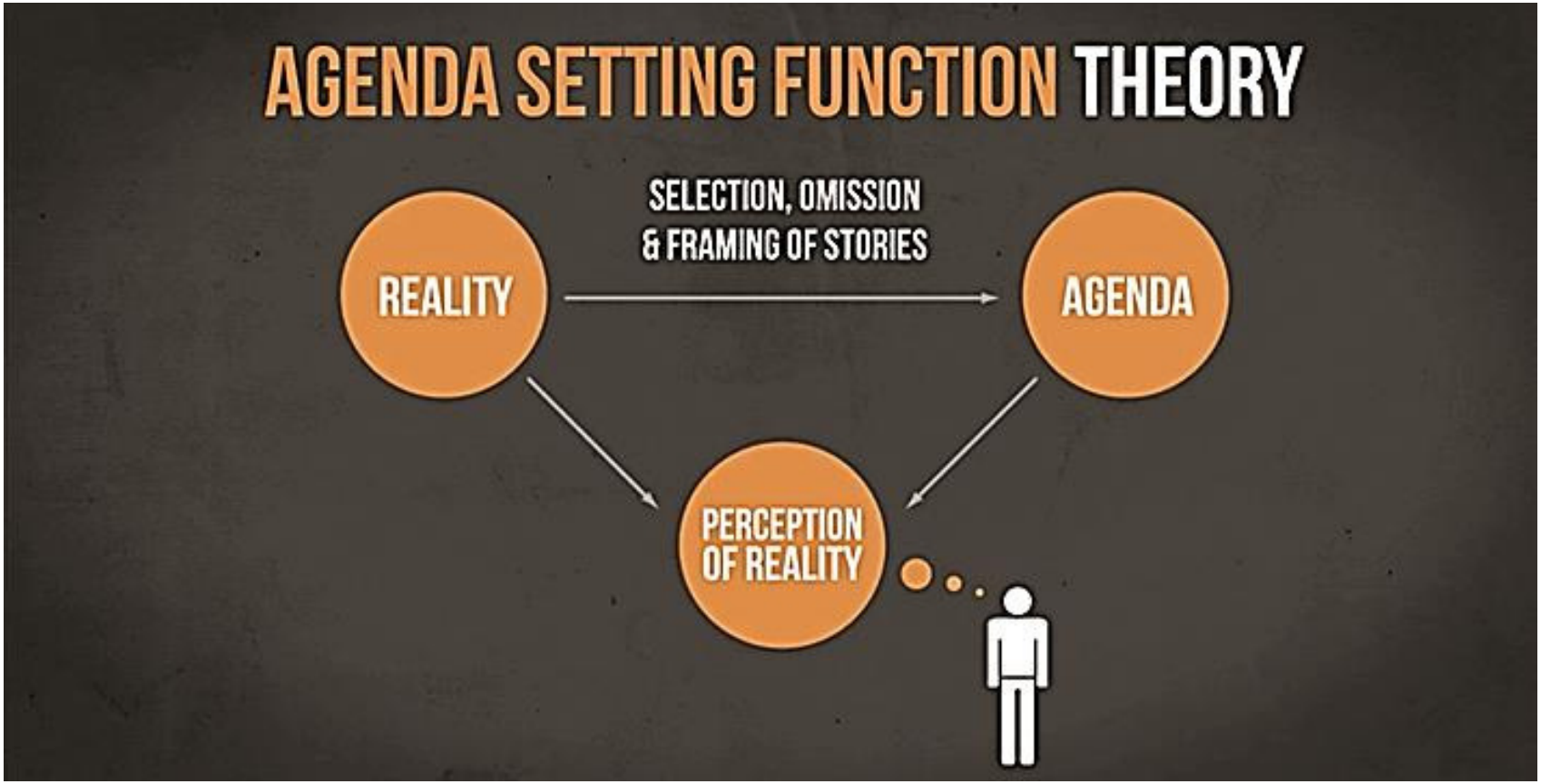
Sistema automàtic d'aprenentatge de preferències personals sobre notícies classificades en seccions i definides per paraules clau

Jordi Pueyo Busquets
Enginyeria en Informàtica

Tutor:
David Isern Alarcón

11 de juny del 2014

Introducció



Font: lessonbucket.com

Problema a resoldre

Disposant d'un **corpus d'articles** de qualsevol temàtica:

- Cada article ha de pertànyer a una **secció principa**l i tenir associat un **conjunt de *tags***
- Simulació de les tries d'un lector entre un conjunt de documents: **perfil ideal**
- Aprenentatge no supervisat de les accions d'un lector: **perfil evolutiu**
- Mesurar com de bé s'ha efectuat l'aprenentatge: **distància entre perfils**

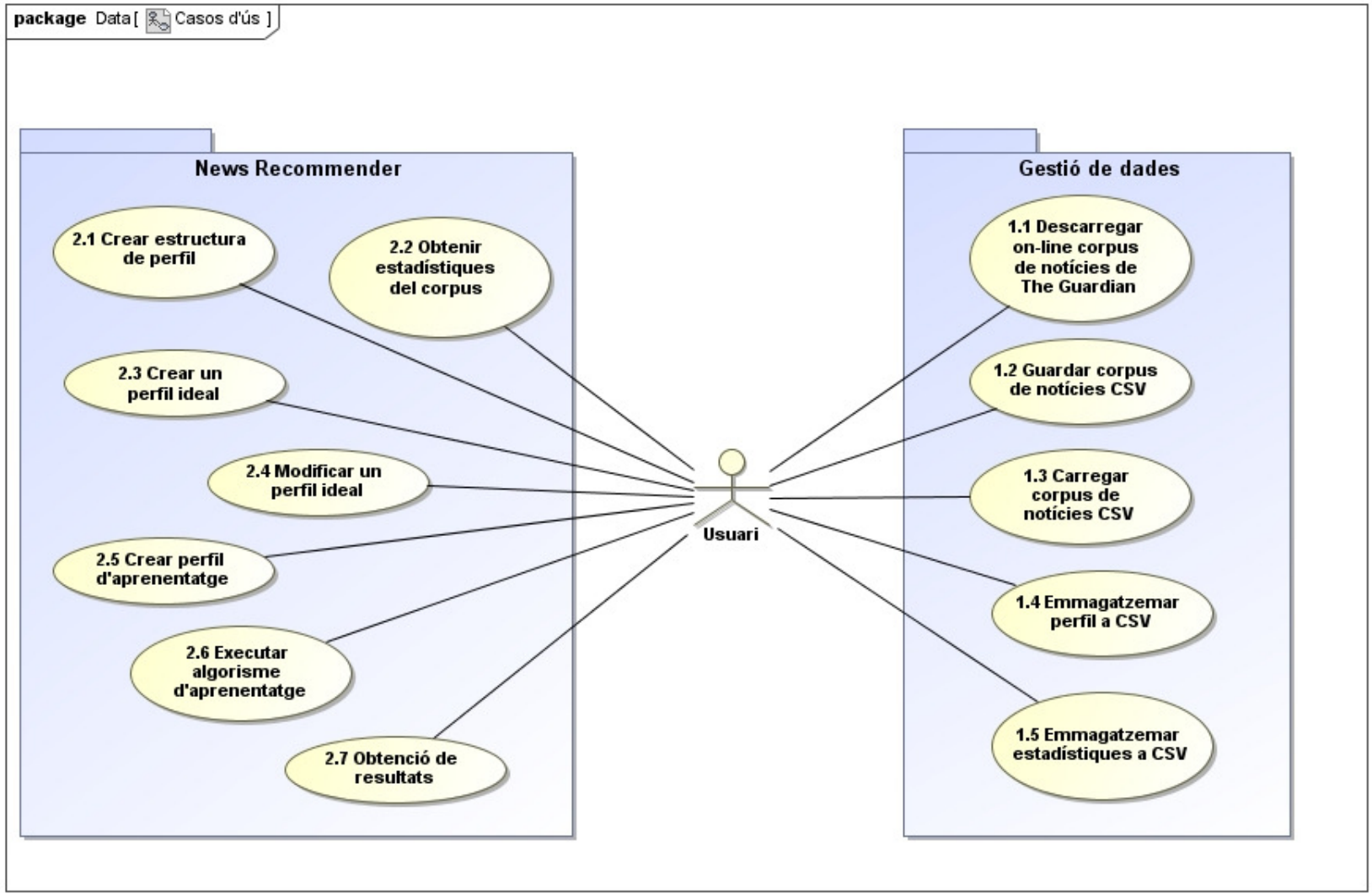
Enfocament i mètode seguit

- Dos corpus de notícies del diari '**The Guardian**' d'unes 6.000 notícies cadascun
- En cada iteració de l'algorisme tractarem 15 notícies
- A cada article se li assignarà un ***rating*** segons el perfil ideal i un altre segons l'evolutiu
- Amb les diferències entre l'elecció dels dos perfils, podrem fer que **el perfil evolutiu vagi aprenent**

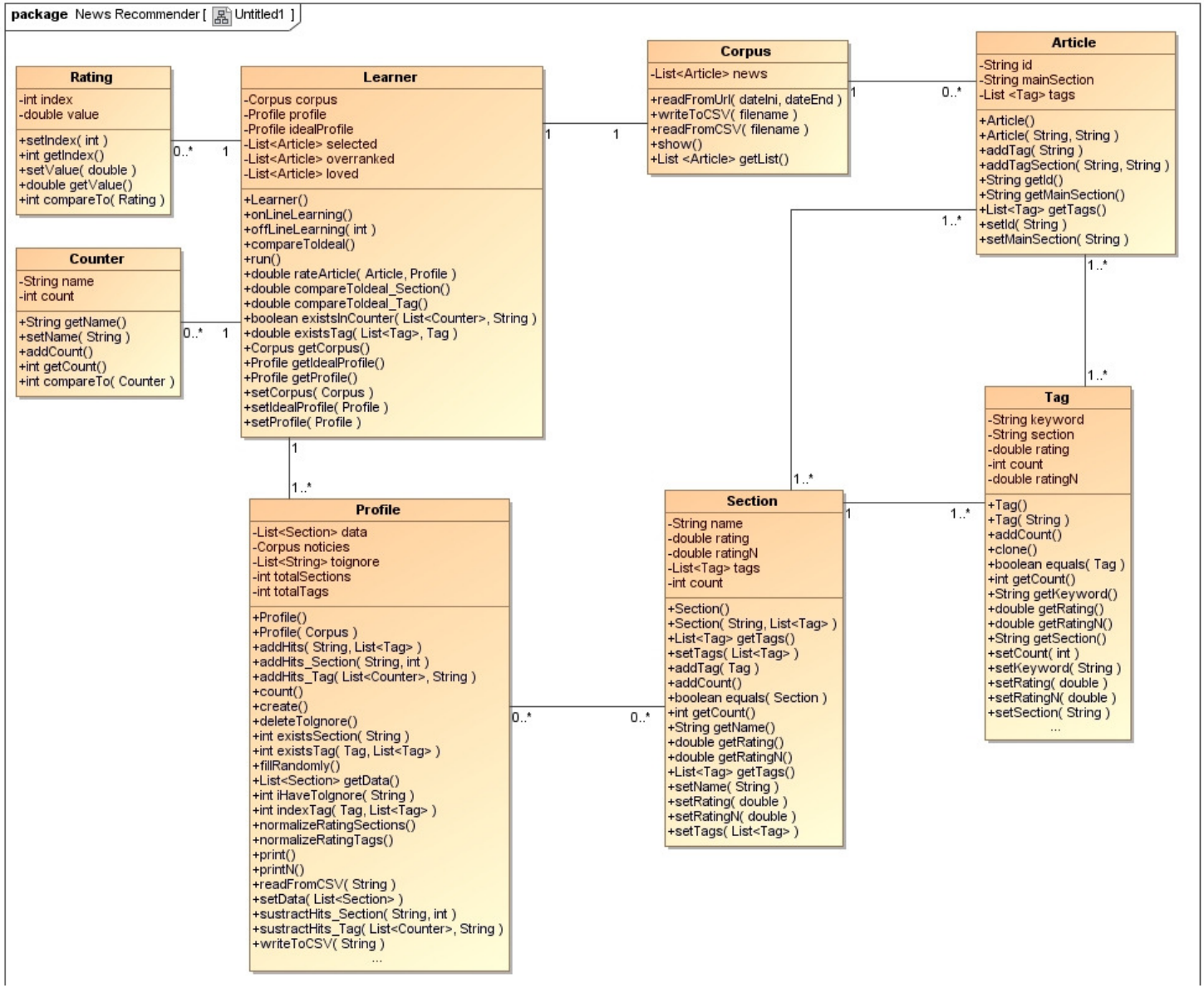
Fases del projecte

- **ANÀLISI I DISSENY**
- **IMPLEMENTACIÓ**
- **ESTUDI DE DIVERSOS SUPÒSITS**
- **CONCLUSIONS I FUTUR**

Anàlisi i disseny: Casos d'ús



Anàlisi i disseny: Diagrama de classes



Anàlisi i disseny: Algorisme d'aprenentatge

- Un **Profile** està format per una llista de **seccions**, cadascuna de les quals amb un *rating* global
- Cada secció té un conjunt de ***tags*** associats i cadascun d'ells també compta amb el seu propi *rating*

L'algorisme d'aprenentatge seguirà les accions de l'usuari (perfil ideal) i anirà modificant els pesos del perfil evolutiu

Anàlisi i disseny: Algorisme d'aprenentatge

rateArticle (Article, Profile)

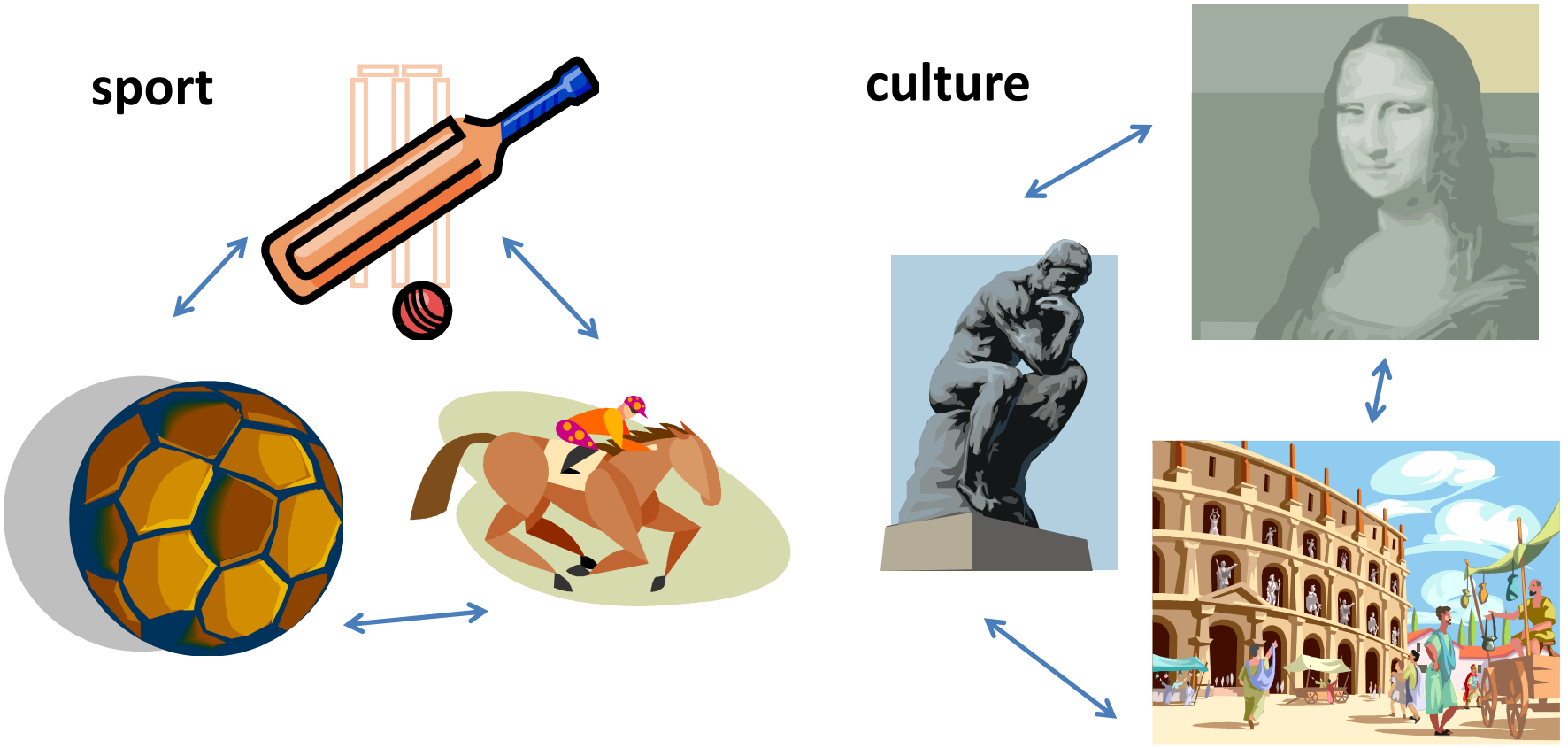
rating = α*valoració_*tags* + β*valoració_secció

Valoració_tags: Agafem tots els tags de la notícia, busquem els pesos que tenen al perfil i els anem sumant

Valoració_secció: Agafem la valoració que la secció té al perfil

Anàlisi i disseny: Algorisme d'aprenentatge

Per què diferenciem la valoració de la secció i la dels tags?



Anàlisi i disseny: Algorisme d'aprenentatge

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 (**loved**), 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

Anàlisi i disseny: Algorisme d'aprenentatge

offLineLearning()

cas (tipus):

overranked:

Fer un recompte de les aparicions de seccions d'entre totes les notícies que hi hagi en aquest moment determinat a la llista overranked

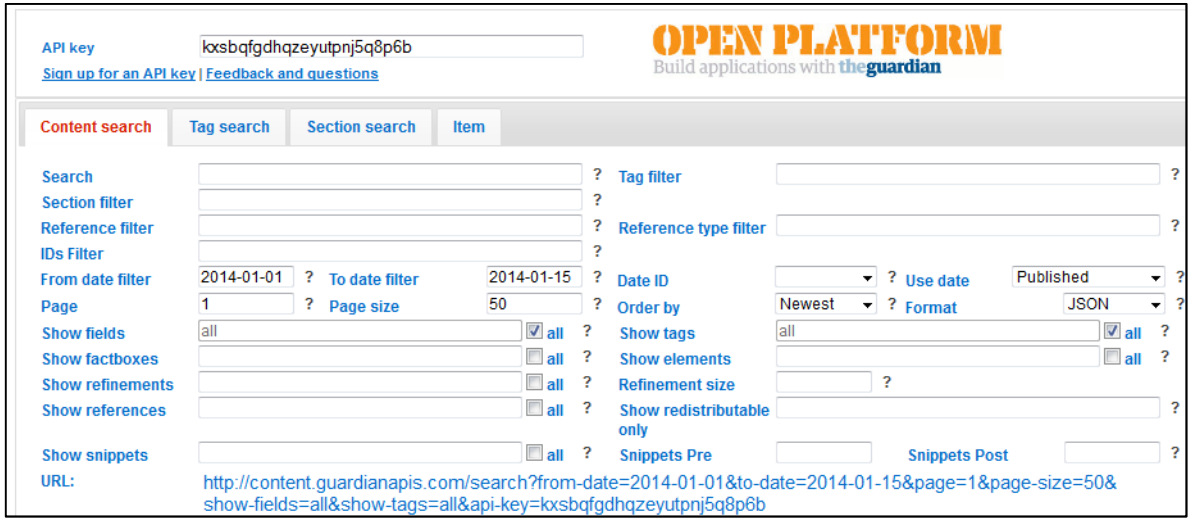
- **Mirar quines seccions superen el PERCENTATGE_SUPERAR (10%)** i esborrar les que no el superin.
- **Ordenar-les** segons el número d'aparicions
- **Per a cada secció d'aquesta llista, restar al perfil evolutiu el nombre de hits que té associats. La magnitud de la resta la marcarà el nombre d'aparicions (superior al 10%).**
- 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.
- **Ens quedarem només amb els més repetits**. NUM_TAGS_OFFLINE* tindrà un valor de deu.
- **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: *(Igual però en positiu)*

fcas

Anàlisi i disseny: Altres

- Valorar l'**eficàcia** de l'aprenentatge: distància euclidiana entre seccions i entre tags (*ratings* normalitzats amb *ranging*)
- Seccions a **ignorar**
- **Repositori** de dades del diari 'The Guardian':



Implementació: 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.**

Implementació: Codi

```
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);
```

Implementació: Codi

```
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;
        }
```

Implementació: Codi

```
//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);
                    }
```

Estudi: Càrrega de dades

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

Estudi: Càrrega de dades

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
==========

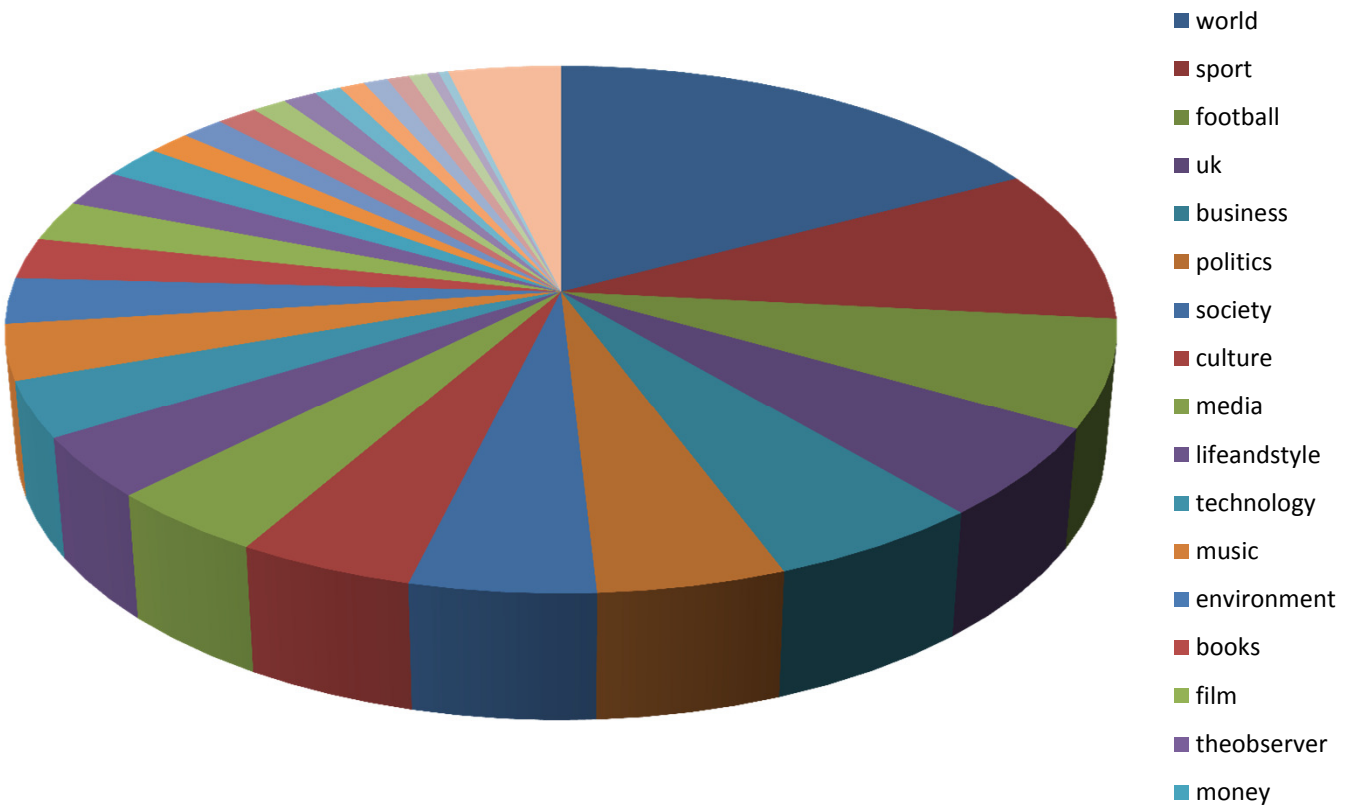
Estudi: Estructura de perfil

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)

Estudi: Corpus1, seccions



Exemple d'aprenentatge

- Creació del **perfil ideal d'una persona de negocis**. Dades aleatòries i següents seccions i *tags* ressaltats:

```
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
```

Exemple d'aprenentatge

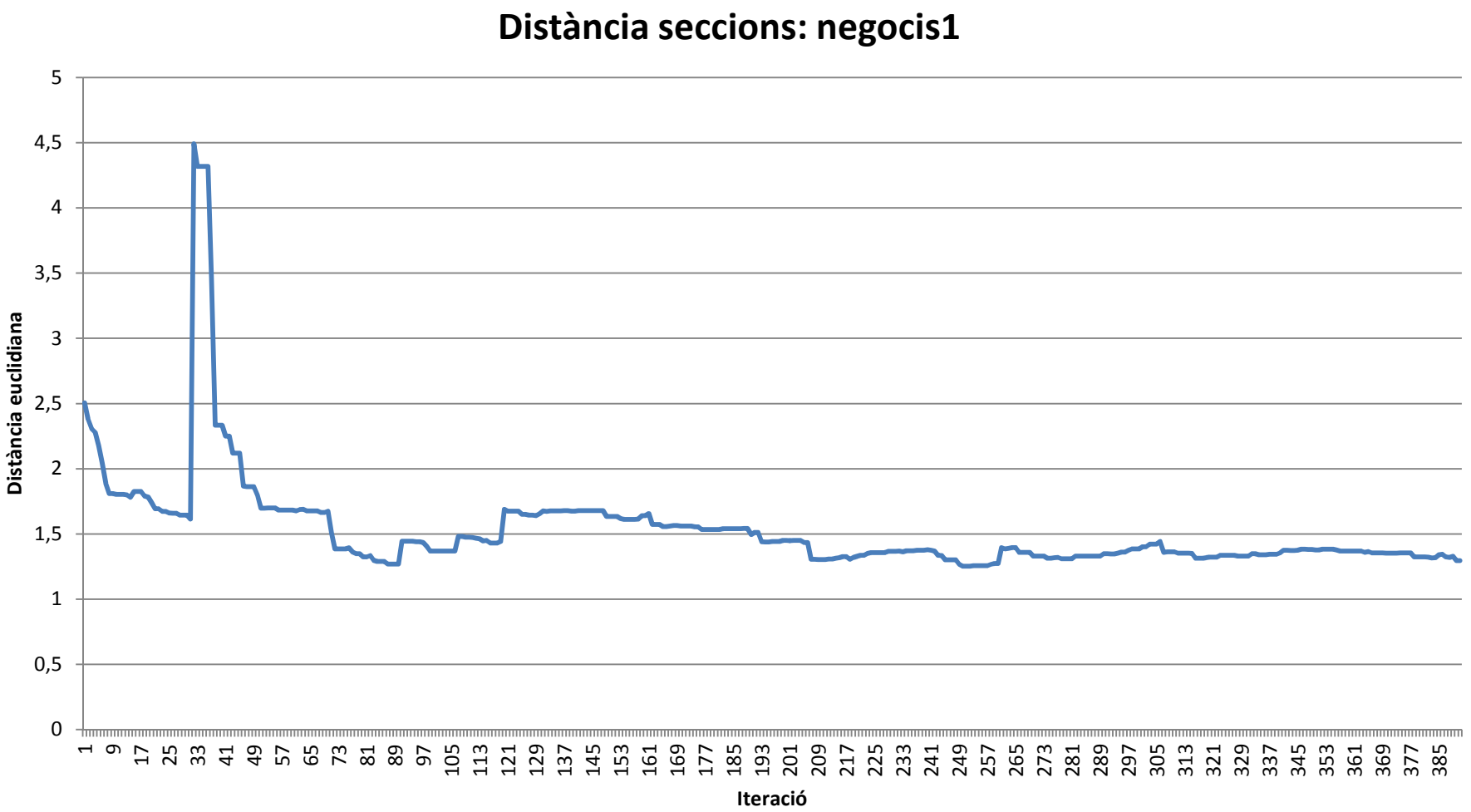
- **Càrrega** del perfil i execució de l'algorisme:

```
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)
```

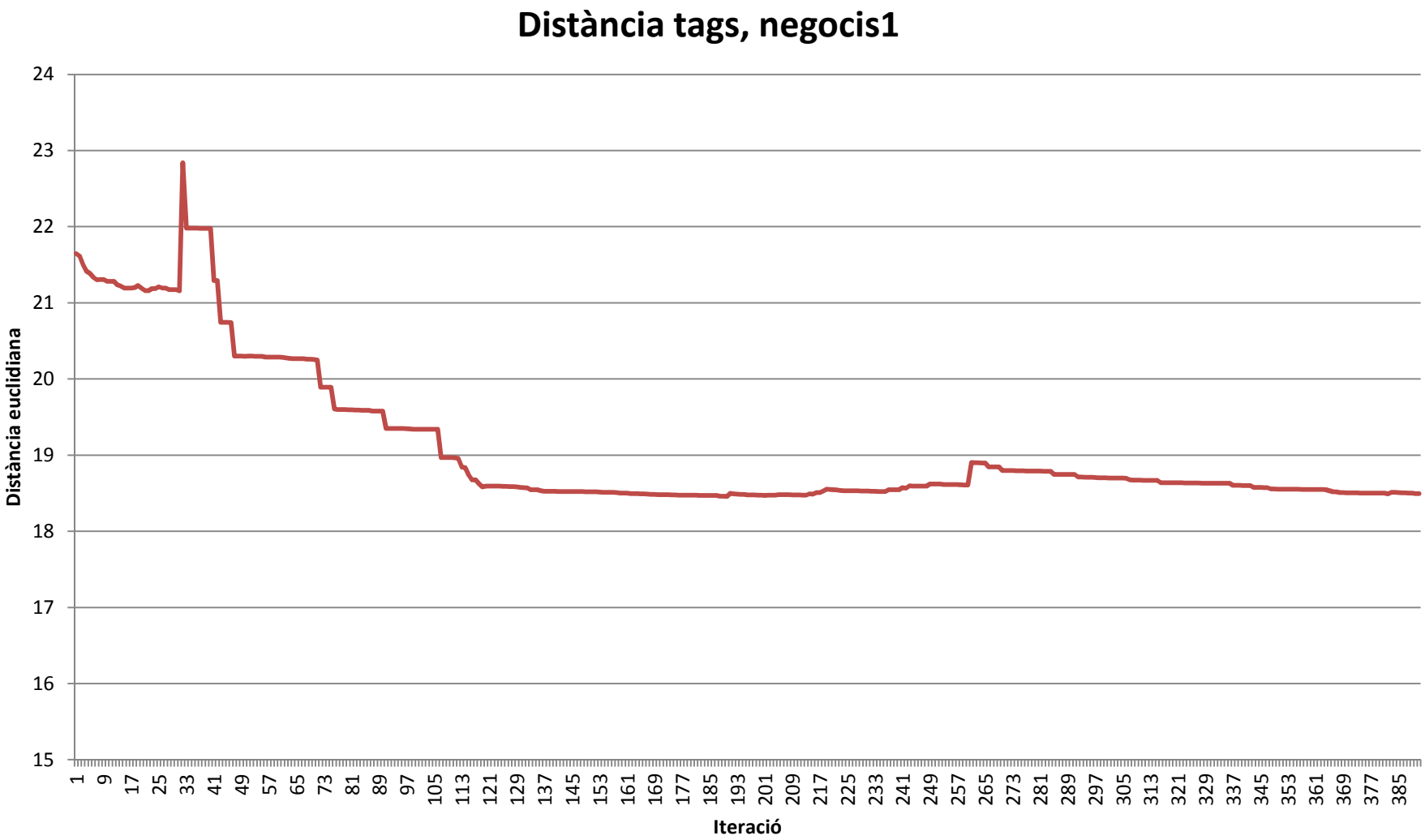
	Distància entre seccions	Distància entre grups de *tags*
Inici	2.504180081383925	21.647480034974407
Final	1.295941043290216	18.493673799014562

```
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
```

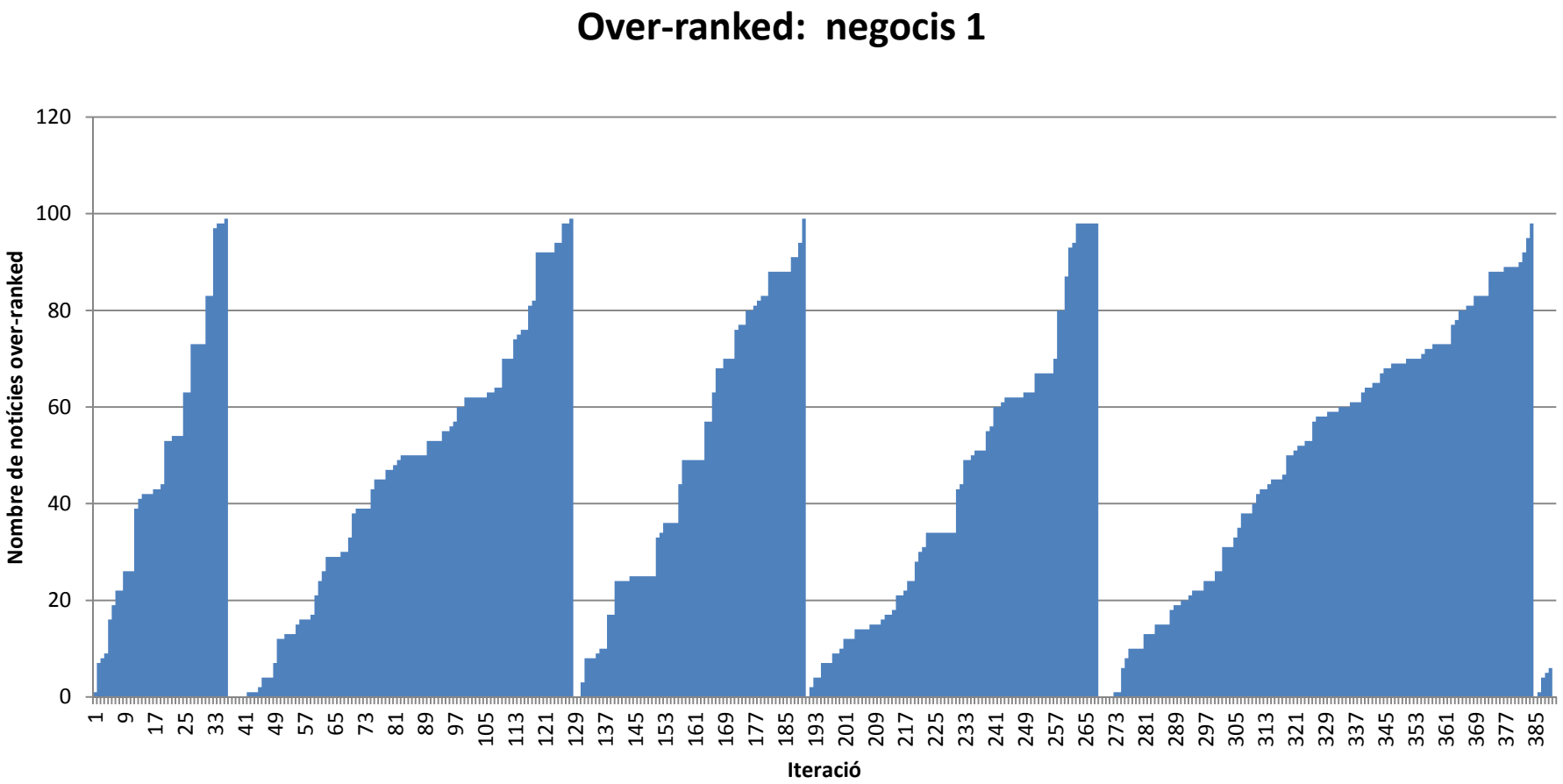
Exemple d'aprenentatge



Exemple d'aprenentatge



Exemple d'aprenentatge



Exemple d'aprenentatge

IDEAL PROFILE
business
technology
world
politics
culture
music
uk
stage
travel
sustainable-business
film
tv-and-radio
environment
artanddesign
lifeandstyle
education
fashion
sport
media
money
society

EVOLUTIVE PROFILE
world
sport
politics
media
society
business
technology
uk
environment
education
commentisfree
culture
money
higher-education-network
tv-and-radio
football
music
artanddesign
global-development
healthcare-network
travel

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

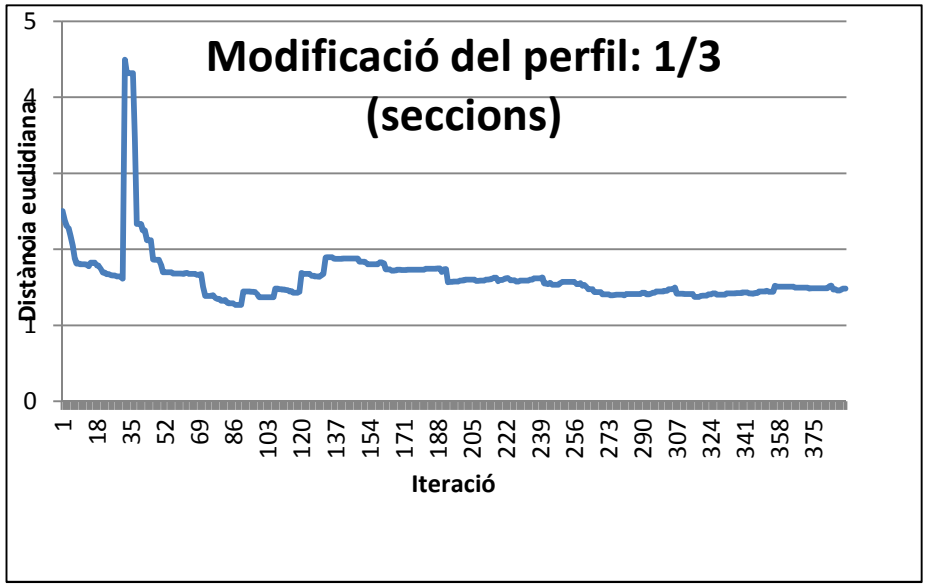
Exemple d'aprenentatge

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

Exemple d'aprenentatge

- **Modificació del perfil ideal** a mitja execució, a 1/3 de les iteracions. Donem pes a 'money' I 'education'

AMB MODIFICACIÓ	SENSE MODIFICACIÓ
politics	world
sport	sport
world	politics
media	media
business	society
technology	business
education	technology
society	uk
money	environment
uk	education
environment	commentisfree
commentisfree	culture
culture	money
global-development	higher-education-network

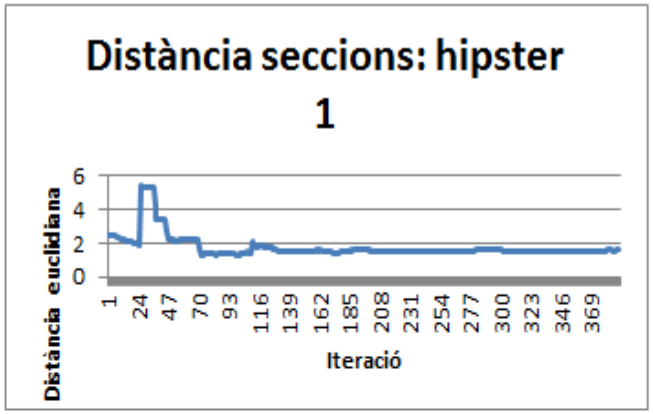


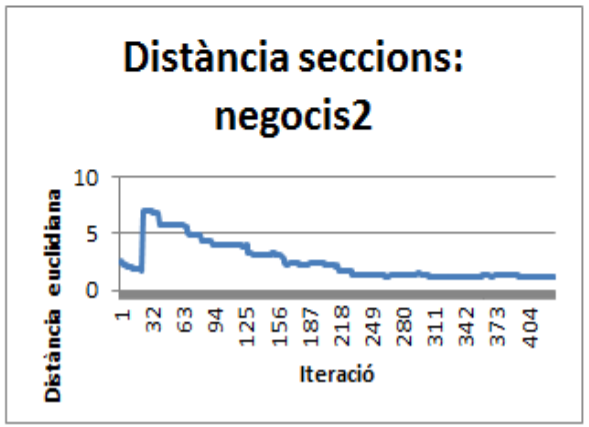
Conclusions

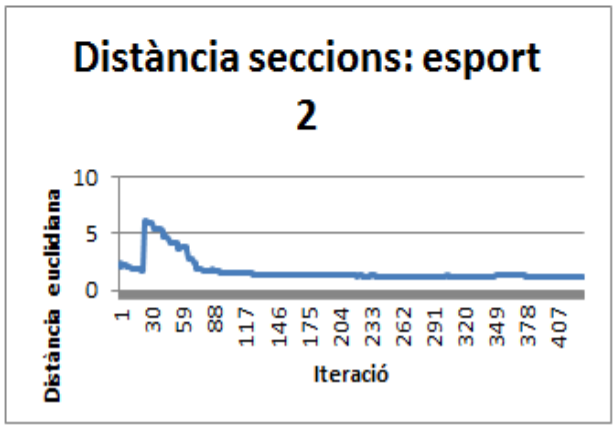
- **Reducció d'un 47% de la distància euclidiana entre seccions**, patró constant

	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

Conclusions





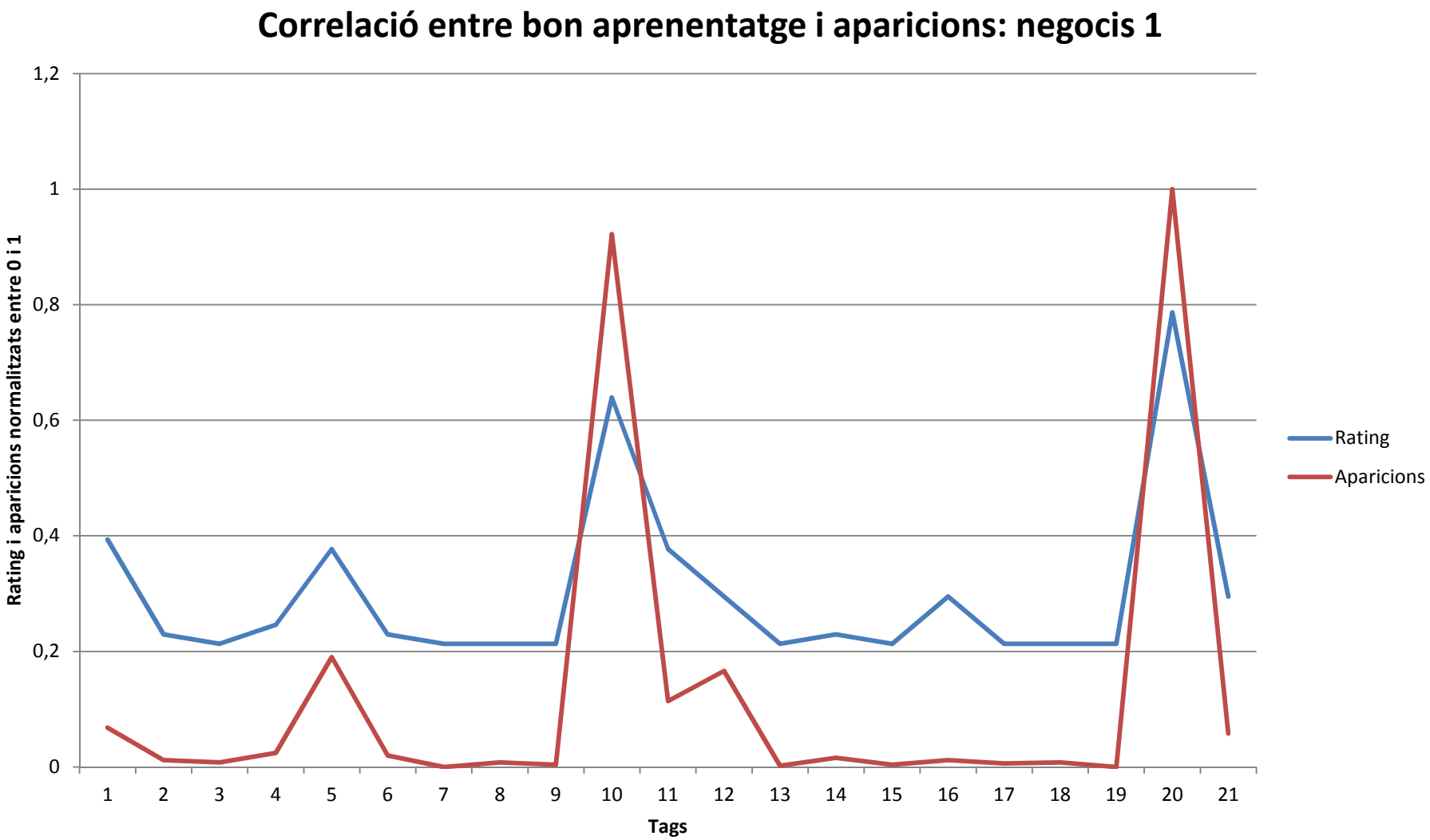


Corba d'aprenentatge que es va repetint. A més a més, el prototipus aprèn bé les seccions destacades al perfil ideal

Conclusions

- **Reducció d'un 14% de la distància euclidiana entre *tags***, patró constant.
- **Percentatge baix**. Cada corpus té més de 4.000 *tags* diferents, mentre que només un centenar de seccions
- Caldria un **procés d'aprenentatge més llarg**
- Bon aprenentatge **proporcional al nombre d'aparicions del *tag***

Conclusions



Conclusions

- Cada vegada costa més omplir el grup **overranked**. L'aprenentatge va millorant
- **Validesa dels paràmetres** de l'aprenentatge offline
- **Possibles millores:** conjunts d'entrenament més grans, modelar el corpus, millorar la interfície, etc.
- Fàcil **escalabilitat** i **reutilització** del codi
- **Aplicació** real

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

jpueyob@uoc.edu