: 0.0
Tags: sustainable-business-fairtrade-partner-zone (0.0)
107.- Nom secció: partner-zone-sas-computacenter/ Rating global: 0.0
Tags: partner-zone-sas-computacenter (0.0)
108.- Nom secció: adam-smith-international-partner-zone/ Rating global: 0.0
Tags: adam-smith-international-partner-zone (0.0)
109.- Nom secció: help/ Rating global: 0.0
Tags: help (0.0)
```

Veiem que s'han creat totes les seccions amb els seus *tags* corresponents i que ens ha deixat totes les valoracions inicialitzades a zero. Amb l'opció 7 del programa podem analitzar les seccions i els *tags* del perfil, ja que ens creen dos fitxers amb el recompte d'aquests objectes:

```
Escull una opció:
7
Prem una tecla per carregar les estadístiques d'aparicions de seccions i tags a l'estructura de
perfil
(...)

100.- Nom secció: partner-zone-path/ Aparicions: 2
Tags: partner-zone-path (1)
101.- Nom secció: social-care-network-skills-for-care-partner-zone/ Aparicions: 3
Tags: social-care-network-skills-for-care-partner-zone (2)
102.- Nom secció: teacher-network-hays-partner-zone/ Aparicions: 3
Tags: teacher-network-hays-partner-zone (2)
103.- Nom secció: direct-line-for-business-partner-zone/ Aparicions: 2
Tags: direct-line-for-business-partner-zone (1)
104.- Nom secció: teacher-network-advertisement-features/ Aparicions: 2
Tags: teacher-network-advertisement-features (1)
105.- Nom secció: housing-network-partner-zone-pinnacle/ Aparicions: 2
Tags: housing-network-partner-zone-pinnacle (1)
106.- Nom secció: sustainable-business-fairtrade-partner-zone/ Aparicions: 2
Tags: sustainable-business-fairtrade-partner-zone (1)
107.- Nom secció: partner-zone-sas-computacenter/ Aparicions: 2
Tags: partner-zone-sas-computacenter (1)
108.- Nom secció: adam-smith-international-partner-zone/ Aparicions: 2
Tags: adam-smith-international-partner-zone (1)
109.- Nom secció: help/ Aparicions: 2
```

```
Tags: help (1)
============================
TOTAL SECCIONS: 110
TOTAL TAGS: 4303
Les dades s'han emmagatzemat als fitxers seccions_count.csv i tags_count.csv
```

Anem a analitzar les dades d'aquests dos fitxers per veure de quines mostres disposa l'estructura de perfil creada amb el corpus1 i la creada amb el corpus2

Aquest és el pes de les seccions del corpus1 mostrat de manera gràfica *[Fig. 4a]*

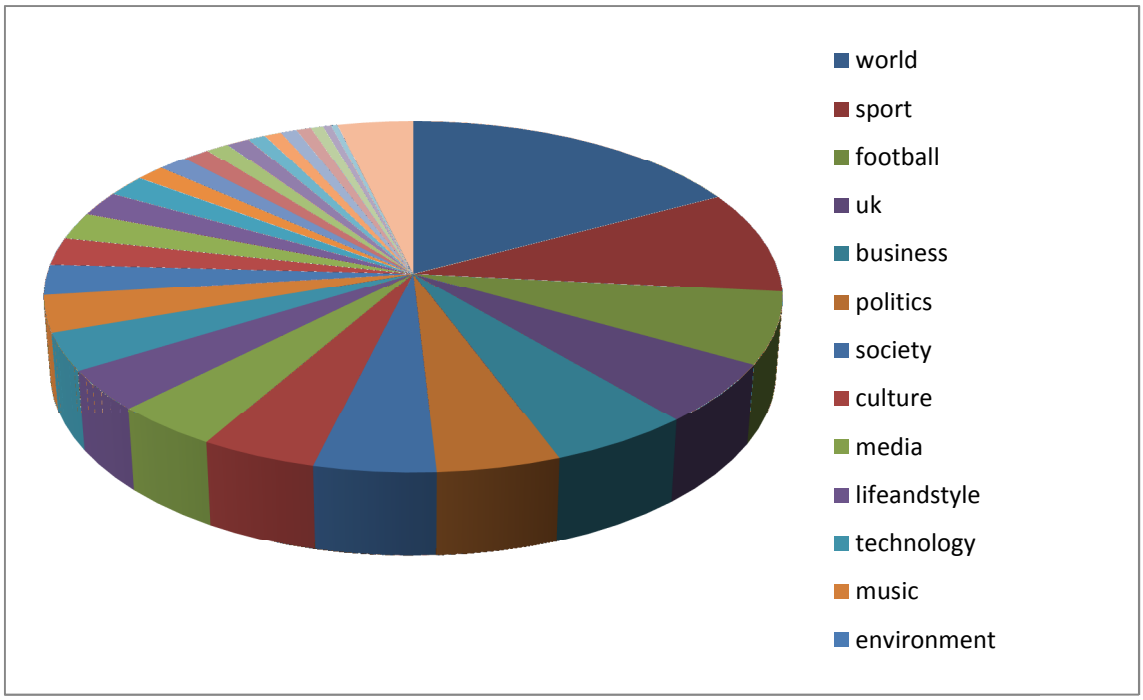


Fig. 4a

En números, tenim aquí la relació de seccions i nombre d'aparicions. Al final s'ha afegit 'other' per encabir la resta de seccions, que es poden consultar de manera completa al fitxer s*eccions_count_corpus1_ordenat* que trobem a la carpeta 'ranking_docs' del projecte:

world	6283
sport	3412
football	2402
uk	2032
business	2002
politics	1748
society	1719
culture	1640
media	1497
lifeandstyle	1368
technology	1259

music	1235
environment	1020
books	917
film	901
theobserver	822
money	723
education	524
commentisfree	522
tv-and-radio	498
science	417
artanddesign	414
stage	323
travel	307
fashion	295
law	271
global-development	224
sustainable-business	145
small-business-network	116
other	1372

Les estadístiques de l'estructura de perfil del corpus2 *[Fig. 4b]* són bastant similars (seccions_count_corpus2_ordenat):

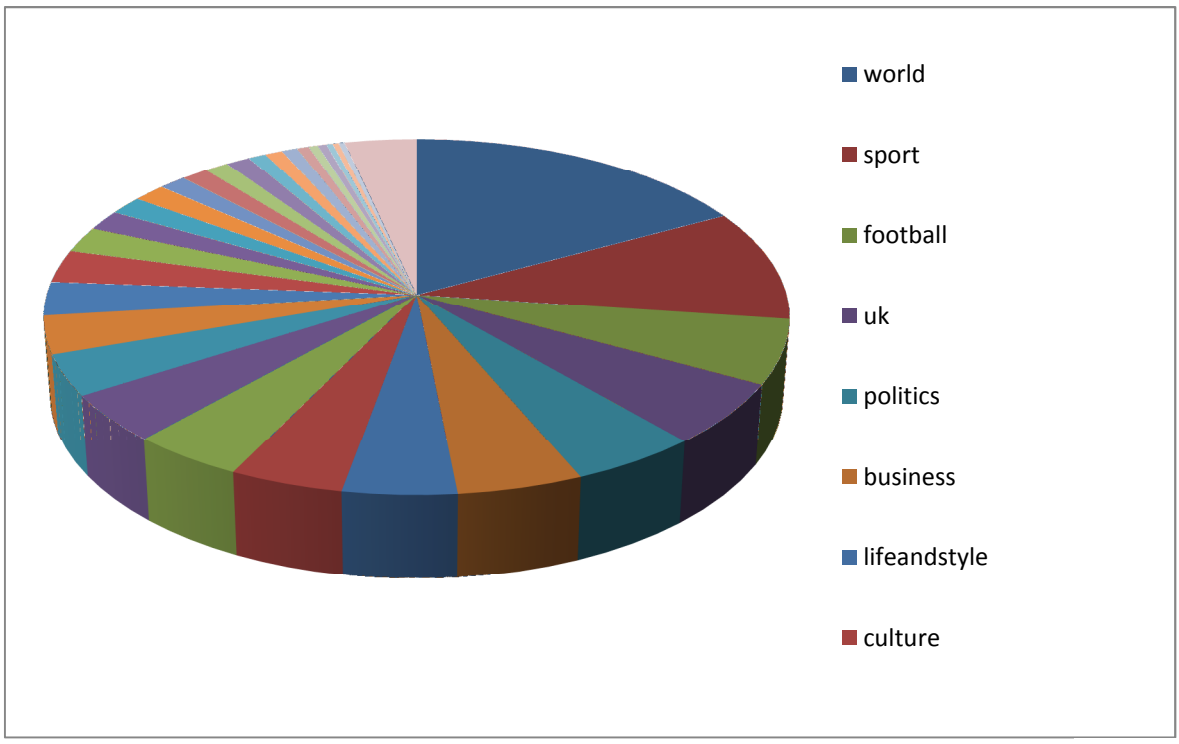


Fig. 4b

world	6653
sport	3912
football	2267

uk	2187
politics	1929
business	1890
lifeandstyle	1713
culture	1711
media	1709
society	1689
environment	1442
technology	1341
music	1183
books	1158
film	898
education	704
money	649
artanddesign	629
science	526
commentisfree	507
stage	461
tv-and-radio	460
travel	339
fashion	338
law	300
global-development	225
small-business-network	181
sustainable-business	162
higher-education-network	123
global-development-professionals-network	119
media-network	112
other	1369

Veiem que tant en el corpus1 (primera quinzena de gener) com en el corpus2 (primera quinzena de febrer) les seccions que presenten més notícies són world, sport, football i uk. Veiem, comparant els dos casos, que capgiren les seves posicions business i politics i que, per exemple, lifeandstyle té més aparicions al febrer que no pas al gener.

En els fitxers *tags*_count_corpus1_ordenat i *tags*_count_corpus2_ordenat s'han recollit també els rànquings d'aparició de cada *tag*.

4.3 Inicialització d'un perfil ideal

Per inicialitzar un perfil ideal, cal seleccionar l'opció 8 que ens permet omplir-lo amb dades aleatòries i, llavors, guardar-lo en un fitxer CSV.

```
================================================================
                        NEWS RECOMMENDER
================================================================
1.  Descarregar online corpus de notícies de The Guardian
2.  Guardar corpus de notícies carregat en un fitxer CSV
3.  Carregar un corpus de notícies des de CSV
4. Emmagatzemar perfil a un CSV
5. Carregar perfil de CSV
6.  Crear estructura de perfil
7.  Obtenir estadístiques del corpus
8.  Crear un perfil ideal amb dades aleatòries
9.  Crear perfil aprenentatge
10. Executar algorisme aprenentatge
Escull una opció:
8
S'ha creat un perfil ideal amb dades aleatòries
```

4.4 Estudi dels valors dels paràmetres

Com s'ha anat comentant en les seccions anteriors, l'algorisme compta amb diversos valors que es poden parametritzar. En aquesta secció es pretén fer una justificació dels valors que s'han escollit per defecte i provar si altres combinacions possibles ens ofereixen millors resultats. Anem a analitzar doncs, els següents paràmetres:

- ALFA i BETA
- OVER_RANKED-SIGNIFICATIU / LOVED_SIGNIFICATIU
- PERCENTATGE_SUPERAR
- NUM_TAGS_OFFLINE
- UP_LOVED / DOWN_OVER_RANKED

La variació d'un d'aquests paràmetres en les següents proves suposarà la congelació de tots als altres en els seus valors originals, excepte en el cas que es descobreixi que un altre valor funciona millor i es decideixi establir com a valor predeterminat.

4.4.1 ALFA i BETA

Recordem que els valors de BETA i ALFA ens servien, per una banda, per donar pes a la valoració d'una notícia d'acord amb la seva secció principal i, per l'altra, d'acord amb els *tags* que la representaven:

$$\textit{rating} = \alpha * \text{valoració\_}\textit{tags} + \beta * \text{valoració\_secció}$$

Inicialment es va pensar que els *tags* donaven més informació sobre el *rating* de la notícia perquè n'hi havia més quantitat i la descrivien de manera més precisa, motiu pel qual se li va donar un 20% més de pes. En aquest apartat es variaran aquests valors per veure com es comporten els resultats en combinacions diferents a la inicial. Veurem que descobrirem que l'escenari B dóna millors resultats que l'inicial. Les proves que es duran a terme són les següents:

	Alfa (*tags*)	**Beta (secció principal)**
Escenari A	0,6	0,4
Escenari B	0,5	0,5
Escenari C	0,4	0,6

Escenari A *(ALFA=0,6; BETA=0,4)*

Els valors de les distàncies entre seccions i *tags* a l'inici i al final de l'execució de l'algorisme són els següents.

	Distància entre seccions	**Distància entre grups de *tags***
Inici	2.504180081383925	21.647480034974407
Final	1.3410420778192813	18.900561177553882

La corba d'aprenentatge que obtenim és la següent *[Fig. 4c]* :

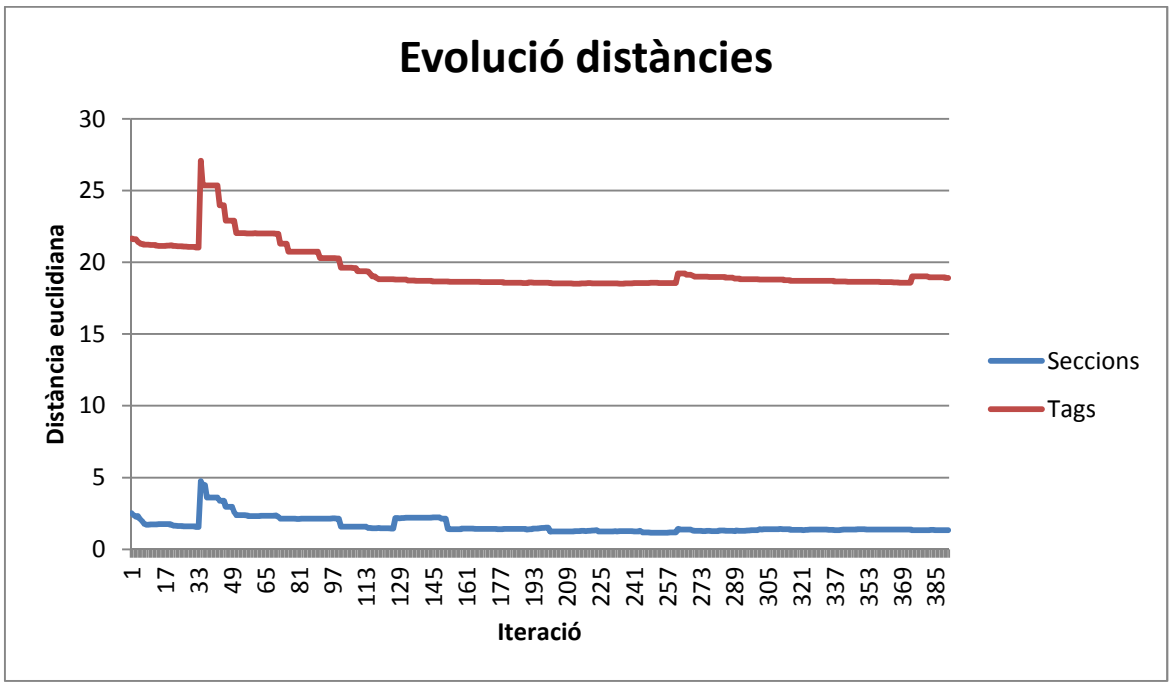


Fig. 4c

A banda d'un petit desajustament inicial i alguna irregularitat molt puntual, veiem que les distàncies, tant entre seccions com entre *tags*, tendeixen a reduir-se.

Escenari B *(ALFA=0,5; BETA=0,5)*

Els valors de les distàncies entre seccions i *tags* a l'inici i al final de l'execució de l'algorisme són els següents.

	Distància entre seccions	Distància entre grups de *tags*
Inici	2.504180081383925	21.647480034974407
Final	1.295941043290216	18.493673799014562

La corba d'aprenentatge que obtenim és la següent *[Fig 4d]* :

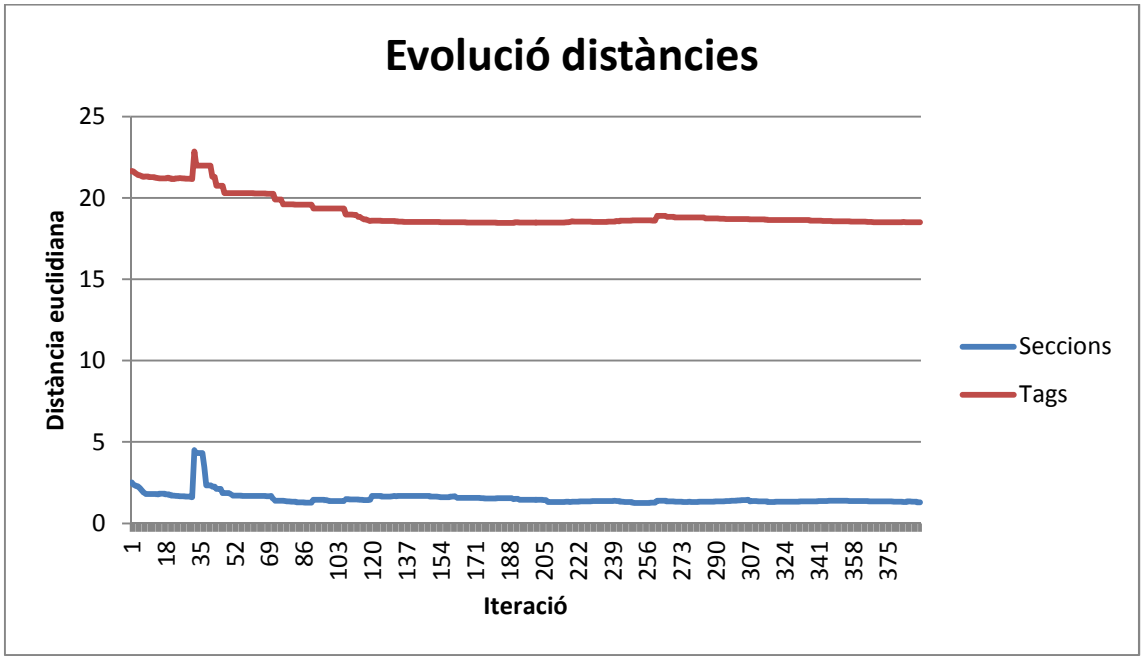


Fig. 4d

Hem vist que les distàncies finals són millors que en l'escenari original, així com també obtenim una corba d'aprenentatge més suau. Així mateix, podem veure que l'ajustament a l'inici de l'algorisme tampoc és tan brusc com en l'escenari A.

Escenari C *(ALFA=0,4; BETA=0,6)*

Els valors de les distàncies entre seccions i *tags* a l'inici i al final de l'execució de l'algorisme són els següents.

	Distància entre seccions	Distància entre grups de *tags*
Inici	2.504180081383925	21.647480034974407
Final	1.3722317176484498	18.508164640250033

La corba d'aprenentatge que obtenim és la següent *[Fig. 4e]* :

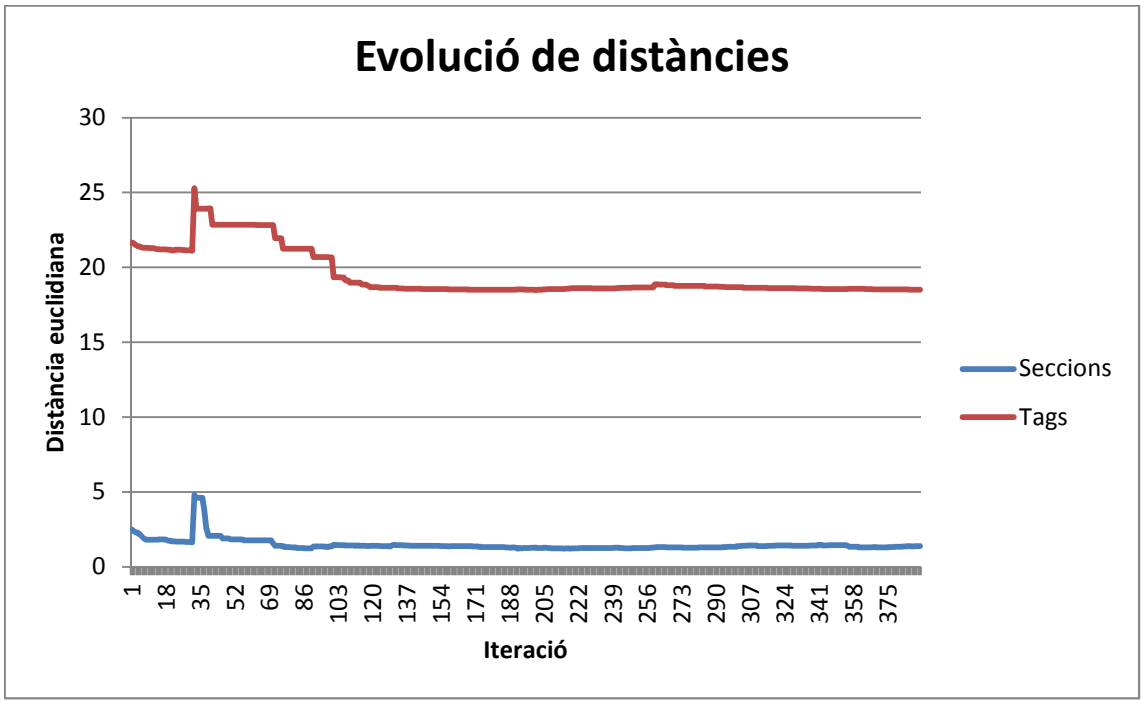


Fig. 4e

Veiem que és una corba d'aprenentatge acceptable però que sobretot la distància entre seccions augmenta molt respecte els altres escenaris.

Conclusió

Recapitulant, tenim aquests tres resultats:

	Distància entre seccions	Distància entre grups de *tags*
Distàncies inicials	2.504180081383925	21.647480034974407
Distàncies finals (A)	1.3410420778192813	18.900561177553882
Distàncies finals (B)	1.295941043290216	18.493673799014562
Distàncies finals (C)	1.3722317176484498	18.508164640250033

Veiem que l'escenari B té un comportament millor que l'original (A), ja que disminuïm significativament la distància entre seccions partint de la mateixa situació inicial. Aquest valor es veu decrementat en 0,05 punts i, per tant, ens proporciona un millor aprenentatge. Passa el mateix amb la distància entre grups de *tags*, que obtenim 0,41 punts de diferència. Convé notar que en aquest cas la diferència és major perquè treballem amb unitats de distància més grans.

Com a conclusió, podem dir que adoptarem com a valors per defecte d'ALFA i BETA els que s'han provat a l'escenari B.

4.4.2 OVER_RANKED_SIGNIFICATIU / LOVED_SIGNIFICATIU

Aquests dos paràmetres serveixen per determinar quan els repositoris d'articles sobrevalorats (*overranked*) i l'historial de notícies escollides pel perfil ideal (*loved*) tenen prou informació com per poder analitzar si troben algun patró entre les notícies acumulades per poder actualitzar el perfil evolutiu amb prou coneixement de causa.

Com que treballem amb repositoris de dades de 6.000 notícies aproximadament, es va considerar com a hipòtesi inicial que grups de 100 notícies serien suficients per poder treure conclusions i una xifra de fàcil tractar a nivell de percentatges. Pel cas de les notícies loved, cal que passin força iteracions per tal d'arribar a aquesta quantitat, si bé en el cas d'overranked s'hi arriba bastant més ràpid i sobretot al principi, en què l'elecció dels articles és encara imprecisa.

Farem proves amb escenaris que varien el valor d'aquest paràmetre. Veurem que s'obtenen uns bons resultats amb les altres proves, però que el valor que dóna millors resultats és l'original. Les proves que s'han fet són les següents:

	OVER_RANKED_SIGNIFICATIU	LOVED_SIGNIFICATIU
Escenari A	100	100
Escenari B	250	250
Escenari C	50	50

Escenari A *(OVER_RANKED_SIGNIFICATIU=LOVED_SIGNIFICATIU=100)*

Els valors de les distàncies entre seccions i *tags* a l'inici i al final de l'execució de l'algorisme són els següents.

	Distància entre seccions	Distància entre grups de *tags*
Inici	2.504180081383925	21.647480034974407
Final	1.295941043290216	18.493673799014562

La corba d'aprenentatge que obtenim és la següent *[Fig 4f]* :

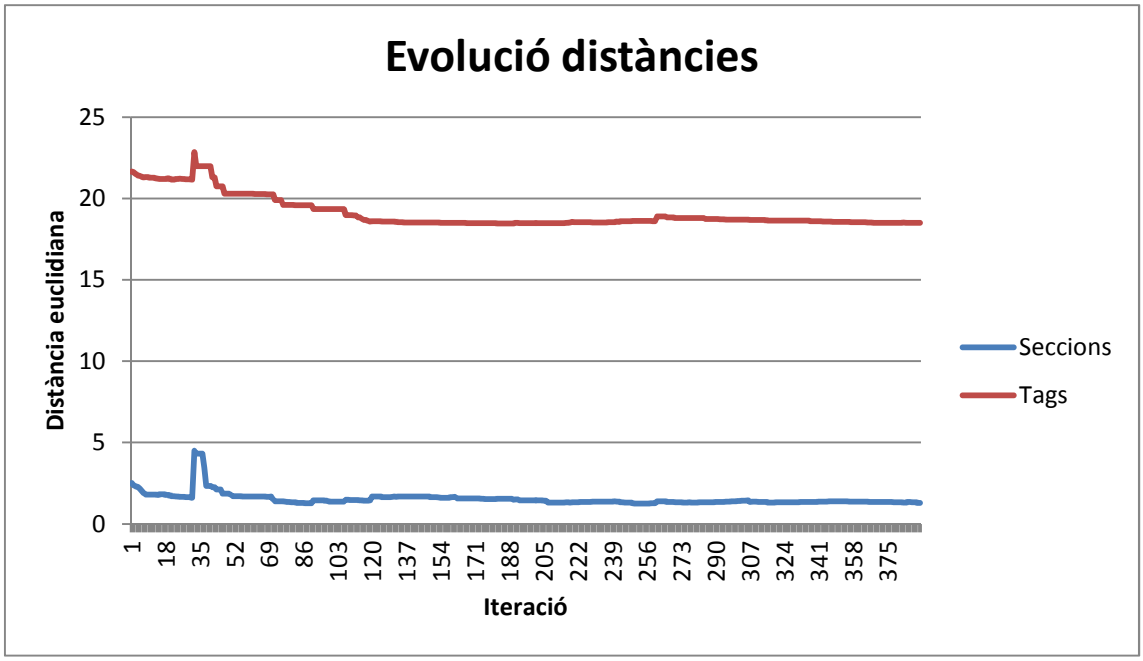


Fig. 4f

És un resultat que ja ens és familiar perquè es correspon a l'escenari B del punt 4.4.1: el que ens ha aportat fins ara un millor resultat.

Escenari B *(OVER_RANKED_SIGNIFICATIU=LOVED_SIGNIFICATIU=250)*

Els valors de les distàncies entre seccions i *tags* a l'inici i al final de l'execució de l'algorisme són els següents.

	Distància entre seccions	**Distància entre grups de *tags***
Inici	2.504180081383925	21.647480034974407
Final	1.3538100472928993	18.95990031837242

La corba d'aprenentatge que obtenim és la que podem verue a la *Fig. 4g*.

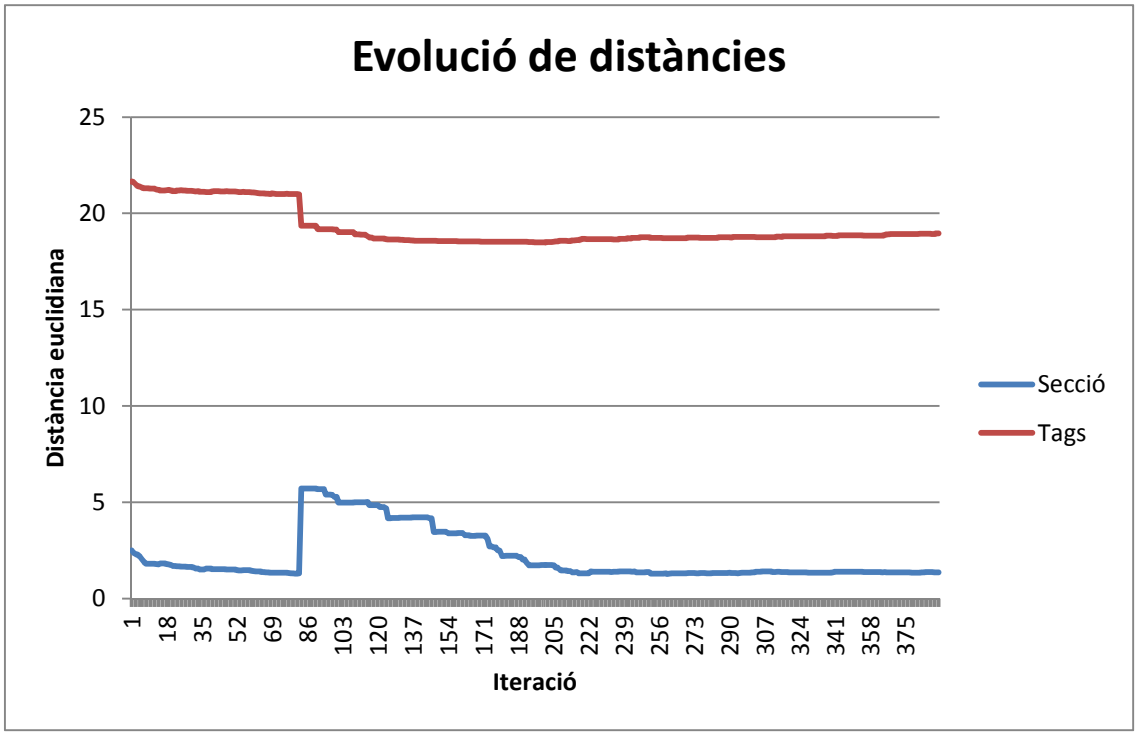


Fig. 4g

Veiem que perdem molta qualitat en l'aprenentatge i que la corba es torna molt més irregular. La dels *tags*, fins i tot té tendència a anar pujant.

Escenari C *(OVER_RANKED_SIGNIFICATIU=LOVED_SIGNIFICATIU=50)*

Els valors de les distàncies entre seccions i *tags* a l'inici i al final de l'execució de l'algorisme són els següents.

	Distància entre seccions	**Distància entre grups de *tags***
Inici	2.504180081383925	21.647480034974407
Final	1.3218826022276033	19.09275200509113

La corba d'aprenentatge que obtenim és la que podem veure a la *Fig. 4h*.

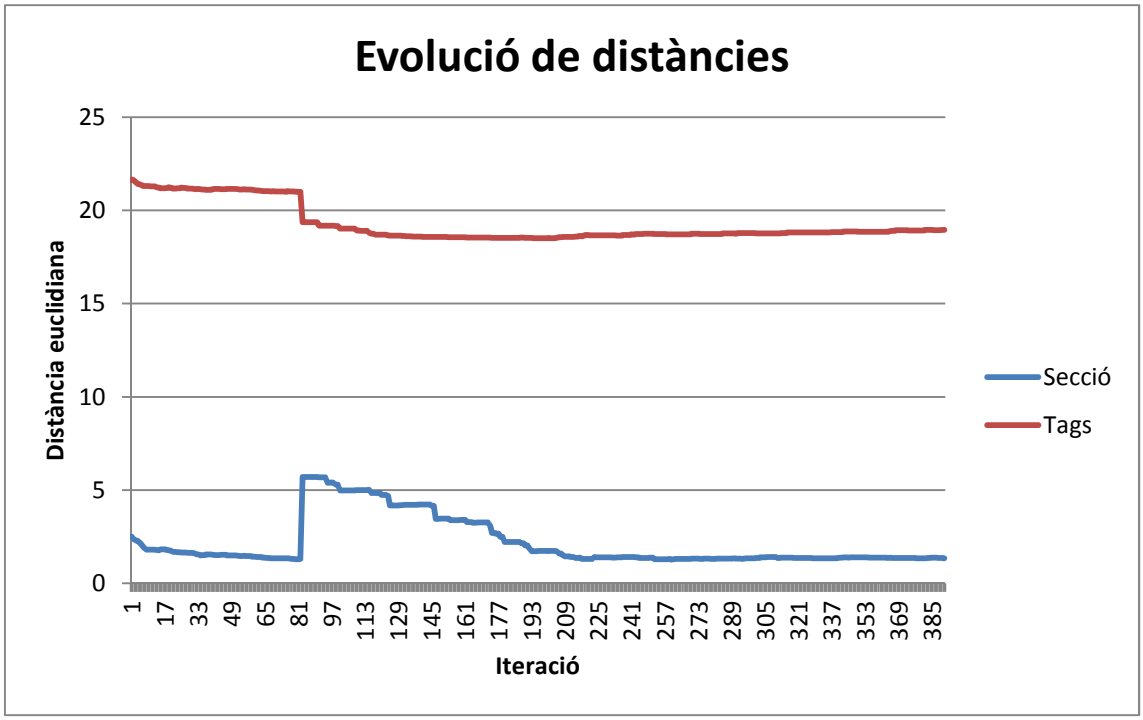


Fig. 4h

Veiem que, com en l'escenari B, no obtenim una bona corba. A nivell de seccions té un funcionament suficient però a nivell de *tags* acaba sense una tendència a decrementar.

Conclusió

Recapitulant, tenim aquests tres resultats:

	Distància entre seccions	Distància entre grups de *tags*
Distàncies inicials	2.504180081383925	21.647480034974407
Distàncies finals (A)	1.295941043290216	18.493673799014562
Distàncies finals (B)	1.3538100472928993	18.95990031837242
Distàncies finals (C)	1.3218826022276033	19.09275200509113

Veiem que si pugem el valor establert en la hipòtesi inicial (escenari B) ens empitjora l'aprenentatge, tant a nivell de *tag* com a nivell de seccions. Passa el mateix disminuint el paràmetre (escenari B). Per tant, prendrem com a valor per defecte l'original i no efectuarem cap canvi després d'aquest joc de proves.

4.4.3 PERCENTATGE_SUPERAR

Quin és el nombre mínim de notícies que han d'aparèixer als grups d'overranked o loved per tal que es tinguin en compte a l'offline learning? Aquest valor l'estableix el paràmetre PERCENTATGE_SUPERAR, que inicialment s'ha establert al 10% (0,1). Juguem amb corpus que compten amb poc més de 100 seccions. Per tant, s'ha considerat que, si en un grup de 100 notícies (OVER_RANKED_SIGNIFICATIU / LOVED_SIGNIFICATIU) n'hi ha com a mínim 10 de la mateixa secció, es tracta d'un grup prou representatiu. Si realment s'acumula un percentatge com aquest de notícies de la mateixa secció es pot parlar d'un patró de comportament. Per validar o descartar aquest valor, estudiarem els següents escenaris:

	PERCENTATGE_SUPERAR	
Escenari A	0,10 (10%)	*(significatiu)*
Escenari B	0,25 (25%)	*(molt significatiu)*
Escenari C	0,05 (5%)	*(poc significatiu)*

Escenari A *(PERCENTATGE_SUPERAR=0,10)*

Els valors de les distàncies entre seccions i *tags* a l'inici i al final de l'execució de l'algorisme són els següents.

	Distància entre seccions	Distància entre grups de *tags*
Inici	2.504180081383925	21.647480034974407
Final	1.295941043290216	18.493673799014562

La corba d'aprenentatge que obtenim és la que podem veure a la *Fig. 4i*.

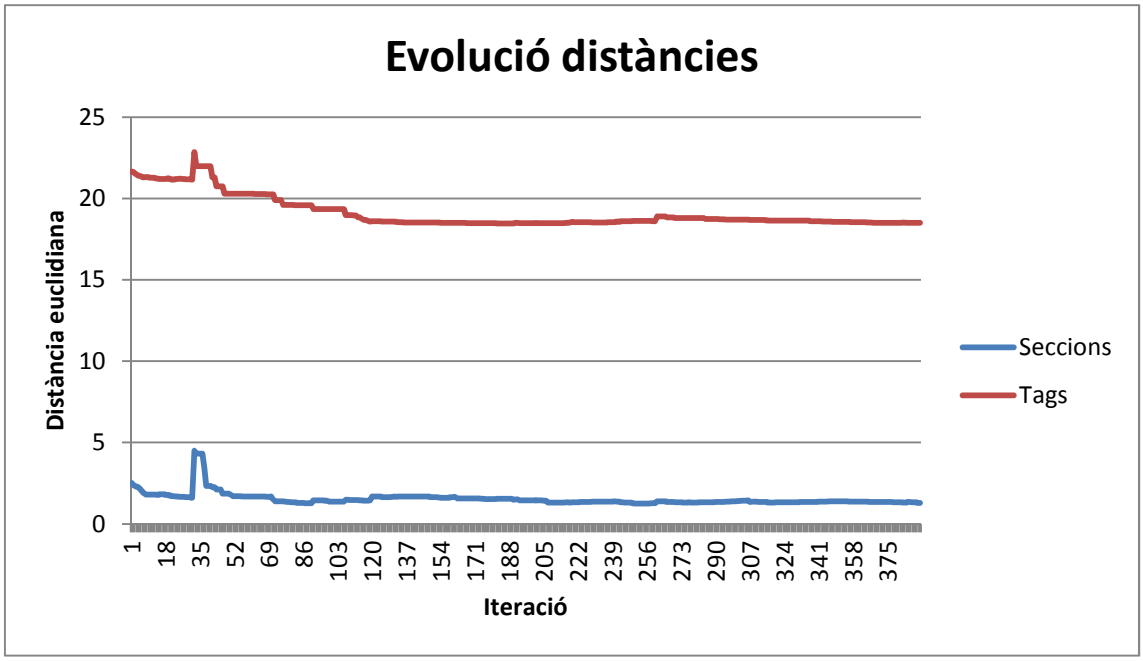


Fig. 4i

És un resultat que ja ens és familiar perquè es correspon a l'escenari B del punt 4.4.1: el que ens ha aportat fins ara un millor resultat.

Escenari B *(PERCENTATGE_SUPERAR=0,25)*

Els valors de les distàncies entre seccions i *tags* a l'inici i al final de l'execució de l'algorisme són els següents.

	Distància entre seccions	Distància entre grups de *tags*
Inici	2.504180081383925	21.647480034974407
Final	1.3079006280345853	18.89457849397519

La corba d'aprenentatge que obtenim és la que podem veure a la *Fig. 4j*.

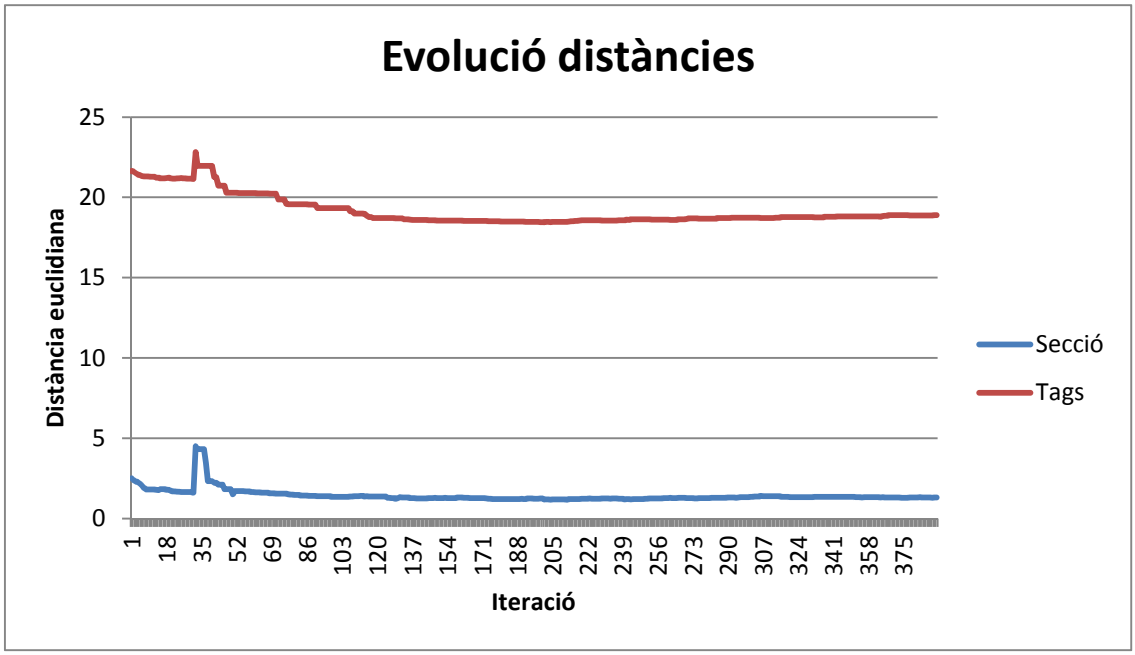


Fig. 4j

Veiem que és una corba acceptable però que no millora els resultats de l'escenari A.

Escenari C *(PERCENTATGE_SUPERAR=0,05)*

Els valors de les distàncies entre seccions i *tags* a l'inici i al final de l'execució de l'algorisme són els següents.

	Distància entre seccions	Distància entre grups de *tags*
Inici	2.504180081383925	21.647480034974407
Final	1.4950488626754355	18.785224894463777

La corba d'aprenentatge que obtenim és la que podem veure a la *Fig. 4k*.

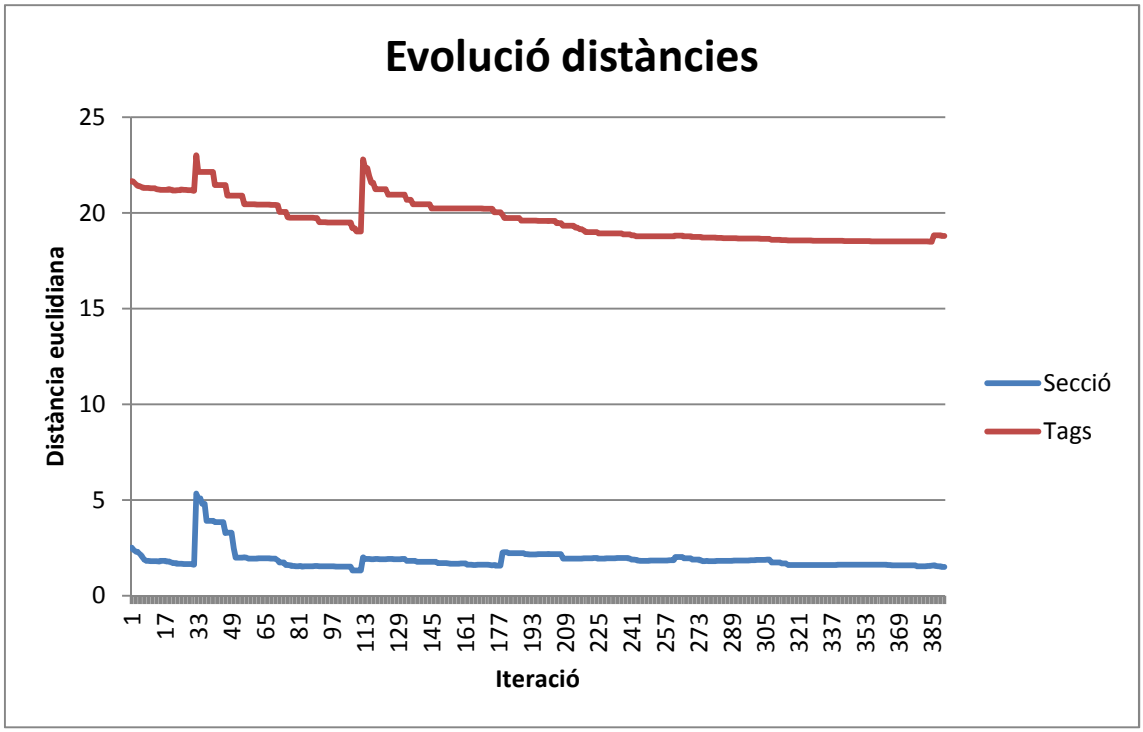


Fig. 4k

És una corba bastant acceptable tot i certes irregularitats però no ens millora l'escenari A i la distància final augmenta molt a nivell de secció.

Conclusió

Recapitulant, tenim aquests tres resultats:

	Distància entre seccions	Distància entre grups de *tags*
Distàncies inicials	2.504180081383925	21.647480034974407
Distàncies finals (A)	1.295941043290216	18.493673799014562
Distàncies finals (B)	1.3079006280345853	18.89457849397519
Distàncies finals (C)	1.4950488626754355	18.785224894463777

Veiem que amb PERCENTATGE_SUPERAR=0,25 l'algorisme es comporta prou bé malgrat que les distàncies augmenten lleugerament. Té sentit perquè es considera que els grups de notícies que superen aquest percentatge són una mostra molt significativa, ja que supera el valor establert inicialment. No obstant això, no s'arriba al mateix nivell d'aprenentatge. Podria ser perquè es perden matisos en no tenir en compte els grups de notícies d'entre el 10 i el 25%.

En canvi, si agafem un percentatge poc significatiu (escenari C) veiem que l'aprenentatge empitjora considerablement. Per tant, continuarem amb el valor establert en l'hipòtesi inicial sense fer cap modificació.

4.4.4 NUM_TAGS_OFFLINE

De les seccions overranked / loved que en un moment determinat superen el PERCENTATGE_SUPERAR, l'algorisme d'aprenentatge offline en fa un recompte de tots els *tags* que tenen associats. Aquells més repetits són als quals els sumarem o restarem valoració. Per defecte, s'ha establert que són significatius els 10 *tags* amb més valoració però en aquest apartat provarem altres valors per veure com es comporta l'algorisme. Els escenaris a tenir en compte seran els següents:

	NUM_TAGS_OFFLINE
Escenari A	10
Escenari B	20
Escenari C	30

Escenari A *(NUM_TAGS_OFFLINE=10)*

Els valors de les distàncies entre seccions i *tags* a l'inici i al final de l'execució de l'algorisme són els següents.

	Distància entre seccions	Distància entre grups de *tags*
Inici	2.504180081383925	21.647480034974407
Final	1.295941043290216	18.493673799014562

La corba d'aprenentatge que obtenim és la següent *[Fig. 4l]* :

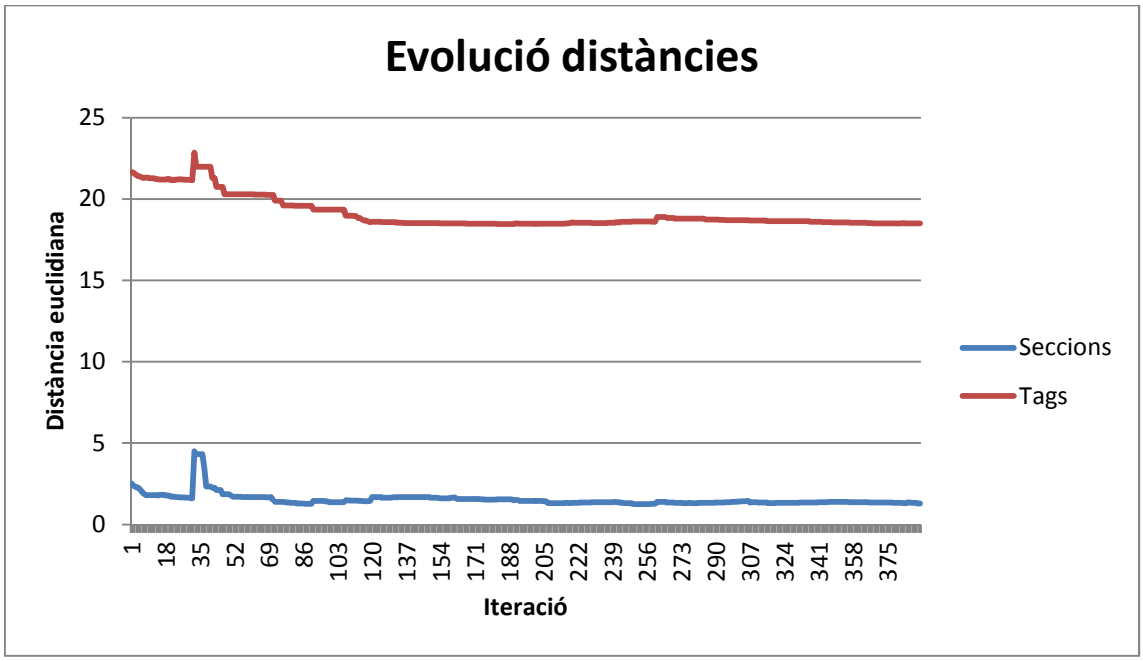


Fig. 4l

És un resultat que ja ens és familiar perquè es correspon a l'escenari B del punt 4.4.1: el que ens ha aportat fins ara un millor resultat.

Escenari B *(NUM_TAGS_OFFLINE=20)*

Els valors de les distàncies entre seccions i *tags* a l'inici i al final de l'execució de l'algorisme són els següents.

	Distància entre seccions	Distància entre grups de *tags*
Inici	2.504180081383925	21.647480034974407
Final	1.292898420258045	18.815334511188443

La corba d'aprenentatge que obtenim és la següent *[Fig. 4m]* :

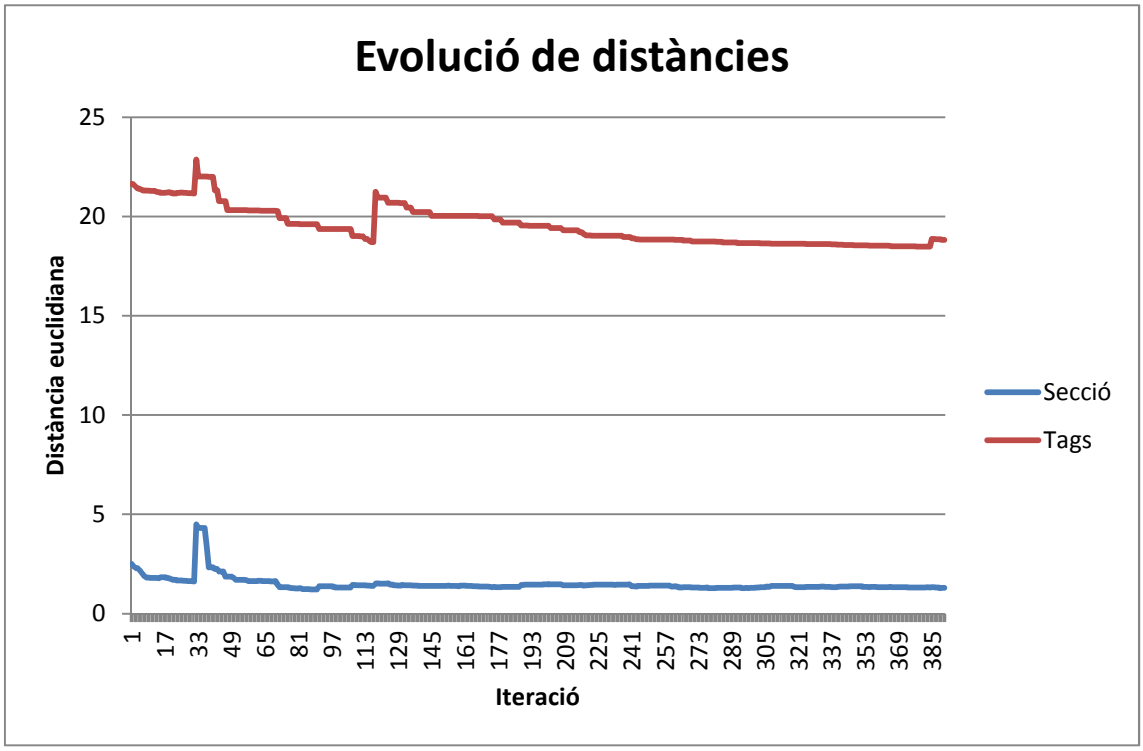


Fig. 4m

Podem observar que aquest escenari ens millora lleugerament l'aprenentatge de la secció, la qual cosa és un fet molt positiu. No obstant això, obtenim una distància entre *tags* superior.

Escenari C *(NUM_TAGS_OFFLINE=30)*

Els valors de les distàncies entre seccions i *tags* a l'inici i al final de l'execució de l'algorisme són els següents.

	Distància entre seccions	Distància entre grups de *tags*
Inici	2.504180081383925	21.647480034974407
Final	1.3523688604552784	18.682097339498185

La corba d'aprenentatge que obtenim és la següent *[Fig 4n]* :

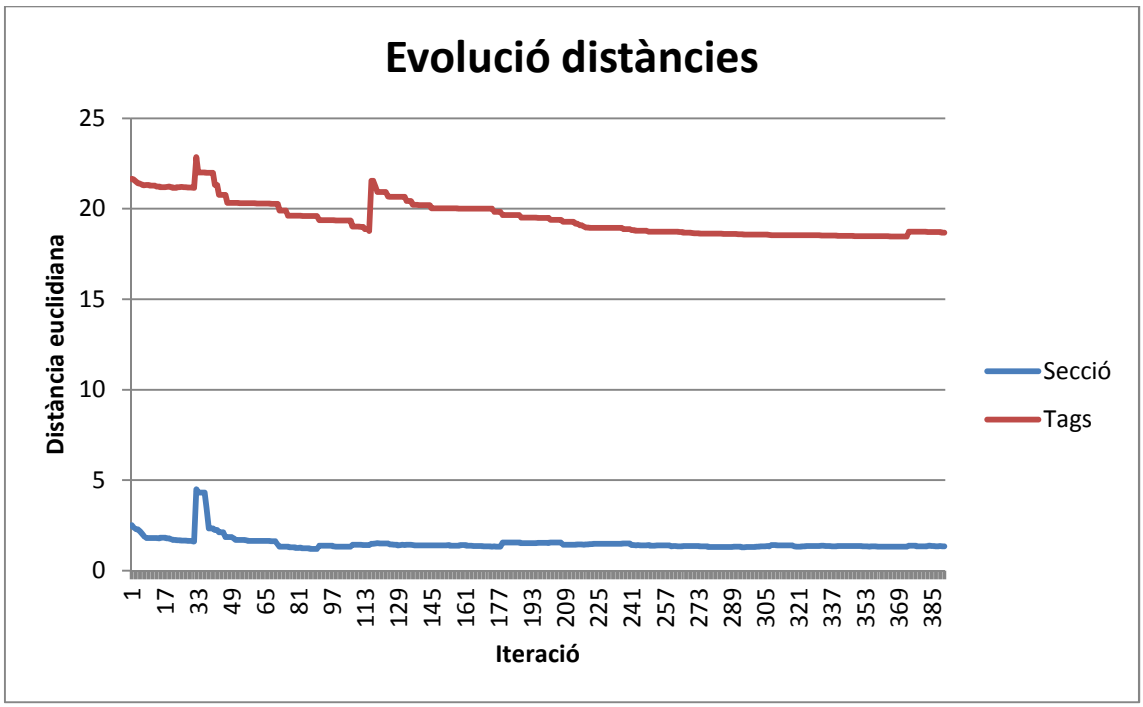


Fig. 4n

Tenim una corba irregular a nivell de *tags* i bastant acceptable a nivell de secció, malgrat que no millora els resultats inicials.

Conclusió

Recapitulant, tenim aquests tres resultats:

	Distància entre seccions	Distància entre grups de *tags*
Distàncies inicials	2.504180081383925	21.647480034974407
Distàncies finals (A)	1.295941043290216	18.493673799014562
Distàncies finals (B)	1.292898420258045	18.815334511188443
Distàncies finals (C)	1.3523688604552784	18.682097339498185

Observem que a nivell de *tags* la distància es mou poquet en els tres escenaris tenint en compte que treballem amb una xifra que frega la vintena. A nivell de seccions veiem que l'escenari B millora l'aprenentatge però la millora global no és prou

significativa com per adoptar el nou paràmetre. Per tant, continuarem amb el valor que s'havia establert a la hipòtesi inicial.

4.4.5 UP_LOVED / DOWN_OVER_RANKED

Quan donem pes a un *tag* a l'aprenentatge online, li afegim una unitat, mentre que el doble a la secció principal perquè té més magnitud global. A l'aprenentatge offline, UPF_OVED i DOWN_OVER_RANKED defineixen les unitats que restem o sumem a les valoracions dels *tags*. Inicialment aquest valor s'ha establert a 5, per marcar la diferència entre l'aprenentatge basat en tan sols un article seleccionat i aquell que ha tingut en compte 100 notícies per intentar buscar un patró. No obstant això, en aquest apartat provarem altres valors per a aquest paràmetre per estudiar com es comporten. Els escenaris que tindrem en compte seran els següents:

	UP_LOVED	DOWN_OVER_RANKED
Escenari A	5	5
Escenari B	10	10
Escenari C	25	25
Escenari D	2	2

Escenari A *(UP_LOVED=DOWN_OVER_RANKED=5)*

Els valors de les distàncies entre seccions i *tags* a l'inici i al final de l'execució de l'algorisme són els següents.

	Distància entre seccions	Distància entre grups de *tags*
Inici	2.504180081383925	21.647480034974407
Final	1.295941043290216	18.493673799014562

La corba d'aprenentatge que obtenim és la següent *[Fig. 4o]* :

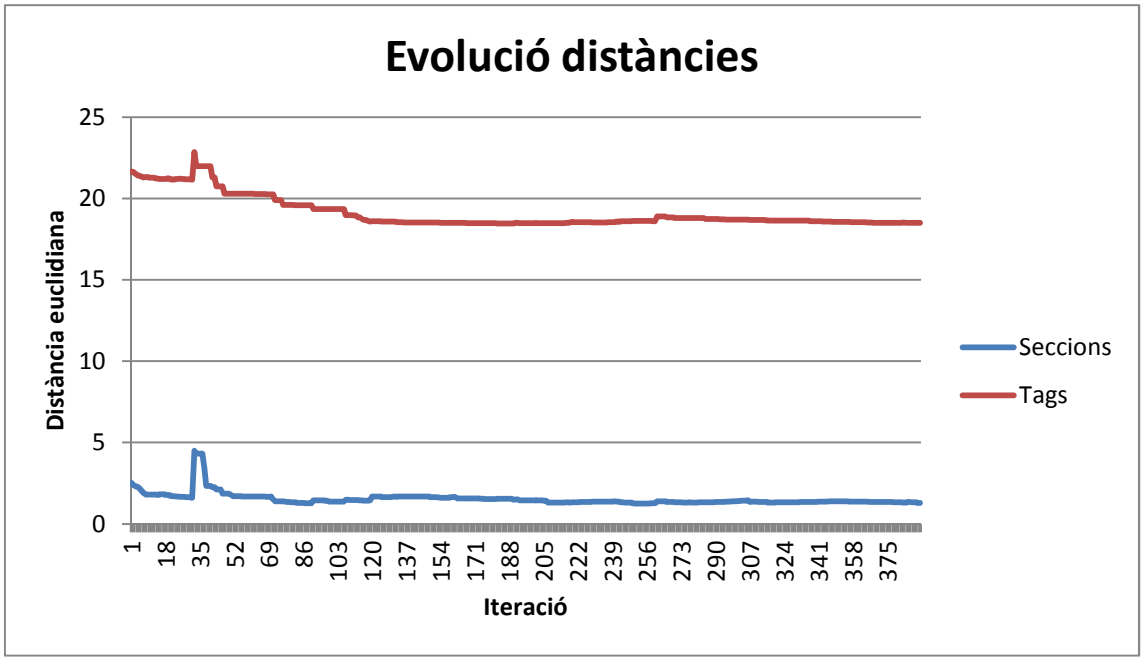


Fig. 4o

És un resultat que ja ens és familiar perquè es correspon a l'escenari B del punt 4.4.1: el que ens ha aportat fins ara un millor resultat.

Escenari B *(UP_LOVED=DOWN_OVER_RANKED=10)*

Els valors de les distàncies entre seccions i *tags* a l'inici i al final de l'execució de l'algorisme són els següents.

	Distància entre seccions	**Distància entre grups de *tags***
Inici	2.504180081383925	21.647480034974407
Final	1.3247634016543783	19.887853950463676

La corba d'aprenentatge que obtenim és la següent *[Fig. 4p]* :

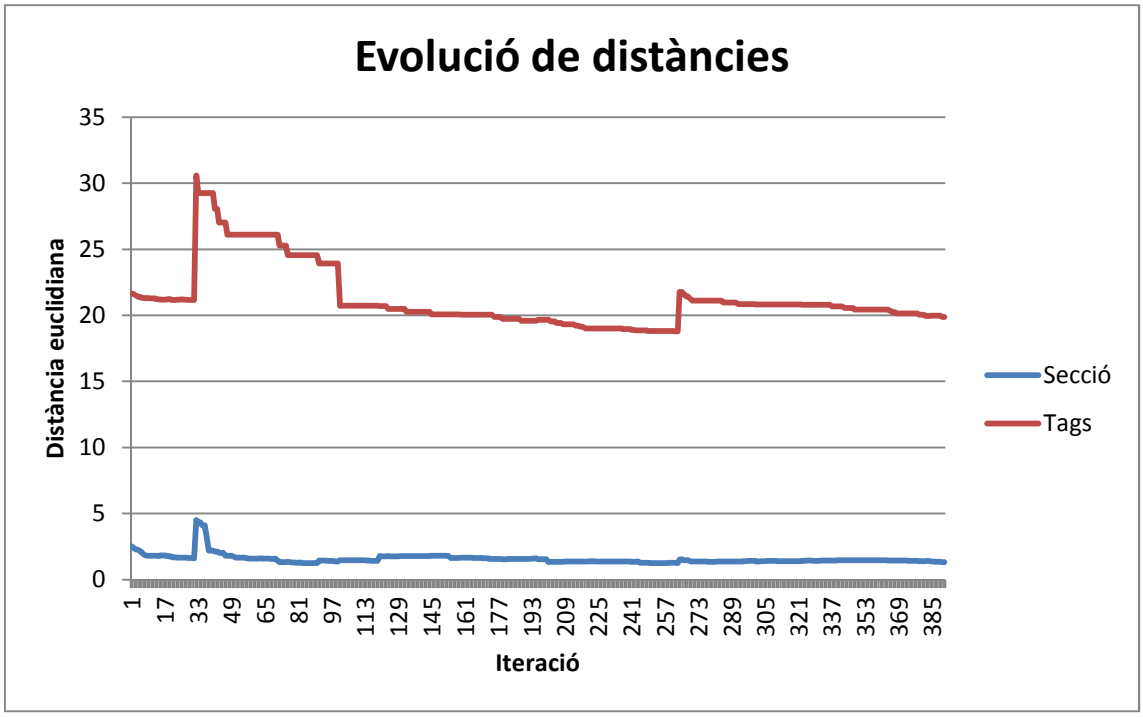


Fig. 4p

Si bé l'aprenentatge de seccions no es veu massa modificat tot i empitjorar el seu resultat, veiem que l'aprenentatge de *tags* esdevé irregular i que, bastant avançat l'algorisme, empitjora la distància, segurament en alguna entrada a l'offline learning.

Escenari C *(UP_LOVED=DOWN_OVER_RANKED=25)*

Els valors de les distàncies entre seccions i *tags* a l'inici i al final de l'execució de l'algorisme són els següents.

	Distància entre seccions	**Distància entre grups de *tags***
Inici	2.504180081383925	21.647480034974407
Final	1.3613197046040058	24.83145405774174

La corba d'aprenentatge que obtenim és la següent *[Fig. 4q]* :

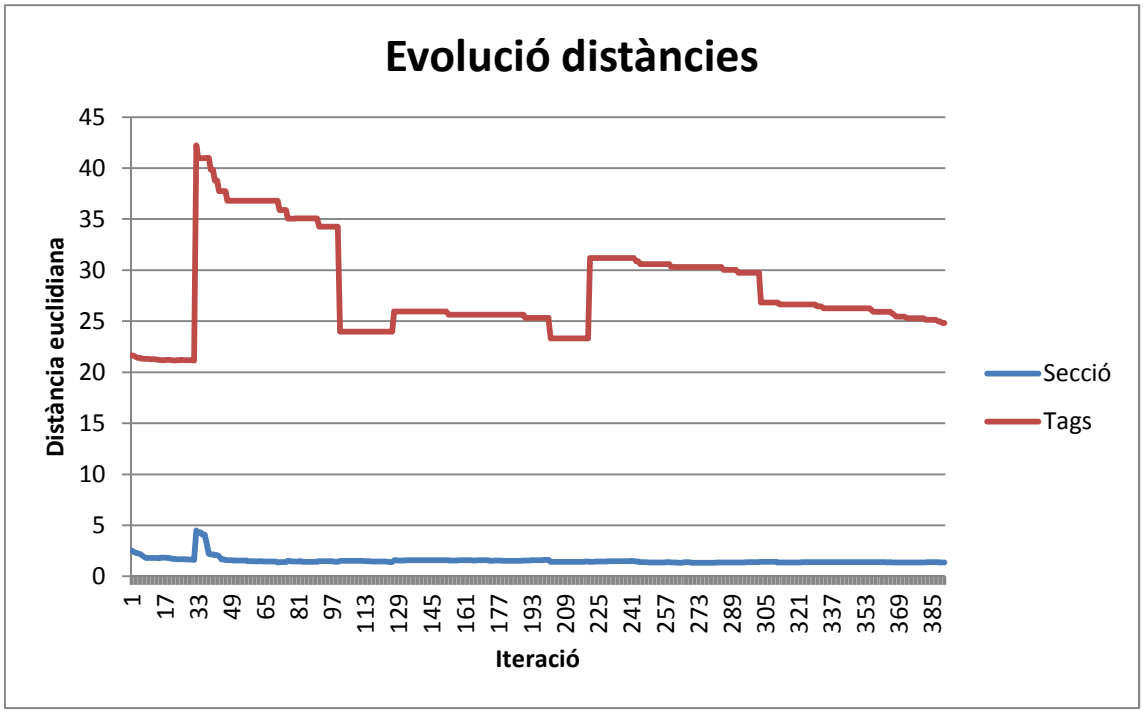


Fig. 4q

Podem comprovar que, amb aquest valor, l'offline learning queda totalment alterat. Farem una prova addicional posant el valor del paràmetre per sota del seu valor habitual per tenir una anàlisi una mica més completa, ja que hem vist que incrementant el valor empitjoren bastant els resultats:

Escenari D *(UP_LOVED=DOWN_OVER_RANKED=2)*

Els valors de les distàncies entre seccions i *tags* a l'inici i al final de l'execució de l'algorisme són els següents.

	Distància entre seccions	**Distància entre grups de *tags***
Inici	2.504180081383925	21.647480034974407
Final	1.3673243313965	18.986759558206476

La corba d'aprenentatge que obtenim és la següent *[Fig. 4r]* :

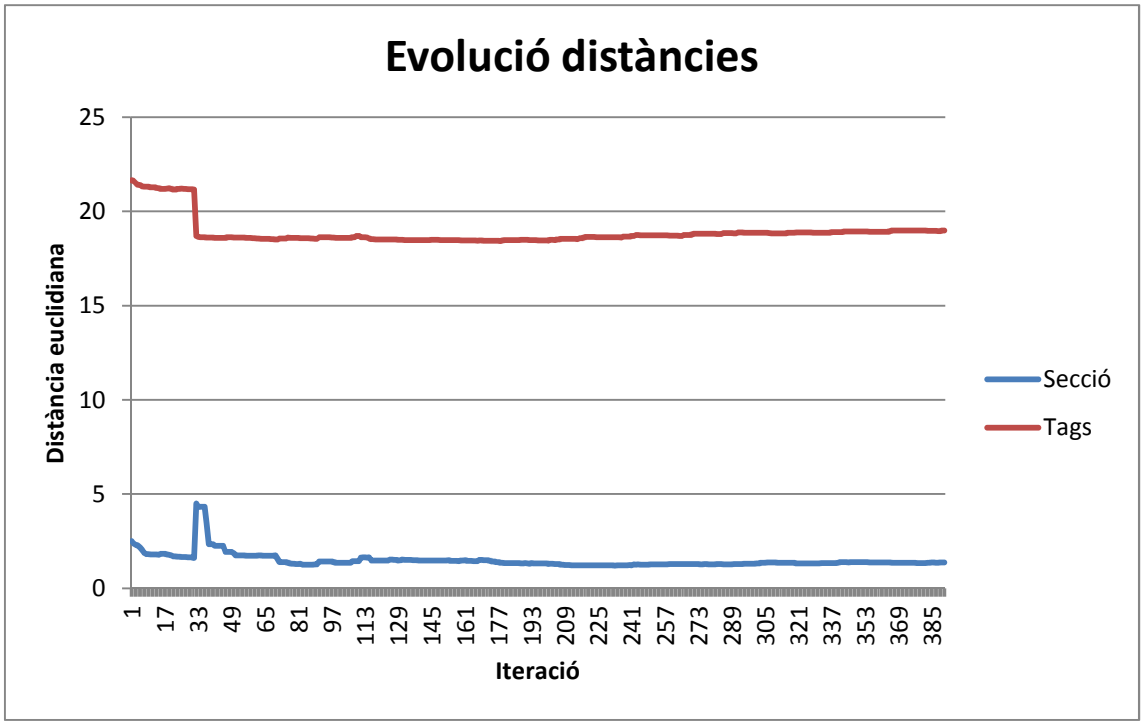


Fig. 4r

Veiem que millora respecte els valors més alts que havíem provat però, de totes maneres, no millorem els resultats obtinguts amb l'escenari A. És important també observar que l'aprenentatge de *tags* no té una evolució favorable en les últimes iteracions.

Conclusió

Recapitulant, tenim aquests tres resultats:

	Distància entre seccions	Distància entre grups de *tags*
Distàncies inicials	2.504180081383925	21.647480034974407
Distàncies finals (A)	1.295941043290216	18.493673799014562
Distàncies finals (B)	1.3247634016543783	19.887853950463676
Distàncies finals (C)	1.3613197046040058	24.83145405774174
Distàncies finals (D)	1.3673243313965	18.986759558206476

Veiem que el valor de l'escenari A funciona i, per tant, mantindrem el paràmetre tal com s'havia definit a la secció d'Anàlisi i disseny. Veiem que a mesura que creix aquest paràmetre, els resultats de l'aprenentatge empitjoren. En l'escenari D també obtenim pitjors resultats.

4.5 Supòsits d'aprenentatge

Anem a provar ara l'algorisme d'aprenentatge partint de diverses situacions. Tindrem en compte tres grans supòsits ja que tres és el nombre mínim que permet avaluar un pratró de funcionament i la reproductibilitat. No obstant això, cadascun d'aquests exemples s'executarà tenint en compte el corpus1 i el corpus2 de notícies amb la qual cosa, finalment, comptarem amb sis execucions diferents.

S'han creat tres perfils: el de la persona de negocis, l'aficionada a l'esport i la hipster. En els apartats dedicats a cadascun d'ells s'indica com s'ha procedit a confeccionar els perfils i s'analitzen els resultats de l'algorisme d'aprenentatge.

Els supòsits en els quals el perfil ideal de la persona varia un cop ja començada execució es veuran en el següent apartat tractats de manera separada.

4.5.1 Persona de negocis

En aquest apartat crearem el perfil ideal que modelarà un usuari que anomenarem "home o dona de negocis". Donarem pes a diversos temes de la secció de tecnologia de 'The Guardian', així com també a la de notícies internacionals. El màxim pes el tindrà la secció de negocis, on donarem valor a una sèrie de paraules clau en concret. També ressaltarem la secció de política. Totes aquestes dades s'introduiran fent una modificació –a través d'un editor de textos- d'un perfil ideal que s'ha creat de manera aleatòria amb el NewsRecommender.

Les seccions i *tags* a destacar són els de la següent llista. A les seccions els hem donat diversos pesos, tots més alts que el màxim pes que hi havia en el perfil aleatori. Als *tags* seleccionats els hem donat el valor màxim normalitzat (1). Per començar, treballarem amb el corpus1 de notícies.

```
SECCIÓ technology;40;0.0
technology/apple;0.0;1
technology/efinance;0.0;1
technology/series/on-social-media-marketing;0.0;1

SECCIÓ world;40;0.0
world/european-commission;0.0;1
world/us-politics;0.0;1
world/us-political-lobbying;0.0;1

sustainable-business/84antande-markets;0.0;1
sustainable-business/finance;0.0;1
```

```
SECCIÓ business;50;0.0
business/hedge-funds;0.0;1
business/business;0.0;1
business/banking;0.0;1
business/85antander;0.0;1
business/85antander;0.0;1
business/commodities;0.0;1
business/luxury-goods-sector;0.0;1
business/euro;0.0;1
business/gas;0.0;1
business/currencies;0.0;1
business/bank-of-america;0.0;1

SECCIÓ politics;30;0.0
politics/politics;0.0;1
politics/foreignpolicy;0.0;1
politics/blog;0.0;1

tv-and-radio/mad-men-tv-series;0.0;1
```

Un cop fetes les modificacions, carreguem el perfil:

```
================================================================
                        NEWS RECOMMENDER
================================================================
1.  Descarregar online corpus de notícies de The Guardian
2.  Guardar corpus de notícies carregat en un fitxer CSV
3.  Carregar un corpus de notícies des de CSV
4. Emmagatzemar perfil a un CSV
5. Carregar perfil de CSV
6.  Crear estructura de perfil
7.  Obtenir estadístiques del corpus
8.  Crear un perfil ideal amb dades aleatòries
9. Donar pes a seccions i tags del perfil ideal - Boost
9.  Crear perfil aprenentatge
10. Executar algorisme aprenentatge
11. Sortir

Escull una opció:
5
Escull una opció:
14. Carregar el perfil evolutiu
15. Carregar el perfil ideal
15
Escriu el nom de l'arxiu del perfil ideal, sense extensió
profileidealnegocis
Perfil carregat satisfactòriament. Prem una tecla per veure'l

(...)

104.- Nom secció: teacher-network-advertisement-features/ Rating global: 0.0
Tags: teacher-network-advertisement-features (0.8328237364658403)
105.- Nom secció: housing-network-partner-zone-pinnacle/ Rating global: 0.0
Tags: housing-network-partner-zone-pinnacle (0.6833361216775692)
106.- Nom secció: sustainable-business-fairtrade-partner-zone/ Rating global:
0.0
Tags: sustainable-business-fairtrade-partner-zone (0.7127025919529387)
```

```
107.- Nom secció: partner-zone-sas-computacenter/ Rating global: 0.0
Tags: partner-zone-sas-computacenter (0.03279385797060874)
108.- Nom secció: adam-smith-international-partner-zone/ Rating global: 0.0
Tags: adam-smith-international-partner-zone (0.9836004963719329)
109.- Nom secció: help/ Rating global: 0.0
Tags: help (0.2050082957896503)
```

Anem a executar l'algorisme d'aprenentatge:

```
================================================================
                        NEWS RECOMMENDER
================================================================
1.  Descarregar online corpus de notícies de The Guardian
2.  Guardar corpus de notícies carregat en un fitxer CSV
3.  Carregar un corpus de notícies des de CSV
4. Emmagatzemar perfil a un CSV
5. Carregar perfil de CSV
6.  Crear estructura de perfil
7.  Obtenir estadístiques del corpus
8.  Crear un perfil ideal amb dades aleatòries
9. Donar pes a seccions i tags del perfil ideal - Boost
9.  Crear perfil aprenentatge
10. Executar algorisme aprenentatge
11. Sortir

Escull una opció:
10
DISTÀNCIA sec: 2.504180081383925
DISTÀNCIA tag: 21.647480034974407
ITERACIÓ1
Escollida evolutive: 0
Escollida ideal: 6
Mida overranked 6
Mida loved 1
ITERACIÓ2
Escollida evolutive: 10
Escollida ideal: 11
Mida overranked 8
Mida loved 2
ITERACIÓ3
Escollida evolutive: 4
Escollida ideal: 9
Mida overranked 20
Mida loved 3
ITERACIÓ4
Escollida evolutive: 11
Escollida ideal: 5
Mida overranked 21
Mida loved 4
ITERACIÓ5
Escollida evolutive: 12
Escollida ideal: 8
Mida overranked 28
Mida loved 5
ITERACIÓ6
Escollida evolutive: 2
Escollida ideal: 6
```

```
Mida overranked 31
Mida loved 6
ITERACIÓ7
Escollida evolutive: 0
Escollida ideal: 8
Mida overranked 36
Mida loved 7
ITERACIÓ8
Escollida evolutive: 2
Escollida ideal: 2
Mida overranked 36
Mida loved 8
ITERACIÓ9
Escollida evolutive: 4
Escollida ideal: 14
Mida overranked 45
Mida loved 9
ITERACIÓ10
Escollida evolutive: 3
Escollida ideal: 3
Mida overranked 45
Mida loved 10
ITERACIÓ11
Escollida evolutive: 14
Escollida ideal: 14
Mida overranked 45
Mida loved 11
ITERACIÓ12
Escollida evolutive: 9
Escollida ideal: 3
Mida overranked 55
Mida loved 12
ITERACIÓ13
Escollida evolutive: 6
Escollida ideal: 3
Mida overranked 61
Mida loved 13
ITERACIÓ14
Escollida evolutive: 1
Escollida ideal: 9
Mida overranked 63
Mida loved 14
ITERACIÓ15
Escollida evolutive: 5
Escollida ideal: 5
Mida overranked 63
Mida loved 15
ITERACIÓ16
Escollida evolutive: 9
Escollida ideal: 9
Mida overranked 63
Mida loved 16
ITERACIÓ17
Escollida evolutive: 1
Escollida ideal: 7
Mida overranked 64
Mida loved 17
ITERACIÓ18
Escollida evolutive: 5
```

```
Escollida ideal: 9
Mida overranked 65
Mida loved 18
ITERACIÓ19
Escollida evolutive: 9
Escollida ideal: 13
Mida overranked 66
Mida loved 19
ITERACIÓ20
Escollida evolutive: 8
Escollida ideal: 13
Mida overranked 73
Mida loved 20
ITERACIÓ21
Escollida evolutive: 14
Escollida ideal: 14
Mida overranked 73
Mida loved 21
ITERACIÓ22
Escollida evolutive: 13
Escollida ideal: 9
Mida overranked 75
Mida loved 22
ITERACIÓ23
Escollida evolutive: 12
Escollida ideal: 12
Mida overranked 75
Mida loved 23
ITERACIÓ24
Escollida evolutive: 6
Escollida ideal: 8
Mida overranked 76
Mida loved 24
ITERACIÓ25
Escollida evolutive: 8
Escollida ideal: 7
Mida overranked 85
Mida loved 25
ITERACIÓ26
Escollida evolutive: 3
Escollida ideal: 3
Mida overranked 85
Mida loved 26
ITERACIÓ27
Escollida evolutive: 11
Escollida ideal: 12
Mida overranked 95
Mida loved 27
ITERACIÓ28
Escollida evolutive: 5
Escollida ideal: 5
Mida overranked 95
Mida loved 28
ITERACIÓ29
Escollida evolutive: 7
Escollida ideal: 7
Mida overranked 95
Mida loved 29
ITERACIÓ30
```

```
Escollida evolutive: 7
Escollida ideal: 4
Mida overranked 99
Mida loved 30
ITERACIÓ31
Escollida evolutive: 0
Escollida ideal: 2
Entro offline learning over ranked
CONTADOR SECCIÓ world 40
Rating antic39.0 0.975
sustract hits
Rating nou-41.0 0.0
CONTADOR SECCIÓ commentisfree 12
Rating antic11.0 0.6419753086419753
sustract hits
Rating nou-13.0 0.345679012345679
Nombre de seccions a tractar2
****Secció a tractar: world
Tag: world 30
Tag: usa 10
Tag: middleeast 8
Tag: australia 7
Tag: europe-news 6
Tag: asia-pacific 4
Tag: al-qaida 4
Tag: india 4
Tag: us-military 3
Tag: usforeignpolicy 3
entro sustractHits_Tag
Tag a tractarworld
Rating antic6.0 0.6666666666666666
sustract hits tag
Rating nou1.0 0.1111111111111111
Tag a tractarusa
Rating antic2.0 0.2222222222222222
sustract hits tag
Rating nou-3.0 0.0
Tag a tractarmiddleeast
Rating antic0.0 0.25
sustract hits tag
Rating nou-5.0 0.0
Tag a tractaraustralia
Rating antic0.0 0.35714285714285715
sustract hits tag
Rating nou-5.0 0.0
Tag a tractareurope-news
Rating antic1.0 0.42857142857142855
sustract hits tag
Rating nou-4.0 0.07142857142857142
Tag a tractarasia-pacific
Rating antic0.0 0.35714285714285715
sustract hits tag
Rating nou-5.0 0.0
Tag a tractaral-qaida
Rating antic0.0 0.35714285714285715
sustract hits tag
Rating nou-5.0 0.0
Tag a tractarindia
Rating antic0.0 0.35714285714285715
```

```
sustract hits tag
Rating nou-5.0 0.0
Tag a tractarus-military
Rating antic0.0 0.35714285714285715
sustract hits tag
Rating nou-5.0 0.0
Tag a tractarusforeignpolicy
Rating antic0.0 0.35714285714285715
sustract hits tag
Rating nou-5.0 0.0
****Secció a tractar: commentisfree
Tag: commentisfree 12
Tag: series/sadhbh-walshe-on-society-and-justice 1
Tag: series/you-told-us 1
Tag: series/guardian-comment-network 1
entro sustractHits_Tag
Tag a tractarcommentisfree
Rating antic3.0 0.5714285714285714
sustract hits tag
Rating nou-2.0 0.21428571428571427
Tag a tractarseries/sadhbh-walshe-on-society-and-justice
Rating antic0.0 0.35714285714285715
sustract hits tag
Rating nou-5.0 0.0
Tag a tractarseries/you-told-us
Rating antic0.0 0.35714285714285715
sustract hits tag
Rating nou-5.0 0.0
Tag a tractarseries/guardian-comment-network
Rating antic0.0 0.35714285714285715
sustract hits tag
Rating nou-5.0 0.0
Mida overranked 0
Mida loved 31
ITERACIÓ32
Escollida evolutive: 4
Escollida ideal: 11
Mida overranked 2
Mida loved 32
ITERACIÓ33
Escollida evolutive: 0
Escollida ideal: 0
Mida overranked 2
Mida loved 33
ITERACIÓ34
Escollida evolutive: 12
Escollida ideal: 12
Mida overranked 2
Mida loved 34
ITERACIÓ35
Escollida evolutive: 12
Escollida ideal: 12
Mida overranked 2
Mida loved 35
ITERACIÓ36
Escollida evolutive: 3
Escollida ideal: 5
Mida overranked 3
Mida loved 36
```

```
ITERACIÓ37
Escollida evolutive: 1
Escollida ideal: 2
Mida overranked 4
Mida loved 37
ITERACIÓ38
Escollida evolutive: 12
Escollida ideal: 12
Mida overranked 4
Mida loved 38
ITERACIÓ39
Escollida evolutive: 10
Escollida ideal: 10
Mida overranked 4
Mida loved 39
ITERACIÓ40
Escollida evolutive: 0
Escollida ideal: 14
Mida overranked 5
Mida loved 40
ITERACIÓ41
Escollida evolutive: 4
Escollida ideal: 4
Mida overranked 5
Mida loved 41
ITERACIÓ42
Escollida evolutive: 3
Escollida ideal: 12
Mida overranked 6
Mida loved 42
ITERACIÓ43
Escollida evolutive: 5
Escollida ideal: 5
Mida overranked 6
Mida loved 43
ITERACIÓ44
Escollida evolutive: 4
Escollida ideal: 4
Mida overranked 6
Mida loved 44
ITERACIÓ45
Escollida evolutive: 0
Escollida ideal: 12
Mida overranked 7
Mida loved 45
ITERACIÓ46
Escollida evolutive: 6
Escollida ideal: 4
Mida overranked 9
Mida loved 46
ITERACIÓ47
Escollida evolutive: 3
Escollida ideal: 3
Mida overranked 9
Mida loved 47
ITERACIÓ48
Escollida evolutive: 7
Escollida ideal: 7
Mida overranked 9
```

```
Mida loved 48
ITERACIÓ49
Escollida evolutive: 0
Escollida ideal: 9
Mida overranked 12
Mida loved 49
ITERACIÓ50
Escollida evolutive: 2
Escollida ideal: 6
Mida overranked 13
Mida loved 50
ITERACIÓ51
Escollida evolutive: 7
Escollida ideal: 7
Mida overranked 13
Mida loved 51
ITERACIÓ52
Escollida evolutive: 8
Escollida ideal: 3
Mida overranked 14
Mida loved 52
ITERACIÓ53
Escollida evolutive: 8
Escollida ideal: 8
Mida overranked 14
Mida loved 53
ITERACIÓ54
Escollida evolutive: 8
Escollida ideal: 8
Mida overranked 14
Mida loved 54
ITERACIÓ55
Escollida evolutive: 9
Escollida ideal: 5
Mida overranked 16
Mida loved 55
ITERACIÓ56
Escollida evolutive: 8
Escollida ideal: 8
Mida overranked 16
Mida loved 56
ITERACIÓ57
Escollida evolutive: 7
Escollida ideal: 7
Mida overranked 16
Mida loved 57
ITERACIÓ58
Escollida evolutive: 4
Escollida ideal: 4
Mida overranked 16
Mida loved 58
ITERACIÓ59
Escollida evolutive: 3
Escollida ideal: 3
Mida overranked 16
Mida loved 59
ITERACIÓ60
Escollida evolutive: 4
Escollida ideal: 1
```

```
Mida overranked 19
Mida loved 60
ITERACIÓ61
Escollida evolutive: 12
Escollida ideal: 5
Mida overranked 22
Mida loved 61
ITERACIÓ62
Escollida evolutive: 10
Escollida ideal: 11
Mida overranked 24
Mida loved 62
ITERACIÓ63
Escollida evolutive: 4
Escollida ideal: 2
Mida overranked 27
Mida loved 63
ITERACIÓ64
Escollida evolutive: 12
Escollida ideal: 12
Mida overranked 27
Mida loved 64
ITERACIÓ65
Escollida evolutive: 9
Escollida ideal: 9
Mida overranked 27
Mida loved 65
ITERACIÓ66
Escollida evolutive: 10
Escollida ideal: 10
Mida overranked 27
Mida loved 66
ITERACIÓ67
Escollida evolutive: 5
Escollida ideal: 2
Mida overranked 33
Mida loved 67
ITERACIÓ68
Escollida evolutive: 3
Escollida ideal: 3
Mida overranked 33
Mida loved 68
ITERACIÓ69
Escollida evolutive: 12
Escollida ideal: 10
Mida overranked 36
Mida loved 69
ITERACIÓ70
Escollida evolutive: 1
Escollida ideal: 14
Mida overranked 43
Mida loved 70
ITERACIÓ71
Escollida evolutive: 0
Escollida ideal: 4
Mida overranked 44
Mida loved 71
ITERACIÓ72
Escollida evolutive: 7
```

```
Escollida ideal: 7
Mida overranked 44
Mida loved 72
ITERACIÓ73
Escollida evolutive: 2
Escollida ideal: 2
Mida overranked 44
Mida loved 73
ITERACIÓ74
Escollida evolutive: 6
Escollida ideal: 6
Mida overranked 44
Mida loved 74
ITERACIÓ75
Escollida evolutive: 13
Escollida ideal: 11
Mida overranked 46
Mida loved 75
ITERACIÓ76
Escollida evolutive: 11
Escollida ideal: 0
Mida overranked 49
Mida loved 76
ITERACIÓ77
Escollida evolutive: 13
Escollida ideal: 6
Mida overranked 52
Mida loved 77
ITERACIÓ78
Escollida evolutive: 5
Escollida ideal: 5
Mida overranked 52
Mida loved 78
ITERACIÓ79
Escollida evolutive: 4
Escollida ideal: 5
Mida overranked 54
Mida loved 79
ITERACIÓ80
Escollida evolutive: 9
Escollida ideal: 9
Mida overranked 54
Mida loved 80
ITERACIÓ81
Escollida evolutive: 3
Escollida ideal: 0
Mida overranked 55
Mida loved 81
ITERACIÓ82
Escollida evolutive: 5
Escollida ideal: 4
Mida overranked 56
Mida loved 82
ITERACIÓ83
Escollida evolutive: 0
Escollida ideal: 13
Mida overranked 57
Mida loved 83
ITERACIÓ84
```

```
Escollida evolutive: 14
Escollida ideal: 14
Mida overranked 57
Mida loved 84
ITERACIÓ85
Escollida evolutive: 11
Escollida ideal: 11
Mida overranked 57
Mida loved 85
ITERACIÓ86
Escollida evolutive: 10
Escollida ideal: 6
Mida overranked 58
Mida loved 86
ITERACIÓ87
Escollida evolutive: 4
Escollida ideal: 4
Mida overranked 58
Mida loved 87
ITERACIÓ88
Escollida evolutive: 8
Escollida ideal: 8
Mida overranked 58
Mida loved 88
ITERACIÓ89
Escollida evolutive: 9
Escollida ideal: 9
Mida overranked 58
Mida loved 89
ITERACIÓ90
Escollida evolutive: 1
Escollida ideal: 5
Mida overranked 61
Mida loved 90
ITERACIÓ91
Escollida evolutive: 10
Escollida ideal: 10
Mida overranked 61
Mida loved 91
ITERACIÓ92
Escollida evolutive: 0
Escollida ideal: 0
Mida overranked 61
Mida loved 92
ITERACIÓ93
Escollida evolutive: 7
Escollida ideal: 7
Mida overranked 61
Mida loved 93
ITERACIÓ94
Escollida evolutive: 10
Escollida ideal: 6
Mida overranked 62
Mida loved 94
ITERACIÓ95
Escollida evolutive: 5
Escollida ideal: 5
Mida overranked 62
Mida loved 95
```

```
ITERACIÓ96
Escollida evolutive: 3
Escollida ideal: 1
Mida overranked 63
Mida loved 96
ITERACIÓ97
Escollida evolutive: 6
Escollida ideal: 13
Mida overranked 65
Mida loved 97
ITERACIÓ98
Escollida evolutive: 0
Escollida ideal: 11
Mida overranked 68
Mida loved 98
ITERACIÓ99
Escollida evolutive: 13
Escollida ideal: 13
Mida overranked 68
Mida loved 99
ITERACIÓ100
Escollida evolutive: 5
Escollida ideal: 5
Mida overranked 68
Mida loved 100
ITERACIÓ101
Escollida evolutive: 6
Escollida ideal: 6
Mida overranked 68
Mida loved 101
ITERACIÓ102
Escollida evolutive: 9
Escollida ideal: 9
Mida overranked 68
Mida loved 102
ITERACIÓ103
Escollida evolutive: 11
Escollida ideal: 11
Mida overranked 68
Mida loved 103
ITERACIÓ104
Escollida evolutive: 0
Escollida ideal: 0
Mida overranked 68
Mida loved 104
ITERACIÓ105
Escollida evolutive: 0
Escollida ideal: 0
Mida overranked 68
Mida loved 105
ITERACIÓ106
Escollida evolutive: 1
Escollida ideal: 0
Entro offline learning loved
CONTADOR SECCIÓ business 16
Rating antic67.0 1.0
add hits
Rating nou99.0 1.0
CONTADOR SECCIÓ commentisfree 14
```

```
Rating antic3.0 0.030303030303030304
add hits
Rating nou31.0 0.31313131313131315
CONTADOR SECCIÓ technology 11
Rating antic60.0 0.6060606060606061
add hits
Rating nou82.0 0.8282828282828283
CONTADOR SECCIÓ politics 10
Rating antic40.0 0.40404040404040403
add hits
Rating nou60.0 0.6060606060606061
Nombre de seccions a tractar4
****Secció a tractar: business
Tag: business 16
Tag: retail 4
Tag: banking 4
Tag: financial-sector 4
Tag: telecoms 2
Tag: housingmarket 2
Tag: realestate 2
Tag: bankofenglandgovernor 2
Tag: executive-pay-bonuses 2
Tag: royalbankofscotlandgroup 2
entro addHits_Tag
Tag a tractarbusiness
Rating antic13.0 0.8571428571428571
add hits tag
Rating nou18.0 1.0
Tag a tractarretail
Rating antic1.0 0.2608695652173913
add hits tag
Rating nou6.0 0.4782608695652174
Tag a tractarbanking
Rating antic3.0 0.34782608695652173
add hits tag
Rating nou8.0 0.5652173913043478
Tag a tractarfinancial-sector
Rating antic3.0 0.34782608695652173
add hits tag
Rating nou8.0 0.5652173913043478
Tag a tractartelecoms
Rating antic1.0 0.2608695652173913
add hits tag
Rating nou6.0 0.4782608695652174
Tag a tractarhousingmarket
Rating antic1.0 0.2608695652173913
add hits tag
Rating nou6.0 0.4782608695652174
Tag a tractarrealestate
Rating antic1.0 0.2608695652173913
add hits tag
Rating nou6.0 0.4782608695652174
Tag a tractarbankofenglandgovernor
Rating antic1.0 0.2608695652173913
add hits tag
Rating nou6.0 0.4782608695652174
Tag a tractarexecutive-pay-bonuses
Rating antic1.0 0.2608695652173913
add hits tag
```

```
Rating nou6.0 0.4782608695652174
Tag a tractarroyalbankofscotlandgroup
Rating antic1.0 0.2608695652173913
add hits tag
Rating nou6.0 0.4782608695652174
****Secció a tractar: commentisfree
Tag: commentisfree 14
Tag: series/first-thoughts 2
Tag: series/response 1
entro addHits_Tag
Tag a tractarcommentisfree
Rating antic3.0 0.34782608695652173
add hits tag
Rating nou8.0 0.5652173913043478
Tag a tractarseries/first-thoughts
Rating antic2.0 0.30434782608695654
add hits tag
Rating nou7.0 0.5217391304347826
Tag a tractarseries/response
Rating antic1.0 0.2608695652173913
add hits tag
Rating nou6.0 0.4782608695652174
****Secció a tractar: technology
Tag: technology 11
Tag: internet 6
Tag: smartphones 5
Tag: apple 4
Tag: mobilephones 4
Tag: google 4
Tag: software 3
Tag: apps 2
Tag: telecoms 2
Tag: android 2
entro addHits_Tag
Tag a tractartechnology
Rating antic7.0 0.5217391304347826
add hits tag
Rating nou12.0 0.7391304347826086
Tag a tractarinternet
Rating antic4.0 0.391304347826087
add hits tag
Rating nou9.0 0.6086956521739131
Tag a tractarsmartphones
Rating antic4.0 0.391304347826087
add hits tag
Rating nou9.0 0.6086956521739131
Tag a tractarapple
Rating antic5.0 0.43478260869565216
add hits tag
Rating nou10.0 0.6521739130434783
Tag a tractarmobilephones
Rating antic4.0 0.391304347826087
add hits tag
Rating nou9.0 0.6086956521739131
Tag a tractargoogle
Rating antic3.0 0.34782608695652173
add hits tag
Rating nou8.0 0.5652173913043478
Tag a tractarsoftware
```

```
Rating antic2.0 0.30434782608695654
add hits tag
Rating nou7.0 0.5217391304347826
Tag a tractarapps
Rating antic1.0 0.2608695652173913
add hits tag
Rating nou6.0 0.4782608695652174
Tag a tractartelecoms
Rating antic0.0 0.21739130434782608
add hits tag
Rating nou5.0 0.43478260869565216
Tag a tractarandroid
Rating antic1.0 0.2608695652173913
add hits tag
Rating nou6.0 0.4782608695652174
****Secció a tractar: politics
Tag: politics 9
Tag: edmiliband 6
Tag: labour 4
Tag: series/politics-live-with-andrew-sparrow 3
Tag: blog 3
Tag: economy 3
Tag: davidcameron 2
Tag: conservatives 2
Tag: welfare 2
Tag: pmqs 1
entro addHits_Tag
Tag a tractarpolitics
Rating antic12.0 0.7391304347826086
add hits tag
Rating nou17.0 0.9565217391304348
Tag a tractaredmiliband
Rating antic4.0 0.391304347826087
add hits tag
Rating nou9.0 0.6086956521739131
Tag a tractarlabour
Rating antic2.0 0.30434782608695654
add hits tag
Rating nou7.0 0.5217391304347826
Tag a tractarseries/politics-live-with-andrew-sparrow
Rating antic1.0 0.2608695652173913
add hits tag
Rating nou6.0 0.4782608695652174
Tag a tractarblog
Rating antic1.0 0.2608695652173913
add hits tag
Rating nou6.0 0.4782608695652174
Tag a tractareconomy
Rating antic1.0 0.2608695652173913
add hits tag
Rating nou6.0 0.4782608695652174
Tag a tractardavidcameron
Rating antic4.0 0.391304347826087
add hits tag
Rating nou9.0 0.6086956521739131
Tag a tractarconservatives
Rating antic1.0 0.2608695652173913
add hits tag
Rating nou6.0 0.4782608695652174
```

```
Tag a tractarwelfare
Rating antic0.0 0.21739130434782608
add hits tag
Rating nou5.0 0.43478260869565216
Tag a tractarpmqs
Rating antic0.0 0.21739130434782608
add hits tag
Rating nou5.0 0.43478260869565216
Mida overranked 70
Mida loved 0
ITERACIÓ107
Escollida evolutive: 13
Escollida ideal: 13
Mida overranked 70
Mida loved 1
ITERACIÓ108
Escollida evolutive: 2
Escollida ideal: 0
Mida overranked 71
Mida loved 2
ITERACIÓ109
Escollida evolutive: 0
Escollida ideal: 0
Mida overranked 71
Mida loved 3
ITERACIÓ110
Escollida evolutive: 14
Escollida ideal: 12
Mida overranked 77
Mida loved 4
ITERACIÓ111
Escollida evolutive: 6
Escollida ideal: 13
Mida overranked 78
Mida loved 5
ITERACIÓ112
Escollida evolutive: 3
Escollida ideal: 0
Mida overranked 80
Mida loved 6
ITERACIÓ113
Escollida evolutive: 7
Escollida ideal: 14
Mida overranked 88
Mida loved 7
ITERACIÓ114
Escollida evolutive: 5
Escollida ideal: 4
Mida overranked 89
Mida loved 8
ITERACIÓ115
Escollida evolutive: 0
Escollida ideal: 14
Mida overranked 90
Mida loved 9
ITERACIÓ116
Escollida evolutive: 9
Escollida ideal: 9
Mida overranked 90
```

```
Mida loved 10
ITERACIÓ117
Escollida evolutive: 4
Escollida ideal: 6
Mida overranked 94
Mida loved 11
(...)

ITERACIÓ390
Escollida evolutive: 0
Escollida ideal: 1
Mida overranked 13
Mida loved 84
DISTÀNCIA sec: 1.295941043290216
DISTÀNCIA tag: 18.493673799014562
```

Anem a comparar la distància euclidiana entre seccions i *tags* que hi havia al principi amb la que ens retorna l'algorisme al final de la seva execució:

	Distància entre seccions	**Distància entre grups de *tags***
Inici	2.504180081383925	21.647480034974407
Final	1.295941043290216	18.493673799014562

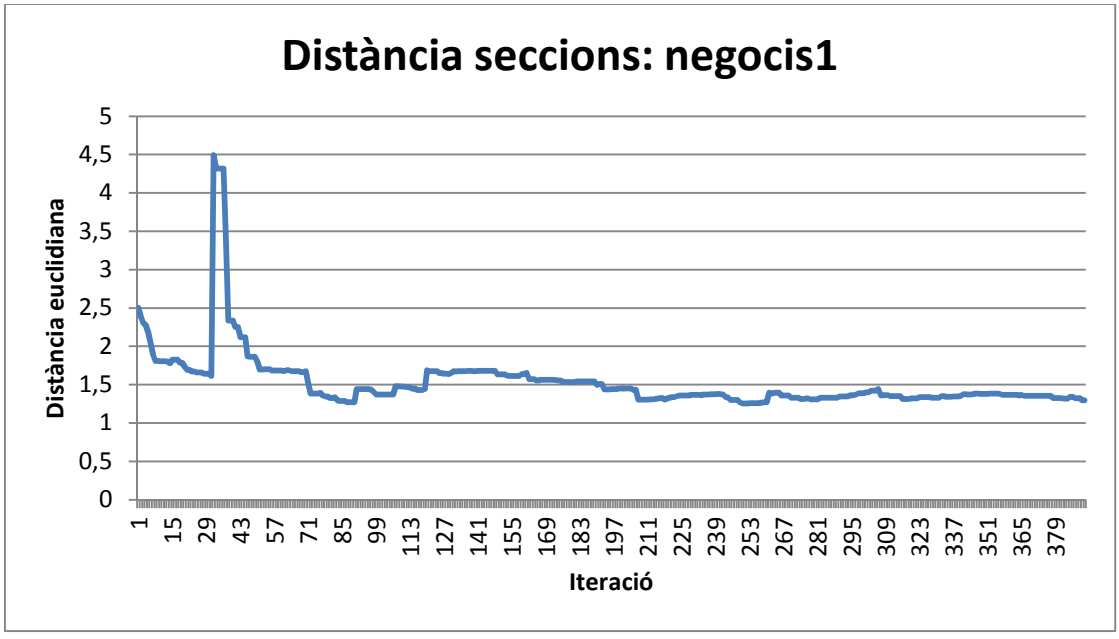


Fig. 4s

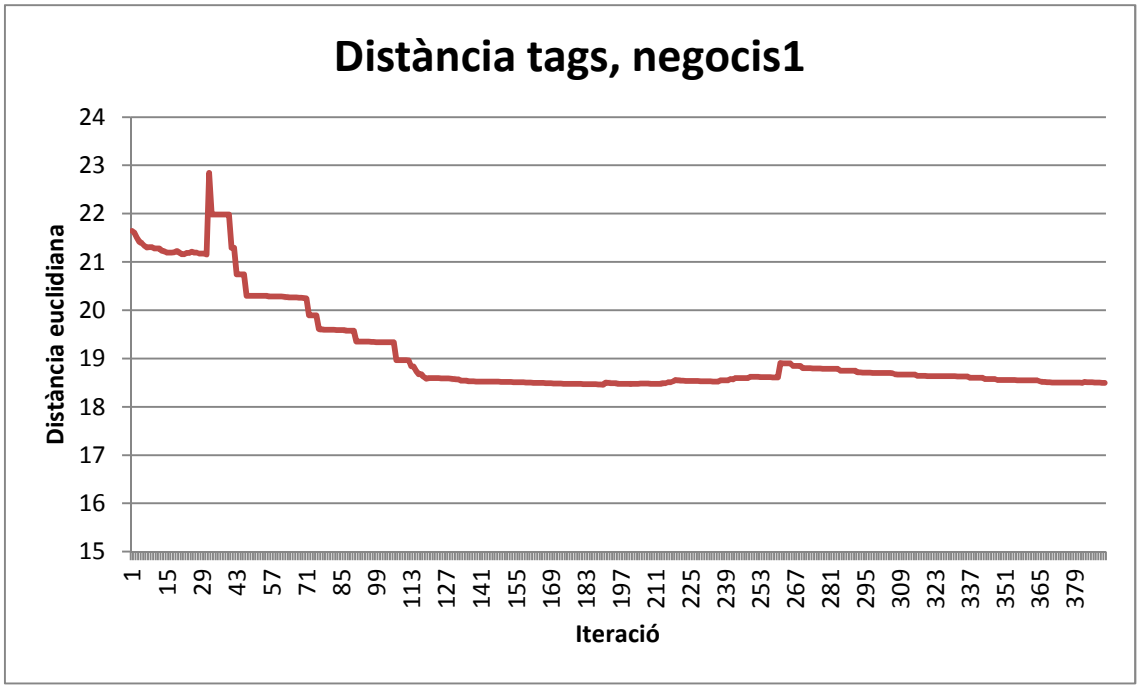


Fig. 4t

Podem veure a *Fig. 4s* i *Fig. 4t* que l'algorisme ha aconseguit que els dos perfils s'acostin després de l'algorisme d'aprenentatge, si bé podem comprovar que ha estat més eficaç l'aprenentatge de seccions –la distància s'ha reduït gairebé a la meitat- que no pas la distància entre *tags* –que s'ha reduït uns tres punts, si bé és cert que en el cas dels *tags* parlem de magnituds de distàncies més grans.

Segurament aquest fet es deu a què els perfils compten amb un nombre molt més elevat de *tags* (4.303 al corpus) que no de seccions (110 al corpus1), la qual cosa fa més difícil la classificació en aquest últim concepte. Probablement, caldria utilitzar un repositori d'articles més gran per millorar els resultats en aquest sentit

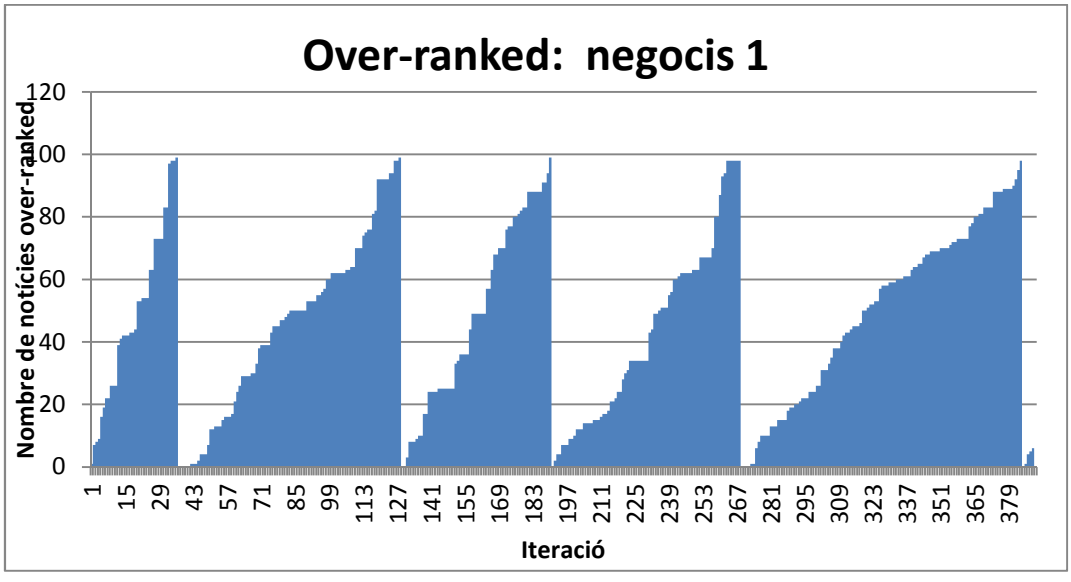


Fig. 4u

En la gràfica anterior –obtinguda a partir de les dades que el programa registra a learner.csv- podem observar un gràfic que ens il·lustra el fet que l'aprenentatge va cada vegada millor a mesura que anem avançant en les iteracions. Es pot apreciar la tendència que cada vegada costa més omplir el llistat de notícies sobrevalorades. Veiem que en les primers iteracions va molt ràpid i que, cap al final, la gràfica s'eixampla, fet que il·lustra que les eleccions cada vegada van essent més encertades i properes al perfil ideal. Es descarten cada cop menys notícies.

Una altra dada a tenir en compte és veure quines són les seccions principals dels articles que més escull el perfil ideal i quines tria el perfil evolutiu al llarg de tot l'algorisme d'aprenentatge. També a través de dades del fitxer learning.csv, podem obtenir la següent taula, en què podem veure els noms de les seccions i el nombre de tries que ha rebut cadascuna d'elles:

IDEAL PROFILE		EVOLUTIVE PROFILE	
commentisfree	69	commentisfree	74
business	38	business	56
media	32	media	38
politics	31	politics	38
world	28	world	28
sport	23	society	19
technology	20	technology	18
environment	15	environment	15
society	15	money	14
uk-news	11	uk-news	14
money	10	sport	14
global-development	9	film	11
higher-education-network	8	lifeandstyle	6
healthcare-network	7	tv-and-radio	6
football	6	football	5
teacher-network	6	global-development	4
tv-and-radio	6	music	3
global-development-professionals-network	5	higher-education-network	3
music	5	books	3
culture	4	artanddesign	3
education	4	culture-professionals-network	2
film	4	culture	2
artanddesign	3	science	2
culture-professionals-network	3	social-care-network	1
lifeandstyle	3	education	1
science	3	fashion	1
books	2	housing-network	1
housing-network	1	law	1
law	2	healthcare-network	1

s*tag*e	2	local-government-network	1
travel	3	crosswords	1
crosswords	1		
fashion	1		
small-business-network	1		
social-care-network	1		
social-enterprise-network	1		
theguardian	1		
theobserver	1		

Podem veure que 'comentisfree' acumula moltes seleccions. Segurament és així perquè els articles encabits en aquesta secció són transversals. Es tracta de l'apartat de discussió de notícies del web de The Guardian. Anem a veure dos exemples d'aquest tipus d'articles:

commentisfree/2014/jan/15/first-world-war-parallels,commentisfree,commentisfree/commentisfree world/firstworldwar politics/michaelgove politics/politics profile/editorial tone/comment tone/editorials theguardian/mainsection/editorialsandreply theguardian/mainsection type/article publication/theguardian

commentisfree/2014/jan/15/labour-bank-reform-comment,commentisfree,commentisfree/commentisfree politics/labour politics/edmiliband business/banking-reform business/banking business/financial-sector politics/politics uk/uk tone/comment tone/editorials profile/editorial theguardian/mainsection/editorialsandreply theguardian/mainsection type/article publication/theguardian

El fet de pertànyer a moltes seccions diferents fa que aquesta secció vagi acumulant molts punts a través de les puntuacions dels *tags*. D'aquesta manera, tot i no ser a les primeres posicions del perfil ideal, capta molta atenció en el moment de calcular la valoració d'un article. Sí que veiem, en canvi, que les altres tries lliguen amb el que hem definit abans al perfil ideal.

En la següent llista podem veure l'ordre de valoracions de les seccions al perfil ideal que havíem definit i com ha quedat aquest en el perfil après. A la tercera columna tenim el rànquing de seccions segons número d'articles al corpus (aparicions). Si un article té un *tag* d'aquella secció, ja es considera com a aparició. Només ensenyem les primeres seccions, ja que en tenim més d'un centenar i ens interessa analitzar amb profunditat les que tenen més valoració en el perfil ideal.

IDEAL PROFILE	EVOLUTIVE PROFILE	NUMBER OF ARTICLES	
business	world	world	6283
technology	sport	sport	3412
world	politics	football	2402
politics	media	uk	2032
culture	society	business	2002
music	business	politics	1748
uk	technology	society	1719
stage	uk	culture	1640
travel	environment	media	1497
sustainable-business	education	lifeandstyle	1368
film	commentisfree	technology	1259
tv-and-radio	culture	music	1235
environment	money	environment	1020
artanddesign	higher-education-network	books	917
lifeandstyle	tv-and-radio	film	901
education	football	theobserver	822
fashion	music	money	723
sport	artanddesign	education	524
media	global-development	commentisfree	522
money	healthcare-network	tv-and-radio	498
society	travel	science	417

Podem observar que, en termes generals, el perfil evolutiu ha après bastant quines eren les seccions preferides del perfil ideal. No surten en el mateix ordre però totes estan entre les vint primeres, que són les que es mostren a la taula. Cinc d'elles (world, politics, business, technology i uk) han quedat molt ben destacades.

No obstant això, veiem que les seccions ressaltades en vermell no han estat tan ben apreses i aquest fet es podria deure a què les seccions que tenen més articles relacionats puguin tenir un lleuger avantatge respecte les altres. Si fem una ullada al rànquing de la tercera columna podem veure que tant sport, media, com society són seccions bastant dominants quant a freqüència d'aparició. No obstant això, val a dir que també eren importants en el perfil ideal i que l'aprenentatge no és gens erroni, ja que estaven entre les vint primeres i hi ha més d'un centenar de seccions.

En la mateixa línia del que s'ha comentat en la distància mesurada en grups de *tags*, aquí tenim com ha anat l'aprenentatge de les etiquetes que s'havien destacat en aquest perfil ideal. Veiem que aquelles que compten amb més aparicions, són aquelles que l'algorisme ha après de la millor manera. S'han ressaltat en vermell i podem veure que s'acosten més a 1 que les altres, que en disten bastant. Sembla, doncs, que l'aprenentatge de *tags* va lligat al nombre d'aparicions, però caldrà anar-ho confirmant.

Aparicions	Valoració
35	technology/apple;11.0;0.39344262295081966
7	technology/efinance;1.0;0.22950819672131148
5	technology/series/on-social-media-marketing;0.0;0.21311475409836064
13	world/european-commission;2.0;0.2459016393442623
96	world/us-politics;10.0;0.3770491803278688
11	world/us-political-lobbying;1.0;0.22950819672131148
1	sustainable-business/emerging-markets;0.0;0.21311475409836064
5	sustainable-business/finance;0.0;0.21311475409836064
3	business/hedge-funds;0.0;0.21311475409836064
462	**business/business;26.0;0.639344262295082**
58	business/banking;10.0;0.3770491803278688
84	business/economics;5.0;0.29508196721311475
2	business/santander;0.0;0.21311475409836064
9	business/commodities;1.0;0.22950819672131148
3	business/luxury-goods-sector;0.0;0.21311475409836064
7	business/euro;5.0;0.29508196721311475
4	business/gas;0.0;0.21311475409836064
5	business/currencies;0.0;0.21311475409836064
1	business/bank-of-america;0.0;0.21311475409836064
501	**politics/politics;35.0;0.7868852459016393**
30	politics/foreignpolicy;5.0;0.29508196721311475
16	politics/blog;2.0;0.2459016393442623

Aquest procés s'ha repetit també utilitzant el corpus2 de notícies. El perfil ideal s'ha modelat de manera molt similar però no és exactament el mateix ja que, de la manera que s'ha dissenyat l'algorisme d'aprenentatge, l'estructura de perfil s'adapta al corpus de notícies que hi ha en cada moment. És a dir, no es pot utilitzar un perfil idènticament igual sobre dos corpus de dades, perquè no forçosament tots dos compten amb les mateixes seccions i *tags*, hi pot haver variacions segons la mostra recollida. Depèn de les notícies publicades en l'interval de temps escollit.

Al corpus2, no hem trobat tres dels *tags* que havíem ressaltat en el perfil d'home o dona de negocis (marcats en vermell). La resta s'han ressaltat de la mateixa manera:

```
SECCIÓ technology;40;0.0
technology/apple;0.0;1
technology/efinance;0.0;1
technology/series/on-social-media-marketing;0.0;1

SECCIÓ world;40;0.0
world/european-commission;0.0;1
world/us-politics;0.0;1
world/us-political-lobbying;0.0;1
```

```
sustainable-business/emerging-markets;0.0;1
sustainable-business/finance;0.0;1

SECCIÓ business;50;0.0
business/hedge-funds;0.0;1
business/business;0.0;1
business/banking;0.0;1
business/economics;0.0;1
business/santander;0.0;1
business/commodities;0.0;1
business/luxury-goods-sector;0.0;1
business/euro;0.0;1
business/gas;0.0;1
business/currencies;0.0;1
business/bank-of-america;0.0;1

SECCIÓ politics;30;0.0
politics/politics;0.0;1
politics/foreignpolicy;0.0;1
politics/blog;0.0;1

tv-and-radio/mad-men-tv-series;0.0;1
```

El resultat de l'aprenentatge és el següent:

	Distància entre seccions	Distància entre grups de *tags*
Inici	2.5799909224646513	22.306064134495458
Final	1.2117846613327106	19.16874776428098

Veiem també que la distància entre seccions s'ha reduït en quasi un 50% mentre que entre *tags* la reducció ha estat de tres punts. Gràficament ho opdem veure a *Fig. 4v* i *Fig. 4w*:

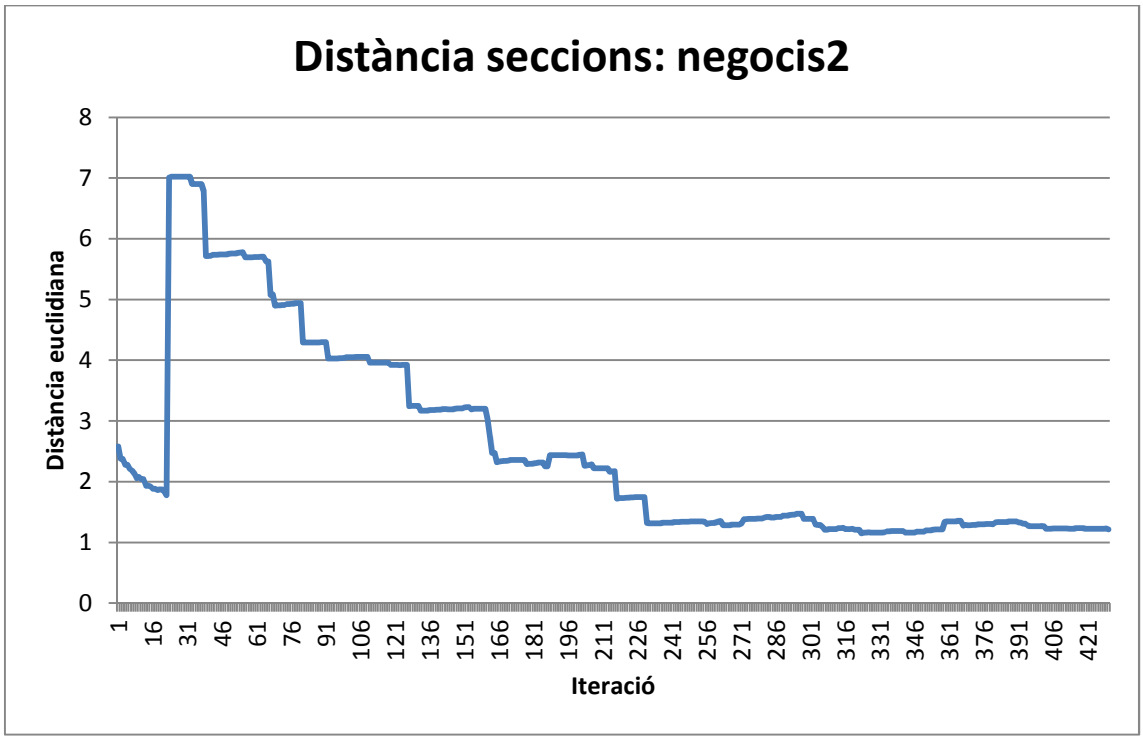


Fig. 4v

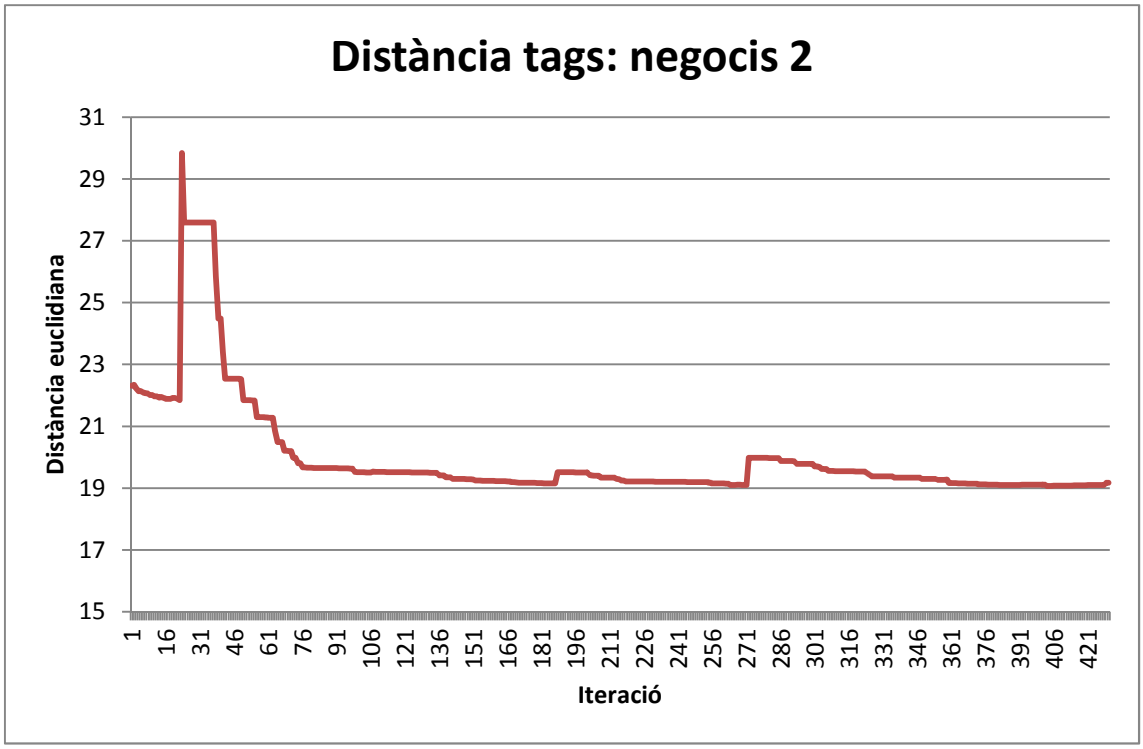


Fig. 4w

Amb el corpus1 havíem obtingut els següents resultats, molt similars.

	Distància entre seccions	Distància entre grups de *tags*
Inici	2.504180081383925	21.647480034974407
Final	1.3410420778192813	18.900561177553882

Veiem que el comportament d'overranked *[Fig. 4x]* també s'assembla bastant al de l'exemple anterior, sobretot en aquest cas al principi, que va molt ràpid i llavors es va relaxant:

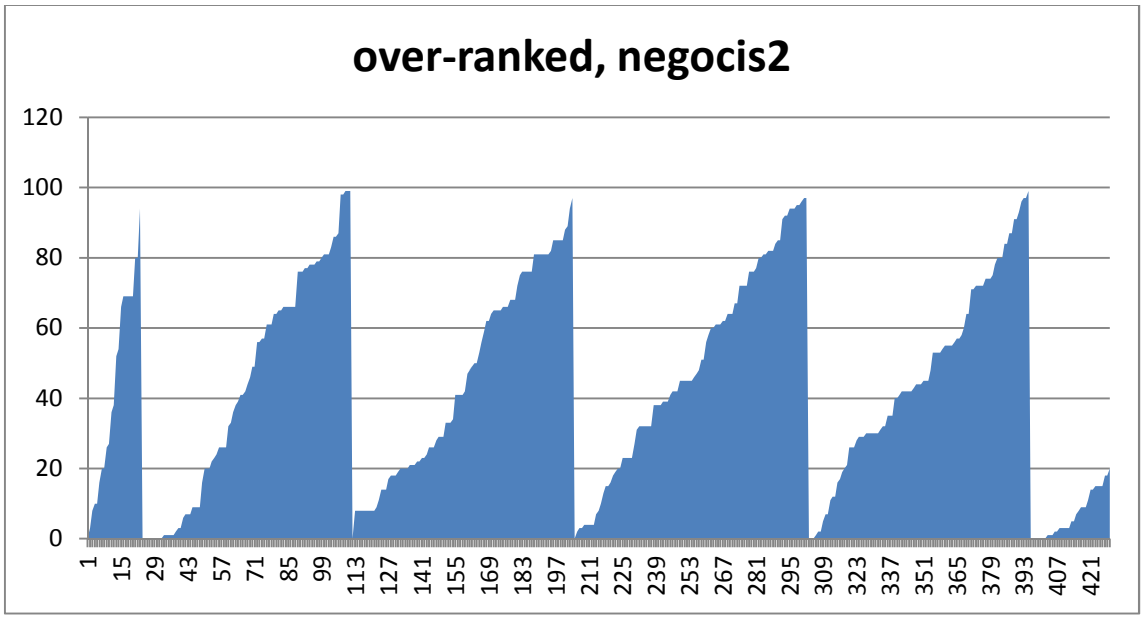


Fig. 4x

Veiem també que, en termes generals, s'han après les seccions que són més importants.

IDEAL PROFILE	EVOLUTIVE PROFILE	NUMBER OF ARTICLES	
business	business	world	6653
world	media	sport	3912
technology	politics	football	2267
politics	world	uk	2187
culture	society	politics	1929
artanddesign	culture	business	1890
tv-and-radio	uk	lifeandstyle	1713
sustainable-business	technology	culture	1711
music	education	media	1709
books	music	society	1689
media	artanddesign	environment	1442
fashion	books	technology	1341
film	film	music	1183
science	environment	books	1158
global-development	higher-education-network	film	898
stage	tv-and-radio	education	704
environment	global-development-professionals-network	money	649
education	global-development	artanddesign	629
sport	*stage*	science	526
travel	law	commentisfree	507
football	teacher-network	*stage*	461

Es repeteix també el comportament pel que fa als *tags*:

Aparicions	Valoració
36	technology/apple;4.0;0.22580645161290322
5	technology/efinance;0.0;0.16129032258064516
4	technology/series/on-social-media-marketing;0.0;0.16129032258064516
8	world/european-commission;1.0;0.1774193548387097
70	world/us-politics;6.0;0.25806451612903225
8	sustainable-business/emerging-markets;0.0;0.16129032258064516
8	sustainable-business/finance;0.0;0.16129032258064516
441	**business/business;37.0;0.7580645161290323**
56	business/banking;12.0;0.3548387096774194
80	business/economics;15.0;0.4032258064516129
1	business/santander;0.0;0.16129032258064516
3	business/commodities;1.0;0.1774193548387097
3	business/euro;0.0;0.16129032258064516

4	business/gas;1.0;0.1774193548387097
19	business/currencies;8.0;0.2903225806451613
544	politics/politics;40.0;0.8064516129032258
43	politics/foreignpolicy;3.0;0.20967741935483872
12	politics/blog;0.0;0.16129032258064516
	tv-and-radio/mad-men-tv-series;1.0;0.1774193548387097

4.5.2 Persona aficionada al cricket i als cavalls

La segona prova la pro*tag*onitzarà un perfil ideal d'una persona amant de l'esport amb especial sensibilitat per al cricket i els cavalls. És per això que donarem un pes especial a la secció d'esport de manera global i, en concret, a aquells *tags* relacionats amb el cricket i les curses de cavalls. Aquesta persona també és amant del futbol, secció a la qual donarem una valoració molt rellevant. També donarem pes a alguns altres *tags* relacionats amb l'esport. Totes aquestes dades s'introduiran fent una modificació –a través d'un editor de textos- d'un perfil ideal que s'ha creat de manera aleatòria amb el NewsRecommender.

Les seccions i *tags* a destacar són els de la següent llista. A les seccions els hem donat diversos pesos, tots més alts que el màxim pes que hi havia en el perfil aleatori. Als *tags* seleccionats els hem donat el valor màxim normalitzat (1).

```
SECCIÓ sport;50;0.0
sport/england-cricket-team;0.0;1
sport/cricket;0.0;1
sport/australia-cricket-team;0.0;1
sport/england-women-cricket-team;0.0;1
sport/australia-women-s-cricket-team;0.0;1
sport/new-zealand-cricket-team;0.0;1
sport/horse-racing;0.0;1
sport/horse-racing-tips;0.0;1
sport/series/talking-horses;0.0;1
sport/british-horseracing-authority;0.0;1

SECCIÓ football;40;0.0
football/football;0.0;1
football/world-cup-2014;0.0;1
football/laligafootball;0.0;1
football/neymar;0.0;1
football/lionel-messi;0.0;1
football/premierleague;0.0;1
football/europeanfootball;0.0;1
football/spain;0.0;1
football/uefa-europa-league;0.0;1
```

```
lifeandstyle/health-and-wellbeing;0.0;1

books/healthmindandbody;0.0;1

public-leaders-network/health-and-social-care;0.0;1

travel/healthandfitness;0.0;1

society/health;0.0;1
```

Un cop fetes les modificacions, carreguem el perfil:

```
================================================================
                        NEWS RECOMMENDER
================================================================
1.  Descarregar online corpus de notícies de The Guardian
2.  Guardar corpus de notícies carregat en un fitxer CSV
3.  Carregar un corpus de notícies des de CSV
4. Emmagatzemar perfil a un CSV
5. Carregar perfil de CSV
6.  Crear estructura de perfil
7.  Obtenir estadístiques del corpus
8.  Crear un perfil ideal amb dades aleatòries
9. Donar pes a seccions i tags del perfil ideal - Boost
9.  Crear perfil aprenentatge
10. Executar algorisme aprenentatge
11. Sortir

Escull una opció:
5
Escull una opció:
14. Carregar el perfil evolutiu
15. Carregar el perfil ideal
15
Escriu el nom de l'arxiu del perfil ideal, sense extensió
profileidealsport1
Perfil carregat satisfactòriament. Prem una tecla per veure'l

(...)

104.- Nom secció: teacher-network-advertisement-features/ Rating global: 0.0
Tags: teacher-network-advertisement-features (0.9148302308560714)
105.- Nom secció: housing-network-partner-zone-pinnacle/ Rating global: 0.0
Tags: housing-network-partner-zone-pinnacle (0.6523553835547434)
106.- Nom secció: sustainable-business-fairtrade-partner-zone/ Rating global:
0.0
Tags: sustainable-business-fairtrade-partner-zone (0.5075720589764134)
107.- Nom secció: partner-zone-sas-computacenter/ Rating global: 0.0
Tags: partner-zone-sas-computacenter (0.6223929583202732)
108.- Nom secció: adam-smith-international-partner-zone/ Rating global: 0.0
Tags: adam-smith-international-partner-zone (0.958100416455814)
109.- Nom secció: help/ Rating global: 0.0
Tags: help (0.4302097569240396)
```

Anem a executar l'algorisme d'aprenentatge:

```
================================================================
```

```
                        NEWS RECOMMENDER
================================================================
1.  Descarregar online corpus de notícies de The Guardian
2.  Guardar corpus de notícies carregat en un fitxer CSV
3.  Carregar un corpus de notícies des de CSV
4. Emmagatzemar perfil a un CSV
5. Carregar perfil de CSV
6.  Crear estructura de perfil
7.  Obtenir estadístiques del corpus
8.  Crear un perfil ideal amb dades aleatòries
9. Donar pes a seccions i tags del perfil ideal - Boost
9.  Crear perfil aprenentatge
10. Executar algorisme aprenentatge
11. Sortir

Escull una opció:
10
DISTÀNCIA sec: 2.373530172548897
DISTÀNCIA tag: 21.912553382012277
ITERACIÓ1
Escollida evolutive: 0
Escollida ideal: 0
Mida overranked 0
Mida loved 1

ITERACIÓ2
Escollida evolutive: 0
Escollida ideal: 13
Mida overranked 13
Mida loved 2
```

```
(...)

ITERACIÓ390
Escollida evolutive: 0
Escollida ideal: 1
Mida overranked 79
Mida loved 90
DISTÀNCIA sec: 1.1318963142745508
DISTÀNCIA tag: 18.771392592218767

====================================
```

Anem a comparar la distància euclidiana entre seccions i *tags* que hi havia al principi amb la que ens retorna l'algorisme al final de la seva execució:

	Distància entre seccions	**Distància entre grups de *tags***
Inici	2.373530172548897	21.912553382012277
Final	1.1318963142745508	18.771392592218767

Podem veure que l'algorisme ha aconseguit que els dos perfils s'acostin després de l'algorisme d'aprenentatge, en el cas de les seccions a menys de la meitat i a una diferència d'uns tres punts en el cas dels *tags*.

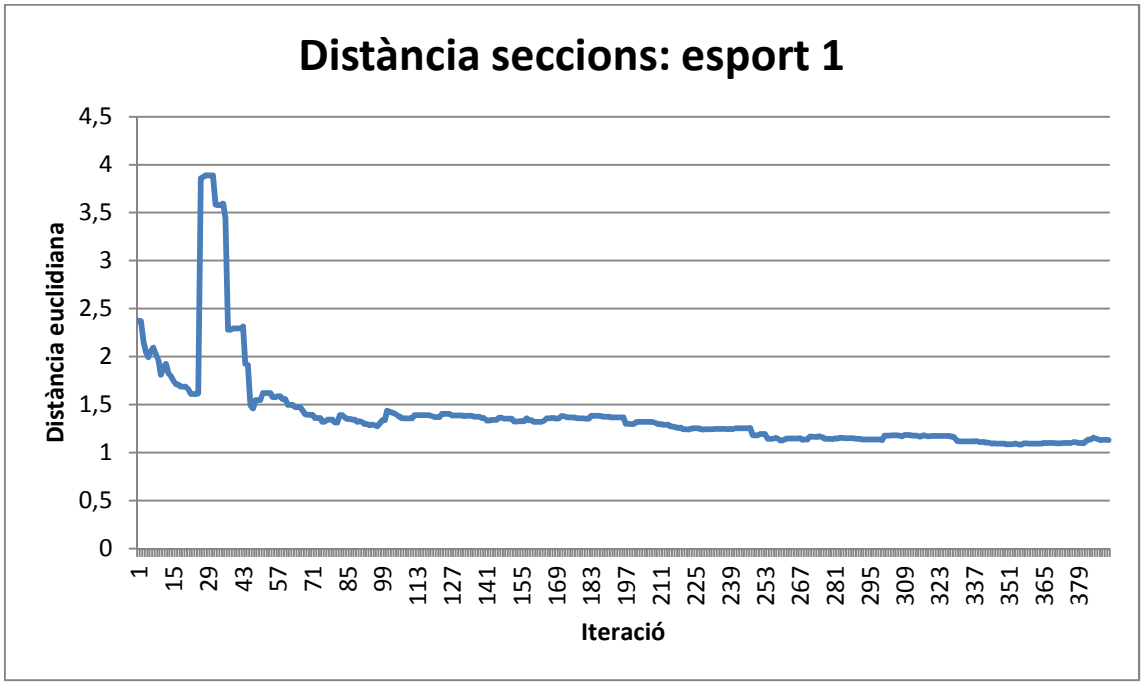


Fig. 4y

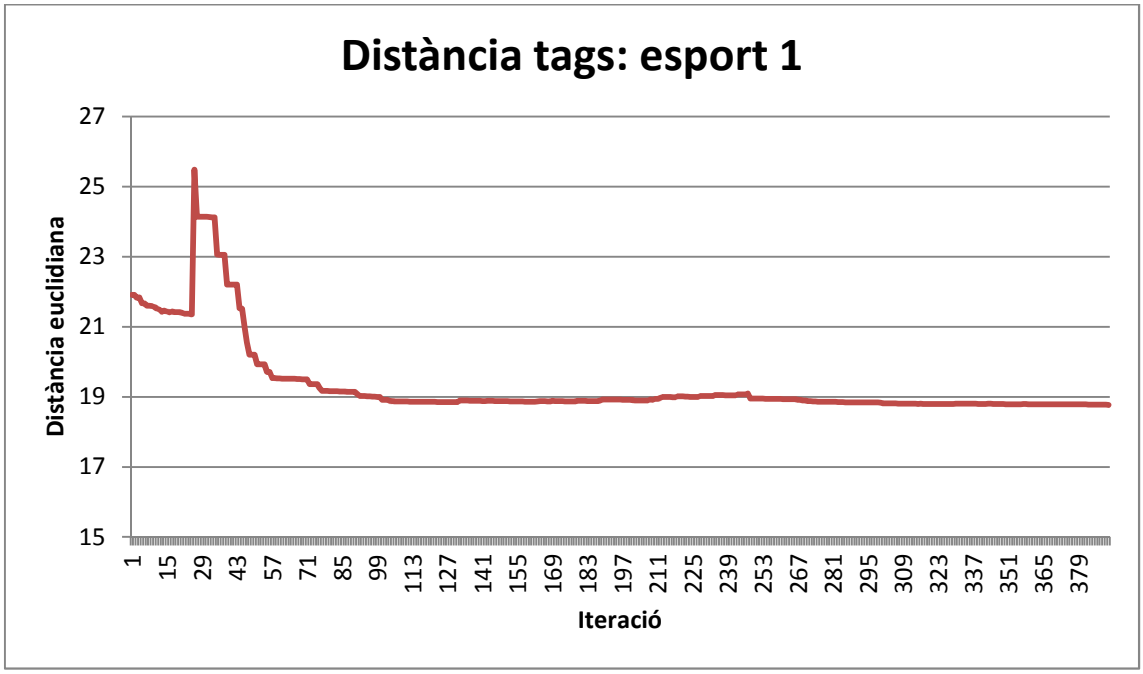


Fig. 4z

Veiem que les dues corbes *[Fig. 4y i Fig. 4z]* tenen tendència a anar-se acostant cada vegada més al perfil ideal malgrat l'ajustament inicial.

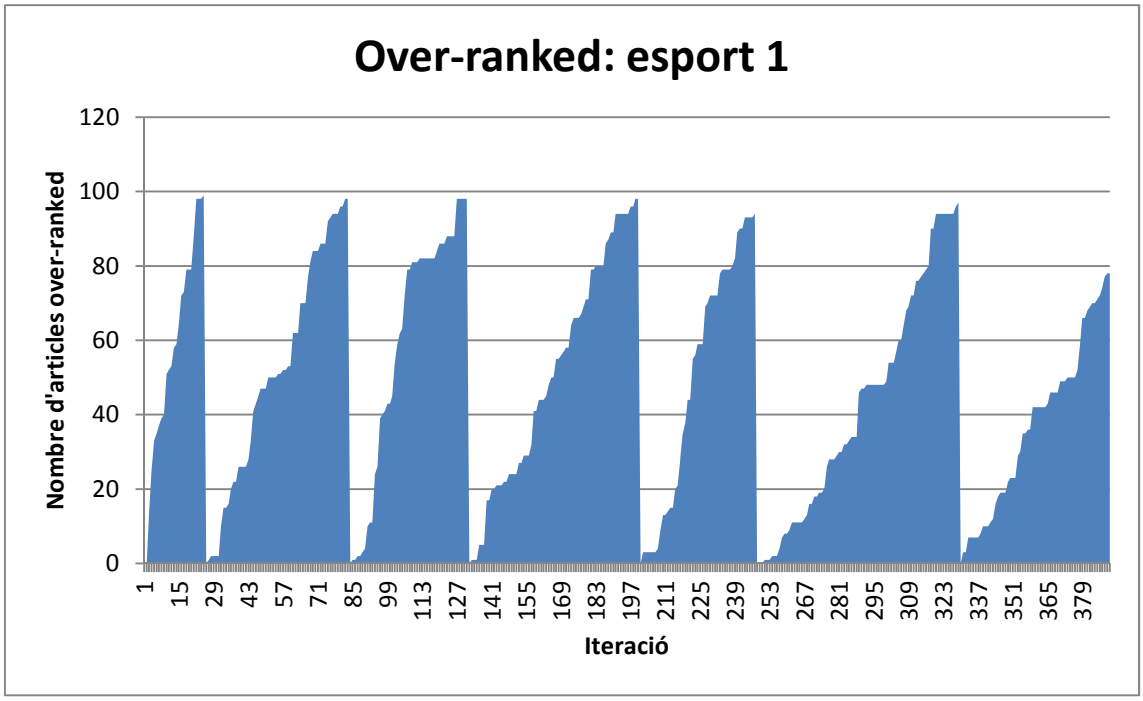


Fig. 4a'

Veiem, una vegada més, a *Fig. 4a'* que el repositori de notícies sobrevalorades té tendència a anar-se omplint de manera més lenta en les últimes iteracions. Veiem ara les tries de seccions que han fet ambdós perfils al llarg de totes les iteracions:

IDEAL PROFILE		EVOLUTIVE PROFILE	
sport	70	commentisfree	74
commentisfree	67	sport	72
football	28	society	36
society	23	politics	31
world	19	world	21
politics	19	business	21
environment	14	environment	22
uk-news	13	media	21
business	10	film	14
film	12	uk-news	13
media	9	football	10
lifeandstyle	8	money	10
money	8	technology	5
technology	8	lifeandstyle	5
global-development	7	global-development	4
higher-education-network	6	music	3
fashion	6	science	3
music	6	tv-and-radio	3
social-care-network	5	culture	2
artanddesign	5	higher-education-network	2

science	4	books	2
voluntary-sector-network	4	fashion	2
healthcare-network	4	law	2
law	2	healthcare-network	1
culture-professionals-network	3	travel	1
sustainable-business	3	global-development-professionals-network	1
books	3	social-care-network	1
culture	3	housing-network	1
global-development-professionals-network	1	voluntary-sector-network	1
housing-network	1	artanddesign	1
travel	2	local-government-network	1
public-leaders-network	2	crosswords	1
social-enterprise-network	2		
theobserver	2		
tv-and-radio	1		
local-government-network	1		
teacher-network	1		
crosswords	1		

Notem que 'sport' i 'commentisfree' lideren el rànquing en tots dos casos. La primera una de les preferides pel perfil ideal i la segona, una secció transversal com ja s'ha comentat en l'apartat anterior. D'altra banda, veiem que 'football' també acomula força seleccions.

En la següent llista podem veure l'ordre de les valoracions de les seccions al perfil ideal que havíem definit i com ha quedat aquest en el perfil après, les que compten amb una millor puntuació ja que no es mostren totes. A la tercera columna tenim el rànquing de seccions segons número d'articles al corpus (aparicions).

IDEAL PROFILE	EVOLUTIVE PROFILE	NUMBER OF ARTICLES	
sport	sport	world	6283
football	society	sport	3412
politics	world	football	2402
world	politics	uk	2032
environment	football	business	2002
fashion	uk	politics	1748
film	environment	society	1719
media	commentisfree	culture	1640
tv-and-radio	technology	media	1497
uk	media	lifeandstyle	1368
society	lifeandstyle	technology	1259
lifeandstyle	culture	music	1235
education	music	environment	1020
books	business	books	917

business	money	film	901
technology	film	theobserver	822
commentisfree	law	money	723
science	education	education	524
music	artanddesign	commentisfree	522
money	higher-education-network	tv-and-radio	498
culture	fashion	science	417
travel	global-development	artanddesign	414

En general podem dir que aprèn bé les seccions. En el cas de 'sport', veiem com el perfil evolutiu la té considerada a la primera posició. Quan a 'football', també veiem que està bastant amunt. En vermell veiem algunes imperfeccions en les primeres posicions tot i que, cal destacar, que gairebé totes apareixen en els dos rànquings, encara que sigui en posicions diferenciades.

Anem a veure com ha funcionat l'aprenentatge dels *tags*:

Aparicions	Valoració
79	sport/england-cricket-team;22.0;0.47761194029850745
113	sport/cricket;25.0;0.5223880597014925
72	sport/australia-cricket-team;24.0;0.5074626865671642
7	sport/england-women-cricket-team;2.0;0.1791044776119403
8	sport/australia-women-s-cricket-team;2.0;0.1791044776119403
1	sport/new-zealand-cricket-team;0.0;0.14925373134328357
51	sport/horse-racing;3.0;0.19402985074626866
26	sport/horse-racing-tips;2.0;0.1791044776119403
13	sport/series/talking-horses;2.0;0.1791044776119403
2	sport/british-horseracing-authority;0.0;0.14925373134328357
540	football/football;19.0;0.43283582089552236
21	football/world-cup-2014;0.0;0.14925373134328357
12	football/laligafootball;1.0;0.16417910447761194
1	football/neymar;0.0;0.14925373134328357

6	football/lionel-messi;0.0;0.14925373134328357
102	football/premierleague;3.0;0.19402985074626866
34	football/europeanfootball;2.0;0.1791044776119403
1	football/spain;0.0;0.14925373134328357
1	football/uefa-europa-league;0.0;0.14925373134328357
56	lifeandstyle/health-and-wellbeing;7.0;0.2537313432835821
7	books/healthmindandbody;1.0;0.16417910447761194
2	public-leaders-network/health-and-social-care;0.0;0.14925373134328357
3	travel/healthandfitness;0.0;0.14925373134328357
155	society/health;19.0;0.43283582089552236

Podem veure en aquest exemple que no es veu una relació tan clara com l'anterior entre el bon aprenentatge dels *tags* i el nombre d'aparicions d'aquests. Les tres millors valoracions ronden els 0,5 punts i no són els *tags* que tenen més aparicions. Per exemple, football/football en té moltes més i no els supera en puntuació.

Aquest procés s'ha repetit també utilitzant el corpus2 de notícies. El perfil ideal s'ha modelat de manera molt similar però no és exactament el mateix ja que, de la manera que s'ha dissenyat l'algorisme d'aprenentatge, l'estructura de perfil s'adapta al corpus de notícies que hi ha en cada moment. És a dir, no es pot utilitzar un perfil idènticament igual sobre dos corpus de dades, perquè no forçosament tots dos compten amb les mateixes seccions i *tags*, hi pot haver variacions segons la mostra recollida.

Una vegada més, els *tags* ressaltats són els següents:

```
SECCIÓ sport;50;0.0
sport/england-cricket-team;0.0;1
sport/cricket;0.0;1
sport/australia-cricket-team;0.0;1
sport/england-women-cricket-team;0.0;1
sport/australia-women-s-cricket-team;0.0;1
sport/new-zealand-cricket-team;0.0;1
sport/horse-racing;0.0;1
sport/horse-racing-tips;0.0;1
sport/series/talking-horses;0.0;1
sport/british-horseracing-authority;0.0;1

SECCIÓ football;40;0.0
football/football;0.0;0.6835688449942317
```

```
football/world-cup-2014;0.0;1
football/laligafootball;0.0;1
football/neymar;0.0;1
football/lionel-messi;0.0;1
football/premierleague;0.0;1
football/europeanfootball;0.0;1
football/spain;0.0;1
football/uefa-europa-league;0.0;1

lifeandstyle/health-and-wellbeing;0.0;1

books/healthmindandbody;0.0;1

public-leaders-network/health-and-social-care;0.0;1

travel/healthandfitness;0.0;1

society/health;0.0;1
```

El resultat de l'aprenentatge és el següent:

	Distància entre seccions	**Distància entre grups de *tags***
Inici	2.374873074503141	21.772033380401965
Final	1.204759286984887	18.659963074507342

Gràficament, veiem *Fig. 4b'* i *Fig. 4c'*:

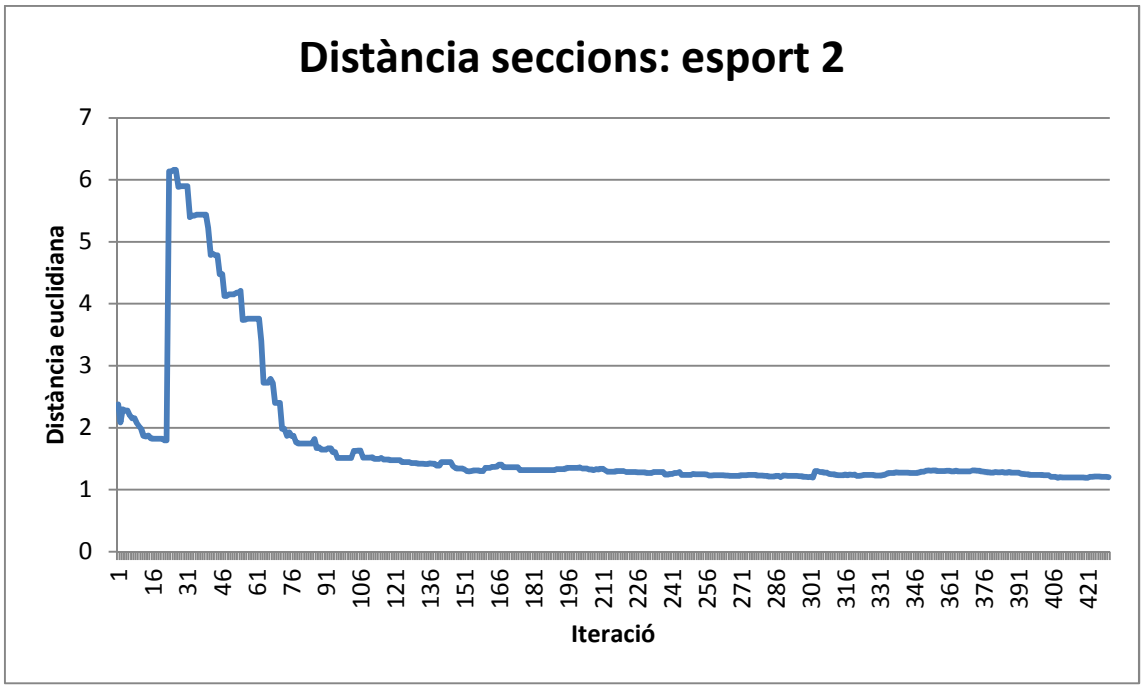


Fig. 4b'

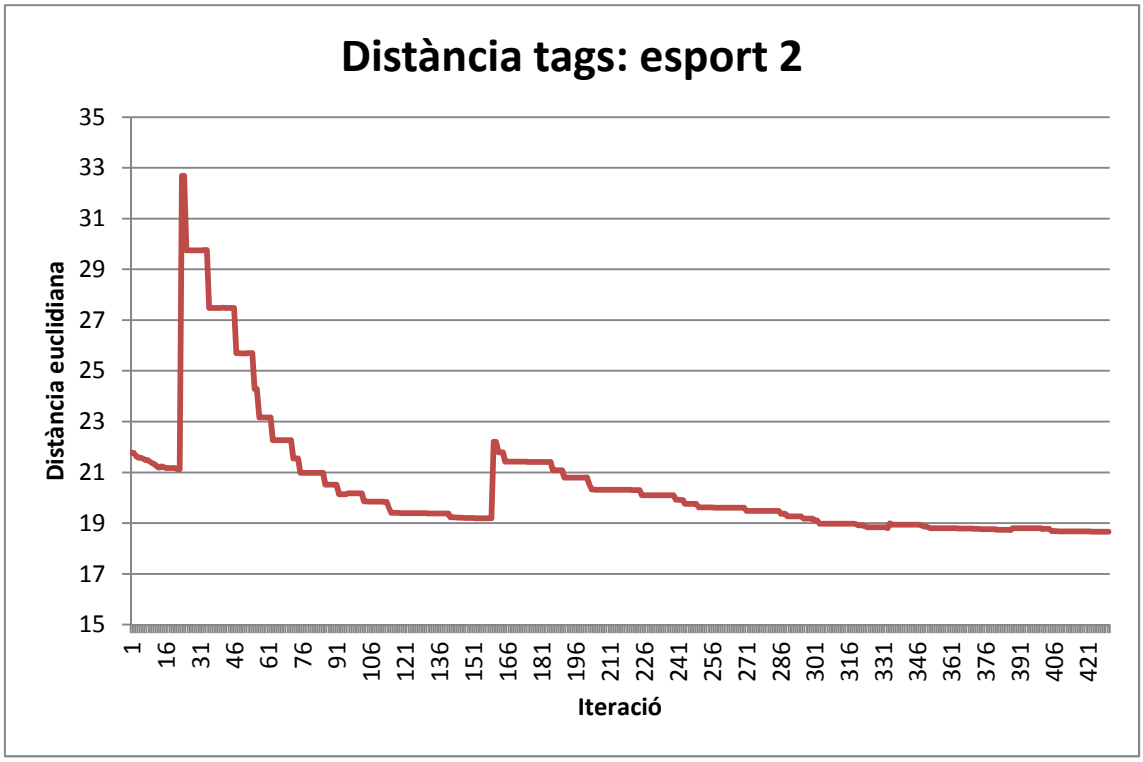


Fig. 4c'

Mentre que el resultat obtingut amb el corpus1 era el següent, molt similar:

	Distància entre seccions	**Distància entre grups de *tags***
Inici	2.373530172548897	21.912553382012277
Final	1.1318963142745508	18.771392592218767

Fem una ullada ara a com queda la gràfica dels overranked *[Fig. 4d']* :

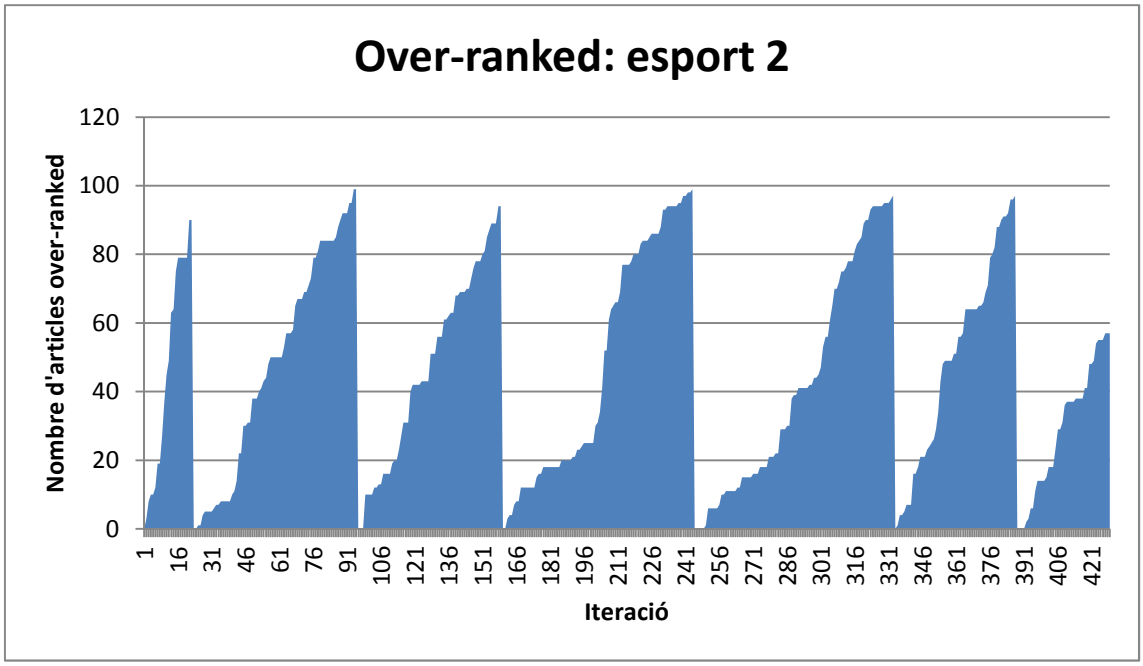


Fig. 4d'

Veiem que, en termes generals, es compleix el fet que el grup tarda més en créixer i ho fa de manera més lenta a mesura que avancen les iteracions. Veiem com s'ha dut a terme l'aprenentatge de les seccions:

IDEAL PROFILE	EVOLUTIVE PROFILE
sport	society
football	sport
film	football
education	world
tv-and-radio	media
money	culture
music	technology
business	uk
law	business
sustainable-business	tv-and-radio
culture	film
society	environment
environment	politics
fashion	higher-education-network
stage	lifeandstyle
science	education
lifeandstyle	books
media	artanddesign
higher-education-network	science
books	music
technology	local-government-network
politics	healthcare-network

NUMBER OF ARTICLES	
world	6283
sport	3412
football	2402
uk	2032
business	2002
politics	1748
society	1719
culture	1640
media	1497
lifeandstyle	1368
technology	1259
music	1235
environment	1020
books	917
film	901
theobserver	822
money	723
education	524
commentisfree	522
tv-and-radio	498
science	417
artanddesign	414

Fem una ullada ara a com han quedat els *tags* que havíem ressaltat en el perfil evolutiu.

Aparicions	Tag
56	sport/england-cricket-team;4.0;0.23333333333333334
95	sport/cricket-the-old-batsman-blog;0.0;0.16666666666666666
25	sport/australia-cricket-team;2.0;0.2
4	sport/england-women-cricket-team;0.0;0.16666666666666666
3	sport/australia-women-s-cricket-team;0.0;0.16666666666666666
4	sport/new-zealand-cricket-team;0.0;0.16666666666666666
53	sport/horse-racing;3.0;0.21666666666666667
30	sport/horse-racing-tips;3.0;0.21666666666666667
15	sport/series/talking-horses;3.0;0.21666666666666667
5	sport/british-horseracing-authority;0.0;0.16666666666666666
520	**football/football;16.0;0.43333333333333335**
29	football/world-cup-2014;4.0;0.23333333333333334
17	football/laligafootball;7.0;0.2833333333333333
3	football/neymar;0.0;0.16666666666666666
8	football/lionel-messi;-2.0;0.13333333333333333
154	football/premierleague;7.0;0.2833333333333333
40	football/europeanfootball;2.0;0.2
6	football/spain;0.0;0.16666666666666666
4	football/uefa-europa-league;0.0;0.16666666666666666
70	lifeandstyle/health-and-wellbeing;10.0;0.3333333333333333
5	books/healthmindandbody;0.0;0.16666666666666666
3	public-leaders-network/health-and-social-care;0.0;0.16666666666666666
5	travel/healthandfitness;1.0;0.18333333333333332
185	**society/health;28.0;0.6333333333333333**

Veiem que, en general, no s'obté un bon aprenentatge dels *tags* i que els dos que s'acosten més a 1 en disten bastant.

4.5.3 Hipster

En el moment de l'elaboració d'aquest projecte, la cultura hipster urbana arrasa a les ciutats, allunyada dels corrents predominants, que aquest grup social considera 'mainstream'. En aquest apartat hem intentat crear un perfil d'usuari que reflecteixi

aquesta realitat. S'ha donat un pes especial al cinema, a l'art i al disseny, a la moda, la cultura i la música.

Les seccions i *tags* a destacar són els de la següent llista. A les seccions els hem donat diversos pesos, tots més alts que el màxim pes que hi havia en el perfil aleatori. Als *tags* seleccionats els hem donat el valor màxim normalitzat (1). Treballem per començar amb el corpus1:

```
SECCIÓ film;60;0.0
film/filmblog;0.0;1
film/martinscorsese;0.0;1
film/woodyallen;0.0;1
film/quentintarantino;0.0;1
film/periodandhistorical;0.0;1
film/documentary;0.0;1
film/audrey-tautou;0.0;1
film/film-criticism;0.0;1
film/alfredhitchcock;0.0;1
film/silent-film;0.0;1

SECCIÓ artanddesign;60;0.0
artanddesign/photography;0.0;1
artanddesign/art;0.0;1
artanddesign/posters;0.0;1
artanddesign/series/pictures-from-the-past;0.0;1
artanddesign/video-art;0.0;1
artanddesign/design;0.0;1
artanddesign/freud;0.0;1
artanddesign/banksy;0.0;1
artanddesign/lichtenstein;0.0;1
artanddesign/streetart;0.0;1


SECCIÓ fashion;30;0.0
fashion/fashion;0.0;1
fashion/series/vintage-years;0.0;1
fashion/fashion-blog;0.0;1

uk/london-underground;0.0;1
culture/lena-dunham;0.0;1
culture/zombies;0.0;1
music/electronicmusic;0.0;1
music/musicblog;0.0;1
music/urban;0.0;0.1
music/arcticmonkeys;0.0;1
music/indie;0.0;1
music/clubs;0.0;1
music/the-1975;0.0;1
music/series/newbandoftheday;0.0;1
music/music-festivals;0.0;1
music/stan-getz;0.0;1
```

Un cop fetes les modificacions, carreguem el perfil:

```
================================================================
                        NEWS RECOMMENDER
================================================================
1.  Descarregar online corpus de notícies de The Guardian
2.  Guardar corpus de notícies carregat en un fitxer CSV
3.  Carregar un corpus de notícies des de CSV
4. Emmagatzemar perfil a un CSV
5. Carregar perfil de CSV
6.  Crear estructura de perfil
7.  Obtenir estadístiques del corpus
8.  Crear un perfil ideal amb dades aleatòries
9. Donar pes a seccions i tags del perfil ideal - Boost
9.  Crear perfil aprenentatge
10. Executar algorisme aprenentatge
11. Sortir

Escull una opció:
5
Escull una opció:
14. Carregar el perfil evolutiu
15. Carregar el perfil ideal
15
Escriu el nom de l'arxiu del perfil ideal, sense extensió
profilealeatorihipster1
Perfil carregat satisfactòriament. Prem una tecla per veure'l

(...)

101.- Nom secció: social-care-network-skills-for-care-partner-zone/ Rating
global: 0.0
Tags: social-care-network-skills-for-care-partner-zone (0.7199288700917582)
102.- Nom secció: teacher-network-hays-partner-zone/ Rating global: 0.0
Tags: teacher-network-hays-partner-zone (0.05431543944826289)
103.- Nom secció: direct-line-for-business-partner-zone/ Rating global: 0.0
Tags: direct-line-for-business-partner-zone (0.2886875455331993)
104.- Nom secció: teacher-network-advertisement-features/ Rating global: 0.0
Tags: teacher-network-advertisement-features (0.5301581783988504)
105.- Nom secció: housing-network-partner-zone-pinnacle/ Rating global: 0.0
Tags: housing-network-partner-zone-pinnacle (0.22708371441698338)
106.- Nom secció: sustainable-business-fairtrade-partner-zone/ Rating global:
0.0
Tags: sustainable-business-fairtrade-partner-zone (0.29410418805256533)
107.- Nom secció: partner-zone-sas-computacenter/ Rating global: 0.0
Tags: partner-zone-sas-computacenter (0.13403971548438087)
108.- Nom secció: adam-smith-international-partner-zone/ Rating global: 0.0
Tags: adam-smith-international-partner-zone (0.33125520565312694)
109.- Nom secció: help/ Rating global: 0.0
Tags: help (0.33880629852876154)
```

Anem a executar l'algorisme d'aprenentatge:

```
================================================================
                        NEWS RECOMMENDER
================================================================
1.  Descarregar online corpus de notícies de The Guardian
2.  Guardar corpus de notícies carregat en un fitxer CSV
3.  Carregar un corpus de notícies des de CSV
4. Emmagatzemar perfil a un CSV
5. Carregar perfil de CSV
```

```
6.  Crear estructura de perfil
7.  Obtenir estadístiques del corpus
8.  Crear un perfil ideal amb dades aleatòries
9. Donar pes a seccions i tags del perfil ideal - Boost
9.  Crear perfil aprenentatge
10. Executar algorisme aprenentatge
11. Sortir

Escull una opció:
10
DISTÀNCIA sec: 2.4639632933773807
DISTÀNCIA tag: 21.912613703483814
ITERACIÓ1
Escollida evolutive: 0
Escollida ideal: 0
Mida overranked 0
Mida loved 1

ITERACIÓ2
Escollida evolutive: 0
Escollida ideal: 11
Mida overranked 11
Mida loved 2

(...)

ITERACIÓ389
Escollida evolutive: 12
Escollida ideal: 12
Mida overranked 86
Mida loved 88
ITERACIÓ390
Escollida evolutive: 0
Escollida ideal: 0
Mida overranked 86
Mida loved 89
DISTÀNCIA sec: 1.5791946016179153
DISTÀNCIA tag: 18.789256609695535
```

Anem a comparar la distància euclidiana entre seccions i *tags* que hi havia al principi amb la que ens retorna l'algorisme al final de la seva execució:

	Distància entre seccions	**Distància entre grups de *tags***
Inici	2.4639632933773807	21.912613703483814
Final	1.5791946016179153	18.789256609695535

Veiem que la reducció de la distància entre seccions ha estat significativa tot i que no tant com en les altres proves que havíem dut a terme fins ara. La distància entre el grup de *tags* sí que segueix la tendència general.

Fixem-nos en les corbes d'aprenentatge a *Fig. 4e'* i *Fig 4f'*:

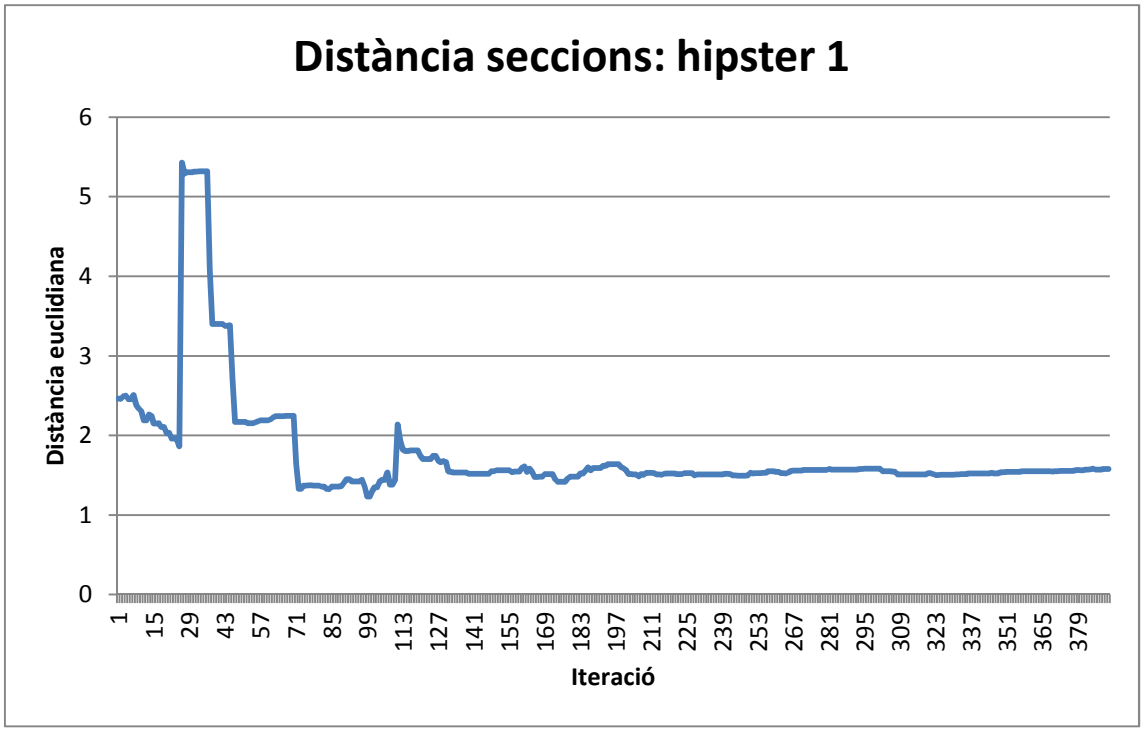


Fig. 4e'

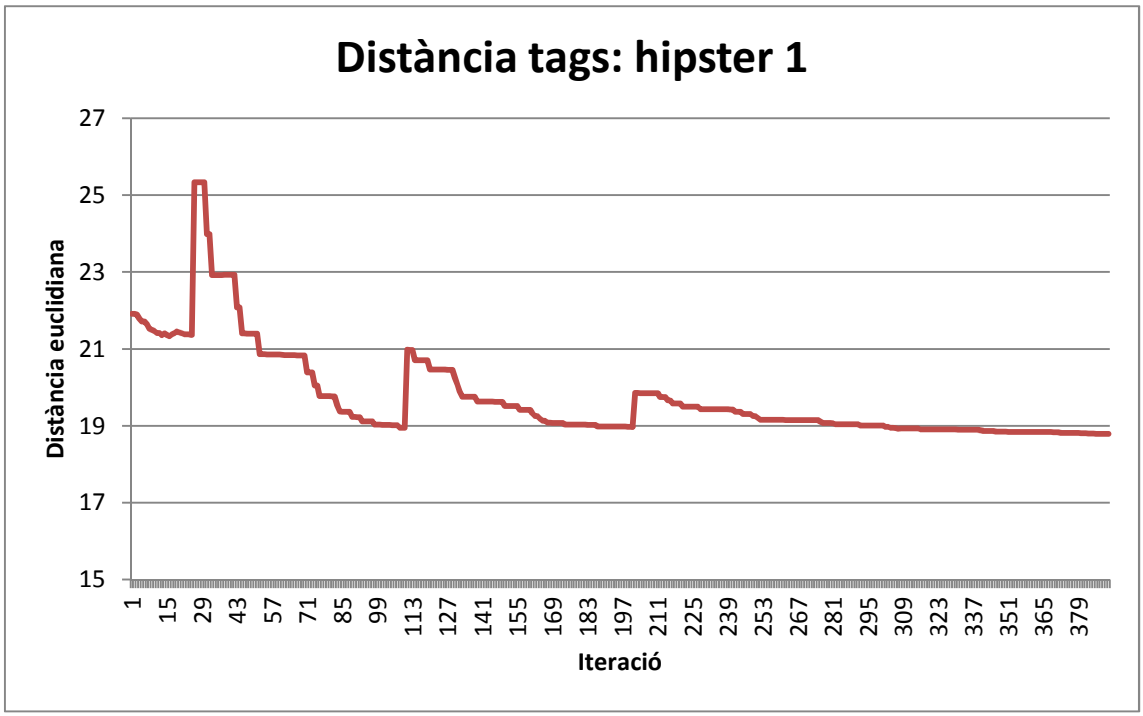


Fig. 4f'

Veiem que els gràfics segueixen la tendència general, de la mateixa manera que el de la mida del conjunt de notícies overranked *[Fig 4g']* :

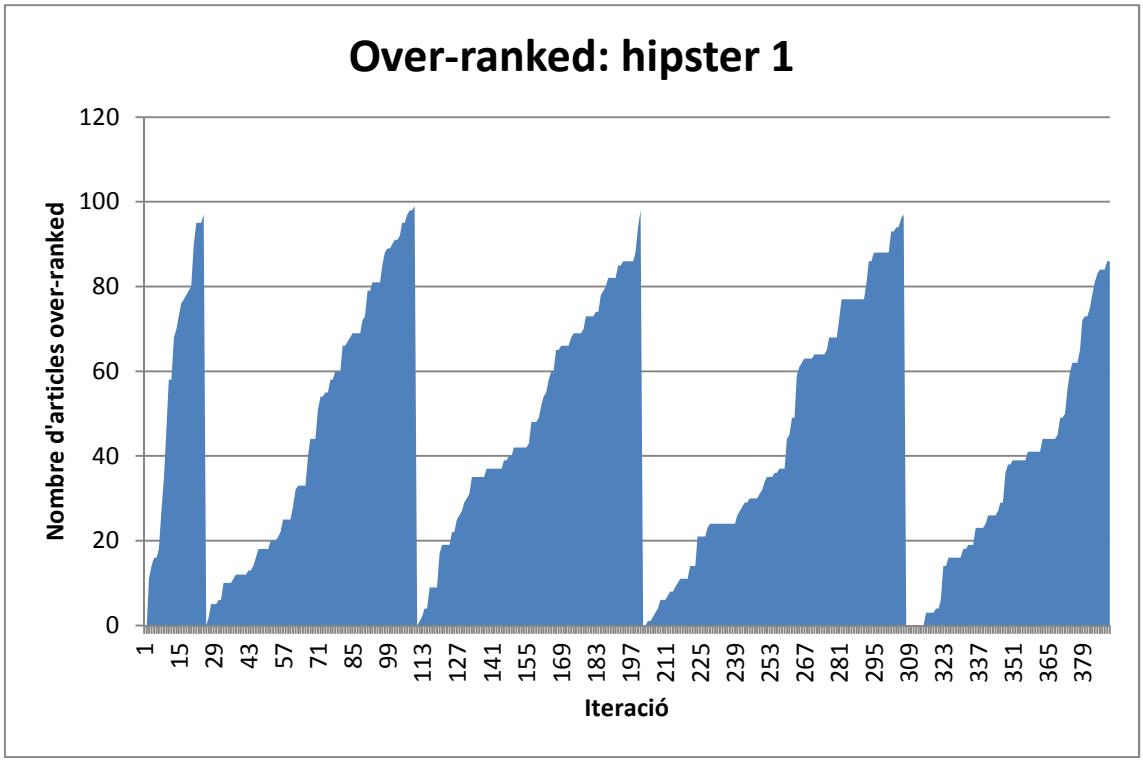


Fig. 4g'

Quant al nombre de seleccions de cada secció que cadascun dels perfils va fent al llarg de les iteracions, obtenim el següent:

IDEAL PROFILE		EVOLUTIVE PROFILE	
commentisfree	47	commentisfree	63
politics	33	politics	44
music	28	society	34
society	24	film	25
film	23	business	23
environment	22	environment	24
business	15	world	20
artanddesign	18	music	22
world	16	media	18
uk-news	13	sport	12
culture	14	uk-news	12
money	11	culture	11
media	11	money	8
childrens-books-site	11	artanddesign	8
sport	7	football	7
law	6	technology	7
technology	7	lifeandstyle	7
stage	7	books	7
football	5	global-development	4
social-care-network	5	science	4
lifeandstyle	5	higher-education-network	3
global-development	5	tv-and-radio	3

healthcare-network	5	travel	2
books	4	theobserver	2
travel	4	fashion	2
culture-professionals-network	4	law	2
science	4	childrens-books-site	2
tv-and-radio	4	*stage*	2
fashion	4	news	0
higher-education-network	2	theguardian	0
social-enterprise-network	3	global-development-professionals-network	1
housing-network	1	culture-professionals-network	1
education	2	social-care-network	1
teacher-network	2	housing-network	1
local-government-network	2	local-government-network	1
theguardian	0	crosswords	1
global	1		
voluntary-sector-network	1		
theobserver	1		
crosswords	1		

Veiem que es repeteix el fenomen de 'commentisfree' però que, a grans trets, coincideixen les seccions que acumulen més nombre de seleccions en amdós perfils. En les primeres posicions podem veure 'film' i 'music', que eren dues de les seccions ressaltades, si bé 'fashion' i 'artandesign' queden una mica més avall.

En la següent llista podem veure l'ordre de valoracions de les seccions al perfil ideal que havíem definit i com ha quedat aquest en el perfil après. A la tercera columna tenim el rànquing de seccions segons el número d'articles al corpus (aparicions).

IDEAL PROFILE	EVOLUTIVE PROFILE	NUMBER OF ARTICLES	
film	politics	world	6283
artanddesign	society	sport	3412
culture	music	football	2402
music	culture	uk	2032
fashion	world	business	2002
tv-and-radio	uk	politics	1748
society	environment	society	1719
education	sport	culture	1640
science	artanddesign	media	1497
travel	film	lifeandstyle	1368
media	media	technology	1259
business	technology	music	1235
politics	money	environment	1020
sport	commentisfree	books	917
football	lifeandstyle	film	901

world	tv-and-radio	theobserver	822
stage	books	money	723
books	education	education	524
uk	stage	commentisfree	522
technology	law	tv-and-radio	498
environment	social-care-network	science	417
money	business	artanddesign	414
law	science	fashion	295
lifeandstyle	childrens-books-site		
global-development	healthcare-network		
sustainable-business	football		
commentisfree	travel		
public-leaders-network	culture-professionals-network		
higher-education-network		global-development	
small-business-network	social-enterprise-network		
global-development-professionals-network		uk-news	
healthcare-network	fashion		

Veiem que les seccions més importants s'aprenen de manera correcta i que, en canvi, 'fashion' apareix bastant avall en el rànquing. Podria ser que fos perquè només compta amb 295 articles al rànquing i que aquest factor hagi alterat l'aprenentatge, com ja hem comentat en algunes ocasions.

Quant a l'aprenentatge de *tags*, comptem amb les següents dades:

Aparicions	**Tag**
17	film/filmblog;1.0;0.18421052631578946
15	film/martinscorsese;-3.0;0.13157894736842105
7	film/woodyallen;1.0;0.18421052631578946
2	film/quentintarantino;0.0;0.17105263157894737
6	film/periodandhistorical;0.0;0.17105263157894737
16	film/documentary;7.0;0.2631578947368421
1	film/audrey-tautou;0.0;0.17105263157894737
5	film/film-criticism;1.0;0.18421052631578946
1	film/alfredhitchcock;0.0;0.17105263157894737
1	film/silent-film;0.0;0.17105263157894737
67	artanddesign/photography;4.0;0.2236842105263158

62	**artanddesign/art;7.0;0.2631578947368421**
2	artanddesign/posters;1.0;0.18421052631578946
10	artanddesign/series/pictures-from-the-past;2.0;0.19736842105263158
2	artanddesign/video-art;1.0;0.18421052631578946
14	artanddesign/design;0.0;0.17105263157894737
1	artanddesign/freud;1.0;0.18421052631578946
1	artanddesign/banksy;0.0;0.17105263157894737
1	artanddesign/lichtenstein;0.0;0.17105263157894737
1	artanddesign/streetart;0.0;0.17105263157894737
99	fashion/fashion;2.0;0.19736842105263158
2	fashion/series/vin*tag*e-years;0.0;0.17105263157894737
10	fashion/fashion-blog;0.0;0.17105263157894737
2	uk/london-underground;0.0;0.17105263157894737
9	culture/lena-dunham;1.0;0.18421052631578946
1	culture/zombies;0.0;0.17105263157894737
15	**music/electronicmusic;8.0;0.27631578947368424**
50	**music/musicblog;12.0;0.32894736842105265**
15	music/urban;0.0;0.17105263157894737
4	music/arcticmonkeys;1.0;0.18421052631578946
33	**music/indie;10.0;0.3026315789473684**
5	music/clubs;2.0;0.19736842105263158
1	music/the-1975;0.0;0.17105263157894737
10	**music/series/newbandoftheday;8.0;0.27631578947368424**
3	music/music-festivals;1.0;0.18421052631578946
1	music/stan-getz;0.0;0.17105263157894737

Veiem que els valors màxims que obtenim disten molt d'1 i que ronden el 0,3. En tots els casos es tracta de *tags* amb força repeticions si bé n'hi ha que en tenen més i no aconsegueixen tants bons resultats. En general, veiem que tenim amb *tags* que no tenen massa representació al corpus, uns quants d'ells compten amb tan sols una aparició. Potser perquè als hipsters no els agraden les coses que són 'mainstream'?

Tot aquest procés s'ha repetit també utilitzant el corpus2 de notícies. El perfil ideal s'ha modelat de manera molt similar però no és exactament el mateix ja que, de la manera

que s'ha dissenyat l'algorisme d'aprenentatge, l'estructura de perfil s'adapta al corpus de notícies que hi ha en cada moment. És a dir, no es pot utilitzar un perfil idènticament igual sobre els dos corpus de dades, com ja hem comentat anteriorment.

Al corpus2 no hem trobat alguns dels *tags* que s'havien ressaltat en el perfil hipster (marcats en vermell). La resta s'ha ressaltat de la mateixa manera:

```
SECCIÓ film;60;0.0
film/filmblog;0.0;1
film/martinscorsese;0.0;1
film/woodyallen;0.0;1
film/quentintarantino;0.0;1
film/periodandhistorical;0.0;1
film/documentary;0.0;1
film/audrey-tautou;0.0;1
film/film-criticism;0.0;1
film/alfredhitchcock;0.0;1
film/silent-film;0.0;1

SECCIÓ artanddesign;60;0.0
artanddesign/photography;0.0;1
artanddesign/art;0.0;1
artanddesign/posters;0.0;1
artanddesign/series/pictures-from-the-past;0.0;1
artanddesign/video-art;0.0;1
artanddesign/design;0.0;1
artanddesign/freud;0.0;1
artanddesign/banksy;0.0;1
artanddesign/lichtenstein;0.0;1
artanddesign/streetart;0.0;1


fashion;30;0.0
fashion/fashion;0.0;1
fashion/series/vintage-years;0.0;1
fashion/fashion-blog;0.0;1

uk/london-underground;0.0;1

culture;50;0.0
culture/lena-dunham;0.0;1
culture/zombies;0.0;1

music/electronicmusic;0.0;1
music/musicblog;0.0;1
music/urban;0.0;0.1
music/arcticmonkeys;0.0;1
music/indie;0.0;1
music/clubs;0.0;1music/the-1975;0.0;1
music/series/newbandoftheday;0.0;1
music/music-festivals;0.0;1
music/stan-getz;0.0;1
```

El resultat de l'aprenentatge és el següent:

	Distància entre seccions	Distància entre grups de *tags*
Inici	2.3566700377156473	22.16267560598956
Final	1.3822702333876835	19.351168482248582

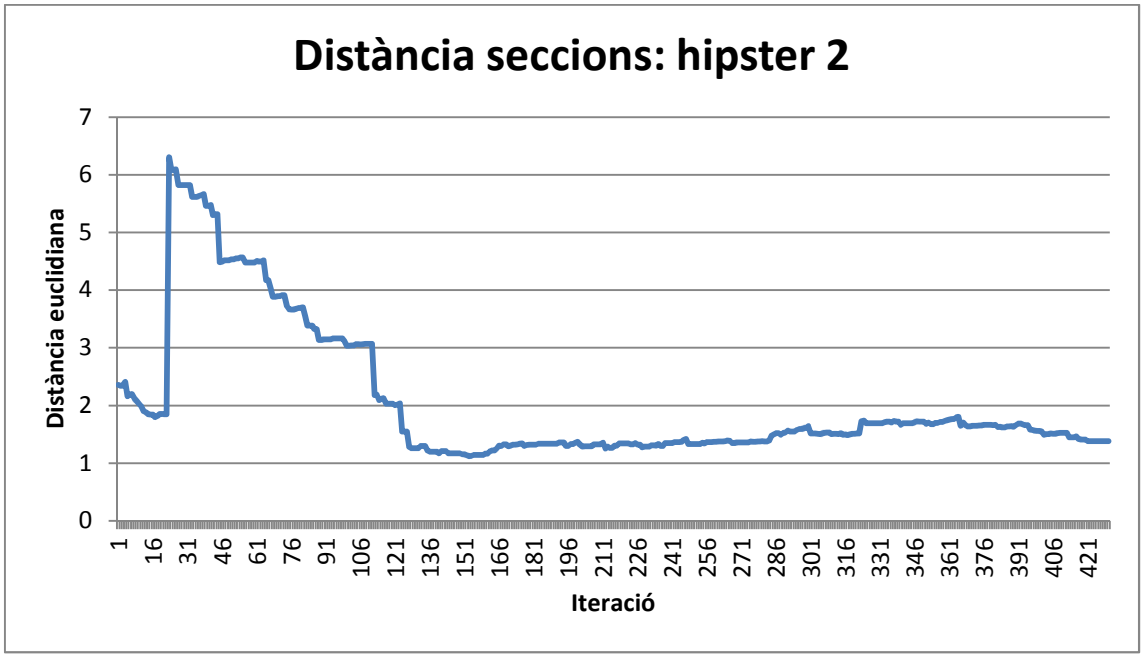


Fig. 4h'

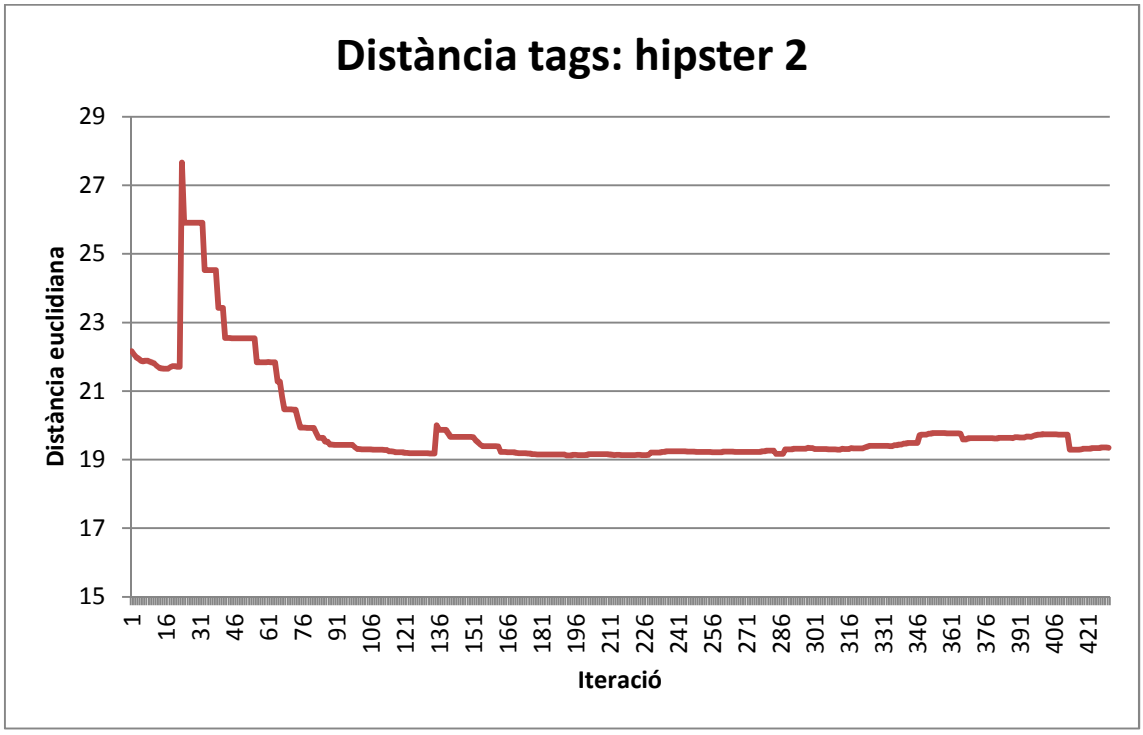


Fig. 4i'

Veiem que aquests gràfics [*Fig. 4h'* i *Fig 4i'*] segueixen el patró que hem anat veient des de la primera prova. Recordem que amb el corpus1 havíem obtingut els següents resultats, una mica pitjors en termes de distància entre seccions:

	Distància entre seccions	Distància entre grups de *tags*
Inici	2.4639632933773807	21.912613703483814
Final	1.5791946016179153	18.789256609695535

Fem una ullada a com queda la gràfica dels overranked a *Fig. 4j'*:

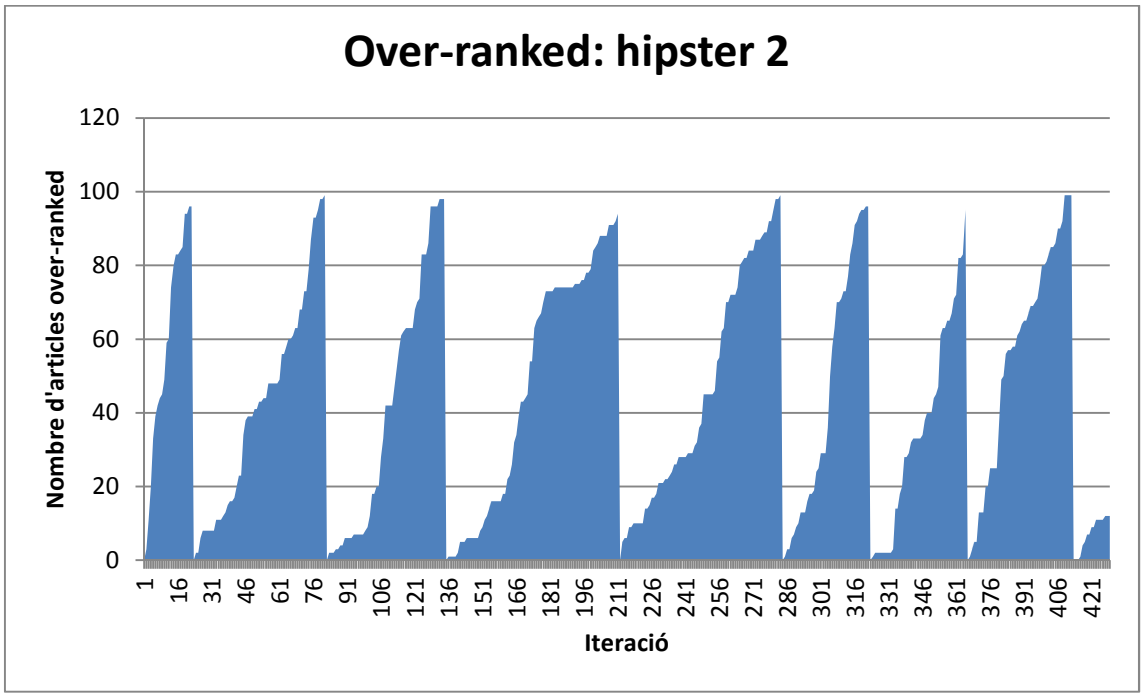


Fig. 4j'

Anem a veure ara com ha anat l'aprenentatge de les seccions:

IDEAL PROFILE	EVOLUTIVE PROFILE	NUMBER OF ARTICLES	
artanddesign	sport	world	6283
film	culture	sport	3412
culture	film	football	2402
fashion	media	uk	2032
football	world	business	2002
business	uk	politics	1748
environment	politics	society	1719
technology	society	culture	1640
sport	business	media	1497
books	technology	lifeandstyle	1368
uk	football	technology	1259
stage	artanddesign	music	1235
science	music	environment	1020
lifeandstyle	higher-education-network	books	917
sustainable-business	global-development-professionals-network	film	901

tv-and-radio		environment		theobserver	822
media		lifeandstyle		money	723
global-development		tv-and-radio		education	524
higher-education-network		commentisfree		commentisfree	522
travel		*stage*		tv-and-radio	498
music		education		science	417
politics		global-development		artanddesign	414
law		books			
education		money			
money		culture-professionals-network			
world		local-government-network			
society		science			
commentisfree		law			
global-development-professionals-network		public-leaders-network			
public-leaders-network		fashion			

Veiem que a grans trets les seccions més importants del perfil ideal s'aprenen si bé hi torna a haver un desajustament amb la secció 'fashion', així com amb 'world' i 'politics', que apareixen en posicions bastant separades en ambdós perfils.

Fem una ullada ara a l'aprenentatge de *tags*:

Aparicions	**Tag**
33	**film/filmblog;9.0;0.21686746987951808**
3	film/martinscorsese;0.0;0.10843373493975904
14	film/woodyallen;8.0;0.20481927710843373
3	film/periodandhistorical;0.0;0.10843373493975904
3	film/film-criticism;0.0;0.10843373493975904
106	**artanddesign/photography;13.0;0.26506024096385544**
73	**artanddesign/art;12.0;0.25301204819277107**
1	artanddesign/posters;0.0;0.10843373493975904
9	artanddesign/series/pictures-from-the-past;1.0;0.12048192771084337
4	artanddesign/video-art;2.0;0.13253012048192772
18	artanddesign/design;2.0;0.13253012048192772
4	artanddesign/freud;1.0;0.12048192771084337
2	artanddesign/streetart;0.0;0.10843373493975904
120	fashion/fashion;7.0;0.1927710843373494

2	fashion/series/vin*tag*e-years;0.0;0.10843373493975904
16	fashion/fashion-blog;1.0;0.12048192771084337
29	uk/london-underground;5.0;0.1686746987951807
3	culture/lena-dunham;0.0;0.10843373493975904
23	music/electronicmusic;3.0;0.14457831325301204
54	music/musicblog;4.0;0.1566265060240964
14	music/urban;2.0;0.13253012048192772
3	music/arcticmonkeys;1.0;0.12048192771084337
29	music/indie;4.0;0.1566265060240964
7	music/clubs;0.0;0.10843373493975904
10	music/series/newbandoftheday;2.0;0.13253012048192772
9	music/music-festivals;1.0;0.12048192771084337

En general, podem veure que obtenim valors molt baixos i que els *tags* millor apresos no assoleixen el 0,3.

4.6 Cap al perfil canviant

Tots els supòsits que hem tingut en compte en les proves de l'algorisme disposaven d'un perfil ideal fix i d'un perfil evolutiu que s'anava acostant cada vegada més al primer en cadascuna de les operacions. Hem vist com funciona el procés d'aprenentatge del prototip elaborat en aquest projecte de final de carrera.

Com s'ha plantejat a la introducció, es vol provar com es comporta l'algorisme en cas que el perfil ideal es vegi modificat al llarg de l'aprenentatge. En aquest apartat es farà una prova de com funcionaria l'algorisme en cas de voler adoptar també aquesta funcionalitat.

Farem una petita modificació del codi i observarem com es comporta davant d'un canvi en el perfil. Per fer-ho ens centrarem en el supòsit d'aprenentatge 4.5.1, el de la persona de negocis tenint en compte el corpus 1. Compararem els resultats que ja tenim amb els dos escenaris següents:

1. Modificació del perfil ideal (donar pes a una secció en concret) just després del primer terç de les iteracions de l'algorisme.

2. Modificació del perfil ideal (donar pes a una secció en concret) just després del segon terç de les iteracions de l'algorisme.

Per tant, haurem de fer una nova versió del perfil ideal que, en un cas, es canviarà per l'original en un punt inicial de l'algorisme i, en el segon cas, quan ja estigui molt avançat. En el perfil modificat, donarem pes a aquestes seccions i *tags*:

```
SECCIÓ money;60;0.0
money/money;0.0;1
money/banks;0.0;1
money/shares;0.0;1
money/currentaccounts;0.0;1
money/pay;0.0;1
money/work-and-careers;0.0;0.1
money/renting;0.0;1
money/maternitypaternityrights;0.0;1
money/blog;0.0;1
money/savings;0.0;1
money/savings-rates;0.0;1

SECCIÓ education;50;0.0
education/schools;0.0;1
education/teaching;0.0;1
education/studenthousing;0.0;1
education/primary-schools;0.0;1
education/studenthealth;0.0;1
education/series/how-i-became-a-teacher;0.0;1
education/educationalbooks;0.0;1
```

Primer de tot haurem d'aconseguir una fotografia del perfil ideal en els dos punts d'aturada que hem mencionat, en acabar el primer terç de l'execució i en acabar el segon terç. Per fer-ho, introduirem el codi temporal que podem veure a la *Fig. 4h'* (que es deixarà més endavant comentat per tal de poder-lo consultar). Com que treballem amb un corpus que itera 389 vegades, agafarem aproximadament la 130a iteració per fer la fotografia del 1/3 i la 260 per fer la del 2/3:

```
if (i==130) {
    System.out.println("**************Fotografio el perfil ideal a 1/3 de l'execució");
          idealProfile.writeToCSV("profileidealnegocis_primerterc.csv");
}

if (i==260) {
    System.out.println("**************Fotografio el perfil ideal a 2/3 de l'execució");
          idealProfile.writeToCSV("profileidealnegocis_segonterc.csv");
}
```

Fig. 4h'

Primer terç

Agafarem el fitxer obtingut en la fotografia del 1/3 i el modificarem manualment donant els pesos que hem comentat a les seccions 'money' i 'education'. Una vegada fet, afegirem aquest codi [*Fig. 4i'*] per tal de fer el canvi del perfil en el punt indicat:

```
if (i==130) {
    System.out.println("**************Fotografio el perfil ideal a 1/3 de l'execució");
        // idealProfile.writeToCSV("profileidealnegocis_primerterc.csv");
        idealProfile.readFromCSV("profileidealnegocis_primerterc_mod.csv");
        idealProfile.normalizeRatingSections();
        idealProfile.normalizeRatingTags();
}
```

Fig. 4i'

Executem l'algorisme d'aprenentatge fent el canvi de perfil a 1/3 de l'execució i obtenim els següents resultats finals:

	Distància entre seccions	**Distància entre grups de *tags***
Inici	2.504180081383925	21.647480034974407
Final	1.4834490889877054	18.980631716309883

Vegem les corbes d'aprenentatge a *Fig. 4j'* i *Fig. 4k'*:

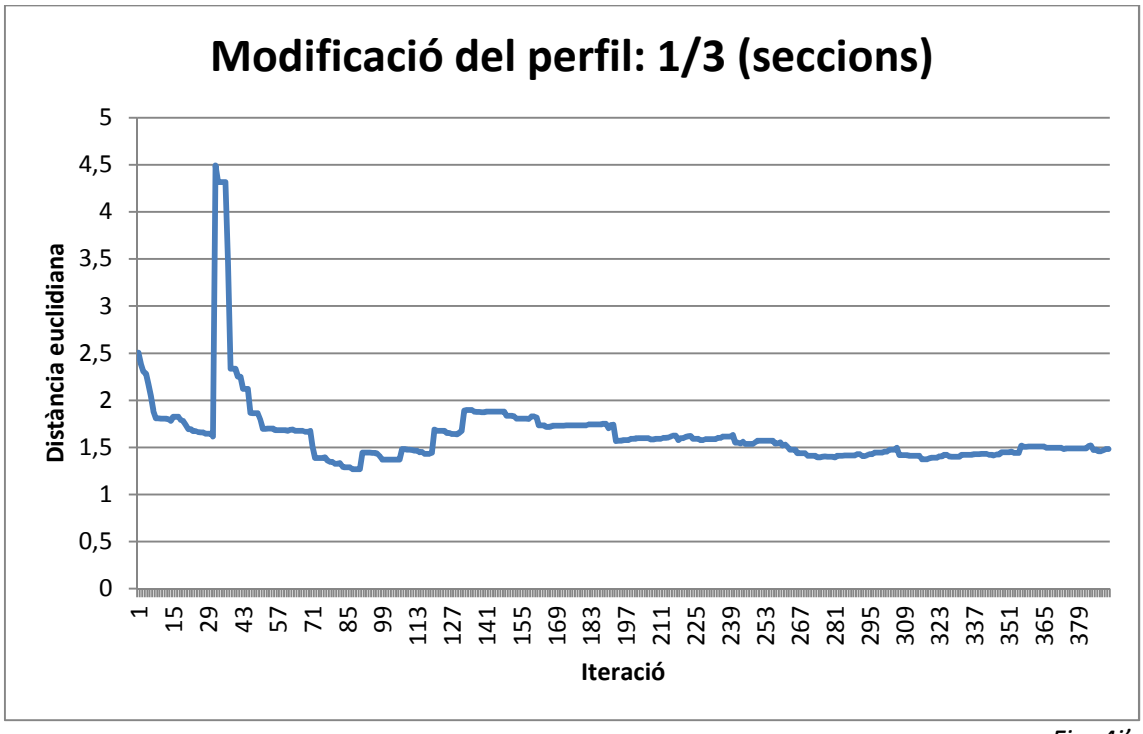


Fig. 4j'

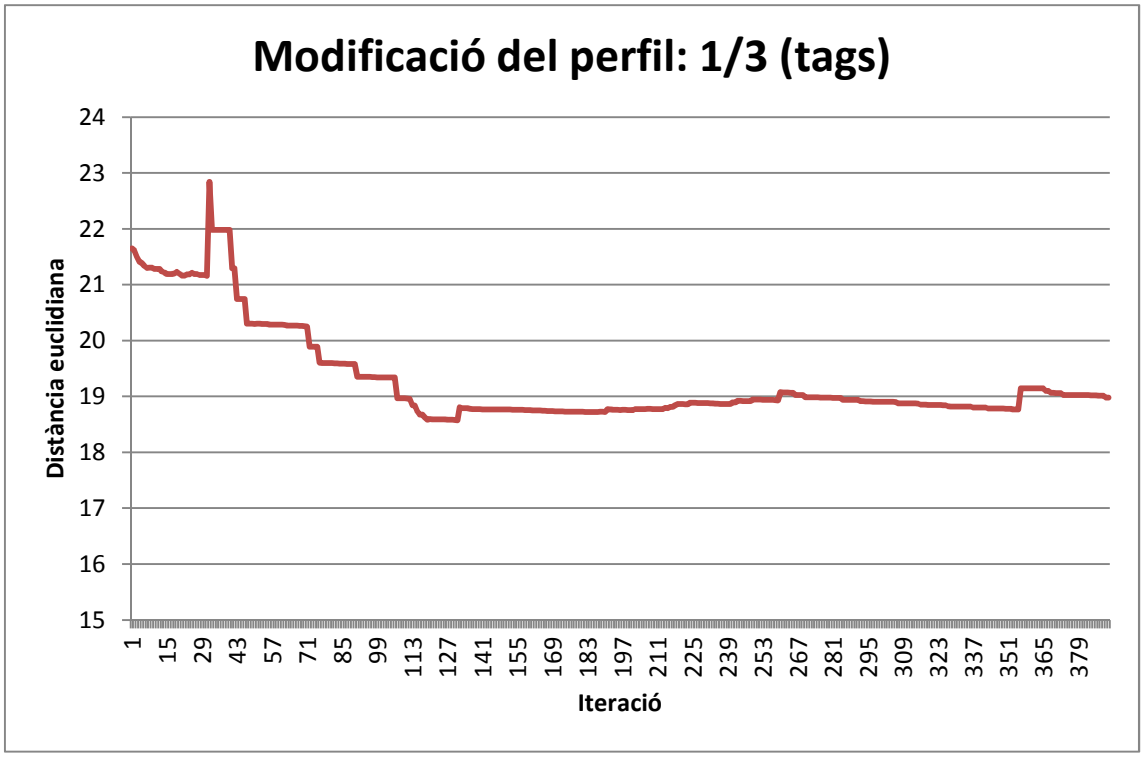


Fig. 4k'

Recordem que aquests eren els resultats seguint el procediment normal, sense modificar el perfil ideal en la mateixa execució:

	Distància entre seccions	**Distància entre grups de *tags***
Inici	2.504180081383925	21.647480034974407
Final	1.295941043290216	18.493673799014562

Veiem que a un terç de l'execució l'algorisme té una espècie de "desaprenentatge" perquè li canviem pesos d'algunes seccions i *tags* i, per tan, augmenta la distància euclidiana entre els dos perfils. Per acabar amb l'exemple, anem a veure com ha funcionat l'aprenentatge de seccions i el comparem amb el perfil evolutiu de l'execució estàndard d'aquest supòsit:

AMB MODIFICACIÓ
politics
sport
world
media
business
technology
education
society
money
uk

SENSE MODIFICACIÓ
world
sport
politics
media
society
business
technology
uk
environment
education

environment	commentisfree
commentisfree	culture
culture	money
global-development	higher-education-network
music	tv-and-radio
tv-and-radio	football
healthcare-network	music
football	artanddesign
lifeandstyle	global-development
artanddesign	healthcare-network

Podem observar que en el perfil evolutiu modificat, les seccions 'education' i 'money' s'han après tot i introduir-se amb l'execució ja començada, si bé no han aconseguit superar les que ja estaven ressaltades al perfil des de l'inici perquè han estat presents en més iteracions i s'han pogut aprendre millor. No obstant això, veiem que l'aprenentatge s'efectua correctament i que en cas de fer més iteracions, aquestes seccions anirien guanyant pes.

Anem a veure com queda l'aprenentatge introduint els canvis més avançada l'execució de l'algorisme, en acabar el 2/3. Les distàncies obtingudes són les següents:

	Distància entre seccions	**Distància entre grups de *tags***
Inici	2.504180081383925	21.647480034974407
Final	1.4695943979358537	18.73598099864217

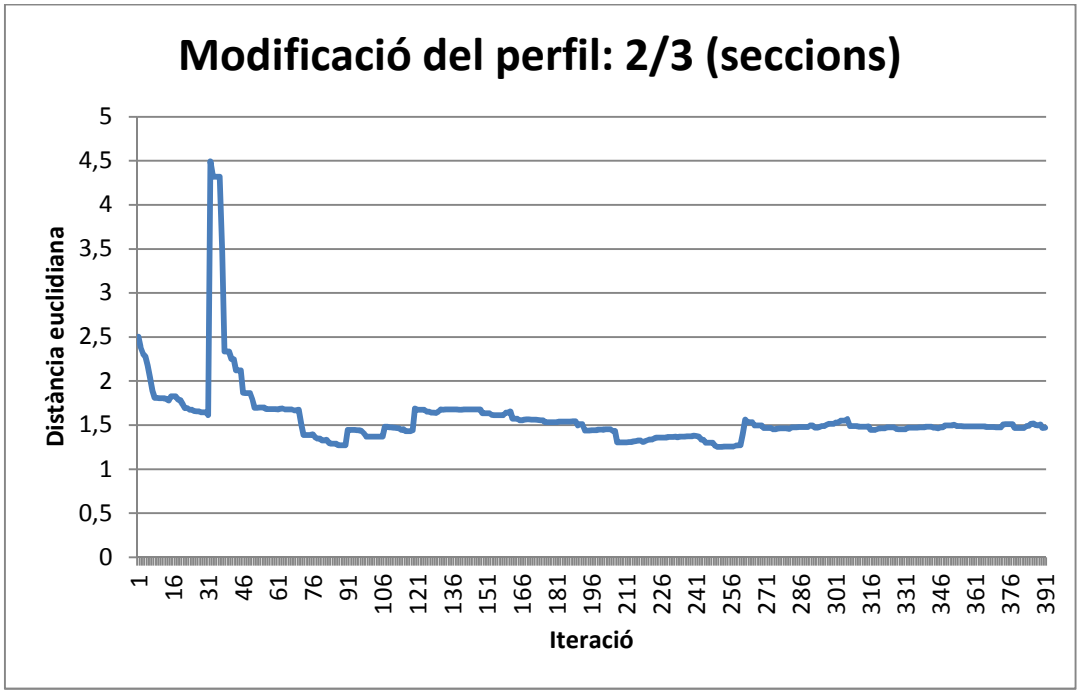


Fig. 4l'

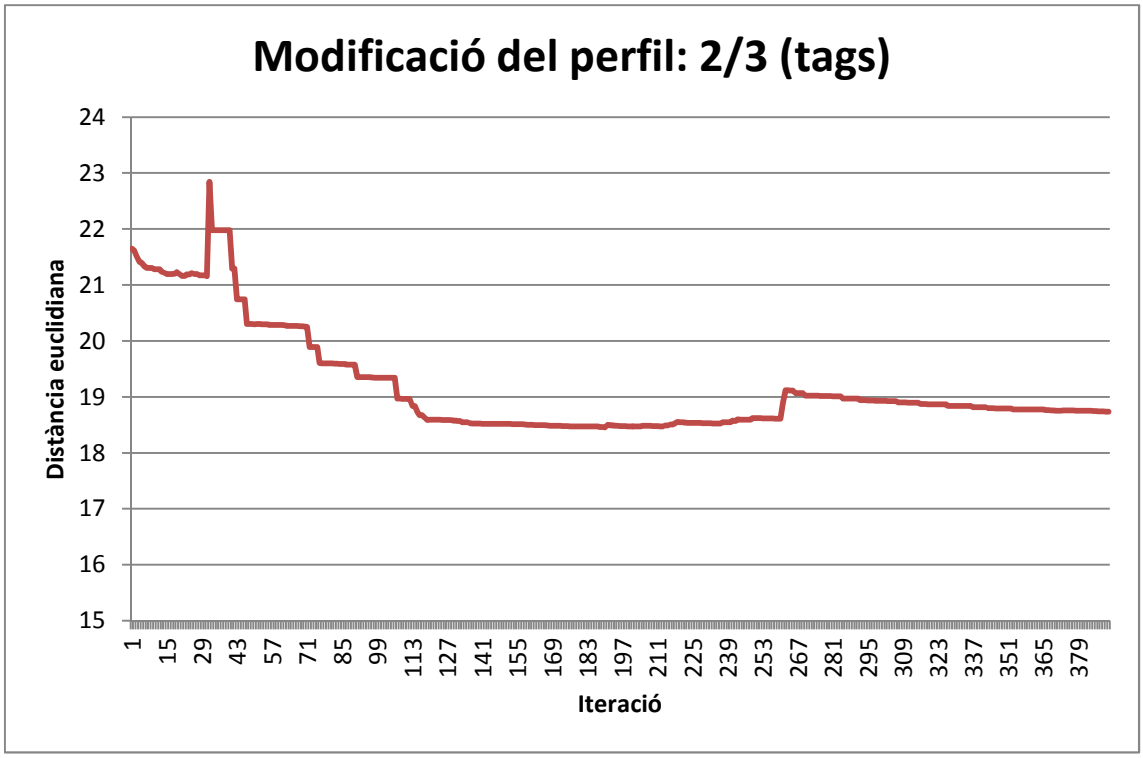


Fig. 4m'

Veiem [*Fig. 4l'* i *Fig. 4m'*] que també hi ha un cert trencament als 2/3 de les iteracions però que llavors l'algorisme continua amb l'aprenentatge. L'ordre de les seccions dels perfils evolutius –aquest i l'original- queda de la següent manera:

AMB MODIFICACIÓ	SENSE MODIFICACIÓ
politics	world
sport	sport
world	politics
business	media
media	society
education	business
technology	technology
society	uk
uk	environment
environment	education
commentisfree	commentisfree
money	culture
culture	money
higher-education-network	higher-education-network
global-development	tv-and-radio
music	football
tv-and-radio	music
healthcare-network	artanddesign
football	global-development
lifeandstyle	healthcare-network

artanddesign travel

Per tant, doncs, l'aprenentatge és correcte ja que totes les seccions ressaltades en l'exemple original i en la modificació apareixen en el rànquing de les 20 seccions més ben valorades pel perfil evolutiu.

5 Conclusions i futur

Aquesta secció servirà per treure conclusions de tots els supòsits d'aprenentatge que s'han tingut en compte a l'apartat anterior així com també per fer una valoració global del procés de disseny i implementació i, en general, de tot el projecte. Primer de tot, farem balanç de l'aprenentatge de seccions i *tags*. Seguidament comentarem temes complementaris a l'aprenentatge, com són les seccions a ignorar i els paràmetres de l'aprenentatge offline. Finalment, es farà una valoració global del projecte i es parlarà de les aplicacions que podria tenir el prototipus desenvolupat, així com també de la seva escalabilitat i reutilització.

5.1 Aprenentatge de seccions

Al llarg de tots els supòsits d'aprenentatge que s'han anat tenint en compte hem anat ja veient que l'aprenentatge de les seccions es feia de manera bastant vàlida. Anem a treure conclusions sobre aquest procés en els punts següents:

5.1.1 Reducció d'un 47% de la distància euclidiana, patró constant

A continuació podem veure un resum de les dades referents a les distàncies euclidianes a l'inici i al final de cada aprenentatge:

	Inici	Final	% Reducció
Negocis c1	2,5041801	1,295941	0,482489
Negocis c2	2,5799909	1,211785	0,530314
Esport c1	2,3735302	1,131896	0,523117
Esport c2	2,3748731	1,204759	0,492706
Hipster c1	2,4639633	1,579195	0,359084
Hipster c2	2,35667	1,38227	0,413465
Mitjana global			0,466862

Amb aquestes dades podem afirmar que el prototipus elaborat redueix de mitjana aproximadament un 47% la distància entre els *ratings* normalitzats de les seccions del perfil ideal i el perfil iteratiu. Per tant, podem dir que l'aprenentatge online i offline duen a terme la seva tasca, aconsegueixen que el perfil evolutiu es vagi assemblant cada vegada més al perfil ideal a través de les valoracions de conjunts d'articles que es van fent en cada iteració de l'algorisme. Es tracta d'un funcionament estable ja que podem veure que, amb una mitjana del 47%, tots els percentatges de reducció de distàncies s'assemblen bastant i, en general, tendeixen a acostar-se al 50%.

Podem veure *[Fig. 5a]* que les corbes d'aprenentatge que aconsegueix l'algorisme en cadascuna de les execucions que s'han dut a terme segueix un patró similar. Això ens permet afirmar que les proves que s'han dut a terme són representatives i que l'aprenentatge supervisat aconsegueix una eficàcia molt similar en cadascun dels supòsits que es tenen en compte a través de la combinació d'aprenentatge online amb offline. Anem a veure a la un mosaic amb totes les corbes d'aprenentatge de seccions per tal de tenir il·lustrada aquesta idea que acabem d'exposar:

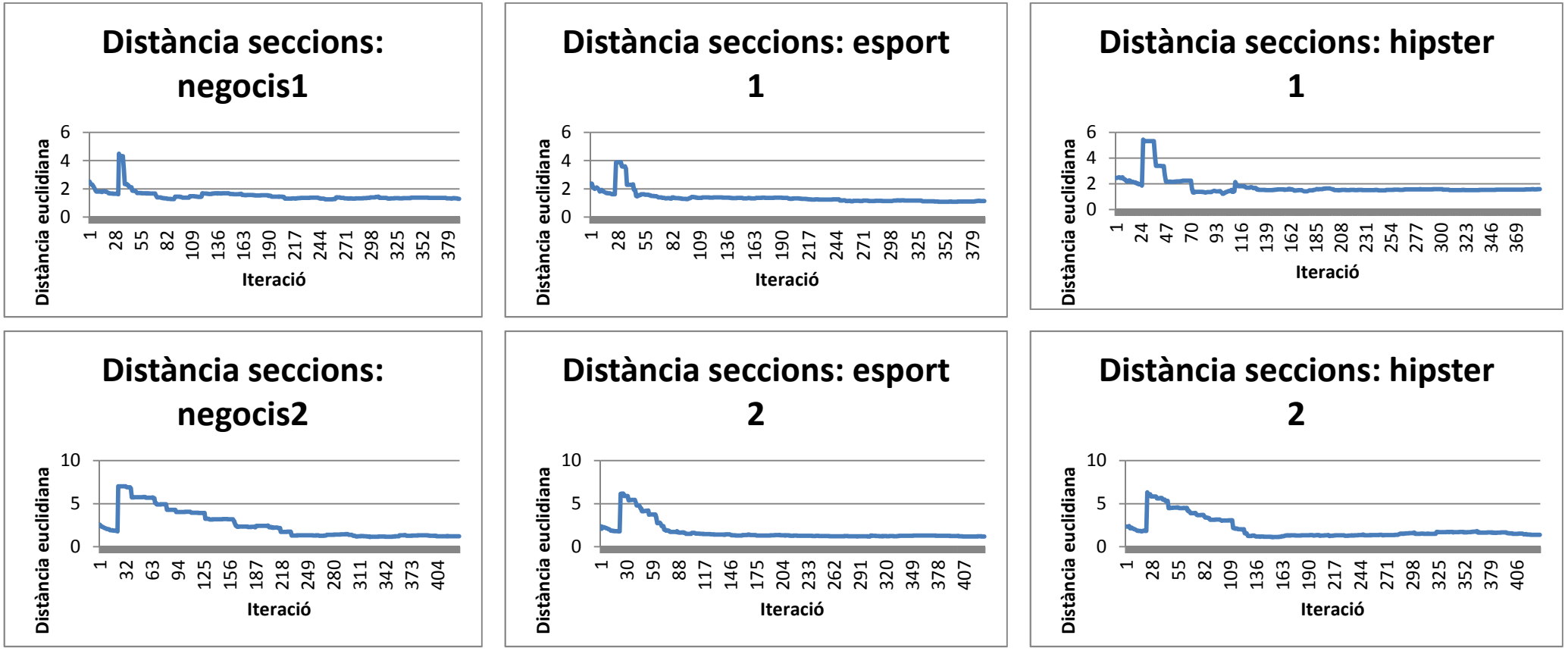


Fig. 5a

5.1.2 Seccions destacades, apreses

L'ordre del rànquing de les seccions més valorades al perfil ideal és similar al que s'obté en el perfil evolutiu una vegada conclosa l'execució de l'algorisme. Hem vist que, excepte algunes petites desviacions, ha estat així en tots els supòsits a estudiar que s'han tingut en compte.

En el perfil de persona de negocis, teníem com a seccions destacades 'business', 'technology', 'world' i 'politics'. Totes elles apareixen a les set primeres posicions del rànquing del perfil evolutiu, juntament amb 'sport', 'media' i 'society', que tot i no ser unes de les seccions més destacades les podem veure en la llista de les primeres vint del perfil ideal del més de centenar de seccions amb què treballàvem.

No s'esperava que l'aprenentatge obtingut amb aquest algorisme fos perfecte, ja que estem treballant amb grans quantitats de dades i el que interessa és obtenir un patró aproximat dels gustos de lectura de l'usuari. Per tant, podem considerar que hem obtingut uns resultats acceptables.

Continuant amb l'exemple dels negocis, en el rànquing del perfil evolutiu obtingut treballant amb el corpus2, també trobem les seccions més destacades entre les vuit primeres posicions, juntament amb 'media', 'society', 'culture' i 'uk'.

L'aparició de seccions que, tot i estar dins la primera vintena de les més valorades del perfil ideal, no eren de les de dalt de tot es pot deure al fet que es tractin seccions i *tags* a la vegada. Per exemple, potser ens apareix 'society' molt amunt perquè algunes notícies de 'world' o 'technology' –seccions molt destacades- estan codificades també amb *tags* pertanyents a 'society'. Aquest fet no devalua l'aprenentatge perquè ja ens interessa que l'algorisme funcioni així.

L'aparició de les seccions destacades al perfil ideal al perfil evolutiu es repeteix, com hem vist, al llarg de tots els exemples, a excepció del perfil 'hipster' que no aconsegueix aprendre tant bé com les altres la secció 'fashion', si bé també és cert que li havíem donat menys valoració que 'film' i 'artanddesign', les altres ressaltades.

Un altre patró que es va repetint és que algunes d'aquestes notícies que apareixen a les primeres posicions del perfil evolutiu tot i no estar especialment destacades és el fet que, generalment, són seccions que compten amb un nombre molt gran de notícies al corpus. És un patró, però, que no sempre es compleix i no tenim dades suficients per establir-lo com a conclusió definitiva.

Per exemple, amb el perfil d'esports, tenim destacades les seccions de 'sport' i 'football' que són la segona i la tercera secció amb més notícies del corpus1. Tot i això, 'society' supera la secció 'football', tot i estant a la posició 11 del perfil ideal i comptant amb molts menys articles que 'sport' i 'football'.

No obstant això, sí que és cert que el perfil hipster aprèn les seccions 'world' i 'uk' amb molta valoració, mentre que no apareixen en el seu top ten del perfil ideal. En canvi, són de les seccions que més articles acumulen. Dóna la sensació que alguna relació hi ha amb le nombre d'aparicions de les notícies però no hi ha un patró constant que ens permeti afirmar-ho del cert amb les dades de les quals disposem.

A tall de resum, però, podem dir que la correspondència entre valoracions de seccions al perfil ideal i al perfil après compleix amb les expectatives que s'havien fixat a l'inici del projecte i se'n fa una valoració molt positiva.

5.2 Aprenentatge de *tags*

Tot i que la valoració de l'aprenentatge de seccions s'ha pogut considerar com a molt vàlid, ja hem anat veient al llarg de tots els supòsits d'aprenentatge que no s'han obtingut els mateixos resultats en l'aprenentatge de *tags*. Anem a treure conclusions sobre aquest procés en els punts següents:

5.2.1 Reducció mitjana d'un 14%

A continuació podem veure un resum de les dades referents a les distàncies euclidianes a l'inici i al final de cada aprenentatge:

	Inici	Final	% Reducció
Negocis c1	21,64748	18,49367	0,145689
Negocis c2	22,306064	19,16875	0,140649
Esport c1	21,912553	18,77139	0,14335
Esport c2	21,772033	18,65996	0,142939
Hipster c1	21,912614	18,78926	0,142537
Hipster c1	22,162676	19,35117	0,126858
Mitjana global			0,140337

Veiem que la mitjana de percentatge de reducció de la distància de *tags* entre el perfil ideal i el perfil evolutiu és molt constant. Això és un punt a favor perquè indica que

l'algorisme funciona de la mateixa manera siguin quines siguin les dades amb què treballi.

No obstant això, una mitjana d'un 14% de reducció no arriba al que s'esperava a l'inici del projecte, ja que és un percentatge molt reduït. L'aprenentatge de seccions –n'hi ha aproximadament un centenar de diferents- s'ha dut a terme de manera molt més correcta i dóna la sensació que l'aprenentatge de *tags* no es fa tant bé perquè hi ha molta més diversitat. Cada corpus en té més de 4.000 de diferents.

Les seccions compten amb prou aparicions com per anar modelant el perfil evolutiu de manera correcta ja que n'hi ha un centenar aproximadament i comptem amb uns corpus de 6.000 notícies. En canvi, tot apunta que hi ha massa diversitat de *tags* i que faria falta un procés d'aprenentatge més llarg per tal de poder-los aprendre millor. No obstant això, la reducció és evident en tots els supòsits amb els quals hem treballat, però caldria millorar aquest prototipus en un futur per tal d'incrementar aquest percentatge almenys fins a l'obtingut a nivell de secció.

5.2.2 Lluny del *rating* ideal i proporcional a nombre d'aparicions

Al llarg de totes les proves, quan hem anat mirant l'aprenentatge dels *tags* als quals havíem assignat el màxim *rating* normalitzat al perfil ideal (1) hem vist que la majoria d'ells no s'acostaven ni de bon tros a aquesta xifra. La hipòtesi ha estat pensar que, com més vegades apareix un *tag* al llarg de l'aprenentatge, millor s'aprèn. Per exemple, a les proves del perfil de negocis sobre el corpus 1 teníem com a *tags* més ben apresos 'business/business' (0,63 punts de *rating*) i 'politics/politics' (0,78 punts de *rating*). Coincidia que eren els *tags* que més apareixien en el corpus i, per tant, en el procés d'aprenentatge. Hem fet una gràfica *[Fig. 5b]* per demostrar aquesta correlació. S'han agafat, per una banda, els *ratings* normalitzats entre 0 i 1 i també s'han normalitzat de la mateixa manera el nombre d'aparicions:

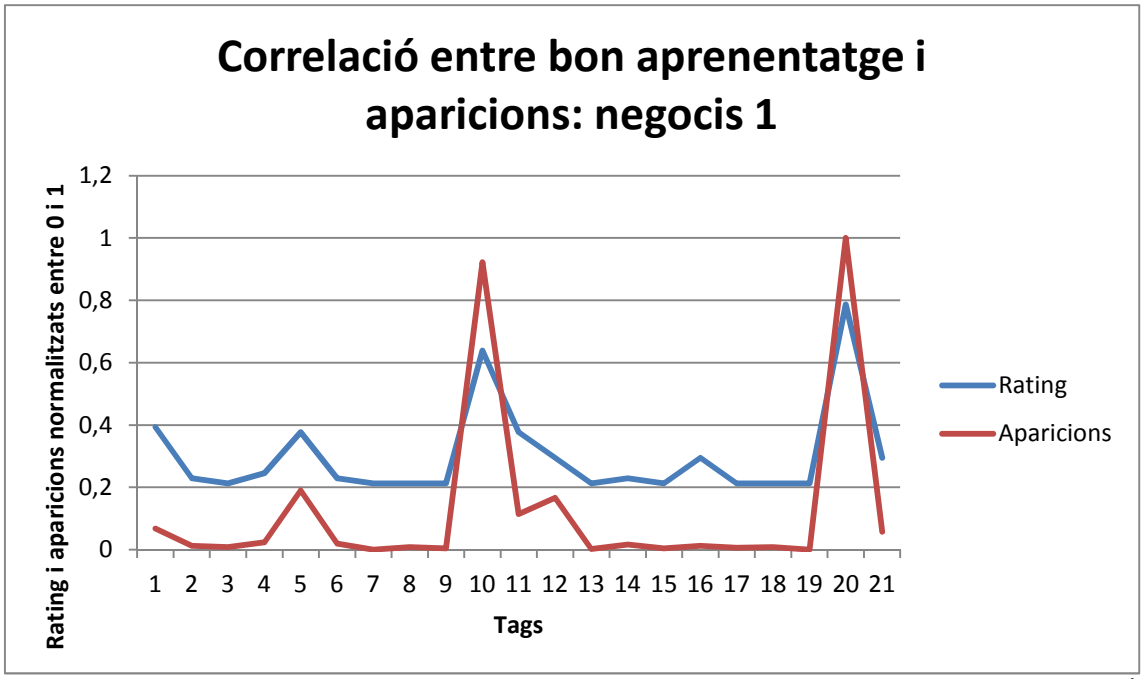


Fig. 5b

Podem observar que la proporcionalitat entre els dos paràmetres és bastant evident. El comportament es repeteix en la mateixa prova feta sobre el corpus 2 però cal afegir un paràmetre a l'anàlisi. Els dos *tags* més valorats són els que compten amb més aparicions al llarg del corpus d'entre el grup de *tags* analitzats –els que s'han ressaltat al perfil ideal- però també pertanyen a dues de les seccions més populars, de les que compten amb més notícies al llarg del corpus.

Anem a repetir la gràfica sobre el perfil d'esports *[Fig. 5c]*. Ens fixem també només en el corpus1 ja que les dues execucions donen resultats bastant similars:

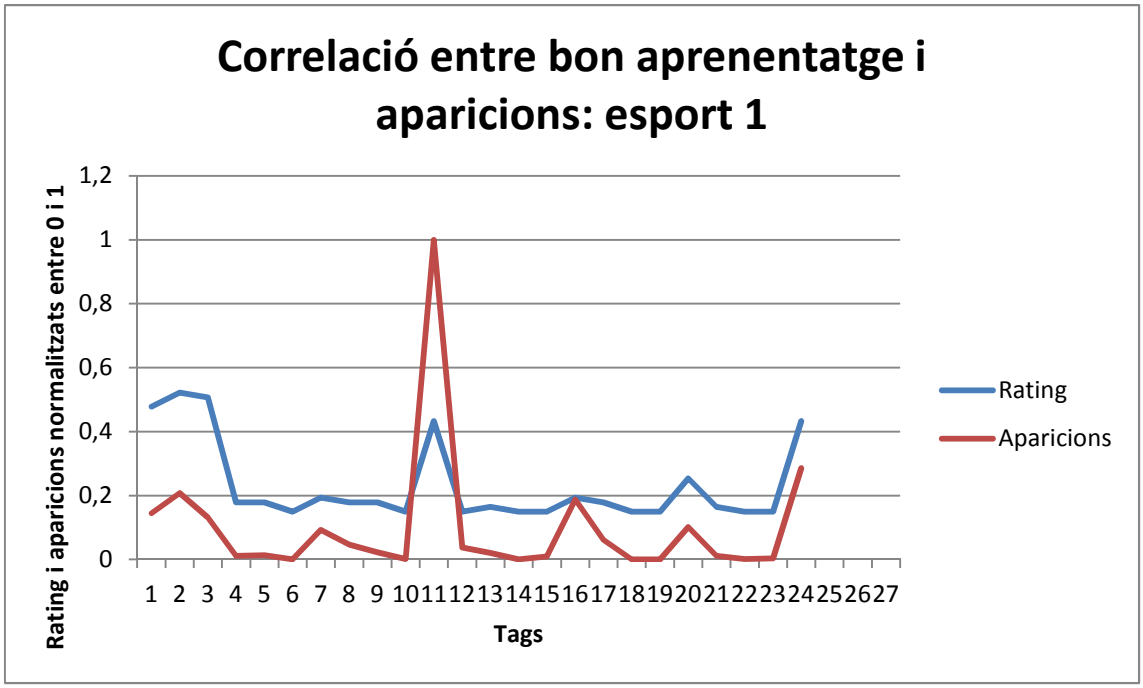


Fig. 5c

Veiem que el patró es va confirmant. Hi ha una certa proporcionalitat òbvia. No obstant això, hi ha una excepció amb el *tag* 'football/football'. Obté un *rating* alt (0,43) però compta amb moltíssimes aparicions (540). És el pic més alt que es pot veure a la sèrie en vermell del gràfic. Malgrat això, es veu superat pels *tags* 'sport/england-cricket-team' (0,47 de *rating* i 79 aparicions); 'sport/cricket' (0,52 de *rating* i 112) i per 'sport/australia-cricket-team' (0,50 de *rating* i 72 aparicions). Si observem, però, els paràmetres del perfil idel, podem veure que la secció 'sport' compta amb un *rating* de 50 punts, mentre que 'football' n'aconsegueix només amb 40. Aquest fet podria ser la causa per la qual es produeix aquesta desviació.

Amb el perfil de 'hipster' l'anàlisi es complica una mica perquè en general comptem amb *tags* que tenen molt poques aparicions en el corpus. El gràfic que s'obté és el que podem veure a la figura *Fig. 5d* (segons corpus1):

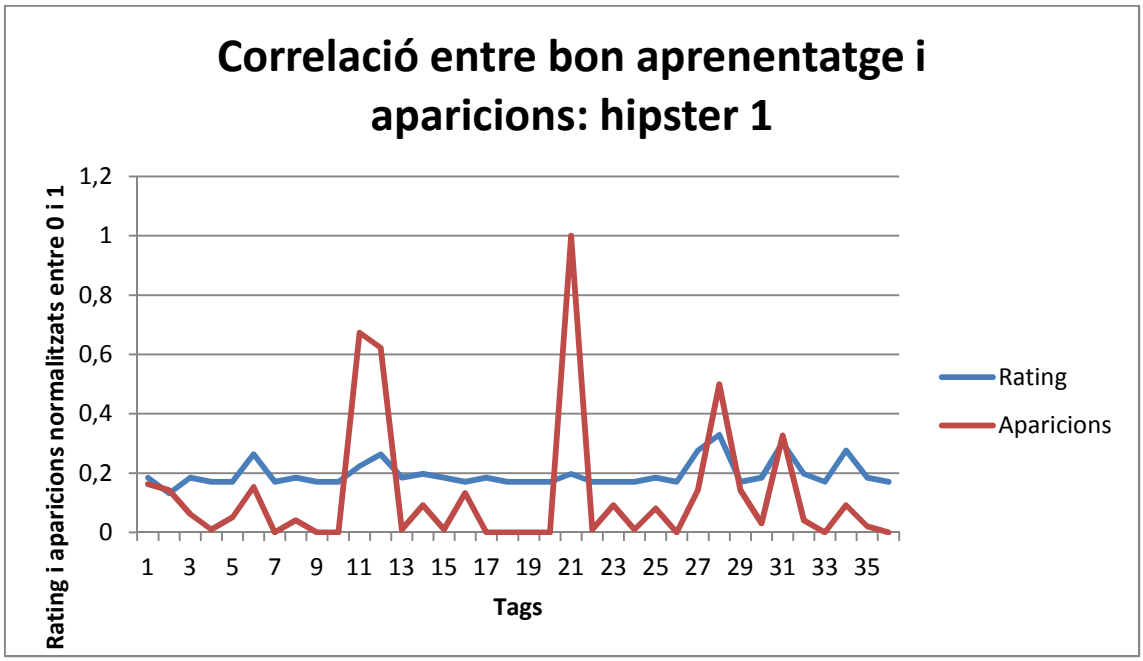


Fig. 5d

Veiem que s'expressa encara una certa proporció però que queda molt més desdibuixada que en els altres casos perquè no comptem amb *tags* tan, per dir-ho d'alguna manera, populars. El que té més aparicions és 'fashion/fashion' (99) però el seu *rating* és molt baix, de 0,19. En canvi 'music/musicblog' obté 0,33 punts amb 50 aparicions. Dóna la casualitat que 'fashion' no ha estat una secció ben apresa en el procés d'aprenentatge, tal com hem comentat en les conclusions del punt anterior. Val a dir, però, que 'fashion' tenia una puntuació de 30 que distava del 60 de 'film' i 'artanddesign'.

A tall de conclusió, podem dir que a grans trets hi ha una proporció entre el *rating* i les aparicions dels *tags*. Els més ben apresos solen ser aquells que tenen més presència en el corpus i que, a la vegada, compten amb el *rating* de secció més elevat. No obstant això, l'aprenentatge es du a terme de manera molt modesta perquè tot apunta a què l'algorisme no es topa prou vegades amb els '*tags*' al llarg de l'aprenentatge i, per dir-ho d'alguna manera, no té prou oportunitats per assignar-los un '*rating*' més alt. En properes fases de recerca sobre aquest prototip s'hauria d'analitzar què passa si el procés es repeteix en corpus de notícies més grans, on cada *tag* compti amb un nombre molt més representatiu d'aparicions.

5.3 Altres consideracions sobre l'aprenentatge

En aquesta secció s'han recollit comentaris addicionals sobre el procés d'aprenentatge. Concretament, es volen valorar les gràfiques obtingutdes sobre la llista de notícies overranked, els paràmetres fets servir en l'aprenentatge offline, el fet que la secció 'commentisfree' aparegui com a líder en les tries de notícies i el funcionament de l'aprenentatge quan duem a terme una modificació en el perfil ideal una vegada ja iniciada l'execució de l'algoritme.

5.3.1 Evolució de l'overranked

Hem anat veient en els diversos supòsits d'aprenentatge que, com a tendència general, el llistat de notícies overranked cada vegada es va omplint de manera més lenta a mesura que va avançant l'aprenentatge. Per aquest motiu, l'hem considerat al llarg de totes les proves com un indicatiu de l'evolució de l'aprenentatge.

5.3.2 Paràmetres aprenentatge offline

A l'apartat 4.4.4 s'han estudiat quin eren els valors dels paràmetres de l'aprenentatge offline amb els quals l'algorisme funcionava millor. Una vegada fetes totes les proves hem pogut veure que ofereixen uns resultats estables i en la mateixa línia en tots els supòsits provats. Es confirma, doncs, la seva validesa òptima tenint en compte les combinacions que s'han pogut provar en el marc d'aquest projecte.

5.3.3 Seccions a ignorar

A 'apartat 2.4.5 s'han establert quines eren les seccions que l'algorisme ignoraria perquè podien interferir a l'aprenentatge. Són aquelles que defineixen el gènere periodístic de cada text. Aquestes seccions gairebé codifiquen tots els articles i es va veure que podrien influir negativament a l'aprenentatge ja que, podien sortir com a molt ben classificades les seccions 'Tone' o 'Type' que, en el fons, ho engloben tot

Al llarg de totes les proves que s'han realitzat s'ha pogut veure que la secció 'comentisfree' –dedicada a continguts on es fa una crida a la participació de l'audiència- té un comportament similar a les que s'havia decidit ignorar. En el seu moment no es va decidir tractar-la com una secció a ignorar perquè engloba un tipus de notícies particulars a diferència de 'Tone' i 'Type'.

Tot i tenir el convenciment que sigui una secció amb entitat pròpia, estaria bé, en pròximes fases del projecte, estudiar què passa amb l'aprenentatge quan no es té en compte, ja que, a tall de resum, podem veure que era una de les seccions més escollides pel perfil ideal i l'evolutiu al llarg de l'aprenentatge, ja que recull *tags* de moltes seccions diferents:

5 seccions més escollides del perfil de negocis:

IDEAL PROFILE		EVOLUTIVE PROFILE	
commentisfree	69	commentisfree	74
business	38	business	56
media	32	media	38
politics	31	politics	38
world	28	world	28

5 seccions més escollides del perfil esportiu:

IDEAL PROFILE		EVOLUTIVE PROFILE	
sport	70	commentisfree	74
commentisfree	67	sport	72
football	28	society	36
society	23	politics	31
world	19	world	21

politics	19	business	21

5 seccions més escollides del perfil hipster:

IDEAL PROFILE		EVOLUTIVE PROFILE	
commentisfree	47	commentisfree	63
politics	33	politics	44
music	28	society	34
society	24	film	25
film	23	business	23

5.3.4 Sobre el perfil canviant

La secció 4.6 ens ha servit per demostrar que l'algorisme desenvolupat serviria també per aprendre un perfil canviant que vagi variant al llarg del temps. Molt útil perquè els gustos i interessos de les persones van evolucionant al llarg dels mesos i, en general, de la vida. El fet que l'algorisme sigui capaç de sumar i restar pesos a les seccions i les etiquetes que la representen fa que això sigui possible. No obstant això, hem vist que en fer variacions en el perfil ideal, l'algorisme necessitaria més iteracions de les que hem configurat en aquest prototipus per acabar d'aprendre bé el perfil. Seria de fàcil arreglar en futures versions del programa, en les quals se'l podria dotar de conjunts d'entrenament més grans.

5.4 Possibles millores

Vistes totes les conclusions que s'acaben d'exposar en els punts anteriors podem concloure que futures possibles millores per refinar l'algorisme implementat serien les següents:

- Fer proves utilitzant més corpus de notícies i més grans per tal de comptar amb conjunts d'entrenament de més magnitud. Així es podria veure si l'aprenentatge de seccions continua estabilitzat al 47% de reducció de distància o bé es millora.
- El percentatge de reducció d'un 14% en l'aprenentatge de *tags* s'hauria de millorar. En la mateixa línia que el punt anterior, estaria bé provar com continua l'aprenen*tag*e amb un conjunt d'entrenament més gran. Hem

comprovat que el bon aprenentatge va lligat al nombre d'aparicions d'un *tag*. Per tant, amb un conjunt d'entrenament més gran, aquests s'haurien d'aprendre millor.

- Hem vist que les seccions amb més notícies del corpus tendeixen a obtenir *ratings* molt alts, sovint superant aquelles seccions més valorades en el perfil ideal. Per tal d'evitar-ho, estaria bé trobar la manera de desvincular aquesta relació. Es podria pensar en agafar corpus de notícies amb un nombre limitat d'articles per a cada secció però això no modela la realitat i és precisament per aquest motiu que en aquest projecte s'ha agafat una mostra de 15 dies sense cap mena de filtre. Potser es podria afegir alguna mena de factor corrector a l'algorisme per controlar aquesta anomalia ja que adaptar el corpus aniria en contra de l'aplicació real de l'algorisme.
- Fer més proves amb noves combinacions dels paràmetres de l'aprenentatge offline.
- Provar què passa sense tenir en compte la secció 'commentisfree' com a secció principal.
- Poder crear perfils ideals sense haver d'editar els fitxers CSV a mà.
- Millorar la interacció i la usabilitat del programa, ja que s'ha elaborat una interfície molt bàsica per poder realitzar les diferents proves.
- Millorar el programa per al tractament d'un perfil ideal que va evolucionant al llarg de les iteracions.

5.5 Escalabilitat i reutilització

Com s'ha comentat al llarg de tota la memòria, l'algorisme d'aprenentatge ha estat dissenyat per poder-se adaptar a altres repositoris de documents, més enllà de les notícies, tot i que la implementació s'hagi aplicat a un cas concret, a l'anàlisi d'articles del diari britànic 'The Guardian'.

Per poder-lo utilitzar en altres conjunts de documents caldria que aquests es poguessin classificar en grups (les nostres seccions principals) i que, a més a més, cada article vingués definit per un conjunt de paraules clau, cadascuna d'elles també associada a una secció. Partint d'aquesta base podríem treballar amb documents mèdics, històrics, etcètera.

En el programa desenvolupat caldria dur a terme tota una sèrie de modificacions, tant a nivell de disseny com d'implementació. Els canvis més significatius s'haurien de fer a la classe Corpus, que és la que modela l'entrada i sortida de les mostres de documents. Per fer-ho senzill i poder deixar gairebé totes les altres classe iguals, estaria bé implementar la lectura d'aquesta classe de manera que convertís el format de la font de dades en el que hem utilitzat per treballar amb The Guardian. És a dir, que les dades de cada document es poguessin tractar de manera d'obtenir-les amb el format amb el que hem treballat:

- `ID (URL): society/2014/jan/15/philippines-child-sexual-abuse-inquiry`
- `Secció principal: society`
- `Tags: society/childprotection society/children society/social-care society/society world/philippines world/asia-pacific world/world uk/police uk/uk tone/news profile/conalurquhart type/article`

Un cop aconseguit això, ja tindríem la informació de la nova font de dades amb el format vàlid per poder reutilitzar el codi i caldria modificar els següents aspectes:

1. Seccions a ignorar. Mirar si la nova font de dades té seccions d'aquest tipus i actualitzar la llista de la classe Profile:

```
private List<String> toignore = Arrays.asList("type","tone", "profile", "theguardian", "publication", "theobserver");
```

2. Caldria treure estadístiques del nou corpus. A la classe Profile tenim el paràmetre MAX_RANDOM_TAG establert a 38 perquè era la mitjana de número de *tags* per notícia dels dos corpus amb què hem treballat. S'hauria d'actualitzar aquest valor a la nova realitat.
3. Modificar el menú principal del programa i modelar nous perfils ideals per dur a terme les proves.
4. Canviar el nom de la classe 'Article' per un que s'adaptés a la nova font de dades, si és que no identifica prou bé els continguts. En alguns casos aquest nom funcionaria igual. Per exemple, si parlem d'articles científics.

A nivell d'escalabilitat, per seguir treballant amb 'The Guardian', no caldria fer modificacions en el codi per poder treballar amb nous corpus i permetria també agafar conjunts d'entrenament molt més grans, ja que el menú principal ens ofereix la possibilitat de descarregar-nos totes les notícies del diari britànic entre dues dates determinades.

5.6 Aplicacions

Després de futures investigacions i una vegada madurat del tot, l'algorisme desenvolupat en aquest projecte de final de carrera es podria utiltizar en qualsevol aplicació de lectura de documents, segurament pensada per executar-se via web o en un telèfon mòbil, que anés fent un seguiment d'eleccions de continguts per part de l'usuari, fent-hi clic o seleccionant-los en una pantalla tàctil. Aquestes tries serien les que nosaltres hem modelat com a perfil ideal, que vindria a ser el cervell de l'usuari i per això hem insistit en què aquest perfil hauria de poder ser canviant al llarg del temps.

En el cas concret en el qual ens hem centrat, podríem pensar en una utilització de l'algorisme al web de 'The Guardian 'o en una aplicació per a mòbil que mostrés les notícies d'aquest rotatiu. L'algorisme podria servir per crear un portal personalitzat per a cada usuari que aniria essent cada vegada més acurat a mesura que anés passant el temps i l'usuari anés acumulant més eleccions de notícies. També es podria fer servir de manera que, sense sortir de la portada habitual del diari, hi hagués un espai reservat per a notícies recomanades.

5.7 Sobre l'experiència

Ha estat molt interessant poder treballar en aquest projecte de final de carrera per l'interès que suscita a l'autor el binomi que formen la informàtica i el periodisme. Endinsar-se en els algorismes d'aprenentatge aplicats a aquest sector s'ha considerat molt positiu per tal de poder treballar en una aplicació de la intel·ligència artificial que té un objectiu molt clar. En l'era d'Internet i en què cada vegada tenim més informació, ens interessa que els ordinadors ens ho donin tot una mica mastegat. Aquest algorisme va en aquesta línia i el fet de veure-hi una possible futura aplicació molt pràctica ha estat molt motivant.

Com a colofó dels estudis d'Enginyeria en Informàtica, ha estat bé poder fer un projecte de principi a fi, dins de les limitacions temporals d'un semestre, que fan difícil poder complir amb totes les expectatives inicials. No obstant això, es considera que s'ha abarcat el tema proposat de manera prou extensa i la valoració final de l'assignatura és postiva perquè permet posar en pràctica els coneixements adquirits al llarg de la carrera i de la trajectòria personal. És una bona possibilitat per fer una síntesi de tot plegat.

6 Bibliografia

- L. Marin, D. Isern, A. Moreno, Dynamic adaptation of numerical attributes in a user profile, Applied Intelligence 39 (2) (2013), 421-437.
- L. Marin, A. Moreno, D. Isern, Automatic preference learning on numeric and multi-valued categorical attributes, Knowledge-Based Systems 56 (1) (2014), 201–215.
- The Guardian, Open Plattform. Build applications with The Guardian (2013), URL: http://www.theguardian.com/open-platform (darrer accés 1/5/2014).

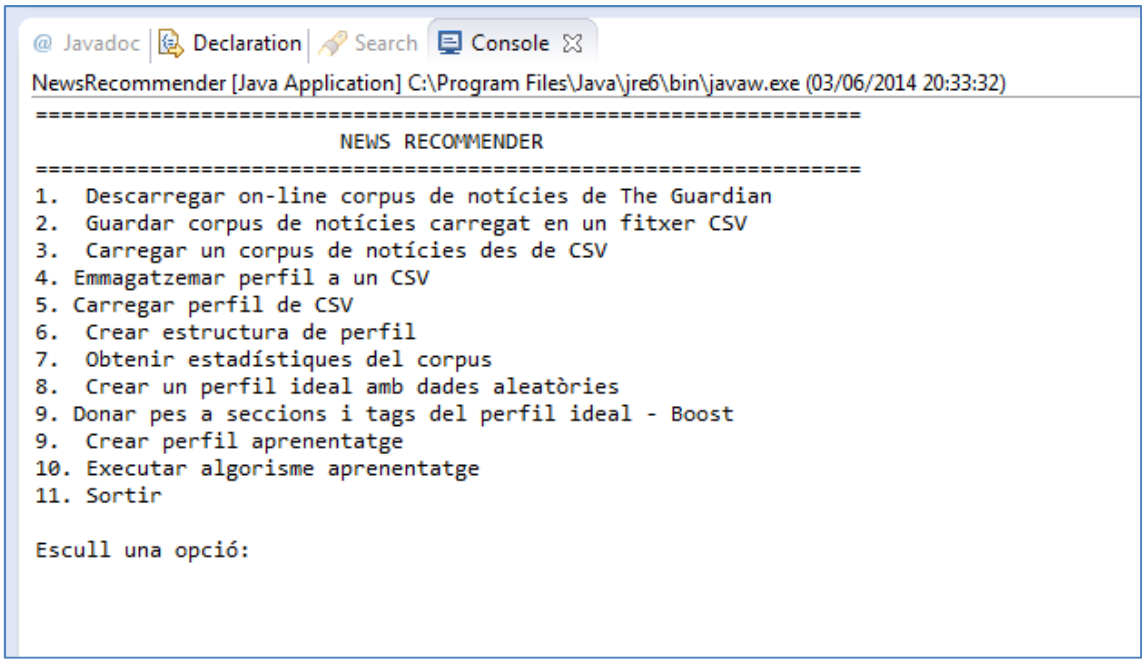
7 Annexos

En aquest apartat s'inclouen continguts que, tot i no formar part rigorosament de la memòria, poden ajudar el lector a entendre millor la feina feta al llarg de tot el projecte.

7.1 Manual d'instruccions

La interacció amb el programa fruit d'aquest projecte és molt simple i no està dirigida a un usuari final perquè es tracta d'una simulació pensada a efectes de recerca sobre un bon mètode de recomanació de continguts. Al llarg de la memòria ja s'ha pogut veure com funcionava però s'ha cregut adient fer una resum del funcionament deslligat a cap exemple en concret.

En accedir al programa, podem veure via terminal d'Eclipse el següent menú principal:



Expliquem un a un el funcionament de cadascuna de les opcions:

1. Descarregar online un corpus de notícies de The Guardian. Aquesta funcionalitat ens demana una data d'inici i una data de finalització de la mostra de notícies que volem agafar. Una vegada especificada, el programa carrega

via JSON a través de l'API del diari totes les notícies compreses dins del rang especificat i les guarda en un objecte de la classe Corpus.

2. Guardar un corpus de notícies carregat en un fitxer CSV: Emmagatzema el corpus carregat a través de les opcions 1. o 2. del programa en un fitxer CSV. El programa ens deixa escollir el nom i el fitxer queda guardat al directori arrel de la carpeta de l'aplicació Java.

3. Carregar un corpus de notícies des d'un CSV: Estableix com a corpus de notícies l'emmagatzemat en el fitxer CSV del qual ens demana el nom. Carrega un corpus de notícies emmagatzemat prèviament amb l'opció 2. del programa, tot i que també carregaria un fitxer elaborat per altres mitjans sempre que seguís el format correcte.

4. Emmagatzemar un perfil a un CSV: Emmagatzema un perfil en un fitxer CSV amb el nom que li especifiquem i el prefix 'profile'. Quan seleccionem aquesta opció ens apareix un segon menú que ens deixa triar si volem emmagatzemar el perfil ideal que hi ha carregat en aquell moment determinat o bé l'evolutiu. Quan s'emmagatzema un perfil, se'ns generen dos fitxers al directori arrel del programa:

    a. *profileNOM.csv*: És l'estructura de perfil que llegeix el programa i que hem d'utlitzar en cas de voler-lo carregar més endavant.
    b. *profileNOM2.csv*: És l'estructura de perfil però sense la informació dels *tags*, només les seccions. Es crea per poder analitzar amb més comoditat l'aprenentatge de les seccions.

5. Carregar un perfil d'un CSV: Carrega un perfil prèviament emmagatzemat en un fitxer CSV amb l'opció 4. o elaborat per altres mitjans seguint el mateix format. Ens deixa especificar si volem carregar el perfil del fitxer com a ideal o com a evolutiu.

6. Crear estructura de perfil: Aquesta opció no ens demana res i s'executa automàticament. És important executar-la sempre després de la càrrega d'un nou corpus al programa o, altrament, no funcionaria. Fa un escaneig de totes les seccions i *tags* presents al corpus per crear una estructura de perfil en blanc.

7. Obtenir estadístiques del corpus: Crea els fitxers *tags_*count.csv i *seccions_count.csv* al directori arrel del programa amb les estadístiques d'aparicions de *tags* i seccions en el corpus que tenim carregat.

8. Crear un perfil ideal amb dades aleatòries: Omple el perfil ideal de la classe Learner amb dades aleatòries.

9. Donar pes a seccions i *tags* del perfil ideal: Aquesta opció només informa que en aquesta versió del programa aquest procés s'ha de fer de manera manual tal com s'ha recollit a la memòria.

10. Executar algorisme d'aprenentatge: Una vegada tinguem el corpus carregat, l'estructura de perfil creada i el perfil ideal inicialitzat, podem seleccionar aquesta opció i s'executa l'algorisme d'aprenentatge. Ens informa per pantalla de tot el procés (cal anar prement una tecla per avançar en les teracions) i crea diversos fitxers amb dades de log que s'han utilitzat per elaborar els gràfics de la memòria:

    a. *learning.csv* té 5 columnes. La primera indica el nombre d'iteració, la segona el nombre de notícies 'overranked' que hi ha acumulades en aquella iteració. El mateix a la tercera columna però fent referència a les noticies 'loved'. Les altres dues columnes corresponen a la secció a la qual pertany l'article triat en aquella iteració. La primera fa referència al perfil ideal i la segona a l'evolutiu.
    b. *distanciaSeccio.csv*. Conté una columna amb el número d'iteració i una amb la distància euclidiana entre seccions en aquell moment determinat.
    c. *distanciaTags.csv*. Conté una columna amb el número d'iteració i una amb la distància euclidiana entre *tags* en aquell moment determinat.

11. Sortir del programa: Surt del programa.

Ales carpetes *estudiPerfils* i *estudiParàmetres* es poden recuperar els fitxers que s'han fet servir a l'apartat de supòsits d'aprenentatge.

7.2 Estadístiques dels corpus utilitzats

7.2.1 Corpus 1

S'han agafat totes les notícies del web de The Guardian compreses entre l'1 i el 15 de gener del 2014. En total, n'hi ha 5837. Fitxer: *corpus1.csv*.

Taula amb el nom de les 110 seccions indexades i el nombre d'aparicions de cadascuna d'elles. S'entén com a aparició el nombre de notícies que estan codificades en la secció en concret, sigui la principal o la d'un tag que la representa:

world	**6283**
sport	3412
football	2402
uk	2032
business	2002
politics	1748
society	1719
culture	1640
media	1497
lifeandstyle	1368
technology	1259
music	1235
environment	1020
books	917
film	901
money	723
education	524
commentisfree	522
tv-and-radio	498
science	417
artanddesign	414
stage	323
travel	307
fashion	295
law	271
global-development	224
sustainable-business	145
small-business-network	116
media-network	88
crosswords	79
healthcare-network	78
teacher-network	77
news	75
local-government-network	74

global-development-professionals-network	73
social-care-network	72
higher-education-network	71
public-leaders-network	68
culture-professionals-network	57
social-enterprise-network	49
childrens-books-site	48
housing-network	48
guardian-professional	43
voluntary-sector-network	37
guardian-masterclasses	30
best-awards	29
cities	28
weather	27
women-in-leadership	26
uk-news	20
gnm-press-office	13
extra	10
owntheweekend	9
powwownow-partner-zone	7
global	6
gnmeducationcentre	6
speedo-swim-to-fitness	5
media-network-outbrain-partner-zone	5
data	5
british-council-partner-zone	4
duracell-power-me	4
info	4
media-network-microsoft-partner-zone	4
marketing-agencies-association-partner-zone	4
advertising	3
randstad-partner-zone	3
what-is-nano	3
partner-zone-ageing-population	3
voluntary-sector-network-zurich-partner-zone	3
social-enterprise-partner-zone-the-co-operative	3
local-government-network-serco-partner-zone	3
recruiters	3
efficiency-exchange-partner-zone	3
lifeandhealth	3
intelesant-partner-zone	3
social-care-network-skills-for-care-partner-zone	3
teacher-network-hays-partner-zone	3
university-nottingham-partner-zone	2

salesforce-partner-zone	2
discover-englands-forests	2
voluntary-sector-network-mydonate-partner-zone	2
hyundai-family-adventure	2
living-with-cancer-macmillan-partner-zone	2
malaria-consortium-partner-zone	2
katine	2
eon-energy-partner-zone	2
partner-zone-path	2
direct-line-for-business-partner-zone	2
teacher-network-advertisement-features	2
housing-network-partner-zone-pinnacle	2
sustainable-business-fairtrade-partner-zone	2
partner-zone-sas-computacenter	2
adam-smith-international-partner-zone	2
help	2
small-business-network/powwownow-partner-zone	1
higher-education-network/efficiency-exchange-partner-zone	1
media-network/partner-zone-microsoft	1
global-development-professionals-network/malaria-consortium-partner-zone	1
media-network/marketing-agencies-association-partner-zone	1
healthcare-network/intelesant-partner-zone	1
small-business-network/eon-energy-partner-zone	1
global-development-professionals-network/partner-zone-path	1
social-care-network/skills-for-care-partner-zone	1
teacher-network/hays-partner-zone	1
small-business-network/direct-line-for-business-partner-zone	1
teacher-network/advertisement-features	1
housing-network/partner-zone-pinnacle-PSG	1
sustainable-business/fairtrade-partner-zone	1
local-government-network/partner-zone-sas-computacenter	1
global-development-professionals-network/adam-smith-international-partner-zone	1

Per veure cadascun dels *tags* i les seves aparicions, consultar *tags*_count_corpus1.csv al directori de NewsReccomender (4.303 *tags*).

7.2.2 Corpus 2

S'han agafat totes les notícies del web de The Guardian compreses entre l'1 i el 15 de febrer del 2014. En total, n'hi ha 6432. Fitxer *corpus1.csv*.

Taula amb el nom de les 123 seccions indexades i el nombre d'aparicions de cadascuna d'elles. S'entén com a aparició el nombre de notícies que estan codificades en la secció en concret, sigui la principal o la d'un tag que la representa:

world	**6653**
sport	3912
football	2267
uk	2187
politics	1929
business	1890
lifeandstyle	1713
culture	1711
media	1709
society	1689
environment	1442
technology	1341
music	1183
books	1158
film	898
education	704
money	649
artanddesign	629
science	526
commentisfree	507
stage	461
tv-and-radio	460
travel	339
fashion	338
law	300
global-development	225
small-business-network	181
sustainable-business	162
higher-education-network	123
global-development-professionals-network	119
media-network	112
childrens-books-site	95
healthcare-network	93
cities	84
teacher-network	83

crosswords	80
public-leaders-network	78
local-government-network	78
news	67
voluntary-sector-network	66
culture-professionals-network	63
guardian-professional	54
social-care-network	53
housing-network	43
social-enterprise-network	35
end-fgm	35
guardian-masterclasses	31
uk-news	29
weather	29
women-in-leadership	24
extra	22
info	20
global	13
australia-2014-reasons	10
appetite-for-life	9
gnm-press-office	7
shelf-improvement	7
british-academy-partner-zone	6
open-university-partner-zone	6
global-supply-chains-summit	6
observer-ethical-awards	5
guardian-us-press-office	5
emirates-foundation-partner-zone	5
marketing-agencies-association-partner-zone	5
dulux-lets-colour-awards	4
duracell-power-me	4
partner-zone-college-of-law	4
discover-culture	4
media-network-outbrain-partner-zone	4
lifeandhealth	4
bupa-care-homes	3
voluntary-sector-network-caf-partner-zone	3
university-arts-london-partner-zone	3
partner-zone-ageing-population	3
gnmeducationcentre	3
data	3
adam-smith-international-partner-zone	3
powwownow-partner-zone	3
social-care-network-advertisement-features	3
social-enterprise-partner-zone-the-co-operative	2

salesforce-partner-zone	2
costa-del-sol-always-warm	2
connect4climate-partner-zone	2
british-council-partner-zone	2
what-is-nano	2
living-with-cancer-macmillan-partner-zone	2
university-nottingham-employability	2
gnm-archive	2
higher-education-hea-partner-zone	2
media-network-partner-zone-ebay	2
randstad-partner-zone	2
healthcare-network-advertisement-features	2
help	2
accenture	2
partner-zone-path	2
cardiff	2
direct-line-for-business-partner-zone	2
sustainability	2
efficiency-exchange-partner-zone	2
teacher-network-partner-zone-zurich	2
public-leaders-network-partner-zone-solace	2
dai-partner-zone	2
eon-energy-partner-zone	2
media-network-partner-zone-brand-union	2
clements-partner-zone	2
nihr-crn-partner-zone	2
voluntary-sector-network-advertisement-featur	2
global-development-professionals-network/adam-smith-international-partner-zone	1
healthcare-network/advertisement-features	1
small-business-network/powwownow-partner-zone	1
global-development-professionals-network/emirates-foundation-partner-zone	1
global-development-professionals-network/partner-zone-path	1
media-network/marketing-agencies-association-partner-zone	1
social-care-network/advertisement-features	1
small-business-network/direct-line-for-business-partner-zone	1
higher-education-network/efficiency-exchange-partner-zone	1
teacher-network/partner-zone-zurich	1
global-development-professionals-network/dai-partner-zone	1

small-business-network/eon-energy-partner-zone	1
media-network/partner-zone-brand-union	1
global-development-professionals-network/clements-partner-zone	1
healthcare-network/nihr-crn-partner-zone	1
voluntary-sector-network/advertisement-features	1

Per veure cadascun dels *tags* i les seves aparicions, consultar *tags*_count_corpus2.csv al directori de NewsReccomender (4.456 *tags*).