Universitat Oberta de Catalunya

uoc.edu

Títol: Anàlisi de les dades del Dia Mundial de les Malalties Minoritàries a Twitter.

Autor: Joaquim de Dalmases Juanet
Pla d'estudis: **Màster de Ciència de Dades (Data Science)**
TFM - Àrea 3: Machine learning en medicina

Consultor/a: Laia Subirats Maté / Elisenda Bonet Carne
Professor/a responsable de l'assignatura: Ferran Prados Carrasco

Data Lliurament: 24/06/2020


Aquesta obra està subjecta a una llicència de Reconeixement-NoComercial-SenseObraDerivada 3.0 Espanya de Creative Commons

FITXA DEL TREBALL FINAL

Títol del treball:	*Anàlisi de les dades del Dia Mundial de les Malalties Minoritàries a Twitter.*
Nom de l'autor:	*Joaquim de Dalmases Juanet.*
Nom del consultor/a:	*Laia Subirats Maté.* *Elisenda Bonet Carne.*
Nom del PRA:	*Ferran Prados Carrasco.*
Data de lliurament (mm/aaaa):	*06/2020.*
Titulació o programa:	*Màster de Ciència de Dades (Data Science).*
Àrea del Treball Final:	*Machine learning en medicina.*
Idioma del treball:	*Català.*
Paraules clau	*Malalties minoritàries, aprenentatge no supervisat, Twitter.*

Resum del Treball (màxim 250 paraules): *Amb la finalitat, context d'aplicació, metodologia, resultats i conclusions del treball*

Les malalties minoritàries són un problema social que afecta a drets de caràcter humanitari bàsics com la igualtat social. La majoria de pacients que les pateixen estan desemparats per afrontar-les. Per lluitar contra elles, és necessari definir el suport a la investigació, el desenvolupament de medicaments, la creació de xarxes entre grups de pacients, per tal d'intensificar la lluita, la sensibilització i moltes altres accions socials organitzades.

Amb l'objectiu de reduir l'impacte de les malalties rares en la vida de pacients i familiars, aquest treball caracteritza el contingut de les dades de Twitter captades al voltant del Dia Mundial de les Malalties Minoritàries de l'any 2020 per actuar en aquesta direcció.

Aquesta caracterització de dades socials, es realitzarà des de 2 punts de vista, estructural i d'anàlisi de continguts. S'aplicaran tècniques d'aprenentatge automàtic no supervisat per tal de trobar les comunitats d'usuaris emergents existents en aquesta temàtica.

Disposar de les comunitats d'usuaris ens permetrà donar veu i potenciar tots els aspectes que les uneix i disposar d'un criteri per la presa de decisions envers el seu grau d'atenció actual i el necessari en el futur.

Un altre repte és l'estructuració de mesures pràctiques i recomanacions d'accions envers les conclusions de l'anàlisi orientades al suport als pacients i lluita contra les malalties minoritàries.

Abstract (in English, 250 words or less):

Rare diseases are a social problem that has a significant impact on basic human rights like social equality. Most people are neglected and are alone and helpless to deal with them. To fight against them is necessary to define support for research, drug development, networking among patient groups to intensify rainfall, awareness and a lot of other organized social actions.

Intending to reduce the impact of this kind of disease on the life of every patient and his family, this thesis characterizes Twitter data around World Rare Disease Day in 2020 to act on that address. This characterization of social data will be performed from 2 points of view, structural and content analysis, to get emerging communities of users inside the rare disease topic.

Having emerging user communities allow us to empower and enhance all the aspects that create them and have a criterion for making decisions regarding their current level of attention and what is needed in the future.

Another challenge is the structuring of practical measures and action recommendations towards the conclusions of the analysis aimed at patient support and the fight against minority diseases.

Agraïments

Sobretot al suport de la meva família al complet, que sempre em recolzen en tot.

En general a tot el personal docent, companys i personal de la UOC del Màster de Ciència de Dades, que m'han ajudat a realitzar-lo i gaudir-lo.

Índex

Llista de figures

Llista de taules

1. Introducció

1.1 Context i justificació del Treball.

Aquest document descriu el treball final (TFM) del Màster de Ciència de Dades de la Universitat Oberta de Catalunya (UOC). El treball proposa, analitzar les dades recollides en la xarxa social Twitter, en un període de temps al voltant del Dia Mundial sobre les Malalties Minoritàries o Rares, que es dur a terme el darrer dia del mes de Febrer de cada any.

En els apartats següents es definirà quina és la temàtica i es justificarà el seu interès i/o rellevància, fixant els objectius principals, secundaris i paral·lels, així com els reptes que es volen assolir en la seva realització.

El tipus de dades utilitzades en l'estudi, provinents de la xarxa social Twitter, marcarà la seva metodologia i tecnologia dintre de l'àmbit de la ciència de dades, aprofitant les noves oportunitats que ofereixen l'aprenentatge automàtic i l'anàlisi de dades massives.

El dia internacional de les malalties minoritàries està coordinat mundialment per EURORDIS-Rare Diseases Europe i el seu Consell d'Aliances Nacionals (EURORDIS, 2015) des de 2008. És una aliança no governamental formada per 894 organitzacions de pacients amb malalties rares de 72 països que treballen junts per millorar la vida dels 30 milions de persones que viuen amb una malaltia rara a Europa. L'objectiu principal de la campanya és conscienciar al públic en general i als que prenen decisions sobre les malalties rares i el seu impacte en la vida dels pacients ("Rare disease day", 2020). Les xarxes socials com Twitter es converteixen en un altaveu extraordinari per descobrir quines comunitats de pacients existeixen i quines són les seves demandes.

A Espanya l'associació de caràcter nacional que coordina des de l'any 1999 l'acció local i regional de cada centre orientat a ajudar als pacients amb malalties rares és la Federación Española de Enfermedades Raras (FEDER, 2020) i també forma part de EURORDIS, junt amb la Federació Catalana de Malalties Minoritàries (FECAMM, 2020).

Una malaltia o trastorn es defineix com a minoritari a Europa, quan afecta a menys d'1 en 2000. Hi ha més de 6000 malalties minoritàries. En general, les malalties minoritàries poden afectar 30 milions de ciutadans de la Unió Europea. El 80% de les malalties minoritàries són d'origen genètic, i sovint són cròniques i potencialment mortals.

En aquest context, aquest treball s'orienta a la realització d'una anàlisi que caracteritzi les dades de Twitter recollides des d'un punt de vista estructural, relacional (conjunt d'interaccions existents) i d'anàlisi de continguts. L'objectiu final és aplicar un conjunt de tècniques d'aprenentatge automàtic no supervisat, que ens descobreixin les comunitats o grups afins existents, tendències en les demandes, necessitats emergents i les activitats manifestades pels pacients de les diferents malalties minoritàries. La repercussió final ha de ser potenciar la

seva difusió i amb la intenció de detectar possibles actualitzacions necessàries en el full de ruta establert per les organitzacions creades per donar-los suport.

Descripció de la Proposta

La proposta descrita anteriorment es basa a realitzar un procés de mineria de dades basat en xarxes socials, en concret la xarxa social Twitter durant l'any 2020. En tot aquest procés, s'aplicaran processos d'agregació de dades per anonimitzar les dades caracteritzades i respectar la privacitat de les dades.

La proposta inclou dues vies d'anàlisi:

- *Anàlisi estructural*: S'estudiarà i visualitzarà de manera descriptiva les dades recollides, com per exemple: nombre de tuits per dia, nombre de tuits per hora, grups d'usuaris amb el nombre de tuits més alt, caracterització geogràfica, concentració de tuits entre d'altres que puguin ser d'utilitat per prendre consciència de la magnitud global del conjunt de dades usat en l'anàlisi.

  Es mostraran les distribucions de dades de Twitter obtingudes al llarg del procés de captació per comunitats o clústers detectats. Per l'obtenció dels clústers s'aplicarà tècniques d'aprenentatge no supervisat on els algorismes que es prenen en consideració inicialment són DBscan, K-means i hierarchical clustering, spectral clustering depenent de la seva viabilitat i idoneïtat.

  La particularitat de la detecció de comunitats en xarxes socials resideix en què els usuaris, estan connectats per una xarxa d'enllaços, mentre que en el clustering tradicional, no tenen per què pertànyer a una xarxa (Zafarani et al., 2014).

  A més a més nosaltres treballarem la cerca de comunitats implícites respecte la característica de la temàtica de les malalties minoritàries, són aquelles comunitats en què els membres no en tenen constància de què en formen part, però que són afins per patrons que inicialment poden desconèixer i que intentarem detectar a partir del nostre procés de mineria de dades (Zafarani et al., 2014).

  Reptes possibles en aquest apartat és l'estudi de comunitats a diferents nivells jeràrquics o de comunitats dinàmiques.

  Lideratges, o popularitat i tipologia d'usuaris detectats en les dades (administracions públiques associacions, centres mèdics, etc...) també poden ser considerats com a elements d'anàlisi de les relacions existents.

  Per tal d'avaluar els agrupaments detectats utilitzarem mètriques de qualitat, que intenten capturar el grau de similitud o dissimilitud dels patrons detectats i es valora com és de compacte el clúster (cohesió) i el nivell de separació entre clústers.

- *Anàlisi de continguts*: Per l'anàlisi de continguts inicialment es proposen 4 tipus de mètodes:

    - Generació de núvols de paraules.
    - Comparació de textos
    - Anàlisi de sentiments
    - Anàlisi semàntica: agrupament jeràrquic.

    En aquesta fase s'aplicaran tècniques de NLP (Natural Language Processing), i tècniques d'anàlisi de sentiments basades en la detecció de la polaritat (positivitat o negativitat del missatge enviat). En l'anàlisi i comparació de textos ens caldrà: el preprocessat de textos aplicant tècniques de neteja de contingut innecessari o que pot incloure soroll en l'anàlisi, vectorització (tf-idf), reducció de dimensionalitat (per la visualització i detecció de textos semblants). Un cop obtingut els textos depurats es generaran els núvols de paraules i s'aplicaran algorismes de clustering jeràrquic (anàlisi semàntica) amb diferents tipus d'enllaç, per obtenir elements per comparar paraules més freqüents, temes d'interès, queixes, necessitats demandades.

Motivació Personal

La lluita contra les malalties minoritàries i en definitiva l'ajut a tots els individus que les pateixen i els col·lectius que generen recursos per combatre-les, ja és un factor de motivació amb molt de pes. En el nostre cas, a més a més, afegim la possibilitat de desenvolupar una anàlisi sobre dades massives, en l'àrea del "Social Data Mining" on és especialment motivant, poder combinar diferents tecnologies en un mateix estudi. La xarxa social de Twitter té un potencial increïble per captar i oferir dades sobre les interaccions socials en el món real per poder-les modelitzar sobre molts problemes socials. Aquest projecte és una oportunitat idònia per aplicar els coneixements adquirits durant el Màster de Ciència de Dades, tant l'anàlisi de continguts de text amb tècniques especialitzades NLP, com per l'anàlisi de sentiments. En concret l'estudi mitjançant 'clustering' no supervisat és especialment atractiu per ser encara un camp obert en desenvolupament per la detecció de comunitats socials, ja que la certesa o "ground truth" de com són les comunitats, no la tenim per comparar.

A més a més de la font de dades de Twitter, hem intentat capturar dades en directe dels actors implicats en el món de les malalties minoritàries. En concret, gràcies a una visita guiada pels principals departaments d'investigació de l'Hospital de la Vall d'Hebron (dins del marc d'activitats posteriors a la celebració del dia mundial de les malalties minoritàries), vam tenir l'oportunitat d'interaccionar amb personal del centre mèdic. Va ser molt motivadora, ja que van demostrar molt interès a saber del treball i dels possibles resultats finals. Actualment totes les accions van més encaminades en la funció de difusió i no tant en la captació de les tendències d'opinió, preocupacions, demandes i punt de vista detectable amb l'ús de les noves tecnologies a partir de l'anàlisi de les xarxes socials.

En la visita guiada Figura 1, vam ser conscients de la voluntat actual de difondre l'estat de les malalties minoritàries i la feina que s'està desenvolupament. La celebració de visites com aquesta, demostra la necessitat d'incrementar la interacció entre centres mèdics, pacients i població en general, així com la necessitat de realitzar una anàlisi de dades massives utilitzant Twitter com a xarxa social, envers la lluita contra les malalties minoritàries.

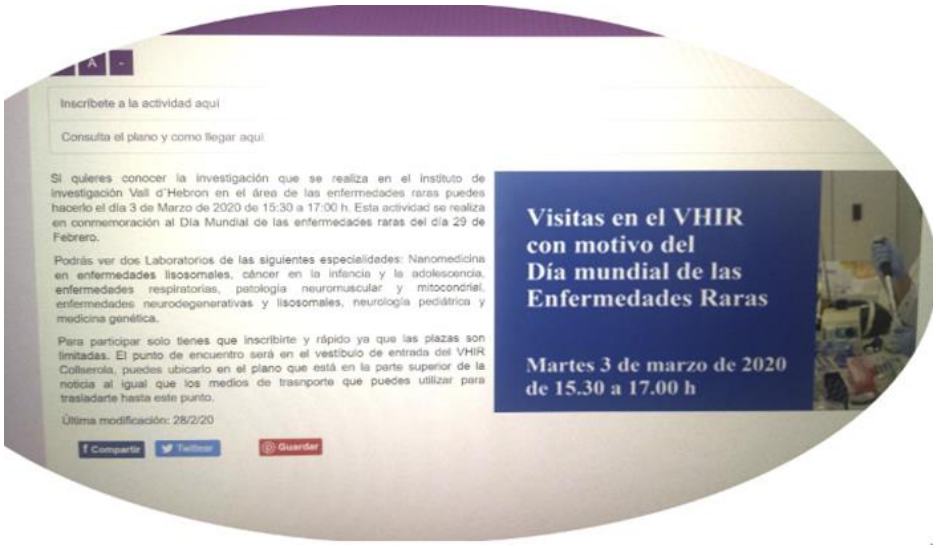


Figura Nº 1: Exemple d'activitat per la difusió de les malalties minoritàries.

Finalment com en tot projecte d'anàlisi de dades massives, la seva visualització i representació també és un repte a assolir com una font més de transmissió de coneixement.

1.2 Objectius del Treball

L'objectiu principal és analitzar el conjunt de dades en el temps capturades de la xarxa social Twitter durant l'any 2020, representatives del dia Mundial de les Malalties Minoritàries aplicant tècniques d'aprenentatge no supervisat.

Per assolir l'objectiu principal cal incidir en els següents aspectes:

1. Dur a terme el procés de captació de dades massives a Twitter.
2. Detectar grups o comunitats d'usuaris.
3. Analitzar i sintetitzar els diferents aspectes derivats dels patrons detectats en els tuits en forma de temes d'interès, demandes detectades, necessitats emergents i actualitzacions dels fulls de ruta de les organitzacions al servei de la lluita contra les malalties minoritàries.
4. Aportar un estudi per consolidar les tecnologies per l'adquisició d'informació social valuosa de les xarxes socials, estudiant la similitud d'usuaris de Twitter, tant des d'un punt de vista estructural, com relacional com d'anàlisi de continguts dels tuits.

5. Implementar un projecte de mineria de dades delimitant i implementant cadascuna de les seves fases, sobre la problemàtica de les malalties minoritàries amb la intenció d'estructurar mesures pràctiques i recomanacions d'accions envers les conclusions de l'anàlisi orientades al suport als pacients i lluita contra les malalties minoritàries.

1.3 Enfocament i mètode seguit.

Com a projecte de mineria de dades, es considera aplicar una metodologia contrastada, àmpliament estudiada i aplicada durant tot el màster, com és la metodologia CRISP-DM. Aquesta metodologia implica la consideració de les següents etapes en el projecte de mineria de dades:

1. Comprensió empresarial o del negoci.
   Explorar com afecta la nostra anàlisi als actors implicats, quina despesa de recursos s'ha de realitzar, el benefici esperat envers la despesa a realitzar. En resum l'impacte social i empresarial que assumim.
2. Comprensió de dades.
   Entendre de quines dades disposem i quines són necessàries. Definició del procés de captació de dades.
3. Preparació o processament de dades.
   Implica totes les tasques necessàries per disposar del dataset òptim per realitzar l'anàlisi.
4. Modelització.
   Procés iteratiu per la selecció de tècniques de modelatge, generació de models, aplicació i avaluació, per trobar el model amb òptims resultats o obtenir informació d'interès d'aquests.
5. Avaluació.
   Avaluació de resultats des d'un punt de vista empresarial.
6. Desplegament o implementació.
   Definir la viabilitat i les formes d'aplicació pràctica dels nostres resultats.

L'estratègia de recerca consisteix a definir un procés de captació de dades massives de la xarxa social Twitter. L'objectiu és disposar d'un 'dataset' corresponent a un període d'uns dies abans i després del dia mundial de les malalties minoritàries, adequat per aplicar les tècniques corresponents als dos components d'estudi definits: estructural i d'anàlisi de continguts. Usant com a referència el model CRISP-DM. Aplicar el cicle d'evolució del model mostrat en la Figura 2, que admet una retroalimentació (feed-back) al final de cada cicle per tal d'afinar cada cop més en la consecució dels objectius.

Figura Nº 2: Metodologia CRISP-DM ("Infografia CRISP-DM", 2020) aplicada a l'anàlisi del DMMM[1].

Les eines previstes inicialment per dur a terme el projecte i que poden complementar-se amb d'altres que es considerin òptimes són les següents:

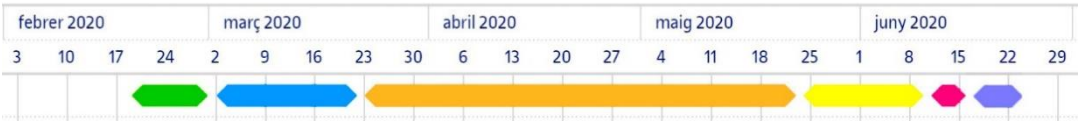
- Captació de tuits de la xarxa social Twitter:
  a. Registre de l'aplicació a Twitter per l'obtenció de credencials.
  b. Llibreria Python Tweepy, per la implementació personalitzada d'un canal de dades (stream) per descarregar els tuits en temps real.
  c. Una alternativa en R podria ser utilitzar la llibreria rtweet.
- Per l'estadística descriptiva, evolució temporal i la geolocalització de dades en mapes R, Python i Leaflet , una llibreria de codi obert en Javascript per a mapes interactius adaptables al mòbil.

[1] Dia Mundial de les Malalties Minoritàries (DMMM).

- Per la generació de núvols de paraules (en anglès 'word clouds') les següents llibreries R: wordcloud, RColorBrewer, wordcloud2.
- Per tasques de NLP (Natural Language Processing): NLTK (plataforma per crear programes Python per treballar amb dades de llenguatge humà) i Text2vec R package per la vectorització de paraules.
- Per realitzar l'anàlisi de sentiment i calcular la polaritat la llibreria de Python TextBlob, que funciona amb compatibilitat amb NLTK.
- Per l'anàlisi semàntica, les llibreries de clustering jeràrquic en R.
- Per comparatives d'algorismes de clustering la llibreria Scikit-Learn (eines d'anàlisi de dades predictives), amb el seu mòdul 'cluster' i les accessibles en la llibreria de Python Scipy en el mòdul scipy.cluster.hierarchy.
- Per les traduccions a un llenguatge homogeni del text, es preveu usar alguna llibreria en Python o R de traducció a l'anglès.

1.4 Planificació del Treball.

La previsió del període de temps per la realització del projecte s'estableix des del 19 de febrer fins al 24 de juny de 2020. Consisteix a cinc fases d'avaluació contínua i una defensa pública final tal com es mostra en la taula 1.



	FASE	Inici	Final
	PAC1 - Definició i planificació del treball final.	19/02/2020	**01/03/2020**
	PAC2 - Estat de l'art o anàlisi del mercat.	02/03/2020	**22/03/2020**
	PAC3 - Disseny i implementació del treball.	23/03/2020	**23/05/2020**
	PAC4 - Redacció de la memòria.	24/05/2020	**10/06/2020**
	PAC5 - Presentació i defensa del projecte.	11/06/2020	**16/06/2020**
	Defensa Pública.	17/06/2020	**24/06/2020**

Taula Nº1: Línia de temps i taula resum de les fases de la planificació del treball final.

1.5 Breu sumari de productes obtinguts.

Els productes obtinguts en aquest treball de màster han estat:

1. Una base de dades documental MongoDB, emmagatzemant tots els tuits capturats en el període definit des del 13 de febrer de 2020 fins al 30 de març del 2020, utilitzant els hashtags:
   #DiaMundialEnfermedadesRaras,
   #RareDiseaseDay, #SomosFEDER, #EnfermedadesRaras,
   #DMEnfermedadesRaras2020, #DM2020.

La base de dades permet consultar qualsevol dada existent en l'estructura de dades definida per Twitter per descriure els tuits. El potencial de la base de dades resideix en disposar de la possibilitat d'extreure dades que no s'ah usat en l'anàlisi d'aquest treball. En l'annex 7.3 es mostra com s'ha fet la còpia de seguretat (en anglès backup) i com es pot restaurar.

El format de presentació és una còpia de seguretat completa de la base de dades documental, BD='DM_MM2020' i Col·lecció='Twitter', comprimit en un fitxer .rar amb mida 42 MB.

2. Un dataset de modelització resultant d'aplicar tasques de preprocessament dels tuits, com la selecció d'atributs d'interès, traducció del text a l'anglès, neteja del text, obtenció de text normalitzat així com tots els atributs descrits en la taula Núm. 3.

   El format de presentació és el d'un full de càlcul Excel extensió .xlsx.

3. L'anàlisi en si, que ens ha permès obtenir les comunitats d'usuaris i les temàtiques més importants existents en els tuits capturats, utilitzant tècniques d'aprenentatge no supervisat, en concret aplicant models creats amb els algorismes KMeans, DBSCAN, i aglomeratiu jeràrquic.

   El format de presentació són documents Jupyter Notebooks per cadascuna de les etapes del procés d'anàlisi descrit en aquest document: captació, processament del contingut dels tuits, generació del dataset de modelització, vectorització i la modelització amb tècniques d'aprenentatge no supervisat. Alternativament s'ofereix el format exportat en html per la seva consulta ràpida.

1.6 Estructura de la resta de document de memòria.

La resta d'aquest document, s'ha estructurat en dos apartats més:

En l'apartat 2 es descriu l'estat de l'art i es presenta el seu marc de referència contextual. S'explica com s'ha efectuat la recerca bibliogràfica per definir l'estat de l'art, i es respon a, per què les xarxes socials poden jugar un paper important per l'anàlisi de les malalties minoritàries i per què Twitter en concret pot facilitar connexions fluides entre els diferents actors implicats. S'ha elaborat una cronologia del material seleccionat, per tots els documents relacionats amb aspectes a tenir en compte com: l'agrupació d'usuaris per disposar de les comunitats d'usuaris per similituds i la detecció de temàtiques generades entre comunitats. A més a més, s'ha cercat l'estat de l'art, per aspectes importants com són: tècniques d'aprenentatge no supervisat, analítica de textos, processament de llenguatge natural, anàlisi de sentiment, geolocalització i mètriques d'avaluació. L'apartat 2 finalitza amb una reflexió de la incidència de tota aquesta tecnologia aplicada sobre la xarxa social Twitter i l'interès a analitzar les dades capturades al voltant del dia mundial de les malalties minoritàries, mitjançant un

diagrama de flux que descriu de quina manera s'ha d'enfocar l'anàlisi elaborat en l'apartat 3.

En l'apartat 3, '*disseny i implementació de l'anàlisi*', seguint la proposta del diagrama de flux de l'apartat 2, es descriu (amb referència la metodologia CRISP-DM), com ha de ser el procés complet seguit per capturar, processar i transformar les dades per poder-les modelitzar i s'implementa cada etapa presentant els resultats.

La majoria de documents complementaris als resultats presentats en aquest document, i els productes generats, resten en un projecte de github que es pot consultar a https://github.com/QuimDJ/TFM_DataScience_UOC.

2. Estat de l'art i marc de referència

Aquest capítol té per objectiu descriure dos aspectes: fer un recull de la literatura d'interès, que serveixi tant de base de coneixement com a punt de partida del treball i establir un marc de referència contextual actual associat al problema plantejat, l'***Anàlisi de les dades a Twitter del Dia Mundial de les Malalties Minoritàries***.

2.1 Procediment de recerca bibliogràfica.

En la realització de l'estat de l'art, s'han revisat múltiples articles i treballs d'investigació procedents de bases de dades de recerca com Google Scholar, Web of Science, i publicacions accessibles d'universitats com Berkeley, Cambridge, Cornell, MIT, Oxford, Standford, etc. que permeten realitzar cerques en línia. Aquest procés inclou la cerca dels termes clau següents: "Social media mining Twitter", "clustering community detection", "unsupervised learning Twitter" i "categorization tuits", "sentiment analysis Twitter" combinats o no amb el terme "rare diseases".

Podem classificar l'estat de l'art avaluat envers les malalties minoritàries en tres grans àrees:

a) Les xarxes socials en l'estudi de recerca i d'investigació.
b) Detecció de comunitats d'usuaris o agrupament d'usuaris.
c) Anàlisi de continguts i anàlisi de sentiment com a part de la mineria de textos i tècniques de processament de llenguatge natural.

2.2 Com poden ajudar les xarxes socials? Per què Twitter?

L'estudi de les xarxes socials (Social Data Mining en anglès) és cada cop una opció més valorada com una forma de disposar de dades en temps real o de manera ràpida per tal de conèixer com són, l'estat d'opinió i els interessos dels seus usuaris. La popularització del seu ús, és el fet que les ha convertit en una eina per la investigació científica. En el cas de Twitter, segons han declarat (Twitter, 2019), existeixen 145 milions diaris i 330 milions d'usuaris actius per mes arreu del món. Tot i la importància d'aquest fet, existeix un conjunt d'aspectes que justifiquen l'ús de les xarxes socials (Merinopoulou, E., et al., 2020):

a. Visibilitat de les malalties rares, suport al procés de desenvolupament de medicaments, presa de decisions i perspectiva del pacient.

   Un dels reptes fixats en la lluita de les malalties minoritàries és el del desenvolupament de medicaments. El fet que no existeixi consciència de les malalties pel fet que els seus pacients són pocs i dispersos geogràficament provoca diagnòstics erronis i llargs temps d'espera en la correcció d'aquests diagnòstics. És necessari comprendre els patrons d'atenció mèdica, la càrrega i les necessitats insatisfetes dels pacients per poder desenvolupar medicaments efectius. A més a més recentment s'ha iniciat un canvi en el

desenvolupament de fàrmacs on es considera vital tenir en compte al pacient. Per tot això, actualment es recomana l'ús de les xarxes socials com a medi vàlid per captar la perspectiva del pacient, com és el cas de la proposta de la guia revisada de l'Administració de Drogues i Aliments dels Estats Units (Drugs and Foods Administration, FDA)[2] del gener de 2019.

b. L'increment de la utilització de les xarxes socials en l'atenció especifica de pacients.

   La necessitat de recerca i compartir de les comunitats de malalties minoritàries ha fet que trobessin també en les xarxes socials un medi ideal per comunicar-se, fer-se conèixer i difondre el seu missatge (Pew Research Center, 2011). Exemples en funcionament són Rareconnect, (RareConnect, s. f.) creada per EURORDIS.

De la mateixa manera, cal tenir en compte els desavantatges a l'hora de fer recerca prenent cura dels següents aspectes: privacitat de les dades i disposar de la quantitat de dades necessària en el cas de les malalties minoritàries.

En aquest estudi, la xarxa social Twitter ha estat seleccionada com a medi on capturar les dades, per ser una font de dades massives de gran abast mundial, per la seva facilitat d'accés, ús i per permetre dissenyar un procés de captura de dades en temps real, basat en dades en streaming. Twitter ofereix als seus usuaris la possibilitat de:

a) Ser escoltats.
b) Satisfer les seves necessitats i curiositats.
c) Fer ús de la xarxa amb facilitat.
d) Disposar d'una comunicació d'interacció ràpida, sense esperes.

Un dels aspectes més importants que ens ha aportat la xarxa social Twitter per la realització d'aquest treball, és la possibilitat de caracteritzar, detectar i modelitzar els seus usuaris en comunitats i extreure profit de la informació derivada de l'anàlisi de continguts dels missatges de text enviats en forma de: temàtiques, comprensió i dimensió d'opinions, denuncies i demandes. En la mesura en què això s'aconsegueix depèn el grau en què podrem contribuir de forma directa en la presa de decisions, la generació de recursos, suport en la seva assistència i lluita envers les malalties minoritàries que pateixen.

2.3 Detecció de comunitats.

L'anàlisi que s'ha realitzat al dia mundial de les malalties minoritàries segueix una metodologia CRISP-DM, basada en dos grans blocs de treball: la detecció de grups d'usuaris similars que anomenarem comunitats (en anglès community detection) i la realització d'una anàlisi de continguts (en anglès content analysis) amb posterior anàlisi de sentiment (en anglès sentiment analysis), que implica

[2] Center for Drug Evaluation and Research (FDA): http://www.fda.gov/regulatory-information/search-fda-guidance-documents/rare-diseases-common-issues-drug-development-guidance-industry-0

l'ús de tècniques de mineria de textos, així com la consideració de moltes tècniques que pertanyen a l'àrea del processament de llenguatge natural PLN (en anglès Natural Language Processing, (NLP)). Tota la recerca d'estat de l'art exposada va encaminada a fer un recull de treballs realitzats en aquestes àrees amb un vincle comú a totes elles, l'entorn de les xarxes socials i en concret de la xarxa social Twitter. Per tant, tots els treballs utilitzen com a dataset, dades no estructurades en format de textos curts capturats a un sistema de missatgeria global mundial d'alta velocitat com Twitter.

També paral·lelament s'ha procurat tenir en compte característiques que en cas d'estar disponibles, són especialment importants com: capacitats de geolocalització inherents a l'origen de dades capturat de Twitter, i les metadades associades als perfils de cada compte d'usuari a Twitter.

Considerables estudis tracten el tema de la detecció de comunitats d'usuaris, per tal de conèixer la seva estructura i comportament. Alguns estudis analitzen el text de cada missatge, però d'altres consideren les dades del perfil (en anglès profile) del compte de Twitter de cada usuari (Sundararaman, D., i Srinivasan, S., 2017), on els autors proposen una metodologia anomenada 'Twigraph' per explorar les connexions entre persones utilitzant els seus perfils de Twitter usant núvols de paraules i nombres de seguidors com a estratègia de versemblança.

En l'estat de l'art per la detecció de comunitats s'ha observat que existeixen diverses vies d'investigació principals que distribuïm en dos grups:

Aplicació directa de la teoria de grafs:

En aquesta categoria l'estructura és una característica principal. Es detecta cada comunitat com un subgraf de la xarxa social global. Sobre un graf dirigit que reflecteix les interaccions de comunicació entre usuaris (tuits i retuits) s'usen diferents mètriques com per exemple, el grau de sortida i entrada de cada node o la centralitat per determinar parts compactes o completament connectades del graf. A (Darmon et al., 2015) es presenta una visió del panorama de les alternatives existents i es descriu que podem caracteritzar les comunitats en els següents tipus: estructural (qui són els teus amics i a qui segueixes), d'activitat (qui comparteix perfils d'activitat similar), de temàtica (qui tracta els mateixos temes) i d'interacció (qui comunica amb qui). A (Abascal-Mena et al., 2015) es modelen els tuits com una xarxa i s'usen fonaments bàsics de l'anàlisi de xarxes socials (SNA) per extreure informació sociosemàntica. A (Dutta, S. et al., 2019) es construeix un primer gràfic de semblança i, després, s'aplica la metodologia de detecció de comunitats en el graf de similitud per tòpics i es representa el resum de temàtiques. També per cada clúster, s'identifiquen tuits importants basats en mesures de centralitat de gràfics diferents, com ara centralitat de grau, centralitat de proximitat, centralitat d'interès, centralitat de vectors autònoms i 'page rank'. Aquests només són dos exemples, però en aquesta categoria entren tots els estudis que donen molta importància a l'estructura de la xarxa i sobretot a l'estudi de subgrafs CLIQUÉ on cada node està connectat a la resta. Una descripció detallada la trobem a (Tang, L., i Liu, H. 2010) i a (Papadopoulos, S. et al., 2012).

Darrerament estan tenint més rellevància altres opcions que consideren grafs múltiplex, que són grafs que mantenen diferents capes de característiques simultàniament per oferir més precisió en l'obtenció de comunitats d'usuaris (Paez, M. S. et al., 2019).

Per acabar aquest apartat comentar altre punt de vista interessant a (Fani et al., 2016) on es tracta de detectar comunitats per temàtiques on els usuaris tenen una tendència temporal similar respecte als seus temes d'interès.

Aplicació d'agrupament no supervisat:

En el moment en què disposem d'una font de dades massives com poden ser les xarxes socials, on les comunitats poden ser molt variades i diferents, altres tipus d'algorismes que no necessiten informació 'a priori' de com són les comunitats, prenen rellevància, a (Han, Y., 2018) ja es fa referència a noves característiques com els interessos dels usuaris i les interconnexions dels nodes que són elements clau per la detecció de comunitats. A més a més, es fa referència al fet que un nou tipus d'algorismes anomenats d'agrupament poden ser usats. Hi ha dos motius clars per aplicar tècniques no supervisades d'agrupament:

a) La quantitat de dades a entrenar és molt gran per un etiquetat manual.
b) La naturalesa de les dades implica l'existència de grups no previstos inicialment que són descoberts en aplicar aquests tipus de tècniques.

A (Han, Y., 2018) s'aplica l'anàlisi de dades al problema d'identificar les comunitats d'usuaris a Twitter, amb l'objectiu d'entendre-les en l'àmbit d'usuari. Es calcula la similitud dels usuaris mitjançant l'anàlisi de continguts textuals: el text de tuit, els URL, els hashtags, la relació de seguiment i la interacció de retuitar. En aquesta línia existeix el treball (Zhang, Y. et al., 2012) que parteix de dues afirmacions:

- (Weng, J., et al. 2010): que afirma que els usuaris de Twitter, no segueixen de manera aleatòria o recíproca, sinó que segueixen un amic quan li interessa els temes que publica l'amic. I a la inversa, l'amic el seguirà perquè troba un interès de similar magnitud al tema.
- (Welch, M. J., 2011): que afirma que el fet de retuitar és una confirmació encara més potent d'interès.

A (Zhang, Y. et al., 2012) es proposa definir un coeficient de similitud global d'usuari compost de l'agregació de 5 coeficients de similitud calculats respectivament sobre el text del tuit, els URL i els hashtags, els seguidors (en anglès followers) i els retuits. Aquest coeficient de similitud global d'usuari és el que finalment s'aplica en els algorismes d'agrupament escollits com a mesura de similitud per la detecció de comunitats.

Per la consideració d'algorismes d'agrupament (en anglès clustering algorithms), a (Alnajran, N. et al., 2017) trobem una completa revisió de treballs d'aplicació de mètodes d'aprenentatge no supervisat sobre dades no estructurades de tuits, i una comparació dels algorismes d'agrupament implementats en cada treball.

Un cop detectades les comunitats d'usuaris, es fa una recerca de temàtiques d'interès. Els algorismes usats en la comparativa són classificats en 4 categories (algorismes de partició, jeràrquics, híbrids i densitat).

Respecte als algorismes basats en partició, K-Means i K-Medoids són els algorismes d'agrupament estricte més populars (Arora et al., 2016). A (Friedemann, 2015) es proposa un enfocament amb K-Means per agrupar clients d'una empresa mitjançant dades de xarxes socials de Twitter.

A (Ifrim et al., 2014) s'aplica agrupament jeràrquic (algorisme aglomerat amb dendrograma tallat a 0,5) per la detecció de temàtiques en fluxos de dades de Twitter (en anglès stream data), basant la seva metodologia en un filtratge agressiu de tuits i temes per eliminar tuits i vocabulari sorollosos, més una agrupació jeràrquica de tuits, retallament dinàmic de dendrogrames i classificació dels clústers resultants per l'obtenció de temes. El mot agressiu fa referència al fet de no considerar els tuits que tenen moltes mencions o hashtags dels usuaris i no tenen prou tokens al text netejat, pel fet de no contenir prou contingut. A (Kaur, 2015) es proposa un enfocament jeràrquic d'agrupament aglomerat i de divisió per crear dinàmicament àmplies categories de tuits similars basats en l'aparició de substantius.

Pel que fa als algorismes per densitat, DBSCAN és el més usat. A (Baralis et al., 2013) es proposa una estratègia multinivell per agrupar dades de text amb una variable de distribució, per descobrir informació cohesionada, usant una mètrica de similitud de cosinus, però es preveu baix rendiment a l'escalar a dades massives. A (De Boom et al., 2015) s'aplica DBSCAN i s'obté un millora en la detecció i agrupament d'esdeveniments mitjançant informació semàntica d'alt nivell utilitzant una matriu de co-ocurrència de hashtags. A (Anumol & Pattani, 2016) s'aplica DBSCAN per fer una anàlisi de sentiment usant com a mètrica de similitud el coeficient de Jaccard.

A (Alnajran, N. et al., 2017) la comparativa es basa en els següents atributs: tipus de mètode d'agrupament, algorisme, nombre de clústers, volum de dades, mesura de distància, atributs de clúster, mètodes d'avaluació i resultats. La conclusió de l'estudi efectuat sobre l'ús de l'aprenentatge no supervisat en la mineria de xarxes socials és que té importants febleses.

A (Vathi et al., 2017) disposem d'una descripció detallada de tot el procés d'anàlisi de dades en tuits, de manera que es defineixen mètriques de similitud d'usuaris per detectar les comunitats d'usuaris. Aquestes mètriques es combinen i s'estableix un coeficient global de similitud d'usuari per detectar les comunitats d'usuaris (usant un mètode d'agrupament no supervisat). Sobre aquestes comunitats se cerquen els temes d'interès existents en cada clúster o grup.

Sobre els diferents tipus de mètodes d'agrupament a (Alnajran, N. et al., 2017) s'indiquen dos aspectes importants: aquells que necessiten per aplicar-los conèixer el nombre de clústers o grups que volem obtenir (p. ex. k-means, k-medoids com algorismes de partionament) dels que no (p. ex. DBSCAN per densitat) i el fet que les instàncies a agrupar siguin elements de tan sols d'un

clúster (partionament estricte) o puguin formar part de més d'un (partionament suau).

A (Gromann, D. et al., 2017) s'estudia que podem analitzar els tuits afegint un altre component. Els tuits i hashtags generalment s'analitzen en l'àmbit de paraula o de missatge, però no en l'àmbit compositiu de paraules concatenades. A més a més es proposa com a millora d'agrupament, un enfocament per a una anàlisi més acurada dels components en hashtags i altre tipus de seqüències de text en els tuits (hipertext). En aquest estudi, s'usen els tipus de paraula clau no només com a indicadors temàtics principals per a tuits de clúster (que és el més comú), sinó aplicats a l'obtenció dels grups d'usuaris.

2.4 Anàlisi de continguts i anàlisi de sentiments.

L'anàlisi de continguts forma part essencial en la detecció de comunitats i aporta nous camins per extreure de la xarxa social valuosos coneixements. Mitjançant aquesta tècnica els usuaris de Twitter poden ser agrupats en diferents comunitats, segons els seus interessos o segons sentiments similars respecte a certs temes. A (Lam, A. J., 2016) s'incorpora l'anàlisi de sentiments per millorar la detecció de comunitats. La proposta consisteix però a fer-ho utilitzant informació contextual com l'estructura de la xarxa social i els contextos de conversa, d'autor i de temes. Aquest treball però està orientat a incloure tota aquesta informació en els pesos de les arestes del graf d'interaccions. Dintre de les possibilitats de l'anàlisi de continguts a (Antelmi, A., 2018) s'introdueix el concepte de contingut generat per usuari (en anglès User Generated Content, UGC) i es proposa un entorn de treball per l'anàlisi dels patrons d'interacció i de comportament d'una comunitat de Twitter.

És molt important però que es tinguin en compte en l'anàlisi de continguts les característiques del text que es fa servir a les xarxes socials. A (Shi, Q. et al, 2018), es fa la diferència entre text i 'text curt' a l'hora d'analitzar-lo. A causa de l'amplada limitada i la llibertat en la construcció de les frases, el text curt és diferent del text normal. Per això es proposa un model anomenat *Enriquiment Conceptual i Semàntic* per la modelització de tema (CSET) combinant el model de temes Biterm (BTM) (un model de temàtica probabilística àmpliament utilitzat i dissenyat per a text breu) amb Probase (una base de coneixement probabilístic a gran escala). CSET és capaç de capturar relacions semàntiques entre paraules per enriquir un text curt. La seva aplicació és la comprensió semàntica de text curt, inclosa la classificació de textos breus i la mesura de la similitud de les paraules en context. A (Zhou, K. et al., 2018) també proposen mètodes que tinguin en compte el fet que el text curt és diferent d'un text d'un document i s'indica que la presència d'unigrames o combinacions de caràcters especials afegiran dificultat a l'obtenció de temes d'interès (en anglès topics). Aquests entrebancs afecten les tècniques tradicionals (Probabilistic Latent Semantic Analysis (PLSA) o Latent Dirichlet Allocation (LDA), precisament per la manca de coincidències.

A (Subirats et al., 2018) es realitza una anàlisi de continguts i de sentiment combinat amb una anàlisi temporal i es mostra que hi ha una alta variació entre la mitjana de polaritat (positivitat o negativitat del missatge enviat) i l'hora del dia.

Un altre aspecte vigent en l'estat de l'art de l'anàlisi de continguts text, i a tenir en compte a l'hora d'analitzar els missatges en xarxes socials, és que Twitter és la xarxa social més estudiada en màrqueting viral i la difusió del rumor és un problema real. A (Serrano, E. et al, 2015) trobem detall sobre aquest problema i es fa una revisió de treballs de recerca que estudien la difusió de rumors a Twitter.

Finalment a (Hu, 2018), es fa un recull de l'estat de l'art de les possibilitats del geoetiquetage (obtenir del text amb tècniques d'anàlisi, la localització geogràfica) que després pot servir per complementar les dades disponibles en una anàlisi, com pot ser el cas de tuits sense referència geogràfica.

2.5 Mètriques.

En cada apartat i operació de l'anàlisi de tuits, ens cal realitzar operacions i avaluar la fiabilitat, precisió i bondat d'aquestes. Les mètriques utilitzades en els estudis de l'estat de l'art estan associades a la similitud entre usuaris o a la qualitat dels clústers o grups detectats com s'ha comentat en apartats anteriors.

Respecte a l'estat de l'art en aquesta qüestió cada estudi proposa la seva alternativa mètrica. Hi trobem varietat de conceptes i mètriques per mesurar-los. Dos exemples són el de '*paraula clau*' (en anglès keyword) i el concepte '*com de clau o important*' o importància (en anglès keyness). En el primer cas obtenim les paraules clau per comparació amb un text estàndard o còrpora, mentre que quan calculem la importància o fem usant la mètrica LL (log-likelihood) o l'estadístic Khi quadrat tal com es descriu a (Gabrielatos, C. i Marchi, A., 2012).

Mètriques relacionades a tenir en compte quan comparem corpus de dades són DIFF (Gabrielatos, C. i Marchi, A., 2012), el factor Bayes (BIC) a (Wilson, A. C., 2013), mida de l'efecte per a la probabilitat de registre (ELL) (Johnston et al., 2006), risc relatiu, relació de registre o la proporció de probabilitats.

Per comparar dos núvols de paraules a (Diakopoulos et al., 2015) s'aconsella usar no més de 500 paraules, però dependrà de les característiques de resolució del dispositiu on es visualitzi.

2.6 Incidència sobre les malalties minoritàries.

A (Davies, W., 2016) es proposa l'ús d'enquestes a les xarxes socials per tractar les malalties minoritàries. (Palomino, M. et al., 2016) estudia l'impacte de les xarxes socials mitjançant l'anàlisi de sentiments amb relació al trastorn per dèficit de natura. (Hand, R. K. et al., 2016) introdueix escepticisme sobre la utilitat de les xarxes socials per connectar pacient i personal mèdic, després de realitzar un estudi de continguts en tuits i recollir informació de professionals.

Respecte a com podem ajudar als afectats amb relació a l'ús de les xarxes socials (Rooney, E., 2016) promou com difondre missatges amb impacte de manera que arribin al màxim nombre de persones. Introdueix el concepte de la narració del cas personal com afectat d'una malaltia minoritària (en anglès rare disease story).

A (Subirats et al., 2018) s'obtenen directrius per maximitzar la repercussió dels posts dels usuaris a les xarxes socials (en aquest cas Facebook) i es dóna suport a les organitzacions de malalties rares a alinear les seves prioritats amb els interessos expressats a les xarxes socials.

A (Tai et al., 2016) es detecta la intenció i la intensitat dels sentiments d'una persona mitjançant l'anàlisi de les seves publicacions en línia.

Finalment, en l'àmbit específic de l'anàlisi de sentiment que abans hem descrit disposem d'aproximacions a les malalties minoritàries envers:

- A capturar l'experiència sobre la malaltia, explicada en tuits per part dels pacients (Greaves, F. et al., 2013).
- A fer servir l'anàlisi del sentiment per avaluar el sistema sanitari públic (Akay, A. et al., 2015).

En la figura 3, podem observar la distribució dels treballs esmentats en aquest capítol, per data de publicació i observem que comencen a tenir notorietat a partir de 2012 i que la producció més significativa de treballs esdevé entre els anys 2015 i 2018.

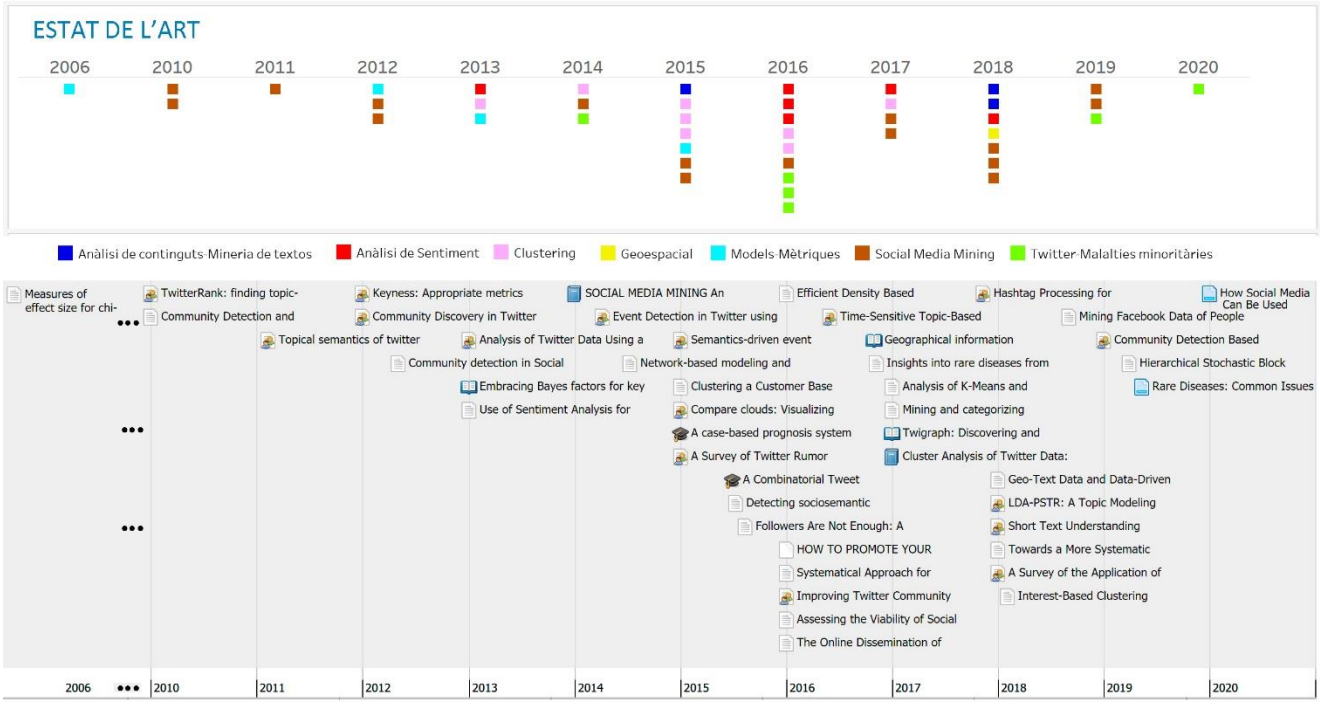


Figura Nº 3: Evolució temporal anual de l'estat de l'art.

A manera de resum en la Figura 4, es mostra el marc conceptual de referència, a tenir present a l'hora d'analitzar les dades de Twitter pel dia mundial de les malalties minoritàries.

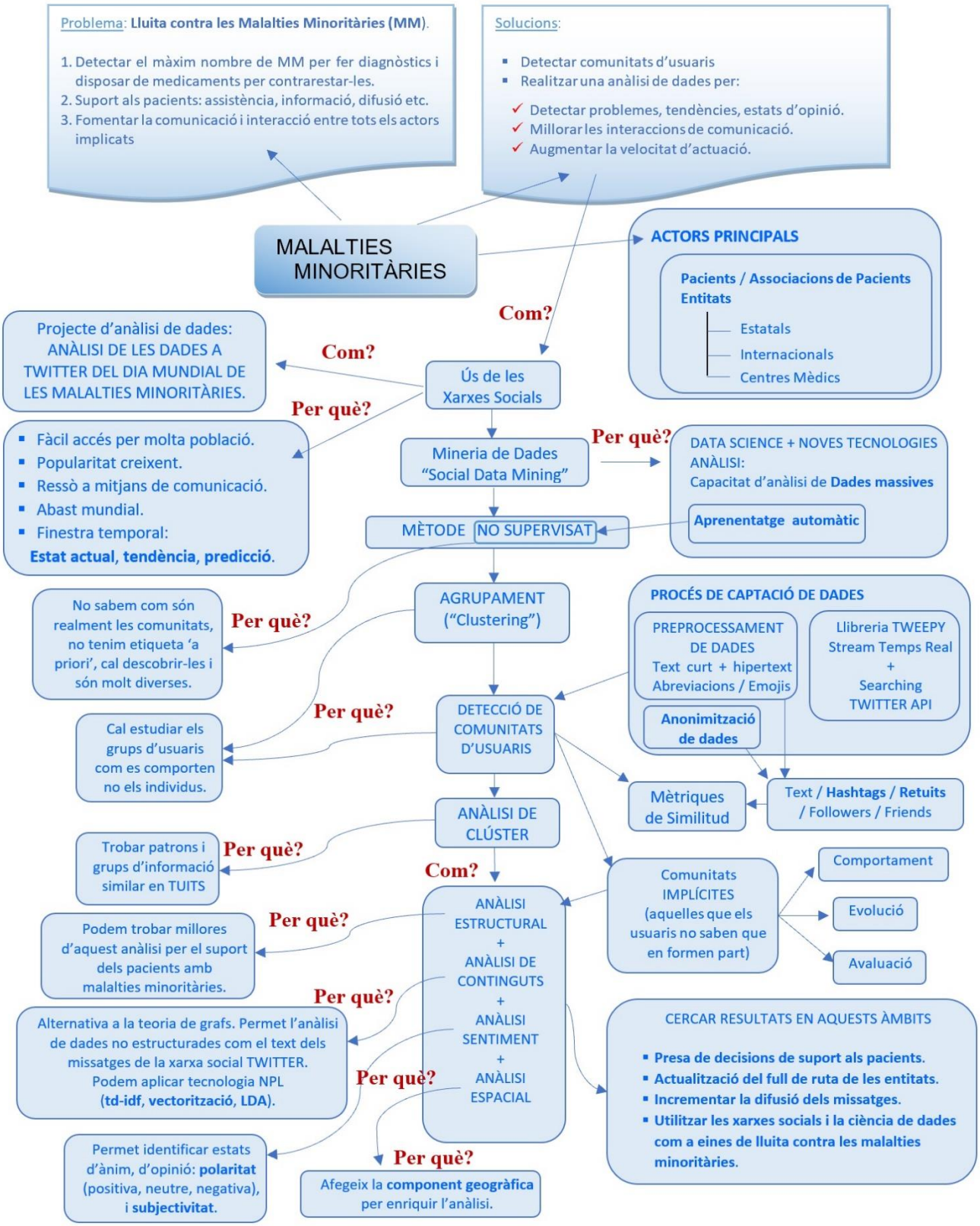


Figura Nº 4: Marc conceptual de referència del treball.

3. Disseny i implementació de l'anàlisi

L'anàlisi proposat està basat en un model CRISP-DM. Segueix tot el cicle de vida de les dades estàndard i executa un flux de processos on és necessari definir un procés de captura de dades, l'emmagatzemament en una base de dades No SQL, i la seva modelització, anàlisi i interpretació. Com s'indica a (Bengfort et al., 2018), estem generant un producte de dades basat en el llenguatge. Utilitzem la dada per generar un producte que aporta nou valor, en el nostre cas però les dades són de tipus text i cal orientar l'anàlisi a una anàlisi de textos, on el llenguatge és la base i el document text, pren la forma d'un tipus de text curt molt específic, molts cops críptic amb abreviacions, ple de simbologia, utilitzem un **tuit de la xarxa social Twitter**. La proposta s'ha basat a dissenyar un flux de processos o 'pipeline' que modelitzi el text dels tuits i generi un coneixement aplicable pel benefici social en el camp de les malalties minoritàries.

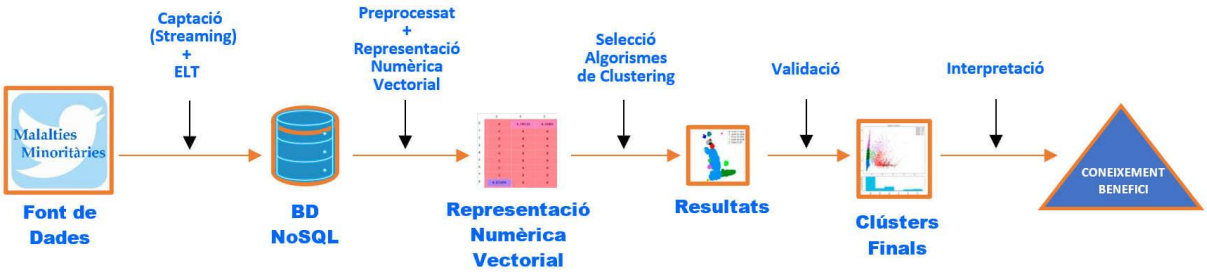


Figura Nº 5: Diagrama de flux global del procés complet d'anàlisi.

3.1 Captació i emmagatzematge de dades.

El primer pas de l'anàlisi consisteix a capturar les dades. S'ha estudiat la millor manera d'obtenir les dades d'entre les diferents opcions possibles. La xarxa social Twitter ofereix un API, per tal de donar accés a fer el tractament de totes les dades. Aquest API està disponible en tres modalitats, de les quals hem escollit l'**estàndard** que és gratuïta. Per un ampli ventall d'eines possibles podem consultar (Social media data collection tools - Social media data collection tools, s. f.). Totes les eines software utilitzen aquest API i per tant requereixen què qui vulgui accedir a disposar dels tuits enviats faci una subscripció en la web de desenvolupament d'aplicacions de Twitter (Twitter Developers, s. f.) i doni d'alta una aplicació. En el procés d'alta, s'especifica la intenció i objectius del projecte de dades. Aquest procés que és validat per integrants de la xarxa social Twitter proporciona un conjunt de claus on els noms només són descriptius **CONSUMER_KEY_KEY, CONSUMER_SECRET_KEY, ACCESS_TOKEN_KEY, ACCESS_TOKEN_SECRET_KEY,** per poder usar qualsevol eina i establir una connexió per la descàrrega dels continguts dels tuits i les seves metadades associades.

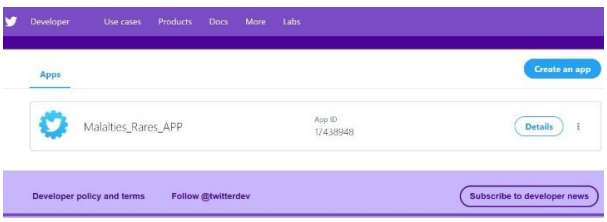


Figura Nº 6: Registre de l'aplicació del procés de captació de tuits

3.1.1 Procés de captació: eina de programari i procediment.

En aquest treball, ens hem decidit per l'ús de l'eina Tweepy (tweepy/tweepy, 2009/2020), que permet definir mitjançant codi Python un procediment de descàrrega utilitzant 'streaming' de dades. Per tant en temps real s'han pogut descarregar els tuits que han usat els hashtags:

#DiaMundialEnfermedadesRaras, #RareDiseaseDay, #SomosFEDER
#EnfermedadesRaras, #DMEnfermedadesRaras2020, #DM2020

Aquests, són els hashtags que s'han considerat més representatius de les possibles comunicacions dels usuaris en la temàtica de la nostra anàlisi sobre el dia mundial de les malalties minoritàries.

El procés de captació s'ha realitzat ininterrompudament durant tot el període de dies entre el 13/02/2020 i el 30/03/2020. El dia mundial de les malalties minoritàries, és va celebrar el 29/02/2020. Cada dia s'ha generat un fitxer text, amb tots els tuits capturats. El format del fitxer ha estat 'JSON Line', on cada tuit queda guardat en una sola línia en format JSON.

En tot el període de captació es va fer un seguiment i monitoratge del procés de captació.

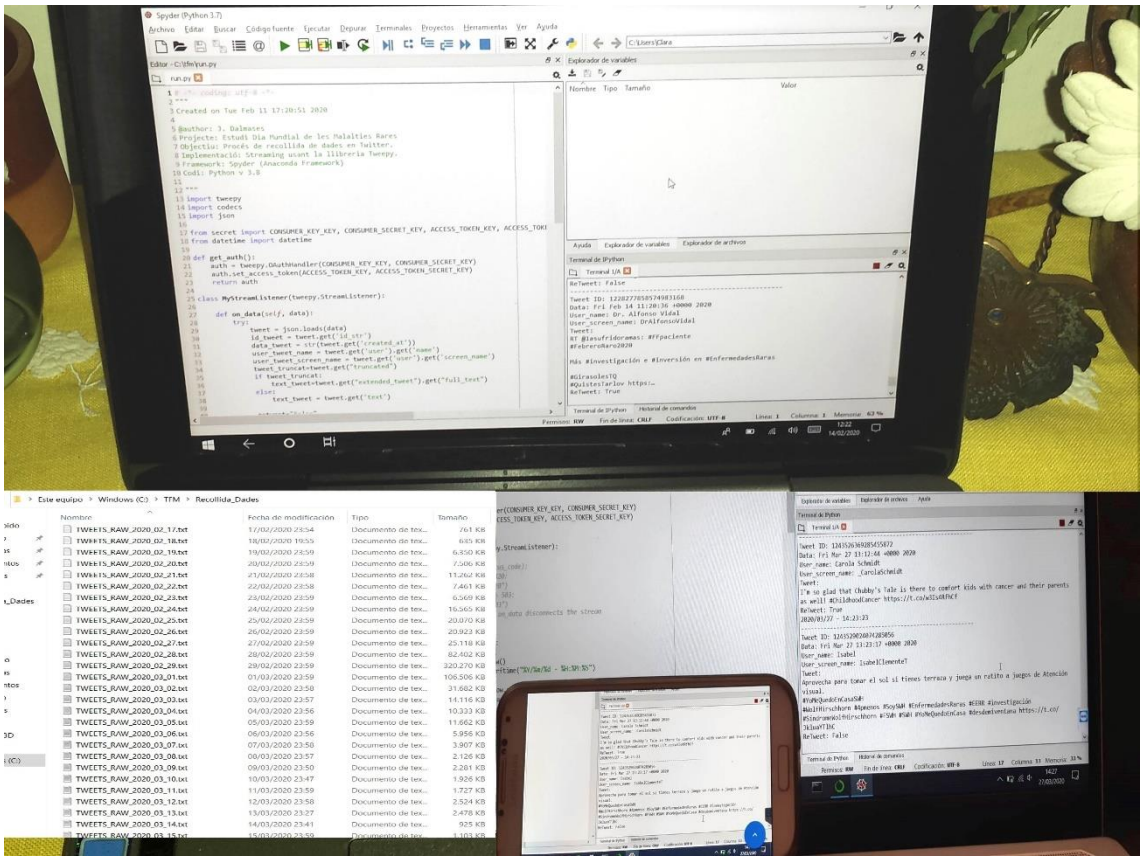


Figura Nº 7: Tres imatges de la monitoratge del procés de captació de dades de la xarxa Twitter.

En la figura 7, es pot observar (sense detall, per mantenir la privacitat de les dades) que es va utilitzar un portàtil amb el framework Anaconda, i l'editor de

Python ‘Spyder’ des d’on es va executar l’script de l’apartat 7.1 dels annexos. En la part de dalt de la figura 7, podem veure com la programació i modificació del script queda en el marc esquerre, i el monitoratge en el marc de la dreta part baixa, on s’anava mostrant en el format breu, els tuits que anaven capturant-se.

El procés de monitoratge va ser constant i a més a més es va realitzar un control remot del portàtil via mòbil, per detectar quan es s’havia produït un error. Els únics errors que es van produir van ser per què la font de Twitter va tallar la connexió per algun motiu.

En la figura 7, també podem observar com el volum dels fitxers capturats va augmentant a mesura que ens apropem al dia 29 de febrer, dia de màxima emissió.

Després de finalitzar la captació es pot afirmar que el software ha funcionat bé i que només cal tenir en compte mantenir una bona relació hashtags seguits/tuits rebuts. Quan el stream de dades resta molt temps inactiu, o està sobresaturat és més susceptible d’errors.

A part de la llibreria tweepy, hem adquirit fonaments per la programació de scripts de (Hawker, 2010).

3.1.2 Base de dades, format i estructura de dades.

Un cop realitzat el procés de captació dels tuits, el següent pas ha consistit a implementar el procés d’emmagatzematge en la base de dades. S’ha utilitzat la base de dades No SQL MongoDB.

Els motius pels quals s’ha seleccionat aquesta base de dades són dos:

1. Que utilitza un format ‘JSON’ amb compressió (‘bjson’) per emmagatzemar les dades, equivalent al format origen dels tuits provinents del API de Twitter.
2. Que és una base de dades documental i ens facilita considerar cada tuit com un document text amb metadades.

Per tal d’implementar la base de dades, s’ha instal·lat el programari ‘servidor’ que deixa un servei executant-se i rebent dades pel port 27017, i per altra banda per la gestió i manipulació de les dades en la part client s’ha utilitzat el client ‘MongoDB Compass’ (figura 8).

En la figura 8, a més a més podem observar les següents dades descriptives:

Nom Base de dades	:	**DM_MM2020**
Nom Col·lecció	:	**Twitter**
Nombre total de tuits	:	**102632**
Volum de la base de dades	:	**698.5 Mb**

L’estructura de cada document és la mateixa que l’objecte ‘tweet’ de l’API de Twitter. Vam optar per una estratègia ELT (Extract Load Transform) i disposar del volum complet de dades en la base de dades i exportar posteriorment les que realment siguin necessàries per a la modelització i anàlisi final.

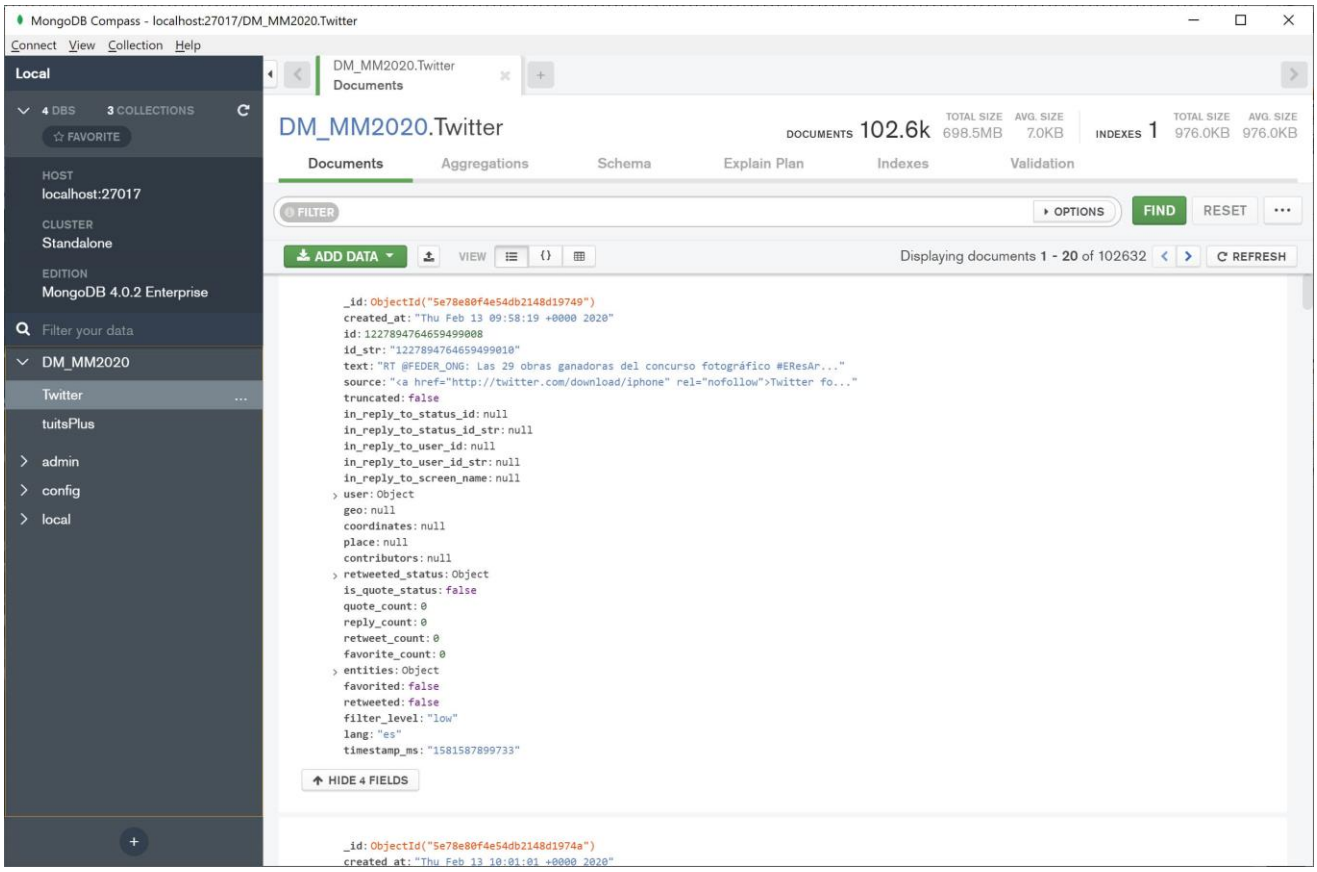


Figura Nº 8: Interfície gràfica (GUI) de l'aplicació MongoDB Compass. Permet organitzar, visualitzar i explorar els documents amb consultes ad hoc.

En la figura 9, mostrem l'estructura completa de cada tuit (objecte Tweet, de Twitter API), que ha estat analitzada al detall per poder extreure cada dada d'interès per generar el dataset tabular que serveix de punt de partida per la 'pipeline' o flux de processos de la modelització i anàlisi aplicats.

Tots els fitxers generats en el procés de captació (47 en total), van ser importats a la base de dades de manera manual, utilitzant l'eina d'importació de MongoDB Compass. D'aquesta manera es va poder controlar amb precisió que tots els tuits capturats fossin inserits sense problemes. Tot i ser molt útil l'eina d'importació, a fet falta controlar amb molta cura dos aspectes:

1. Fraccionar els fitxers de captació perquè no excedissin els 7000 o 8000 tuits. Per mides molt grans el procés no finalitzava i no es podia saber realment quina quantitat de tuits s'havien introduït.
2. Vigilar de no saturar el servidor MongoDB en el nombre d'actualitzacions pendents, per tal de no generar errors en successives importacions. Donant temps al servidor, a realitzar les actualitzacions pendents, es va aconseguir reduir els errors en les importacions.

```
{"_id": identificador únic de l base de dades
"created_at":"Wed Feb 12 09:41:33 +0000 2020",
"id":1227528156287991810,
"id_str":"1227528156287991810",
"text":"Lucha contra las enfermedades raras: https://t.co/Ye13AZzqIJn  29 - 02 - 2020  -  Dia mundial de las enfermedadesu2026 https://t.co/ePjcxTpFA7",
"source":"u003ca href="https://mobile.twitter.com" rel="nofollow"u003eTwitter Web Appu003c/au003e",
"truncated":true,
"in_reply_to_status_id":null,
"in_reply_to_status_id_str":null,
"in_reply_to_user_id":null,
"in_reply_to_user_id_str":null,
"in_reply_to_screen_name":null,
"user":  {"id":1209454411308814336,
          "id_str":"1209454411308814336",
          "name":"QDJ_UOC_24",
          "screen_name":"QdjUoc",
          "location":null,
          "url":null,
          "description":null,
          "translator_type":"none",
          "protected":false,
          "verified":false,
          "followers_count":0,
          "friends_count":0,
          "listed_count":0,
          "favourites_count":0,
          "statuses_count":6,
          "created_at":"Tue Dec 24 12:45:55 +0000 2019",
          "utc_offset":null,
          "time_zone":null,
          "geo_enabled":false,
          "lang":null,
          "contributors_enabled":false,
          "is_translator":false,
          "profile_background_color":"F5F8FA",
          "profile_background_image_url":"",
          "profile_background_image_url_https":"",
          "profile_background_tile":false,
          "profile_link_color":"1DA1F2",
          "profile_sidebar_border_color":"C0DEED",
          "profile_sidebar_fill_color":"DDEEF6",
          "profile_text_color":"333333",
          "profile_use_background_image":true,
          "profile_image_url":"http://pbs.twimg.com/profile_images/1209460351185694720/uiu3w9M5_normal.jpg",
          "profile_image_url_https":"https://pbs.twimg.com/profile_images/1209460351185694720/uiu3w9M5_normal.jpg",
          "default_profile":true,
          "default_profile_image":false,
          "following":null,
          "follow_request_sent":null,
          "notifications":null},
"geo":null,
"coordinates":null,
"place":null,
"contributors":null,
"is_quote_status":false,
"extended_tweet":{"full_text":"Lucha contra las enfermedades raras: https://t.co/Ye13AZzqIJn  29 - 02 - 2020  -  Dia mundial de las enfermedades
rarasn#diamundialenfermedadesrarasnnLas ER o poco frecuentes son aquellas                    que tienen una baja prevalencia en la poblaciu00f3n: nmenos de 5
de cada 10.000 habitantes.n**",
           "display_text_range":[0,280],
           "entities":{"hashtags":[{"text":"diamundialenfermedadesraras",
                                    "indices":[122,150]}],
                       "urls":[{"url":"https://t.co/Ye13AZzqIJ",
                                "expanded_url":"https://enfermedades-raras.org/index.php/enfermedades-raras",
                                     "display_url":"enfermedades-raras.org/index.php/enfeu2026",
                                     "indices":[37,60]}],
                   "user_mentions":[],
                   "symbols":[]}},
"quote_count":0,
"reply_count":0,
"retweet_count":0,
"favorite_count":0,
"entities":{"hashtags":[],
       "urls":[{"url":"https://t.co/Ye13AZzqIJ",
            "expanded_url":"https://enfermedades-raras.org/index.php/enfermedades-raras",
            "display_url":"enfermedades-raras.org/index.php/enfeu2026",
            "indices":[37,60]},
           {"url":"https://t.co/ePjcxTpFA7",
            "expanded_url":"https://twitter.com/i/web/status/1227528156287991810",
            "display_url":"twitter.com/i/web/status/1u2026",
            "indices":[117,140]}],"user_mentions":[],
       "symbols":[]},
"favorited":false,
"retweeted":false,
"possibly_sensitive":false,
"filter_level":"low",
"lang":"es",
"timestamp_ms":"1581500493486"}
```


Figura Nº 9: Estructura completa d'un tuit.

3.1.3 Estudi de l'estructura de dades d'un tuit.

Per tal de tenir coneixement de l'estructura de dades d'un tuit i les seves possibilitats, s'han revisat tots els camps i s'ha seleccionat la següent informació d'interès, llistada en la taula 2:

Camp	Significat
_id	Id únic assignat per la base de dades.
created_at	Data d'enviament del tuit (**UTC time**).
entities.hashtags	Llista de hashtags del tuit.
entities.urls	Llista de Urls, normalment codificades. Si volem podem recuperar l'original sense codificar.
extended_tweet.entities.hashtags.text	Part text del hashtag en cada element de la llista de hashtags pertanyent a un tuit de longitud 280 caràcters continguda a **extended_tweet.entities.hashtags.**
extended_tweet.entities.urls	Llista de 'Urls' contingudes en un tuit de longitud 280 caràcters.
extended_tweet.full_text	En cas de '**Truncated**':true, sabem que aquest camp conté el missatge complet de 280 caràcters, mentre que el camp '**text**', només conté els primers 140.
extended_tweet.user_mentions	Llista de mencions a usuari, contingudes en un tuit de longitud 280 caràcters. Ex: @NomUsuari.
favorite_count	Nombre de 'favorits'.
favorited	Val 'true', quan el tuit ha estat marcat com a favorit per algun usuari (símbol de l'estrella en l'aplicació de Twitter).
id	Id numèric.
id_str	Mateix id però tipus 'String'.
is_quote_status	Si 'true', el tuit s'ha creat **citant** text d'altre tuit. Implica que la informació que desitgem recuperar està al camp **quote_status**. Si 'false' del camp **retweet_status**.
lang	Orientació de l'idioma del text del tuit.
reply_count	Nombre de vegades que ha estat respost
retweet_count	Si l'actual tuit ha estat retuitat, '**retweeted**':true, indica el nombre de vegades.
retweet_status	Aquest camp indica que el tuit és un retuit.
retweeted	Indica si l'actual tuit ha estat retuitat.
text	Contingut text del tuit. Missatge a analitzar.
truncated	Indica si la longitud del text és 140 (false) o 280 (true).
user.favourites_count	Nombre de 'favorits' de l'emissor del tuit.
user.followers_count	Nombre de seguidors de l'emissor del tuit.
user.friends_count	Nombre d'amics de l'emissor del tuit.
user.id	Identificador únic per l'usuari emissor (Numèric).
user.id_str	Identificador únic per l'usuari emissor (String).
user.listed_count	Llista d'usuaris amb els quals té relació.
user.location	Informació accessible puntualment de la base de dades.
user.name	Nom de l'emissor es considera dada privada (no s'usen en l'anàlisi).
user.statuses_count	Nombre de tuits enviats per l'usuari emissor.

Taula Nº2: Selecció d'informació d'interès per l'estudi.

3.1.4 Accés a la base de dades NoSQL de tipus documental.

Alternativament necessitàvem disposar d'una via de connexió cap a la base de dades per fer la gestió de dades massives. El programari utilitzat ha estat la llibreria 'pymongo', que incorpora tota la funcionalitat per connectar, consultar, i actualitzar documents sobre una base de dades MongoDB. Sobretot ha estat útil, per exportar dades en format tabular amb els camps d'interès per efectuar l'anàlisi de dades. Tot el codi programat resta disponible en un projecte de GitHub. En la figura 10, a manera d'exemple es mostra un script d'exemple, utilitzat per generar una còpia de seguretat de la base de dades en tres formats pickle, CSV i Excel:

```
from pymongo import MongoClient
# pprint library is used to make the output look more pretty
#from pprint import pprint
# connect to MongoDB, change the << MONGODB URL >> to reflect your own connection string
#client = MongoClient("mongodb://localhost:7017/?readPreference=primary&appname=py&ssl=false")

# Requires the PyMongo package.
# https://api.mongodb.com/python/current

import json
import pickle
import pandas as pd
import time

# Connexió al servidor de dades
client = MongoClient("mongodb://localhost:27017/?readPreference=primary&appname=py&ssl=false")

# No definim un filtre, perque es un backup de tota la base de dades.
filter={}

# Projecció de camps que volem en la sortida.
project={"_id":1,"contributors":1,"coordinates":1,"created_at":1,"display_text_range":1, \
         "entities":1,"extended_entities":1,"extended_tweet":1,"favorite_count":1,"favorited":1, \
         "filter_level":1,"geo":1,"id":1,"id_str":1,"in_reply_to_screen_name":1,"in_reply_to_status_id":1, \
         "in_reply_to_status_id_str":1,"in_reply_to_user_id":1,"in_reply_to_user_id_str":1,"is_quote_status":1, \
         "lang":1,"place":1,"possibly_sensitive":1,"quote_count":1,"quoted_status":1,"quoted_status_id":1, \
         "quoted_status_id_str":1,"quoted_status_permalink":1,"reply_count":1,"retweet_count":1,"retweeted":1, \
         "retweeted_status":1,"source":1,"text":1,"timestamp_ms":1,"truncated":1}

time_start = time.time()
# Executem la consulta
result = client['DM_MM2020']['Twitter'].find(filter=filter, projection=project)

# 'temps' registra el temps parcial de realitzar la consulta.
temps=(time.time()-time_start)/60
print("#Temps de Consulta sobre tots els tuits de la base de dades:","\n Durada: ",int(temps) if temps>0 else 0,"minut/s ", \
      int((temps-int(temps))*60),"segons.")

time_start = time.time()
# Definició d'un dataframe amb el contingut de la consulta
df = pd.DataFrame.from_records(result)
temps=(time.time()-time_start)/60
print("#Temps de generació del dataframe:","\n Durada: ",int(temps) if temps>0 else 0,"minut/s ", \
      int((temps-int(temps))*60),"segons.")

# Notació del nom del fitxer de backup
nom_fitxer="backup_DM_MM2020"

time_start = time.time()
# Exportació en format pickle:
df.to_pickle('{}.pkl'.format(nom_fitxer))
temps=(time.time()-time_start)/60
print("#Temps d'exportació a format Pickle:","\n Durada: ",int(temps) if temps>0 else 0,"minut/s ", \
      int((temps-int(temps))*60),"segons.")

time_start = time.time()
# Exportació en format CSV:
df.to_csv ('{}.csv'.format(nom_fitxer), index = False, header=True)
temps=(time.time()-time_start)/60
print("#Temps d'exportació a format CSV:","\n Durada: ",int(temps) if temps>0 else 0,"minut/s ", \
      int((temps-int(temps))*60),"segons.")

time_start = time.time()
# Exportació en fitxer Excel:
df.to_excel('{}.xlsx'.format(nom_fitxer), index=0)
temps=(time.time()-time_start)/60
print("#Temps d'exportació a format Excel:","\n Durada: ",int(temps) if temps>0 else 0,"minut/s ", \
      int((temps-int(temps))*60),"segons.")
```

```
#Temps de Consulta sobre tots els tuits de la base de dades:
 Durada:  0 minut/s  0 segons.
#Temps de generació del dataframe:
 Durada:  0 minut/s  18 segons.
#Temps d'exportació a format Pickle:
 Durada:  0 minut/s  14 segons.
#Temps d'exportació a format CSV:
 Durada:  0 minut/s  22 segons.
#Temps d'exportació a format Excel:
 Durada:  1 minut/s  42 segons.
```

Figura Nº 10: Connexió i processament de tuits utilitzant la programació de scripts en Python.

3.2 Generació del dataset de dades.

En l'anàlisi de dades, s'ha considerat cada tuit com la unitat documental. El conjunt dels textos dels tuits i les seves paraules defineixen el nostre 'corpus' de dades. Sobre aquest 'corpus', s'han aplicat les tècniques de preprocessament prèvies a la construcció del model d'anàlisi com: la neteja del text, la seva normalització, tokenització i l'obtenció d'estructures significatives. Per fer-ho el primer pas ha estat generar el dataset de dades. Amb la informació de la taula 2 de base s'ha concentrat tota la informació d'interès per aplicar el procés de modelització. En aquest apartat es descriu totes les etapes d'aquest procés que finalitza amb el primer producte d'aquest treball, el dataset de dades.

3.2.1 Estructura del dataset.

El primer pas, fruit d'haver seleccionat la informació que ens pot fer servei a l'hora d'analitzar el nostre corpus de tuits, ha estat definir el conjunt d'atributs que compondran el dataset final. En la taula 3 es descriu breument cada atribut:

_id	Identificador de tuit únic assignat per la base de dades.
created_at	'Timestamp' (Data/Hora) de creació del tuit.
text_x	Text original del tuit, tal com es va capturar.
text_net	Text format per cadenes de caràcters sense puntuació, 'stopwords', ni dígits numèrics, cadenes amb nombre repetit d'espais en blanc, cadena 'RT' (simbologia de retuit) o '…' (senyal de text tallat).
text_Norm	Text que a més a més d'haver estat netejat, ha estat lematitzat.
diaSem	Nom del dia de la setmana.
dia	Dia en format numèric. Domini [0..31].
mes	Mes en format numèric. Domin [1..12].
yy	Any en format numèric. Es defineix per compatibilitat futura.
hora	Hora del dia en format numèric de dos dígits. Domini [01..24].
minuts	Minuts d'una hora. Domini [00..59].
segs	Segons d'un minut. Domini [00..59].
hashtags	Llista amb el text de cada hashtag inclòs al tuit (sense el caràcter #).
user_mentions	Llista amb el text de cada menció d'usuari (sense el caràcter @).
user_name	Nom de l'usuari emissor.
user_idstr	Identificador text de l'usuari emissor.
user_friends_c	Quantitat d'amics de l'usuari emissor.
user_followers_c	Quantitat de seguidors ('followers') de l'usuari emissor.
user_listed_c	Nombre de llistes d'usuaris definides per l'usuari emissor.
retweet_count	Quantitat de retuits del tuit.
lang	Estimació de l'idioma usat en escriure el text del tuit.
polarity	Coeficient de sentiment del contingut del tuit. Domini [-1..1]. On -1 és negatiu, 0 és neutre i 1 és positiu.
subjectivity	Coeficient que reflecteix el grau d'opinió, emoció o judici existent al tuit. Domini [0..1]. On 0 implica no subjectivitat i 1 subjectivitat màxima.
emojis	Conjunt de símbols de tipus caràcter propis de l'argot de Twitter.
text_y	Traducció a l'anglès el contingut del tuit.

Taula Nº3: Selecció de la capçalera de camps del dataset de modelització.

L'estructura mostrada és l'estructura base dissenyada pel dataset de modelització i en els següents apartats es descriu com s'han implementat les diferents etapes fins a disposar un dataset final.

3.2.2 Procés de traducció de tuits.

En aquest treball s'ha decidit traduir tots els continguts de tuit a l'idioma anglès. El procés de traducció s'ha dissenyat en dos blocs de treball. Un primer bloc processant tots els tuits en altres idiomes diferents del castellà i l'anglès, i un segon bloc processant la traducció del castellà a l'anglès.

Per la traducció de textos disposem de diverses opcions, però poques realment gratuïtes. Cercant recursos per aquesta opció, s'ha decidit finalment que una bona solució era aprofitar el marge de traducció gratuïta dels diferents recursos.

En total, la base de dades emmagatzema 102682 tuits, dels quals 4190 no tenien un idioma definit i finalment els hem descartat. El motiu ha estat que qualsevol mètode de detecció d'idioma automatitzat no ens garantia bons resultats. Sobre la base de 98442 tuits s'ha aplicat el procés de traducció, essent un total de 9195 en idiomes diferents del castellà i l'anglès i un total de 37676 en castellà. Per generar el dataset final s'han afegit els textos dels 51571 restants en anglès.

Per la traducció dels tuits en castellà, s'ha utilitzat el servei d'IBM Watson 'Language Translator' i per la traducció dels tuits en altres idiomes el servei del traductor de Google en la seva modalidat de traducció per documents. En el següent apartat es descriu el procediment de generació del dataset final. En aquest procés es genera per cada tuit en un idioma diferent del castellà, un fitxer excel amb els tuits a traduir amb la següent notació de noms '**altres_id.xlsx'**, on '**id**', és un identificador de país de dos caràcters format estandard usat per l'API de Twitter. Per cada fitxer s'han calculat les traduccions, usant el traductor de Google, i configurant l'idioma adequat de forma manual. El script de generació del dataset final llegeix les traduccions d'aquest fitxer amb la traducció i notació de nom '**altres_id_traduits.xlsx**' i defineix un dataframe amb tots els tuits traduïts, eliminant els repetits. Aquest procediment també s'ha aplicat als tuits en idioma castellà. El procés de traducció està integrat en el procés global de generació del dataset final, on a més a més de fer la traducció, s'extreu la informació d'interès per la modelització.

Per tant queda clar, que en aquest treball, la part de traducció es deixa implementada de manera que qualsevol altra opció de traducció és integrable, si se segueixen els criteris de notació d'arxius.

3.2.3 Tasques de preprocessament: neteja i normalització de dades.

Per tal de poder aplicar tècniques d'aprenentatge no supervisat ha estat necessari, que el text estigui representat de manera numèrica, identificant cada 'token' o alternativament cada sentència amb significació important del contingut del tuit per un vector. D'aquesta forma podem mesurar la proximitat o similitud de les paraules i per extensió dels tuits. Amb la capacitat de decidir la similitud

dels tuits es pot fer la seva agrupació, i identificar cada grup per unes paraules clau o temàtiques de grup. És clar, que ens cal disposar d'un text depurat, on tinguem definides el millor possible les paraules i els grups de paraules o frases que millor representen el significat del tuit. Les tasques de neteja han consistit en:

a) Eliminar les Urls.
b) Substituir seqüències de diversos caràcters en blanc per un de sol.
c) Eliminar els signes de puntuació i caràcters especials com retorns de carro, alimentació de línia, tabuladors, etc.
d) Eliminar els dígits numèrics.
e) Eliminar les 'stop words' (paraules molt freqüents però poc significatives).
f) Eliminar els caràcters @ i # en mencions d'usuari i hashtags. En aquest cas podem escollir també eliminar-los del text.
g) Eliminar seqüències de text pròpies del domini Twitter, com la paraula 'RT' que designa un retuit i el caràcter '…' a final dels tuits de longitud major a la permesa.
h) Extracció i eliminació d'emojis.

Pel procés de normalització, es necessitava reduir les derivacions possibles d'una mateixa paraula troncal, o les inflexions d'una mateixa forma gramatical, en definitiva, aplicar un procés de 'stemming' per obtenir aquella part de la paraula que les representa a totes. Per l'idioma anglès, usant la llibreria **nltk** disposem d'un mòdul per 'stem' que ofereix dues opcions. En aquest treball ens hem decidit per '**LancasterStemmer**', que va ser creat el 1990 i utilitza un enfocament més agressiu que l'altra opció **Porter Stemming Algorithm** que és més antic. El fet de ser més agressiu implica que les paraules troncals tendeixen a ser més curtes.
El fet d'haver aplicat un procés de traducció, ha simplificat el procés de normalització i no ens cal seleccionar opcions per llenguatges diferents de l'anglès.

En la figura 11, es mostra el codi de la funció Python implementada per realitzar totes les tasques descrites. La funció ***NetejaNorm***, aplica totes aquestes operacions sobre una llista de tokens, que són aquells termes representatius del contingut del tuit. NLTK ofereix un 'tokenitzador' específic per l'anàlisi de tuits anomenat ***TweetTokenizer()***.

Integrat en el procés de generació del dataset de modelització, es fa una crida a la funció ***depuraTextTuit(dataset,idioma)***, (mostrada en la figura 12), un cop es disposa de tots els tuits traduïts. Aquesta funció crida a la funció ***NetejaNorm***, per netejar i normalitzar el text. A més a més calcula la polaritat i subjectivitat del text. La llibreria **textblob**, calcula la propietat 'sentiment' d'un text i retorna una estructura de dades, ***Sentiment (polaritat, subjectivitat)***. La puntuació de polaritat és un nombre decimal dins del rang [-1,0, 1.0]. Un valor -1.0 indica una polaritat negativa, un valor 0 neutral i un valor 1.0 una polaritat positiva o a favor del contingut del tuit. La subjectivitat és un nombre decimal dins del rang [0.0, 1.0] on 0.0 és gens objectiu i 1.0 és totalment subjectiu.

```python
def NetejaNorm(text_o,idioma='english', stem=0, meta=0):
    text=""
    text_normalitzat =""
    # Calculem tokens
    tweet_tokenizer = TweetTokenizer()
    tokens = tweet_tokenizer.tokenize(text_o)
    # Crea un element anomenat 'stemmer' que selecciona l'arrel comuna de la paraula.
    # evitant al text diverses formes d'una mateixa paraula.
    # Ex: cantar és 'stem' o troncal de cantat, cantada, cantaria
    if stem==0:
        stemmer = LancasterStemmer()
    else:
        stemmer = PorterStemmer() # Opció menys agressiva i més antiga.
    # Processa cada token del text i aplica totes les regles de netejat
    # i finalment aplica steamming per obtenir el text normalitzat.
    for elem in tokens:
        # Estandaritzem el tex a minúscules.
        elem=elem.lower()
        # Processem sempre i quan no sigui una 'stopword'.
        if elem  not in stopwords.words(idioma):
            # Elimina URL's
            text_net=re.sub(r"\w+:\/{2}[\d\w-]+(\.[\d\w-]+)*(?:(?:\/[^\s/]*))*", '',elem,flags=re.UNICODE)
            # Eliminem la informació de context:
            if meta==0:
                # Elimina hashtags
                text_net=re.sub(r'#([^\s]+)', "", text_net)
                # Elimina referencies a usuari
                text_net=re.sub(r'@([^\s]+)', "", text_net)
            # Obté text alfanumeric (substitueix tot caràcter que no és alfanumèric per "")
            text_net = re.sub('([^\w\'+]|[0-9]+)',' ',text_net).strip()
            # Neteja caracteres en blanc, marques 'rt'
            text_net = re.sub('[\t\n\r\f\v]+','', text_net)
            text_net = re.sub(r' +',' ', text_net,flags=re.UNICODE)
            if text_net.startswith("rt"):
                text_net = text_net.replace('rt','',1)
            # En paraules apostrofades ens quedem amb la part no apostrofada.
            text_net = text_net.split("'")[0]
            # Filtrem les paraules d'un sol caràcter residuals.
            # Reconstruim usant dues llistes, el text del tuit netejat
            # i el text normalitzat.
            if len(text_net)>1:
                if text=="":
                    text = text_net
                    text_normalitzat = stemmer.stem(text_net)
                else:
                    text = text + " " + text_net
                    # Cerquem l'arrel o paurala base en cada mot del text detectat.
                    text_normalitzat = text_normalitzat + " " + stemmer.stem(text_net)
    # Si el tuit en el procés de neteja queda buit
    # el representem per un caràcter en blanc per compatibilitat al exportar a fitxer excel.
    if len(text)==0:
        text = " "
        text_normalitzat = " "
    ftext=text+'@'+text_normalitzat
    return ftext
```

Figura Nº 11: Script Python amb la funció ***NetejaNorm*** (realitza les tasques de neteja del text d'un tuit).

```python
def depuraTextTuit(dataset,idioma):
# Funció que aplica sobre tots els tuits traduits, les regles de
# netejat del text i normalització. A més a més per l'anàlisi de
# sentiment, calcula la polaritat i la subjectivitat del text.

    textClean=[]
    textNorm=[]
    emojis=[]
    polarity = []
    subjectivity = []

    # Processat de tots els tuits del dataset
    # Mostrant seguiment per pantalla, dia a dia.
    time_start0 = time.time()
    time_start = time.time()
    m=1 ; dia=0
    for row in dataset.iterrows():
        text = row[1]['text_y']
        # Eliminem parts del text que no ens aporten significat
        # Elimina stopwords, hashtags i referències a usuaris.
        # Normalitzem el text, per reduir la diversitat de paraules al tuit.
        Netejat = NetejaNorm(text, idioma)
        # Obtenim el text netejat (abans del caràcter @)
        # i el normalitzat (després del caràcter @)
        Netejat=Netejat.split('@')
        textClean.append(Netejat[0])
        # Calcula els emojis continguts al text.
        emojis.append(cerca_emojis(text))
        textNorm.append(Netejat[1])
        # Utilitzem la llibreria 'textblob' per calcular
        # Polaritat i Subjectivitat.
        polarity.append(TextBlob(row[1]['text_y']).sentiment.polarity)
        subjectivity.append(TextBlob(row[1]['text_y']).sentiment.subjectivity)
        if m!=1:
            if row[1]['dia']!=dia:
                temps=(time.time()-time_start)/60
                print("Data: {}/{}/{}".format(dia,mes,yy),"\n#Tuits processats:", \
                      m-1,"\n Durada: ", \
                      int(temps) if temps>0 else 0,"minut/s ", int((temps-int(temps))*60), \
                      "segons.")
                m=1
                time_start = time.time()
        dia=row[1]['dia'] ; mes=row[1]['mes'] ; yy=row[1]['yy']
        m=m+1
    # Temps parcial diari.
    temps=(time.time()-time_start)/60
    print("Data: {}/{}/{}".format(dia,mes,yy),"\n#Tuits processats:",m,"\n Durada: ", \
          int(temps) if temps>0 else 0,"minut/s ", \
          int((temps-int(temps))*60),"segons.")
    # Temps total
    temps=(time.time()-time_start0)/60
    print("#Total de tuits processats:",dataset.shape[0],"\n Durada: ",int(temps) if temps>0 else 0, \
          "minut/s ", \
          int((temps-int(temps))*60),"segons.")

    df_net_norm=pd.DataFrame({'text_net':textClean,'text_Norm':textNorm,'emojis':emojis, \
                              'polarity':polarity,'subjectivity':subjectivity})

    return df_net_norm
```

Figura Nº 12: Script Python: procés de depuració del text i càlcul d'informació de context.

Per finalitzar aquest apartat de processament del text, es calcula un nou atribut a utilitzar com a informació de context, que conté els emojis utilitzats en el tuit. Per obtenir els emoticons o emojis, cal considerar un ampli ventall de conjunts de caràcters que contínuament es van actualitzant. En la figura 13, es mostra el codi utilitzat per la seva detecció, i es pot observar quina codificació s'ha usat per obtenir-los.

```python
RE_EMOJI = re.compile("["
        u"\U0001F600-\U0001F64F"  # emoticons
        u"\U0001F300-\U0001F5FF"  # simbols i pictogrames
        u"\U0001F680-\U0001F6FF"  # simbols de transport i senyalització/icones
        u"\U0001F1E0-\U0001F1FF"  # flags (iOS)
        u"\U00002500-\U00002BEF"  # caràcters xinesos
        u"\U00002702-\U000027B0"
        u"\U00002702-\U000027B0"
        u"\U000024C2-\U0001F251"
        u"\U0001f926-\U0001f937"
        u"\U00010000-\U0010ffff"
        u"\u2640-\u2642"
        u"\u2600-\u2B55"
        u"\u200d"
        u"\u23cf"
        u"\u23e9"
        u"\u231a"
        u"\ufe0f"  # dingbats
        u"\u3030"
        u"\uf7e2"
                      "]+", re.UNICODE)
# Extreiem els emojis/emoticons del text del tuit.
def cerca_emojis(text):
    return RE_EMOJI.findall(text)
```

Figura Nº 13: Codi Python amb la implementació del procés d'obtenció d'emojis del text d'un tuit.

El procés complet de processament de dades aplicat a la base de dades de tuits, queda integrat en el procés de generació del dataset de modelització implementat en el notebook ***Genera_Dataset.ipynb*** inclòs en GitHub.

3.2.4 Procés de generació del dataset usat en la modelització.

El procés de generació del dataset de dades, té per objectiu, realitzar la selecció de dades d'interès per l'anàlisi i disposar-les en un format tabular. Part dels atributs inclosos en el dataset final, són dades directes de l'estructura origen en què s'ha capturat cada tuit (tweet object JSON Line) i part són resultat d'aplicar el processament de dades explicat en l'apartat anterior. En la figura 14, es presenta una mostra del dataset generat, i en la figura 15, un esquema global d'aquest procés.

created_at	text_x	text_net	text_Norm	diaSem	dia	mes	yy	hora	...	user_idstr	user_friends_c	user_followers_c	user_listed_c	retweet_count	lang	polarity	subjectivity	emojis	text_y
Thu Feb 13 09:58:19 +0000 2020	RT @FEDER_ONG: Las 29 obras ganadoras del conc...	winning works photographic contest form travel...	win work photograph contest form travel exhibi...	Thursday	13	2	2020	9	...	4054806561	8	3	0	0	es	0.5	0.75	[]	RT @FEDER_ONG: The 29 winning works of the pho...
Thu Feb 13 10:01:01 +0000 2020	📌 Esta tarde, a partir de las 21:00h, nuevo #C...	afternoon starting new centered	afternoon start new cent	Thursday	13	2	2020	10	...	64270216	853	2104	31	0	es	0.0681818	0.277273	[📌]	📌 This afternoon, starting at 21 :00h, new #Cr...
Thu Feb 13 10:01:31 +0000 2020	RT @FEDER_ONG: Gracias a Emilio Butragueño, Da...	thanks emilio butragueño david de maria soleda...	thank emilio butragueño david de mar soledad g...	Thursday	13	2	2020	10	...	233031604	446	648	9	0	es	0.2	0.2	[]	RT @FEDER_ONG: Thanks to Emilio Butragueño, Da...
Thu Feb 13 10:01:38 +0000 2020	Va a quedar tan bonito @nh487 ...\n\n#FebreroR...	going look beautiful	going look beauty	Thursday	13	2	2020	10	...	237839609	11718	19380	682	0	es	0.85	1	[]	It's going to look so beautiful.\n\n#FebreroRa...
Thu Feb 13 10:02:02 +0000 2020	RT @ERdivulga: todo un honor tener a @PastorAl...	honor collaborating teaching subject physiopat...	hon collab teach subject physiopatholog	Thursday	13	2	2020	10	...	3002280588	236	123	0	0	es	-0.166667	0.333333	[]	RT @ERdivulga: it is an honor to have @PastorA...

Figura Nº 14: Mostra dels cinc primers registres del dataset resultant de la fase de processat.

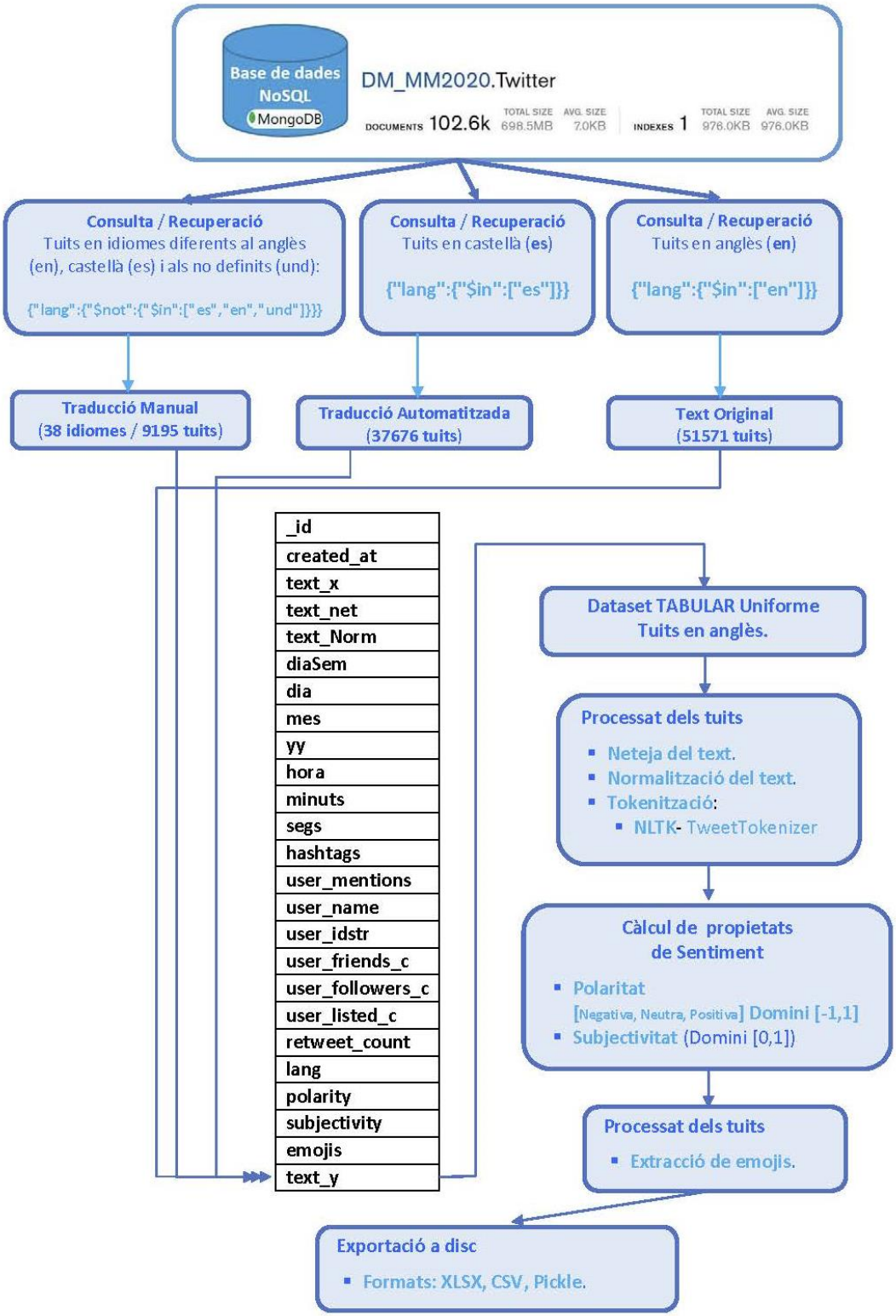


Figura Nº 15: Visió general del processament de dades i generació del dataset de modelització.

3.2.5 Conclusions.

Les conclusions obtingudes després d'aquest primera fase d'anàlisi, són la dificultat de gestionar la immensa variabilitat del text que ens trobem en un tuit. La mateixa filosofia d'un missatge acotat en longitud i la possibilitat de poder referenciar informació externa per estalviar text, pròpia de Twitter, determina la

forma de processar el text, per obtenir un text normalitzat que el representi al màxim. Dos fets són importants la personalització de les funcions de neteja i depuració del text, (en el cas de textos amb un domini especialitzat) i el fet de disposar de bones traduccions del text en altres idiomes. Altres opcions com la de gestionar corpus multi llenguatge poden dependre molt de les llibreries existents i es corre el risc de perdre el control de la fase de processat.

3.3 Anàlisi de les dades.

Un cop definit i preprocessat el conjunt de dades capturat, l'objectiu és aplicar les tècniques d'aprenentatge no supervisat per analitzar i modelitzar el volum de dades. Seguint el model CRISP-DM, s'ha realitzat una exploració inicial de dades i una transformació de la representació de les dades, a un format numèric (en anglès feature engineering') de manera que puguin ser introduïdes en els models no supervisats.

3.3.1. Exploració inicial, anàlisi estructural i visual.

L'exploració inicial s'ha dut a terme sobre un *dataframe* o taula de dades, indexat per la data de cada tuit, per disposar d'una sèrie temporal de tuits i poder generar les gràfiques que ens permetin comprendre de quina manera s'han generat en el temps. En la figura 16, es mostra el dataframe amb el contingut del text ja netejat en la columna '**text**', normalitzat en la columna '**text_norm**' i amb l'usuari de Twitter en l'atribut '**autor**'. Per conservar la privacitat dels usuaris utilitzem un identificador numèric obviant el seu nom i àlias.

```
# Lectura del dataset generat per modelitzar
df=pd.read_excel("DMMM_dataset_Final.xlsx")
```

```
# Llegim els tuits depurats. I generem una llista
# de documents amb el text de cada tuit.
tuits = list(df.text_net)
```

```
tuits[0:5]
```

```
['winning works photographic contest form travelling exhibition tour geography espa',
 'afternoon starting new centered',
 'thanks emilio butragueño david de maria soledad giménez commitment thang thang',
 'going look beautiful',
 'honor collaborating teaching subject physiopathology']
```

```
index_list=[]
for row in df.iterrows():
    index_list.append(pd.to_datetime(row[1]['created_at'], format='%a %b %d %H:%M:%S +0000 %Y'))
df.index = index_list
```

```
tuits=df[['_id','text_net','text_Norm','user_idstr']]
tuits.columns=['_id','text', 'text_norm','autor']
df=[]
```

```
tuits.head(3)
```

	_id	text	text_norm	autor
2020-02-13 09:58:19	5e78e80f4e54db2148d19749	winning works photographic contest form travel...	win work photograph contest form travel exhibi...	4054806561
2020-02-13 10:01:01	5e78e80f4e54db2148d1974a	afternoon starting new centered	afternoon start new cent	64270216
2020-02-13 10:01:31	5e78e80f4e54db2148d1974b	thanks emilio butragueño david de maria soleda...	thank emilio butragueño david de mar soledad g...	233031604

Figura Nº 16: Dataframe de dades usat en l'exploració inicial.

En les successives tres figures es mostra una anàlisi temporal de com s'han anat produint els tuits, en cada minut i hora de cada dia.

En la figura 17, s’observa una visió global de la freqüència d’enviament de tuits per minuts, en tot el període de temps analitzat. Es distingeix els dies on s'han celebrat esdeveniments o bé oberts temes de debat, amb més participació per parts dels usuaris de Twitter. La participació més gran, pertany al dia 29 de febrer, on el màxim nombre de tuits és proper als 140 tuits en un minut, a partir de les 09:00h i durant quasi bé tot el dia. El pic màxim de tot el dia es produeix entre les 14:00h i 15:00h, però en totes les hores es mantenen aquests valors en algun minut.

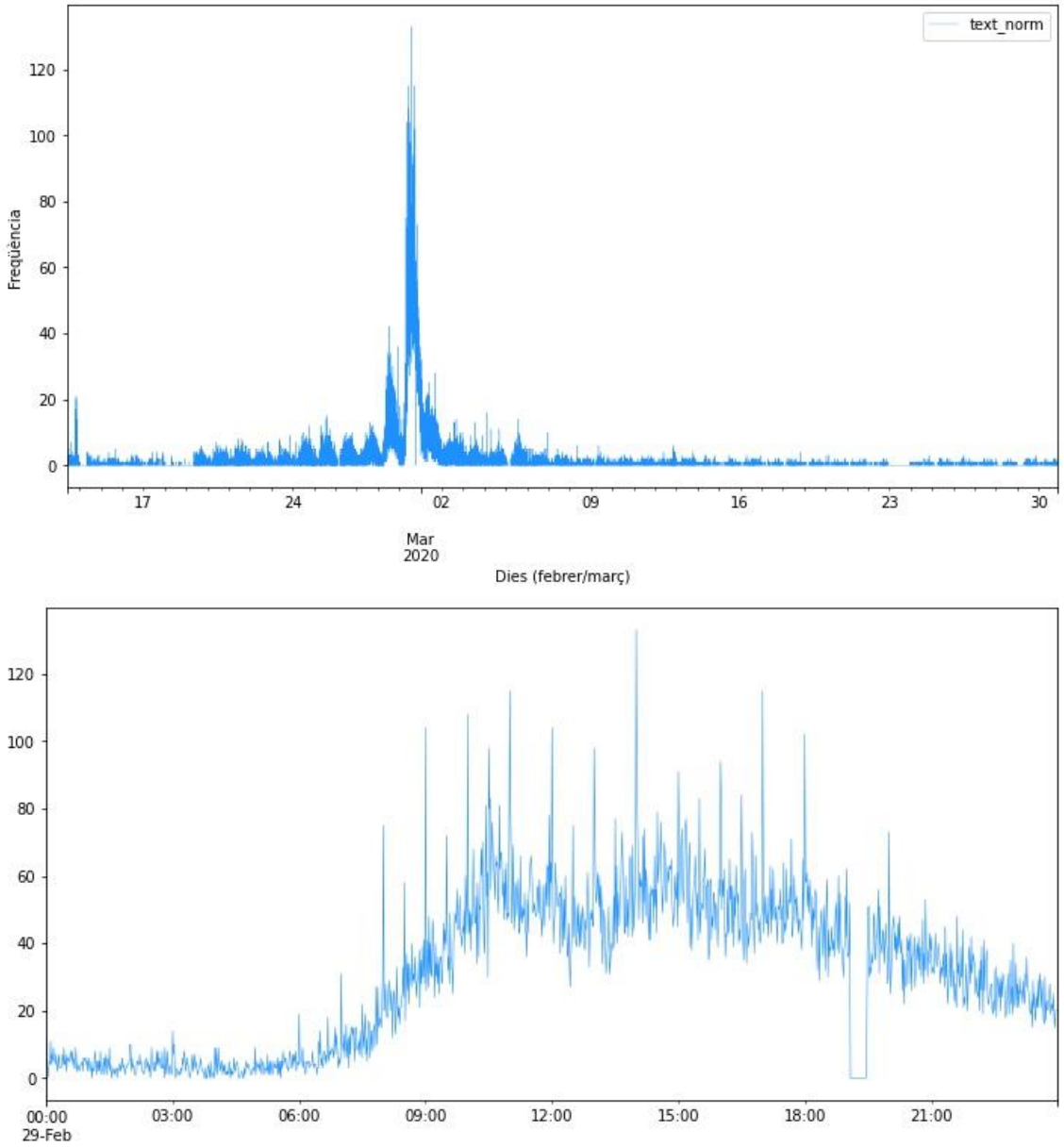


Figura Nº 17: Freqüència minutal diària d’emissió de tuits.

En la part superior de la figura 17, tenim l’histograma de tot el període, mentre que en la part inferior, el detall del dia 29 de febrer.

En la figura 18, es pot observar la freqüència d’enviament diària de tuits.

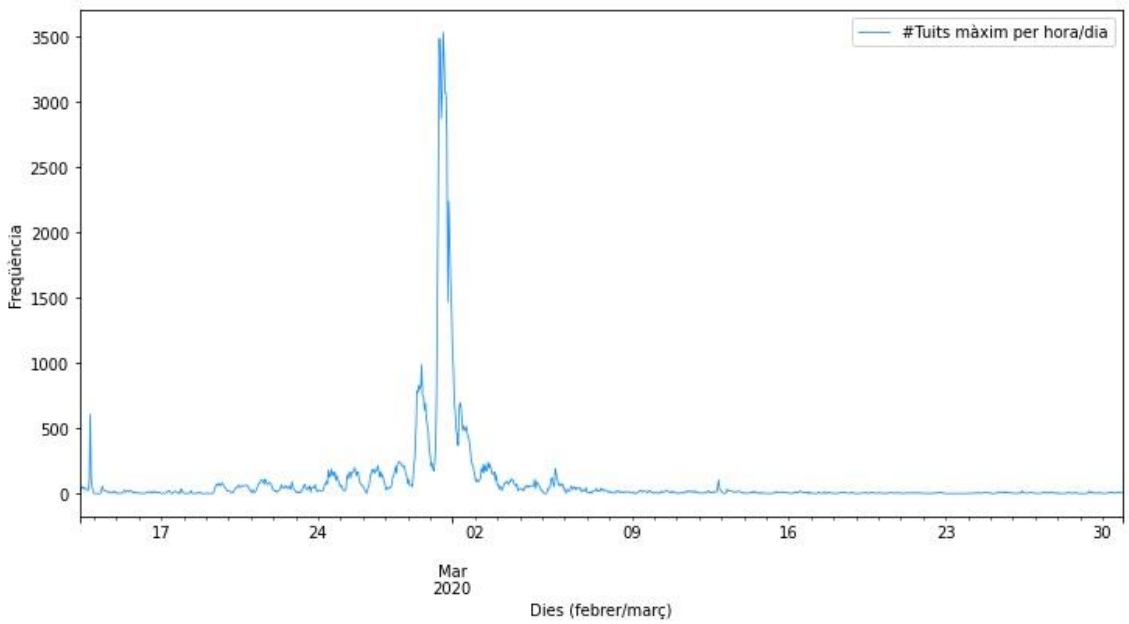


Figura Nº 18: Nombre de tuits diaris.

A l'ampliar el mostreig a hores, s'ha observat què el nombre de tuits màxim per dia arriba a assolir-se també el dia mundial de les malalties minoritàries, en la mateixa franja horària i és de 3530 tuits. Com s'esperava, la progressió d'emissió de tuits a mesura que ens apropem a la celebració del dia mundial de les malalties minoritàries, és creixent fins al dia 29, i decreixent després, reflectint de manera nítida, com durant cada dia hi ha unes hores de més activitat.

Quant a la freqüència horària, és interessant definir quines hores del dia tenen més activitat, per tal de conèixer en quines hores podem aconsellar realitzar la difusió de missatges o aprofitar per proporcionar futurs esdeveniments amb més capacitat d'arribar als usuaris.

En les següents gràfiques es mostra aquesta activitat horària. S'han seleccionat el període de dies del 27 al 29 de febrer i de l'1 al 4 de març. Ja en el procés de captura de dades, coneixíem, que el volum de tuits diari de mitjana en dies normals i llunyans al dia 29, estava per l'ordre de 100 tuits diaris. Per tant l'increment en els dies propers, durant i posteriors al dia 29 el creixement és molt sobtat. Aquest fet ens fa pensar que per les malalties minoritàries hi ha molt interès, i que la participació és molt alta quan els usuaris són animats a participar. Però tanmateix requereix d'una promoció més activa d'esdeveniments que engresquin la participació activa fora d'aquest període, per els usuaris no afectats.

En les següents gràfiques de barres en format horitzontal observem la data i hora de cada dia mostrejat amb l'etiquetatge del nombre de tuits corresponent. Tal com s'ha indicat abans cal observar també en quines hores l'activitat dels usuaris, és més alta. Les dades han estat recollides en format horari UTC (on les sigles UTC signifiquen "Temps Universal Coordinat" [en anglès, "Universal Time Coordinated"], que es coneix també per l'horari en el meridià de Greenwich o

“GMT”), per tant, cal realitzar una conversió d’una hora més en l’horari peninsular, exceptuant els dies posteriors al 27 de març en què pel canvi horari són dues hores més.

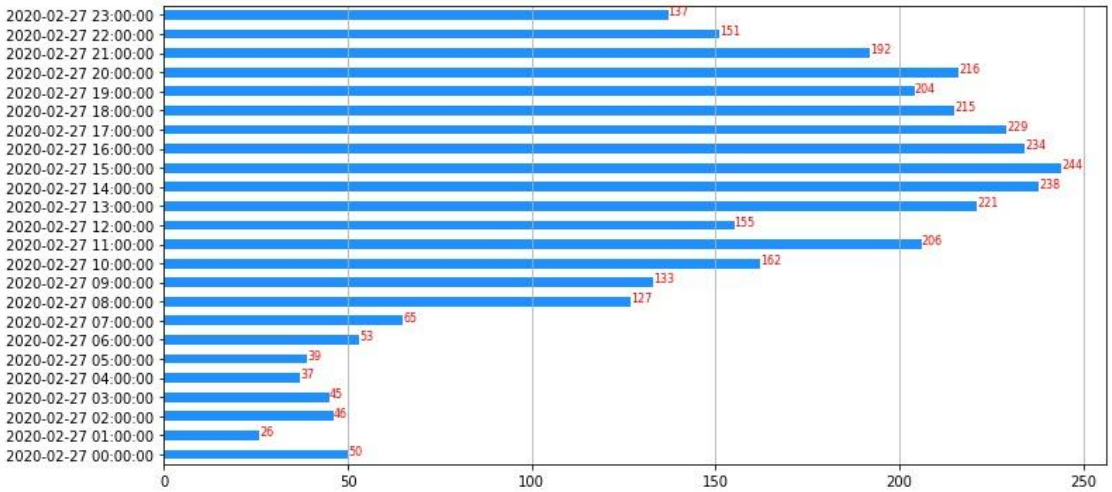


Figura Nº 19: Nombre de tuits per hora durant el dia 27 de febrer.

El dia 27 és el primer dia d’entre els ‘normals’, que a partir de les 09:00h es comencen a enviar tuits amb una freqüència horària comparable a la diària de dies posteriors. A més a més les hores del migdia i tarda són força més actives.

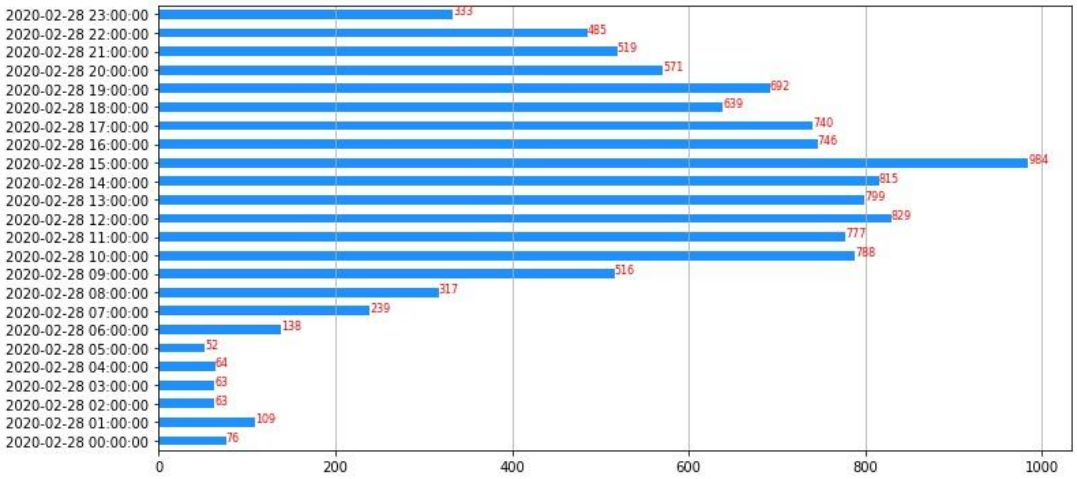


Figura Nº 20: Nombre de tuits per hora durant el dia 28 de febrer.

Durant el dia 28, s’observa que en les primeres hores de la mitjanit, la tendència del dia 27, es trenca i es produeix un descens que no es repeteix en la resta de dies. Però des de bon matí la freqüència d’enviaments és força alta. Passem d’un enviament màxim de 244 tuits en una hora, a triplicar en hores actives aquesta quantitat. És el primer dia que gairebé s’arriba a un màxim de 1000 tuits, quan en un dia normal s’envien 100.

Del processament de dades, sabem que s’envien 11.356 tuits. En idiomes diferents de l’anglès i castellà, s’envien 1168 tuits, en castellà 4117 i en idioma anglès 6171.

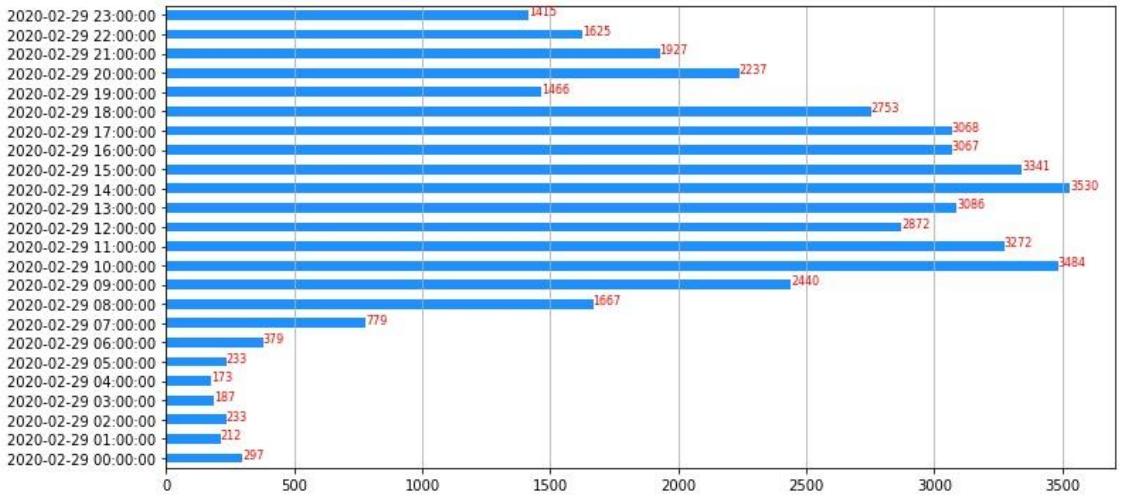


Figura Nº 21: Nombre de tuits per hora durant el dia 29 de febrer.

El dia 29 de febrer, s'assoleix la màxima activitat en totes les hores. Inclòs en les hores de son, es registra una activitat com la d'un dia normal (aproximadament 100 tuits). Aquest fet implica que tota la mostra quedi influenciada per aquest dia, per què l'increment és molt sobtat. El màxim nombre de tuits s'assoleix amb 3530, però com totes les hores són similars, estem en el cas que l'activitat d'enviament és gairebé d'un tuit per segon durant les hores de màxima activitat i de més de 15 tuits cada minut en la resta. Per idiomes, d'un total de 43.750 tuits enviats, en castellà s'envien 14.890, en anglès 24420 i en altres idiomes 4.440.

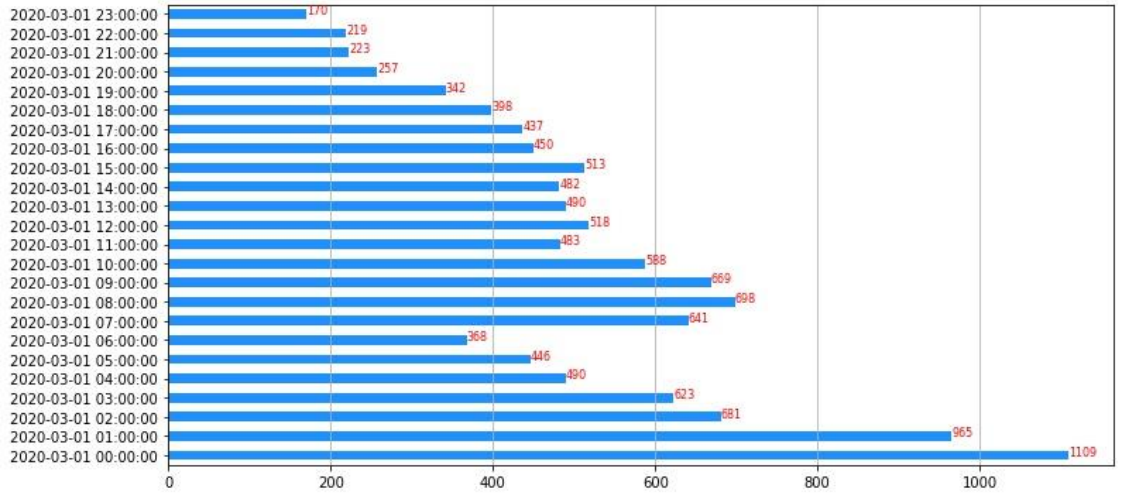


Figura Nº 22: Nombre de tuits per hora durant el dia 1 de març.

El dia 1 de març és un reflex de la des escalada d'activitat. Però és interessant veure com fins a les 06:00h hores de la matinada encara hi ha repercussió de la celebració del dia anterior. Durant tot el dia, gradualment es va reduint el nombre d'enviaments i s'assoleix al final del dia, una certa normalitat situada en 200 tuits (el doble d'enviaments d'un dia normal).

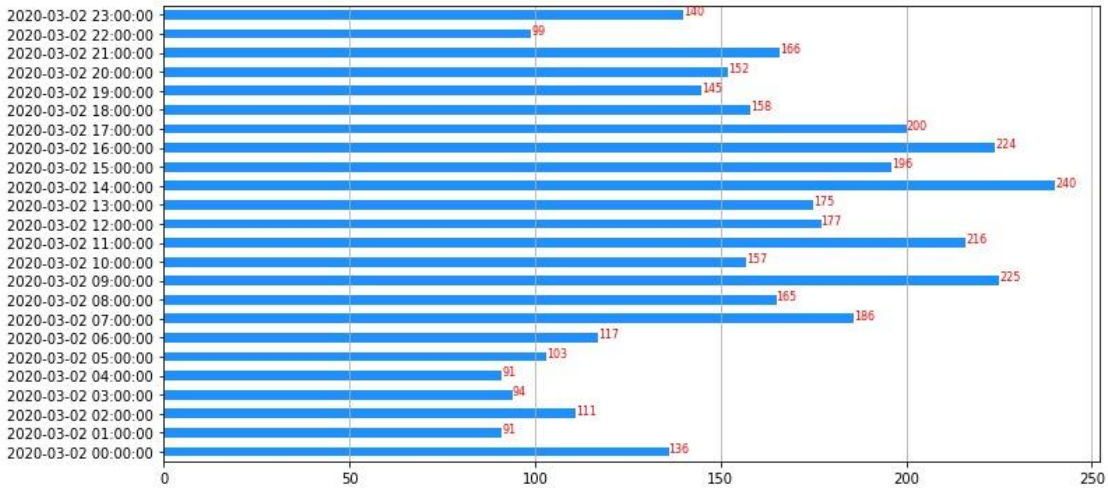


Figura Nº 23: Nombre de tuits per hora durant el dia 2 de març.

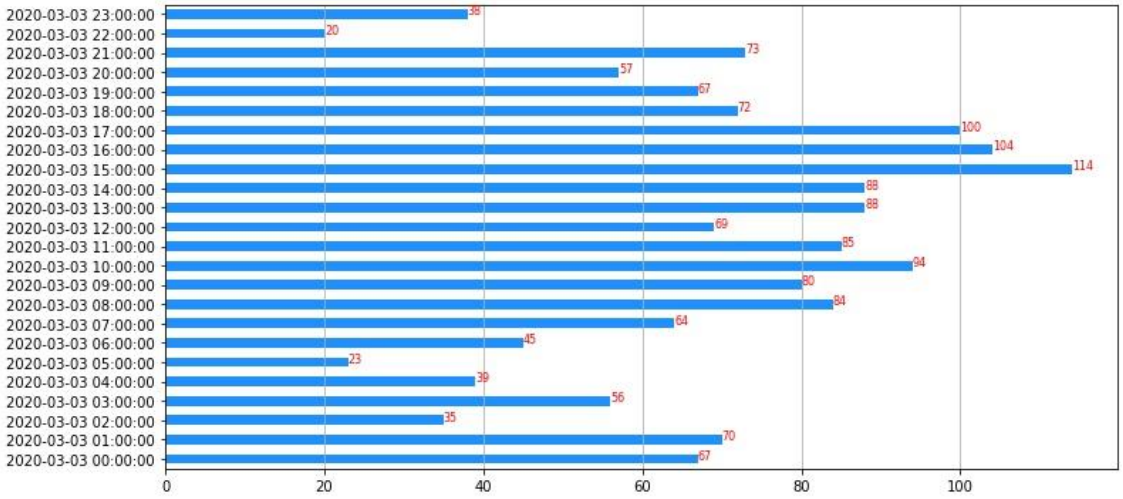


Figura Nº 24: Nombre de tuits per hora durant el dia 3 de març.

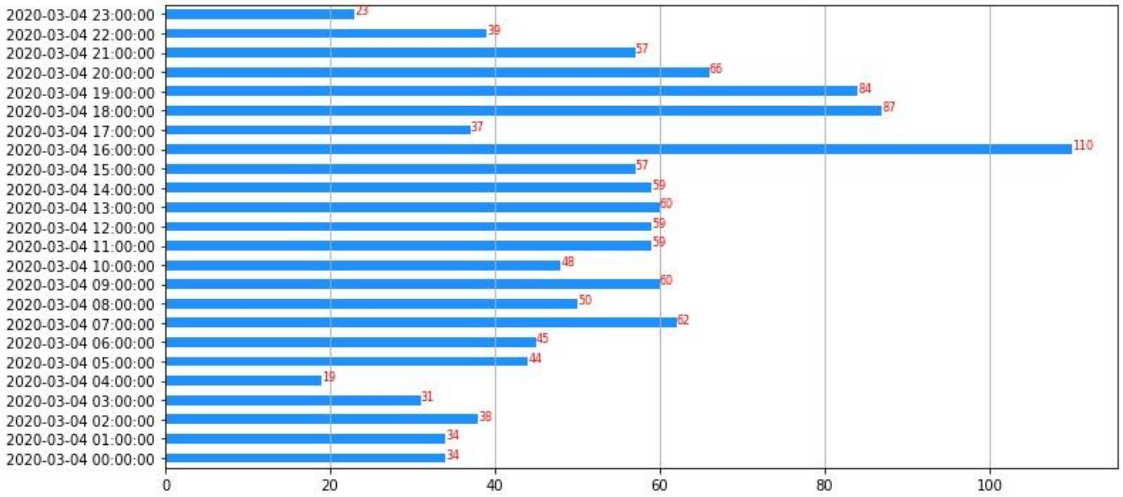


Figura Nº 25: Nombre de tuits per hora durant el dia 4 de març.

Els dies 2,3 i 4 de març destaquen per presentar una freqüència regular d'enviament a gairebé totes les hores del dia. És molt probable que el conjunt

d'activitats efectuades durant aquests dies generi aquest efecte. Tot i haver baixat totalment el nombre de tuits, poden ser els dies en conjunt, en què els tuits reflecteixin un nombre més divers de temàtiques diferents. El dia 13/02, es un cas de molta activitat però per la celebració d'un esdeveniment puntual que va concentrar tota l'activitat.

Una comparativa global la podem acabar fent visualitzant la figura 26, on es pot observar la mitjana, valor mínim, màxim de tuits de cada dia així com la seva distribució, i comparar entre dies.

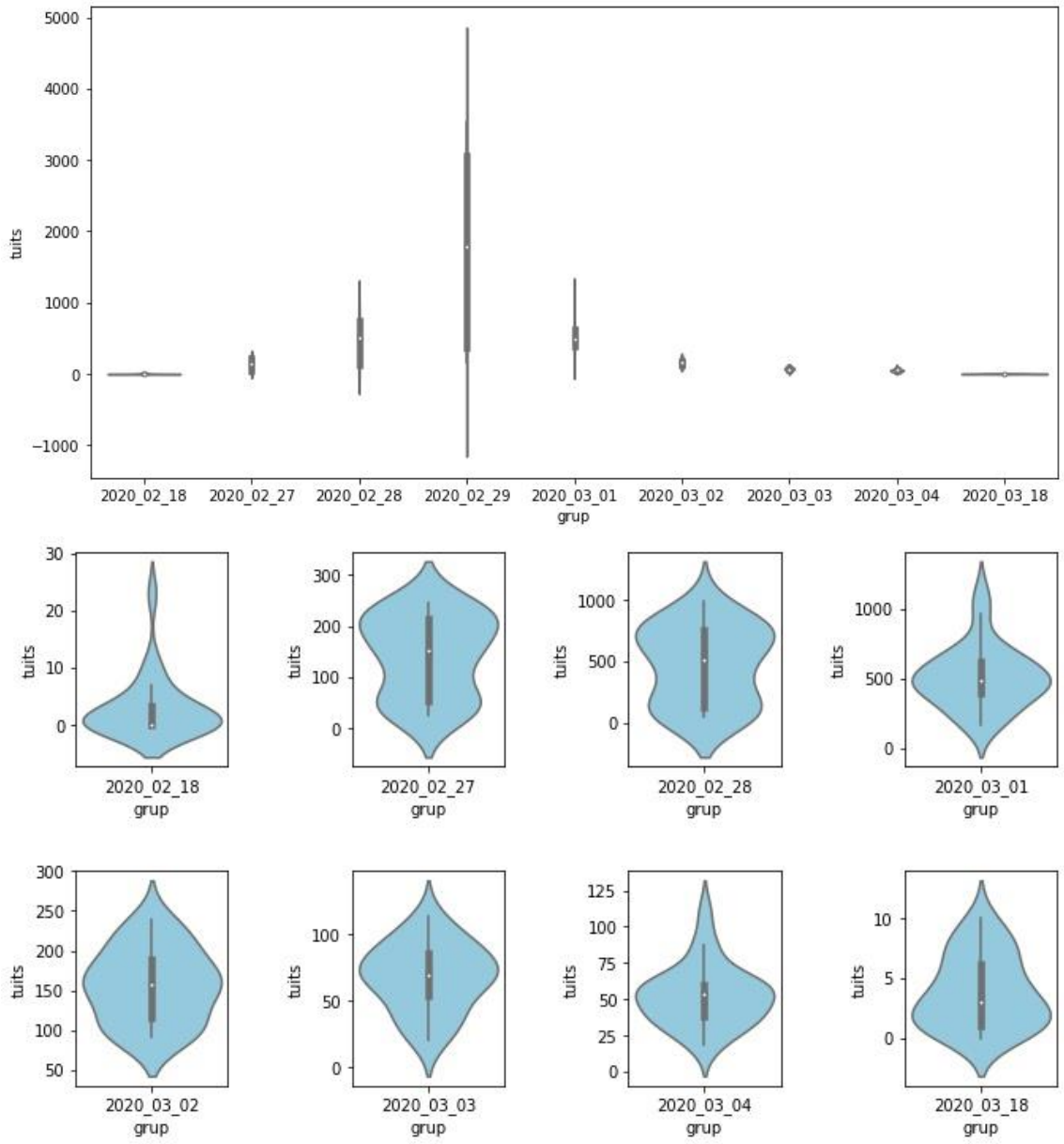


Figura Nº 26: Gràfics tipus violí per la comparativa global en el període dels dies 27,28,29,1,2,3,4.

Una de les maneres més simple d'avaluar el contingut del text, consisteix a fer recompte de les paraules que conté (en anglès 'bag of words'). Per fer-ho s'ha construït un 'corpus' de text format per tots els textos netejats de cadascun dels

tuits. Com en el camp '*text_net*', del dataset de dades per modelitzar disposem del contingut de cada tuit, s'ha calculat la llista de paraules i després utilitzat el codi mostrat en la figura 27 per comptar la freqüència de cada paraula.

```
from nltk.probability import FreqDist
RecompteParaules = FreqDist(word.lower() for tuit in tuits.text for word in tuit.split())
RecompteParaules
```

```
FreqDist({'rare': 38093, 'day': 26268, 'today': 19299, 'disease': 16293, 'diseases': 15709,
          'people': 13588, 'world': 9464, 'us': 8557, 'research': 7850, 'february': 6933, ...})
```

Figura Nº 27: Generació del 'Bag of Words' o recompte d'aparicions al text de cada paraula.

Com en el procés de captació s'ha captat els tuits usant hashtags, hem eliminat aquestes paraules evidents de la llista resultant. En la figura 28, es pot observar en la part superior, la llista dels mots més usats, i en la part inferior un gràfic de tipus 'núvol de paraules' ('word cloud' en anglès) generat amb la llibreria '**wordcloud**', associant la mida de la paraula amb la seva freqüència.

paraula	comptador								
people	13588	living	5071	life	4348	thank	3558	offers	3027
world	9464	patients	4988	many	4288	opportunity	3498		
us	8557	together	4868	know	4243	year	3334		
research	7850	support	4790	one	4066	tomorrow	3169		
awareness	6595	help	4643	raise	3900	story	3124		
million	5868	care	4375	want	3589	ill	3061		

Figura Nº 28: Paraules de màxima freqüència.

En aquest núvol de paraules, es pot observar sense dubte quina és en primera instància, la tendència i les temàtiques presents i que cal tenir en consideració.

Paraules com *persones*, *mundial*, recerca, *milions*, *suport*, consciència, *desitjos*, *tractaments*, *conèixer*, *integració*, *ajuda*, *malalts*, *nens*, *créixer*, *reptes*, *compartir*, *vida*, *genètica*, *esdeveniments*, tenir *cura*, *unió*, *síndrome*, *amics*, *patiment*, *millorar*, apareixen en primer pla. Aquest núvol de paraules confirma moltes de les necessitats actuals de les malalties minoritàries com és la recerca d'ajuda de

moltes persones, per obtenir un diagnòstic, per saber del mal que pateixen (molts cops d'origen genètic) i per trobar després quin tractament pot ser el millor remei.

Per acabar l'exploració de les dades, s'han calculat les associacions entre paraules consecutives al text, o n-grams, un cop depurat el seu text. Aquesta informació permet conèixer no només les n-tuples, sinó també quines són les més freqüents. En el nostre cas, s'han cercat, a efectes de millorar la comprensió del context en què es fa servir cada paraula. No les apliquem en el procés de vectorització per què el nombre de paraules o '*features*' ja és molt elevat. Ens interessa sobretot reduir-lo o optimitzar el conjunt, eliminant les paraules amb poc significat.

En la figura 29 mostrem les associacions existents en el conjunt de dades dels tuits processats, on també s'han eliminat les referències poc significatives, als hashtags de captura.

```
n_grams = collections.Counter(everygrams(tokens, min_len=2, max_len=4))
# Extreiem les 4 posicions inicials per ser paraules: Rare, Disease, Day, World
print(len(n_grams))
n_grams.most_common(25)[4:]
```

```
927585
[(('million', 'people'), 4285),
 (('rare', 'disease', 'day'), 4175),
 (('raise', 'awareness'), 3410),
 (('brothers', 'sisters'), 2997),
 (('today', 'day'), 2940),
 (('us', 'opportunity'), 2936),
 (('care', 'brothers'), 2929),
 (('care', 'brothers', 'sisters'), 2928),
 (('offers', 'us'), 2926),
 (('offers', 'us', 'opportunity'), 2926),
 (('opportunity', 'together'), 2925),
 (('together', 'care'), 2925),
 (('sisters', 'ill'), 2925),
 (('us', 'opportunity', 'together'), 2925),
 (('opportunity', 'together', 'care'), 2925),
 (('together', 'care', 'brothers'), 2925),
 (('brothers', 'sisters', 'ill'), 2925),
 (('offers', 'us', 'opportunity', 'together'), 2925),
 (('us', 'opportunity', 'together', 'care'), 2925),
 (('opportunity', 'together', 'care', 'brothers'), 2925),
 (('together', 'care', 'brothers', 'sisters'), 2925)]
```

Figura Nº 29: Llista de les 25 n-grams més freqüents (n:2-4).

3.3.2. Feature Engineering: Vectorització.

A l'hora de modelitzar necessitem representar les dades del text d'un tuit de forma numèrica. Es continua utilitzant el mateix concepte de recompte de paraules, però s'incorpora l'ús d'un factor de ponderació, que permet identificar les paraules més importants de les molt freqüents però irrellevants. La tècnica de vectorització és ***Tf-idf*** (en anglès, **T**erm **f**requency **i**nverse **d**ocument **f**requency).En aplicar *Tf-idf*, les paraules que es produeixen amb freqüència dins d'un document però no freqüentment dins del corpus, reben una ponderació més alta, ja que es parteix de la suposició que aquestes paraules són més

significatives en relació amb el contingut del document. En realitat, consisteix a fer un recompte de paraules i aplicar una transformació d'escalatge que assigna un factor d'importància a cada paraula. Al final obtenim una matriu normalitzada que descriu numèricament el conjunt de tots els tuits. Aquesta matriu té per files els vectors que representen els tuits i per columnes les paraules del nostre corpus de dades.

S'ha utilitzat la implementació de la llibreria **sklearn**. En aquest procés de transformació del corpus de text a la seva representació numèrica, molts dels paràmetres aplicables en la vectorització *tf-idf* de sklearn, no s'han usat, per ser operacions ja realitzades en la fase anterior de preprocessament. En la figura 30 veiem l'estructura de l'objecte de vectorització, els seus paràmetres i un exemple de visualització d'una part de la matriu escollida aleatòriament. També es pot comprovar, què per reduir al màxim les columnes de la matriu s'ha generat utilitzant el text normalitzat.

```
TfidfVectorizer(analyzer='word', binary=False, decode_error='strict',
                dtype=<class 'numpy.float64'>, encoding='utf-8',
                input='content', lowercase=True, max_df=1.0, max_features=None,
                min_df=1, ngram_range=(1, 1), norm='l2', preprocessor=None,
                smooth_idf=True, stop_words=None, strip_accents=None,
                sublinear_tf=False, token_pattern='(?u)\\b\\w\\w+\\b',
                tokenizer=None, use_idf=True, vocabulary=None)
```

```
matriu_tfidf=tfidf_vect.fit_transform(tuits.text_norm)
matriu_tfidf.shape
```

```
(98433, 14303)
```

```
dfmatriu_tfidf=pd.DataFrame(matriu_tfidf.todense(),columns=paraules)
```

```
dfmatriu_tfidf.iloc[1980:1990,8540:8555]
```

	myth	mytho	mytocondr	myxofibrosarcom	mzahir	má	málaga	mé	médi	médico	mégane	méndez	ménière	mésquemil	méthus
1980	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1981	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1982	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1983	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1984	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1985	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1986	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1987	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1988	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1989	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figura Nº 30: Estructura i exemple de la matriu tf-idf utilitzada per modelitzar.

3.4 Models no supervisats: algorismes d'agrupament.

Modelitzar les dades aplicant aprenentatge no supervisat és l'objectiu fixat. Un cop analitzades les dades i representades numèricament, s'ha procedit a seleccionar diferents modalitats d'algorismes de clustering per crear els respectius models i comparar els resultats. En aquest apartat es presenta per diverses alternatives, la creació del model i optimització de paràmetres, execució, avaluació i visualització dels resultats.

Els algorismes seleccionats han estat: KMeans, DBSCAN, i d'entre els possibles algorismes jeràrquics el de tipus aglomeratiu, en les seves diverses modalitats d'enllaç. Pel que fa als motius de la selecció, l'algorisme KMeans és molt versàtil, i un clàssic en aprenentatge no supervisat i en la comunitat de científics de dades, ha estat seleccionat com un dels més rellevants en aquest tipus d'aprenentatge i en machine learning en general. L'algorisme DBSCAN, s'ha escollit per què ofereix prestacions que es complementen bé amb KMeans: com no necessitar indicar un nombre de clústers inicial i oferir bons resultats quan els clústers en forma, no són convexos, o isotròpics, o presenten formes irregulars que l'algorisme KMeans podria agrupar incorrectament.

Els algorismes jeràrquics ens aporten informació afegida com és la generació d'una estructura jeràrquica paral·lela a la definició de clústers i permeten una anàlisi visual molt intuïtiu. En contra, no tenen un criteri de finalització establert, i cal definir, algun criteri que detecti quan no s'estan generant més grups, o bé finalitzar el procés amb una profunditat o nombre de clústers desitjat.

El tipus d'avaluació aplicat, s'ha basat en analitzar la qualitat dels clústers finals obtinguts, seguint els criteris ideals de forta compactació interior i major distància o separació entre clúster diferents.

Seguint la filosofia CRISP-DM, analitzarem el dataset generat per modelitzar i retro alimentarem l'anàlisi (amb tècniques d'optimització noves) en aquelles febleses del primer anàlisi, per tal d'assolir els objectius finals.

3.4.1 Temptativa inicial d'agrupament amb KMeans i DBSCAN.

En aquesta etapa d'anàlisi, l'objectiu inicial ha estat definir un procediment estable per testejar els diferents algorismes d'aprenentatge no supervisat. Inicialment se'ns van presentar diferents dubtes de com actuar:

1. Utilitzar tot el dataset raw o fer una selecció del dataset?
2. Quina mesura de similitud utilitzar?
3. Com definir el nombre òptim de clústers?
4. Com valorar els resultats obtinguts sense conèixer quina és l'agrupació ideal o si existeix més d'una solució.
5. Intuïm realment que les dades són agrupables?

Per afrontar aquests reptes s'han pres les següents decisions:

1. Fer una prova inicial amb tot el dataset i valorar com procedir.
2. Utilitzar distància euclidiana que és la mesura natural dels algorismes i comparar-la amb la mesura de similitud del cosinus que és la mesura que podria adaptar-se millor al fet d'utilitzar la vectorització, com a tècnica per representar les dades text dels tuits en un espai n-dimensional.
3. Utilitzar diversos criteris com el test del colze i coeficient silueta, per detectar el nombre de clústers òptim pel cas de KMeans.

4. La utilització de la distància de cada punt als n punts més propers en el cas de DBSCAN i l'anàlisi visual del dendrograma pel cas dels algorismes jeràrquics.
5. Analitzar la qualitat dels clústers finals obtinguts, seguint els criteris ideals de forta compactació interior i major distància o separació entre clúster diferents com s'ha comentat anteriorment.
6. A més a més d'avaluar visualment el resultat de les proves amb aquesta primera temptativa mirar la separabilitat de temàtiques.

A continuació es mostra els resultats d'aquesta primera temptativa sobre el conjunt complet de tuits resultat de l'etapa de preprocessament utilitzant l'algorisme Kmeans.

L'algorisme K-Means pertany al conjunt d'algorismes, que requereix el **nombre de clústers** o **grups** en què volem assignar finalment els tuits del dataset. Per esbrinar-ho s'ha aplicat l'anomenada 'regla del colze', basada en cercar per quin nombre de clústers 'k', s'aconsegueix minimitzar de distàncies intragrup i/o maximització de distàncies intergrup (Roig et al., 2017).

En la figura 31, es mostren dues proves realitzades amb la regla del colze, per avaluar el nombre de clústers òptim per aplicar KMeans. La primera, sobre un rang petit de clústers [1..20], i la segona sobre una selecció de valors manual [10, 50, 500, 1000]. En cada gràfica l'eix de les abscisses mostra el valor de k candidat i en l'eix de les ordenades, la mètrica SSE (en anglès Sum Squared Error). El que busquem és un valor de k a partir del qual que ja no millori o es redueixi el valor de SEE o error.

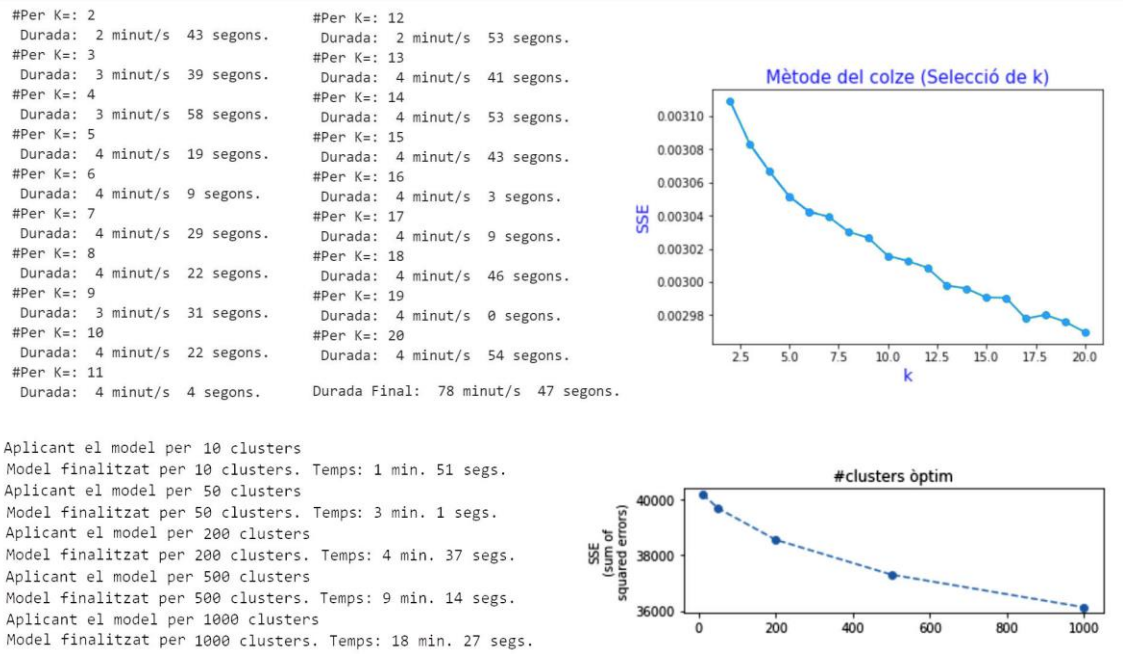


Figura Nº 31: Procés de cerca d'un valor pel # de clústers 'k' òptim.

El resultat de les proves no ens ha resolt completament els dubtes. No obtenim cap 'colze' on el valor de 'k' no faci millorar significativament el valor de la mètrica, Per això seleccionem un k=200 com un valor que pot estar bé inicialment.

Un cop executem l'algorisme de KMeans sobre la matriu *tf-idf*, que representa numèricament el conjunt de tuits disponibles com s'ha demostrat anteriorment, obtenim l'assignació de cada tuit al seu clúster corresponent, per defecte kmeans a la llibreria sklearn utilitza distància euclidiana com a mètrica de comparació.

Per visualitzar els resultats, s'ha aplicat una reducció de dimensionalitat per PCA, (en anglès Principal Component Analysis) sobre la matriu *tf-idf* de 14303 columnes. El paràmetre de nombre de components que cal indicar en l'execució de PCA l'hem establert a 2, per després visualitzar els resultats del clustering en una gràfica de dues dimensions. En la figura 32, es mostra el codi usat i es pot observar els resultats obtinguts.

```
# Apliquem reducció de dimensionalitat
# per millorar els resultats del clustering.

time_start = time.time()
X = matriu_tfidf
# reducció de la dimensionalitat amb PCA:
#pca = PCA(n_components=2)
n_comp=2
print("\nCreant el model i matriu PCA - Nº Components=",n_comp)
pca=delayedsparse.PCA(n_components=n_comp)
X_r_PCA = pca.fit(X).transform(X)
print("Dimensions de les dades reduïdes amb PCA:", np.shape(X_r_PCA))

temps=(time.time()-time_start)/60
print("#Reducció de dimensionalitat (PCA):",np.shape(X_r_PCA),"\n Durada: ",int(temps) if temps>0 else 0,"minut/s ", \
      int((temps-int(temps))*60),"segons.")
```

```
Creant el model i matriu PCA - Nº Components= 2
Dimensions de les dades reduïdes amb PCA: (98433, 2)
#Reducció de dimensionalitat (PCA): (98433, 2)
 Durada:  0 minut/s  15 segons.
```

```
time_start = time.time()
fig, ax = plt.subplots(1, 1, figsize=(8, 6))
# Apliquem KMeans per k=200 sobr la matriu reduïda amb PCA.
kmeans = KMeans(n_clusters=200)
dists = kmeans.fit_transform(X_r_PCA)
# Visualitzem en 2D.
ax.scatter(np.array(X_r_PCA[:,0]), np.array(X_r_PCA[:,1]), c=kmeans.labels_, s=1, cmap='viridis_r')
ax.set_title('Clusters- PCA dim={}'.format(n_comp))
plt.tight_layout()
plt.savefig("kmeans_k200.jpg",format='jpg',bbox_inches='tight')

temps=(time.time()-time_start)/60
print("Visualització de KMeans:",np.shape(X_r_PCA),"\n Durada: ",int(temps) if temps>0 else 0,"minut/s ", \
      int((temps-int(temps))*60),"segons.")
```

```
Visualització de KMeans: (98433, 2)
 Durada:  1 minut/s  30 segons.
```

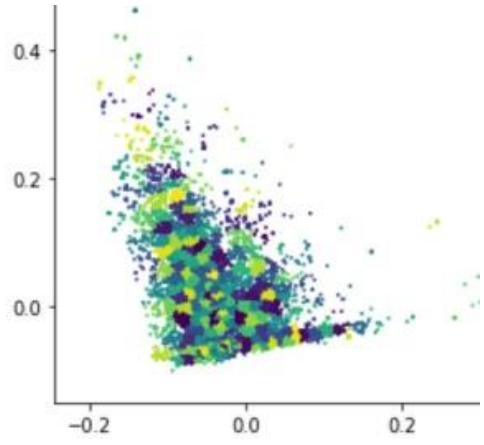


Figura Nº 32: Visualització de l'agrupament de tuits utilitzant l'algorisme KMeans per k=200.

La distribució del nombre de tuits per cada clúster es pot observar en l'histograma de la figura 33.

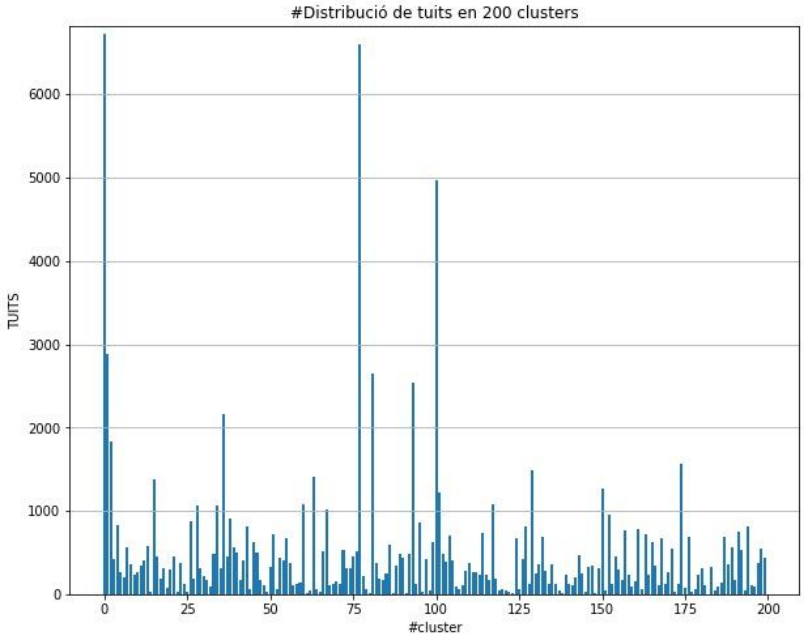


Figura Nº 33: Distribució dels tuits en els clústers obtinguts pel model KMeans.

Un cop detectats els clústers per k=200, ens queda visualitzar el seu contingut text per estudiar quines paraules els componen i per quines temàtiques queden definits, tal com es mostra en la figura 34 (per una mostra de tuits).

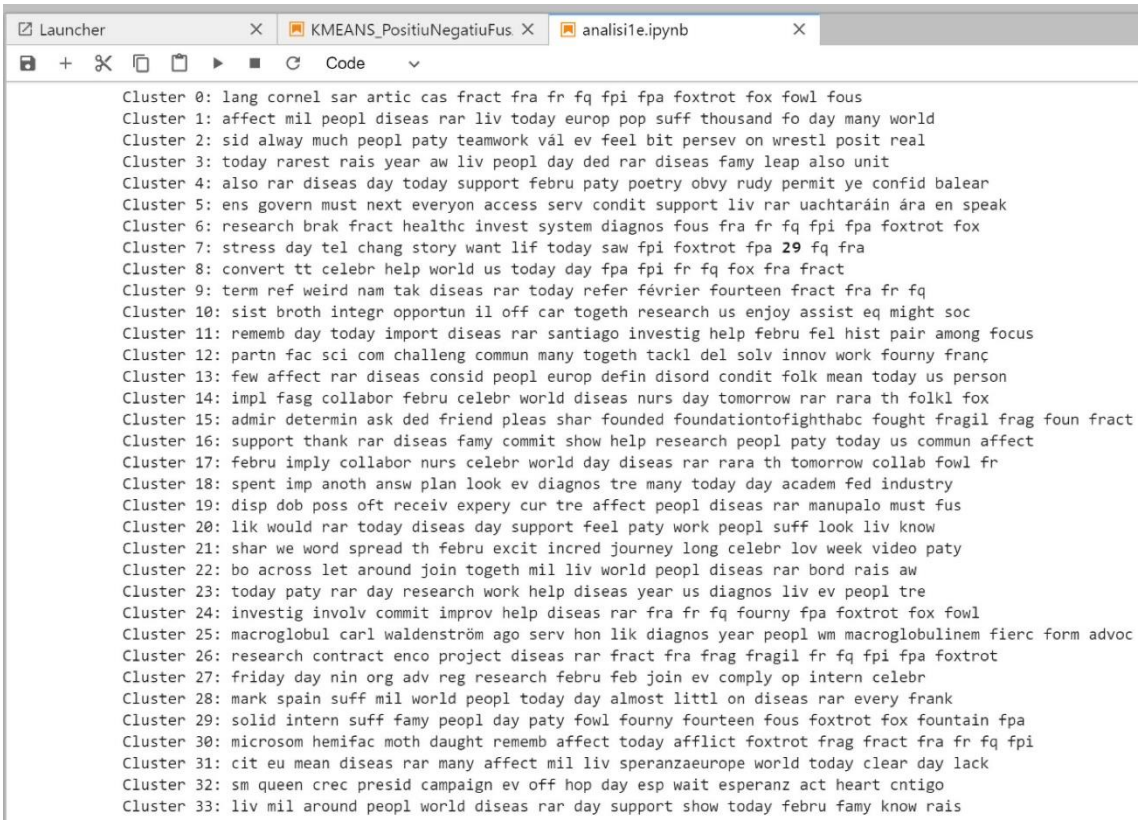


Figura Nº 34: Llistat de clústers resultat d'aplicar l'algorisme de KMeans.

En el cas de DBSCAN no especifiquem el nombre de clústers, però cal indicar el radi de veïnat que volem aplicar (paràmetre **eps**), i el nombre de veïns a considerar en aquest radi (paràmetre **min_samples**). Per aplicar l'algorisme amb

els paràmetres òptims s'ha calculat i representat les distàncies entre tuits per ajudar a identificar el valor òptim de 'eps'. A partir d'aquest valor de eps, empíricament s'ha escollit el valor de 'min_samples'. En la figura 35, observem el càlcul de distàncies per obtenir un valor òptim del paràmetre 'eps'.

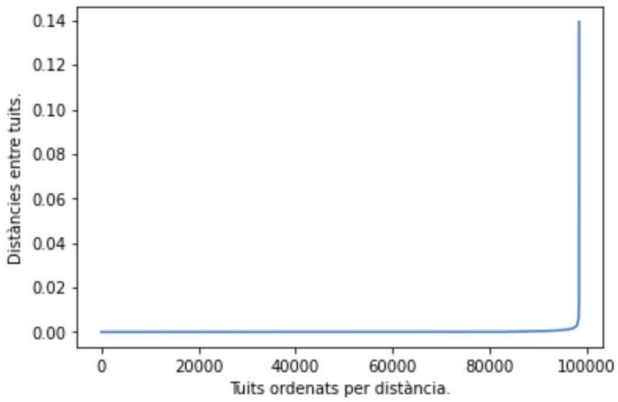


Figura Nº 35: Distàncies entre tuits mitjançant l'algorisme k-NN.

Com es pot observar, existeix un grup nombrós de tuits a molt poca distància i altres grups amb variabilitat de distàncies, perquè la gràfica és estable fins més enllà dels 90.000 tuits a distàncies entre 0 i 0.001 i després s'incrementa a un valor de 0.14. Això implica considerar diferents configuracions de paràmetres per obtenir bons resultats. En la figura 36 es pot observar per tenir una representació de referència dels resultats amb DBSCAN com es poden agrupar els tuits amb una configuració amb eps=0.01 i un valor de min_samples=300. En aquest exemple el nombre de punts considerats com soroll per l'algorisme és alt.

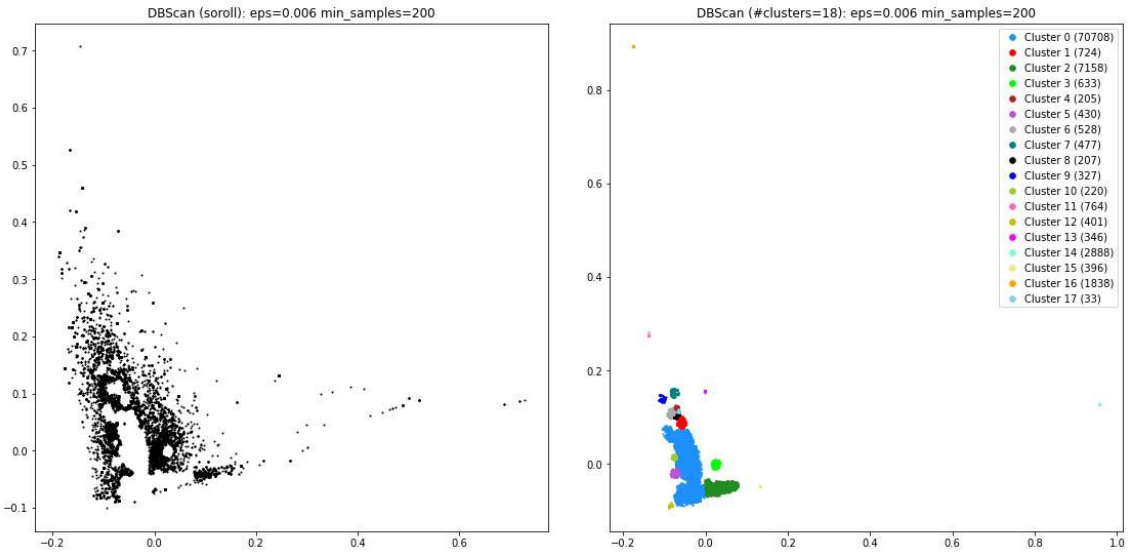


Figura Nº 36: Agrupament per DBSCAN pel dataset de modelització.

En la figura 36 es pot observar com tenim altre cop un clúster amb un major nombre de tuits i amb densitat de tuits diferent. El paràmetre min_samples s'ha ajustat per obtenir un nombre de clústers (18) més identificable que k=200. En la figura 37 es mostren els continguts de cada clúster de tuits.

```
Cluster 0: rare diseases people disease day today world february know patients support million awareness thank tomorrow
Cluster 1: day world today atlas produced mortality interactive wonder argentinians cu scientists every raise february find
Cluster 2: us research rare together people join bo disease let across million care help around diseases
Cluster 3: help donation bpan small could save morgan cure us please life today research visited germany
Cluster 4: day today world initiation adding joined celebration professionals crecer february geographic loud hope statement institutional
Cluster 5: rare refers diseases term weird name take today falling pr strong show time many worldwide
Cluster 6: day today particular people world investiing visibility rare give february need support cancers important celebrated
Cluster 7: day neu sickness nursing today called would little one tell like every calendar years february
Cluster 8: day ro collaborative associations global february important today pair diseases last rare importance bed remember
Cluster 9: marks suffer spain million world day today people rare february year th disease fitting diseases
Cluster 10: disease rare large asking community nord stripes day show support world people diseases million affect
Cluster 11: today rarest year raise living awareness people day catalonia world disease rare federac feder_ong feet
Cluster 12: citizens eu means rare affected many million living disease diseases strong proud show people support
Cluster 13: convert tt helping celebrate us world today day minority must work together february disease felices
Cluster 14: integrate brothers sisters offers ill opportunity care together research us felices felipe feeling feliz felix
Cluster 15: research brake fractionated investment system healthcare feits fees feet fegerec fei feil 29 felices feels
Cluster 16: day stress changed tell story want life today feeling feelings feels fees feet fegerec fei
Cluster 17: dif day cases today pathologies world chosen held considered february celebrated kerstin year objective recognize
```

Figura Nº 37: Contingut dels clústers per l'agrupament de l'algorisme DBSCAN.

En el cas dels algorismes jeràrquics l'execució del càlcul del dendrograma ja ens ha plantejat problemes de memòria intentant representar els dendrogrames de 98.433 elements de partida.

3.5 Millores i procediment final d'anàlisi.

Aquesta primera temptativa ens a permès observar que era necessari aplicar una sèrie de millores per tal d'afinar més els resultats. La quantitat de dades i sobretot el soroll existent per molts tuits, amb poc text o paraules poc significatives, tant correctes com incorrectes sintàcticament o bé moltes paraules amb baixa freqüència d'aparició. En el gràfic final obtingut, observem que les dades realment són potencialment agrupables, però cal intentar cercar les comunitats i temàtiques **més importants** i aconseguir un **nombre de clústers més reduït** amb el risc de perdre alguna temàtica minoritària pel camí. També observem punts, que representen comunitats apartades del nucli central. Alguns d'aquests valors extrems o atípics, els hem eliminat a l'establir millores sobre el dataset de modelització i altres els hem conservat per ser correctes i ho hem tingut en compte. Amb l'experiència d'aquesta temptativa s'han establert les següents consideracions o millores:

1. Sense reduir el nombre de tuits a valorar, minimitzar les instancies o files de la matriu de vectorització tf-idf.
2. Millorar el vocabulari quant a nombre de paraules i qualitat d'aquestes.
3. Intentar sintetitzar les temàtiques i reduir el soroll de tuits curts amb molta informació de context com URL's, hashtags i referències a usuari.

Per la primera millora, s'han fusionat tots els textos de tuits de cada usuari en un únic text. Per la segona, s'han configurat els valors de l'objecte de vectorització per tf-idf **min_df** (nombre mínim d'aparicions requerit) (**max_df** (nombre màxim d'aparicions no excedible) no s'ha modificat). En el nostre cas s'ha establert el nombre mínim d'aparicions d'una paraula en el contingut de text global de tots els tuits de cada usuari, en l'1%. A més a més s'han filtrat paraules que tot i ser correctes aportaven poca informació per ser globals a tots als tuits de totes les comunitats i de totes les temàtiques com per exemple 'febrer', 'rare', o 'disease'. Per la tercera hem exclòs els retweets, que s'ha comprovat que un cop netejats de hashtags, url's i referències d'usuari aporten poc text i repeteixen les temàtiques a detectar.

En la figura 38 es detalla l'organigrama que s'ha definit per mostrar l'estratègia a seguir finalment per realitzar l'anàlisi d'aquest document un cop efectuada aquest primer tanteig d'agrupament.

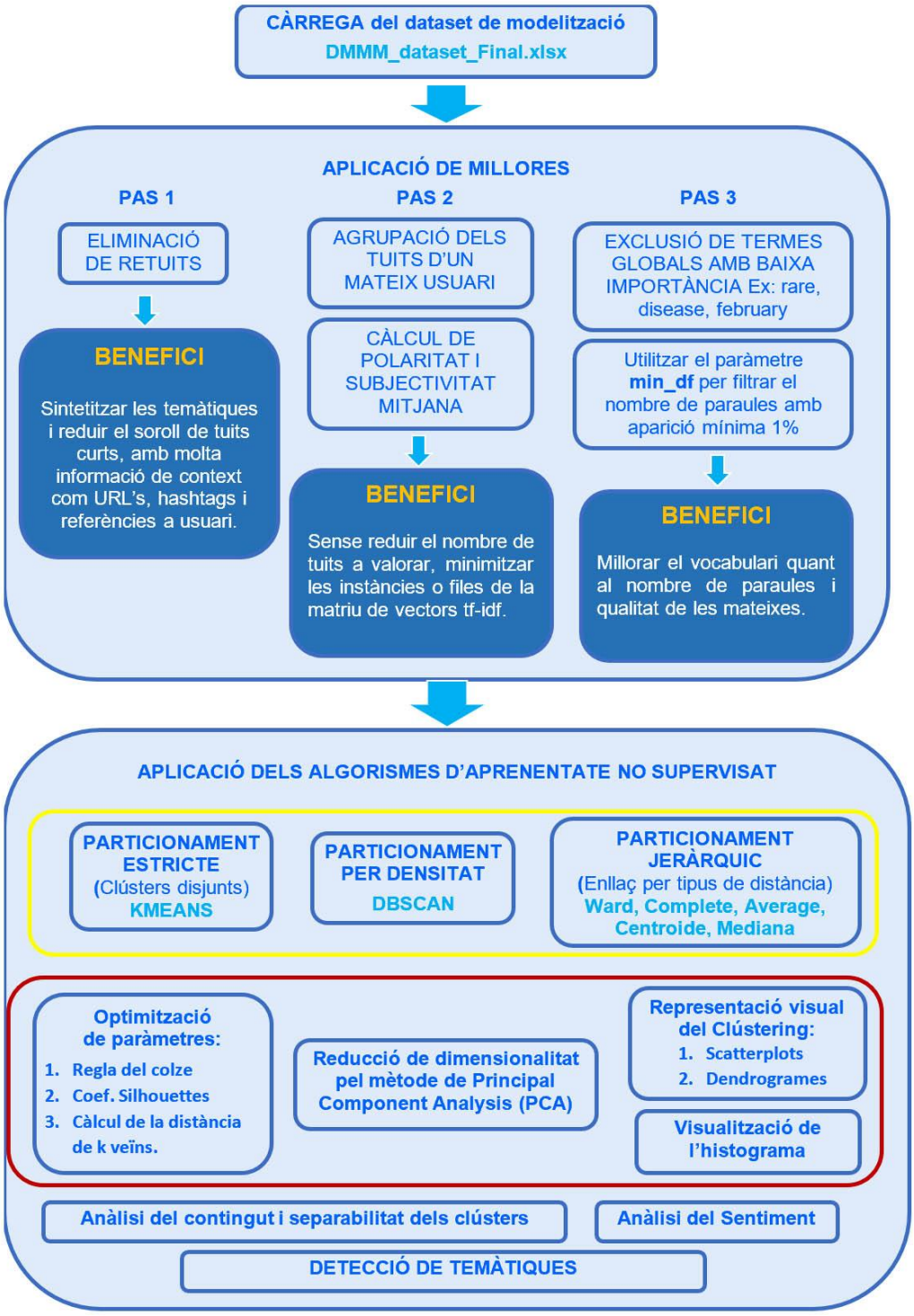


Figura Nº 38: Organigrama del procediment seguit en l'anàlisi de mètodes no supervisats.

3.5.1 Llibreries utilitzades en l'anàlisi.

En la figura 39, es mostra un llistat descriptiu de les llibreries Python usades.

```
import numpy as np
# NumPy és un paquet de processament de matrius de propòsit general.
# Proporciona un objecte de matriu multidimensional d'alt rendiment i eines per treballar amb aquestes matrius.
# És el paquet fonamental per a la computació científica amb Python.

import pandas as pd
# Pandas és L'eina per treballar amb dades tabulars: dades emmagatzemades en fulls de càlcul o bases de dades.
# Permet explorar, netejar i processar dades tabulars usant l'objecte DataFrame.
# Operations Seleccionar, filtrar per files o columnes o per una condition i exportar les dades o visualitzar-les.

from sklearn.feature_extraction.text import TfidfVectorizer
# Sklearn és un paquet/llibreria indicat per aplicar Aprenentatge Automàtic.
# Per l'extracció de característiques s'utilitza l'objecte
# TfidfVectorizer per representar numericament mitjançant vectors dades texte.

from sklearn.cluster import KMeans, DBSCAN
# Per executar les implementacions dels algorismes de KMeans i DBSCAN
# utilitzem el mòdul 'cluster' de la llibreria Sklearn.

from sklearn.neighbors import NearestNeighbors
# Implementació de l'algorisme K-Nearest Neighbors
# per l'òptimització del paràmetre eps en l'execució
# de l'algorisme DBSCAN.

from sklearn.metrics.pairwise import cosine_similarity
# Implentació de la mètrica de similitud de vectors.
# S'ha usat per comparar aquesta mètrica de vectors
# amb la mètrica de distància euclidiana.

from sklearn.cluster import AgglomerativeClustering
# Implementació de l'algorisme jeràrquic de tipus Down Up
# o aglomeratiu. Finalment s'han comparat els resultats amb
# els resultats calculats amb la llibreria SCIPY.

import matplotlib.pyplot as plt
# matplotlib.pyplot permet la visualització gràfica, controlant tots
# els aspectes d'una gràfica des de la plantilla, títols, eixos,
# representació de les dades, reixeta, etiquetes etc....

from matplotlib.ticker import FormatStrFormatter
import matplotlib.ticker as ticker
# Llibreries usades com a complement de la llibreria matplotlib.pyplot.
# Tractament específic dels eixos en la definició i
# visualització de l'histograma resultat d'un agrupament.

import time
# Utilitzat per el càlcul de durada dels processos.

import delayedsparse
# Implementació eficient de matrius disperses per a diverses
# anàlisis de components principals PCA. En concret ho apliquem
# per aplicar PCA a una matriu dispersa resultat de vectoritzar
# amb l'objecte TfidfVectorizer un volum de dades gran.

import scipy.sparse
# Llibreria per la manipulació de matrius disperses
# SciPy 2-D per a dades numèriques.

import re, collections
# Llibreries
# 're': usada per la definició, manipulació i tractament de
# text mitjançant 'expressions regulars'.
# 'collections': usada per el conteig de paraules en textos.
# i per la implementació del 'Ba of Words' (BoW)

from textblob import TextBlob, Word
# Llibreria de soport de l'anàlisi de sentiment
# en paraules i frases, en concret s'ha aplicat al
# càlcul de la polaritat i subjectivitat en textos.

from nltk.tokenize import TweetTokenizer
from nltk.stem import LancasterStemmer, PorterStemmer
from nltk.corpus import stopwords
from nltk.probability import FreqDist
# NLTK: Llibreria formada per un conjunt de mòduls per el
# tractament i manipulació de textos i tractament del llemguatge
# natural en general. En concret s'han usat en la fase de preprocessat.
# TweetTokenizer: Usat per obtenir els elements representatius o tokens.
#                 en concret aquest mòdul és específic per textos de Twitter.
# LancasterStemmer, PorterStemmer: Mòduls valorats per l'operació de
# steaming on s'obté el mot arrel i s'en descarten els derivats d'ell.
# Stopwords: Eliminació de les paraules freqüents por significatives
# sovint usades en el llenguatge per l'unió de frases o de complement
# a substantius i verbs.
# FreqDist: S'ha usat per el càlcul ràpid del conteig
# de paraules o 'Bagg of words' (Bow).

from scipy.cluster.hierarchy import dendrogram, linkage, single, complete, ward
import scipy.cluster.hierarchy as hc
import scipy.spatial.distance as metrica
# Conjunt de llibreries per el càlcul dels diferents
# tipus d'enllaç en L'aplicació del algorisme jeràrquic aglomeratiu
# i les representacions mitjançant un dendrograma associades.

import fastcluster
# Alternativa a slearn al càlcul dels algorismes
# d'agrupament i la seva representació.

from itertools import cycle, islice
# Llibreries optimitzades per la implementació
# d'iteracions eficients.
```

Figura Nº 39: Descripció i ús de les llibreries Python utilitzades

A continuació s'exposen per cada cas d'algorisme les proves efectuades.

3.5.2 Modelització amb l'algorisme KMeans.

Un cop realitzades, la càrrega del dataset i les transformacions indicades anteriorment obtenim un dataset de modelització amb tan sols els camps necessaris per modelitzar i on s'han aplicat totes les millores indicades. Cada instància és un tuit com anteriorment, però ara, també representa a cada usuari i els clústers cercats representaran un cop fet l'agrupament, les comunitats d'usuaris.

El primer pas ha estat aplicar la reducció de dimensionalitat utilitzant el mètode Principal Component Analysis (PCA) per un nombre de components igual a 2. Amb només dues components, es pot visualitzar els resultats en dues dimensions amb comoditat. Tot i que la reducció pot implicar la pèrdua de representativitat. També cal considerar que PCA és un mètode sensible als valors atípics. Com s'havia mostrat en la primera temptativa, cal cercar el valor del 'k' òptim, essent k el nombre de clústers que volem com a resultat del clustering. En la figura 40, es mostren diversos índexs per aquest càlcul, els coeficients de silueta (en anglès Silhouette), Calinski-Carabasz, Davies Bouldin i dos més recents Xu i ZCF (Palacio-Niño & Berzal, 2019). Volem reduir el nombre de clústers i en aquest cas l'anàlisi, ens valida que un bon valor podria estar entre 10 i 18 clústers.

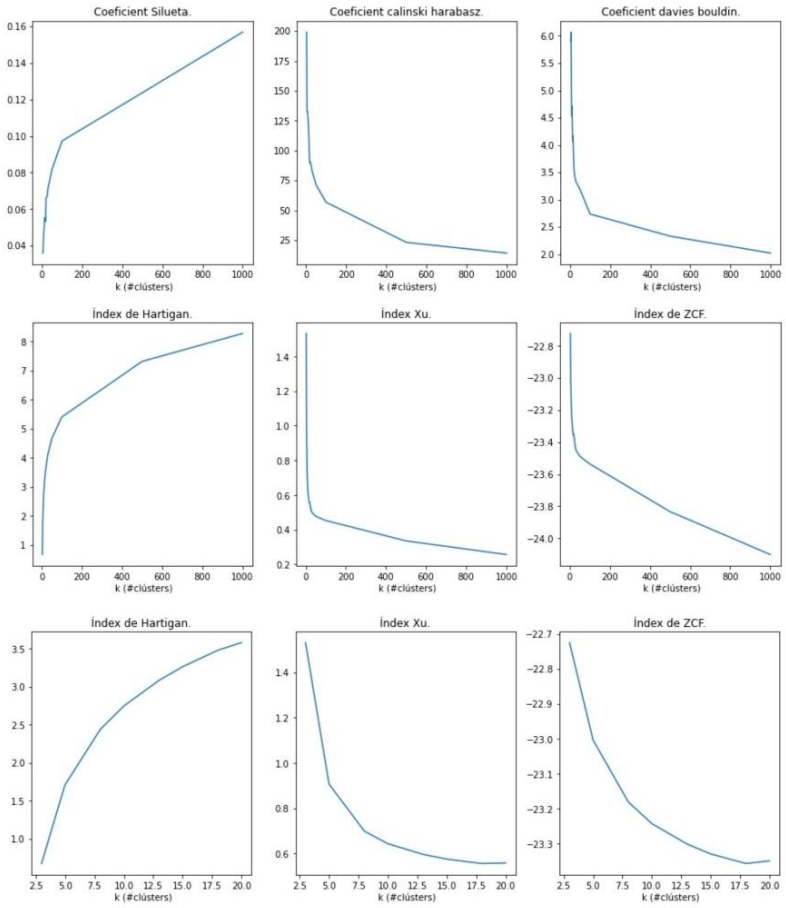


Figura Nº 40: Comparació d'índexs: avaluació del paràmetre òptim pel #clústers amb KMeans.

El coeficient de silueta, quantifica qualitat de l'assignació que s'ha fet a una observació. Es compara la seva similitud amb la resta d'observacions del seu clúster enfront de les dels altres clústers. Pren valors en un rang entre -1 i 1, sent valors alts un indicatiu que l'observació s'ha assignat al clúster correcte. El coeficient de Calinski Harabasz també indica el nombre de clústers òptim en valors alts i el coeficient de Davies Boudin per valors baixos. Els índexs de Hartigan, Xu i ZCF es mostren en les dues darreres fileres. En la primera s'usa un ampli rang de valors fins a 1000 i segona amb valors de detall, sobre els rangs que ens volem centrar entre k=3 i k=20.

Per efectuar l'anàlisi visual entre els diferents valors de k possibles s'han definit tres scripts. El primer aplica el clustering per KMeans, el segon utilitza la reducció de dimensionalitat per PCA per visualitzar el resultat i el tercer complementa la informació gràfica amb l'histograma associat al clustering.
En la figura següent mostrem els tres scripts en l'aplicació per k=10 a manera d'exemple essent el mateix procediment per la resta de valors de k visualitzats posteriorment.

```
# CLUSTERING per KMeans

k_optim = 10
seed=88
X_Norm = X_PCA
time_start = time.time()
print('Aplicant el model per {} clusters'.format(k_optim))
km_model_10 = KMeans(n_clusters=k_optim, n_jobs=-1, random_state=seed)
km_model_10.fit(X_Norm)
temps=(time.time()-time_start)/60
print('Model finalitzat per {} clusters. Temps: {} min. {} segs.'. \
                format(k_optim, \
                       int(temps) if temps>0 else 0, \
                       int((temps-int(temps))*60)))
```

```
Aplicant el model per 10 clusters
Model finalitzat per 10 clusters. Temps: 0 min. 2 segs.
```

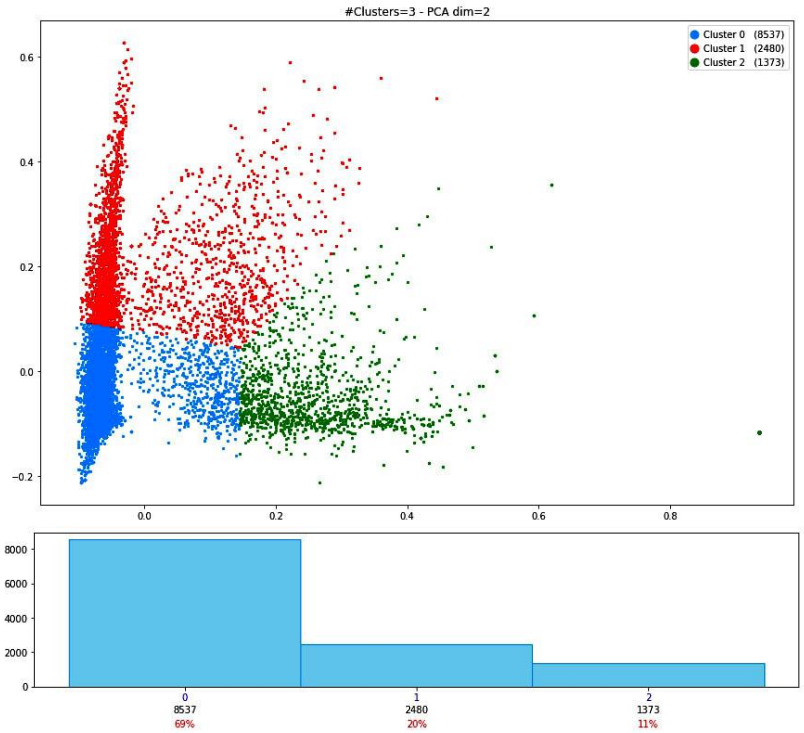
```
# Model en execució actual
model=km_model_10
# Definició del conjunt de colors o paleta de colors per la visualització.
colors = np.array(list(islice(cycle(['dodgerblue', 'red', 'forestgreen',
                                     'lime', 'brown', 'mediumorchid',
                                     'darkgrey', 'teal', 'black',
                                     'blue','yellowgreen','hotpink','y',
                                     'magenta','aquamarine','khaki',
                                     'orange','skyblue','yellow']),
                                      int(len(set(model.labels_))))))
```

```
# Script de Visualització mitjançant scatter plot
# del model per K=10
num_clusters=k_optim
model=km_model_10
elements=np.bincount(model.labels_)
fig, ax = plt.subplots(1, 1, figsize=(12, 8))
cmap = plt.cm.Spectral
norm = plt.Normalize(vmin=0, vmax=num_clusters-1)
for idcluster in range(0,num_clusters):
    ax.scatter(np.array(X_PCA[:,0][model.labels_==idcluster]), np.array(X_PCA[:,1][model.labels_==idcluster]), \
               c=colors[idcluster], norm=norm, cmap=cmap, s=5, label="Cluster "+str(idcluster) + \
               "   ("+str(np.bincount(model.labels_)[idcluster]) + ")")
    ax.set_title('#Clusters={} - PCA dim={}'.format(num_clusters, n_comp))
    ax.legend(handlelength=0.5, markerscale=4)
plt.tight_layout()
```

```
# Representació de l'histograma resultat del CLUSTERING.
plt.figure()
fig, ax = plt.subplots(figsize=(14.5,3))
counts, bins, patches = plt.hist(model.labels_, bins=np.array(range(0,num_clusters+1)), facecolor='skyblue', edgecolor='dodgerblue')
# Configuració pel format dels eixos
ax.xaxis.set_major_formatter(FormatStrFormatter('%0.1f'))
bin_centers = 0.5 * np.diff(bins) + bins[:-1]
# Definició de la posició i etiquetes de l'eix X.
ax.set_xticks(bin_centers)
ax.set_xticklabels(bins, rotation=0,color='b')
# Definició dels valors
for count, x in zip(counts, bin_centers):
    # Etiquetes dels comptadors
    ax.annotate(str(count), xy=(x, 0), xycoords=('data', 'axes fraction'),
        xytext=(0, -20), textcoords='offset points', va='top', ha='center')
    # Etiquetes dels percentatges
    percent = '%0.0f%%' % (100 * float(count) / counts.sum())
    ax.annotate(percent, xy=(x, 0), xycoords=('data', 'axes fraction'),
        xytext=(0, -35), textcoords='offset points', va='top', ha='center',c='r')
plt.show()
print("Nombre de tuits en el clúster més gran: {}".format(np.bincount(model.labels_).max()))
```

Figura Nº 41: Script de càlcul, visualització i histograma per k=10.

Els resultats per k=3 clústers, els podem observar en les figura 42, i es componen de la representació gràfica, l'histograma i la correspondència de comunitats d'usuaris i temàtiques deduïbles de les paraules més significatives contingudes.




```
Cluster 0:
 research thank us patients support one know thanks like many happy tomorrow help life great
Cluster 1:
 people awareness million raise living to support patients one raising wide help affected around know
Cluster 2:
 to one celebrated support know every research us want year make work people celebrate many
```

Figura Nº 42: Resultats de l'execució de l'algorisme KMeans per K=3.

Dels resultats mostrats en la figura 42 observem que per cada clúster obtenim una comunitat d'usuaris i també un conjunt de paraules. Aquestes paraules són les més properes al centroide en cada clúster, i s'han considerat per tant les més importants o representatives, utilitzant-les per extreure la temàtica global.

En la figura 43 es mostra el codi usat per la seva obtenció.

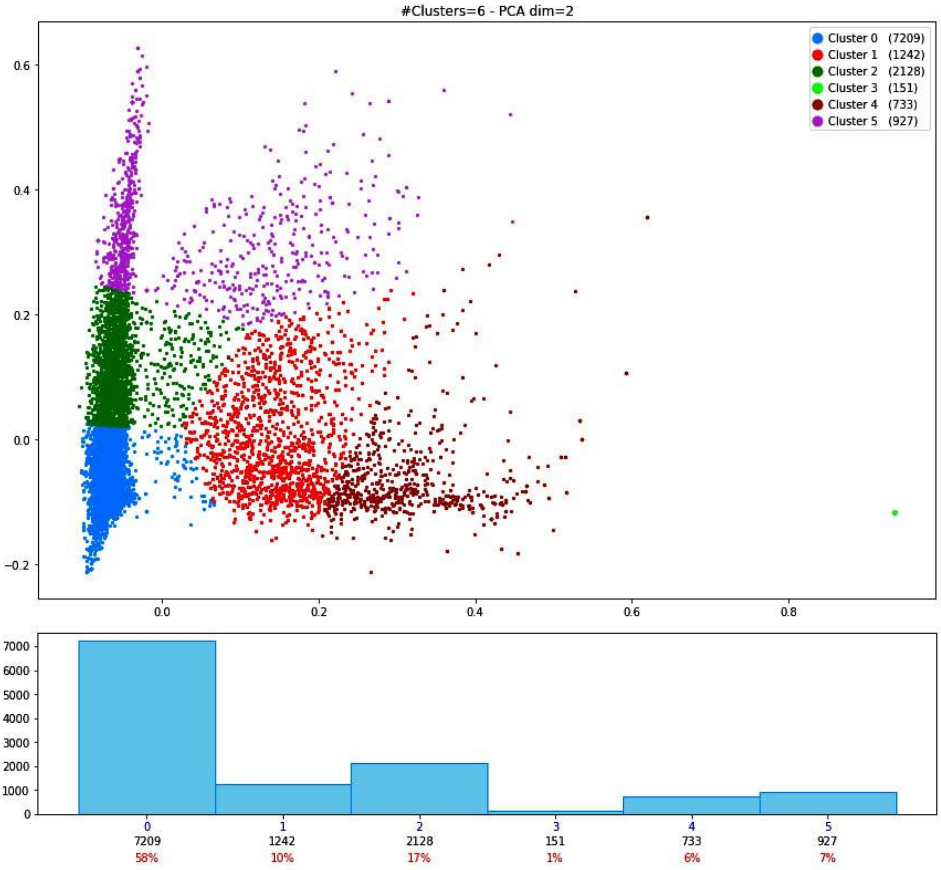
```python
def tokens_mes_propers(model, vectorizer, mat_vect, topk=10):
#    Representació text dels primers-k mots més propers al seu centroide
#    model: model de sklearn escollit.
#    vectorizer: tipus de vectorització (tfidf)
#    topk: nombre de paraules k seleccionades per cluster
nom_model = model.__class__.__name__
paraules = vectorizer.get_feature_names()
if nom_model is 'KMeans':
    grups=np.bincount(model.labels_)
    etq_rellevants = list(set(model.labels_))
    centroides = model.cluster_centers_.argsort()[:,::-1]
    for id_cluster in etq_rellevants:
        matching_rows = np.where(model.labels_ == id_cluster)[0]
        usuaris=[tuits.autor[tuits.index==u] for u in matching_rows]
        print('Cluster {} amb {} usuaris.'.format(id_cluster,grups[id_cluster]))
        for ind in centroides[id_cluster, :topk]:
            print(' {}'.format(paraules[ind]), end='')
        print()
```

Figura Nº 43: Extracció de comunitats d'usuaris i temàtiques amb l'algorisme KMeans.

Aquesta funció utilitza com a paràmetres, el model KMeans calculat, la matriu resultant de la vectorització tf-idf real (sense aplicar la reducció de dimensionalitat per PCA) i un paràmetre per especificar quantes paraules del clúster volem obtenir (pel cas de l'algorisme DBSCAN se creat una de similar).

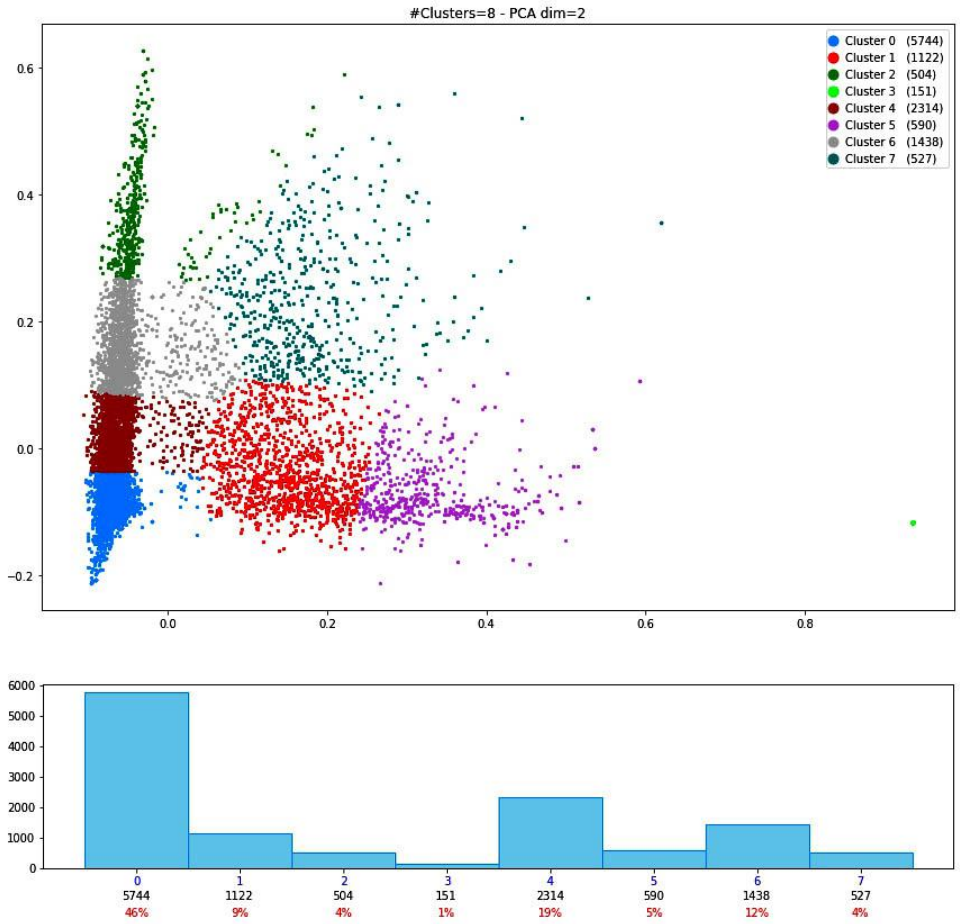
L'anàlisi de comportament dels clústers, s'ha realitzat sobre tots els resultats obtinguts, per poder determinar quina partició ajusta millor les temàtiques.

L'anàlisi visual sobre la representació dels clústers consisteix a identificar cada clúster per un color diferent, i estudiant la seva distribució, forma i la seva magnitud en nombre d'usuaris. Per comparar clústers, s'ha utilitzat l'histograma associat, que disposa en l'eix de les abscisses, de 3 tipus de descripcions: l'identificador de clúster, el nombre d'usuaris i la proporció d'usuaris respecte del total d'usuaris (12.390). A continuació es mostren tots els resultats obtinguts per k=6,8,10,13,15 i al final una valoració global.



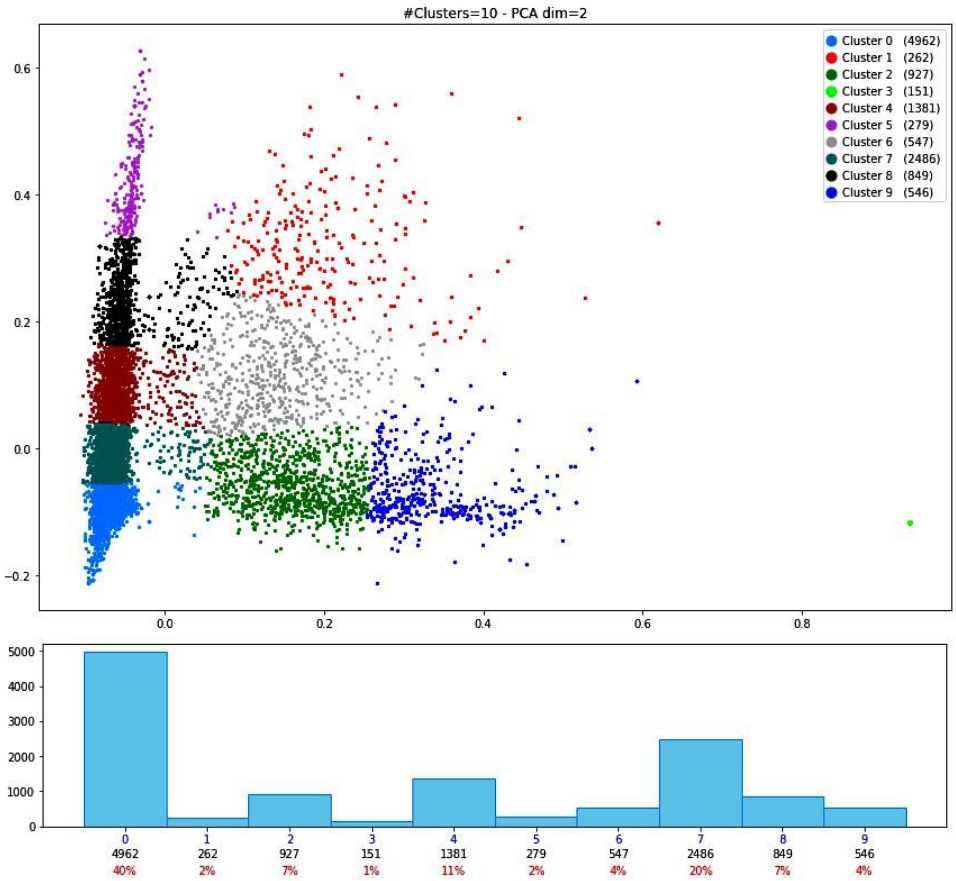

```
Cluster 0:
 thank us one patients support know like many tomorrow happy help great th life every
Cluster 1:
 awareness raise raising to help people patients support impact lives living us families year join
Cluster 2:
 people million living one wide to know around affected support many live suffer affect affects
Cluster 3:
 thanks latest daily news to work thank sharing support great many every much us children
Cluster 4:
 research support to patients treatments care us medical life find help treatment families cure people
Cluster 5:
 to celebrated one support know want us every year celebrate international work many th make
```

Figura Nº 44: Resultats de l'execució de l'algorisme KMeans per K=6.



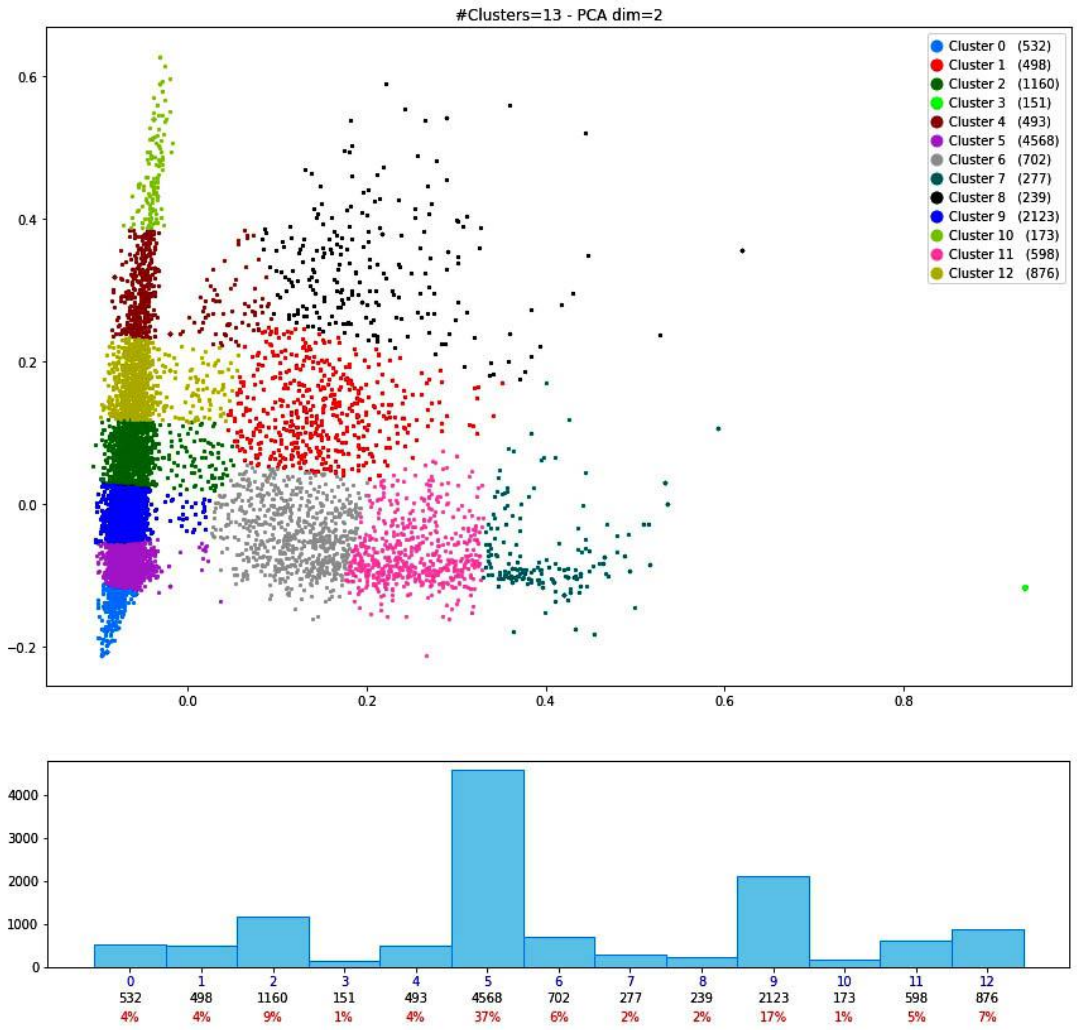

```
Cluster 0:
 thank happy know thanks latest syndrome years like good life th year also love every
Cluster 1:
 awareness raise raising to help people patients impact lives living year families support please around
Cluster 2:
 research support to care treatments medical patients life us find help new treatment cure need
Cluster 3:
 support campaign official event to show people families patients please thank want proud living us
Cluster 4:
 us patients tomorrow many great help new to thank families story work proud like every
Cluster 5:
 to celebrated celebrate us every know work international want make th year learn years also
Cluster 6:
 people million living wide to around affected know live affect suffer many treatment affects genetic
Cluster 7:
 one to people year every know million affects diagnosed many syndrome living us get like
```

Figura Nº 45: Resultats de l'execució de l'algorisme KMeans per K=8.



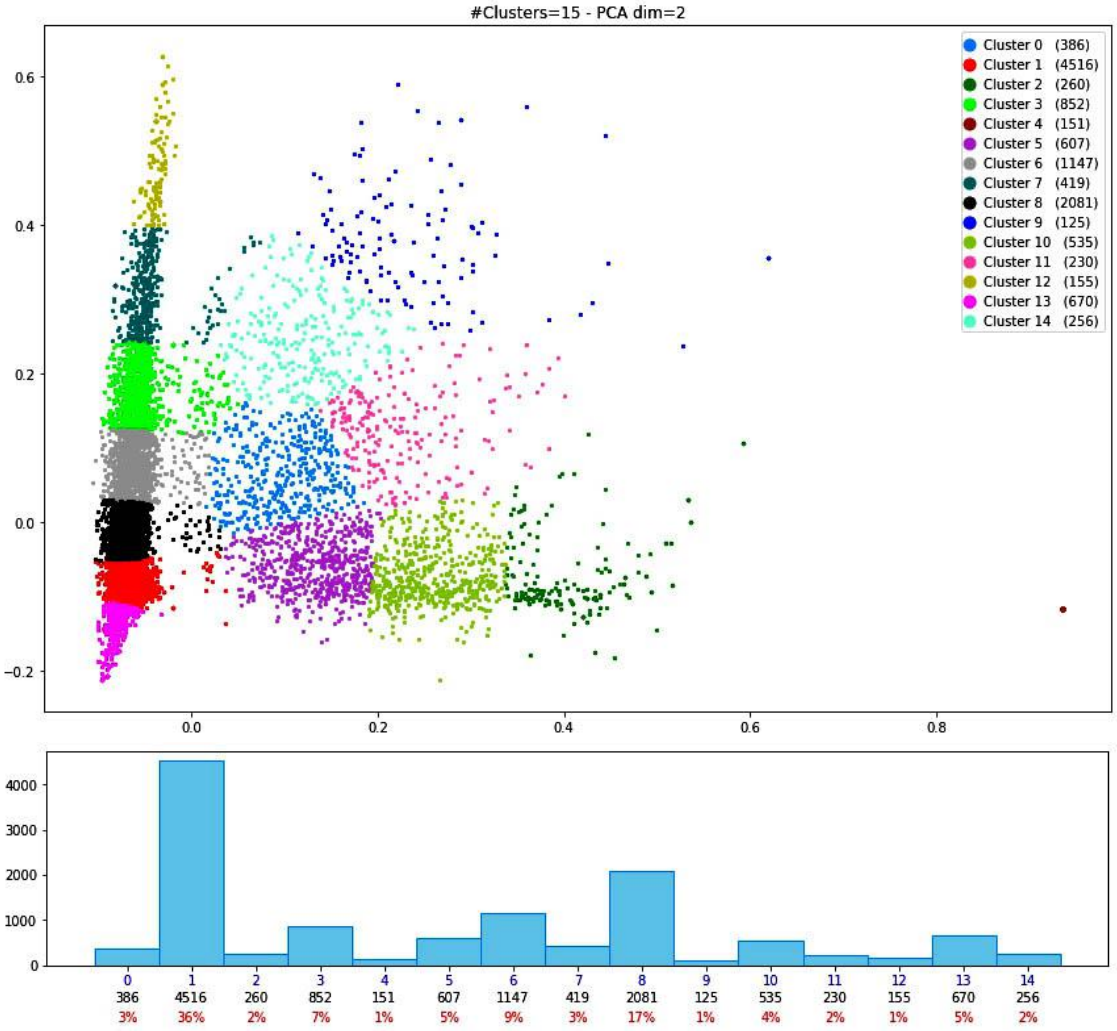

```
Cluster 0:
 thank happy patients many syndrome proud good life love also every work help story year
Cluster 1:
 one to people year every million know affects diagnosed many syndrome us living hope get
Cluster 2:
 know like years to people every life diagnosed many would syndrome diagnosis get still us
Cluster 3:
 latest thanks daily news thank many to special read something research via genetic help syndrome
Cluster 4:
 people million living wide to around affected live affect many suffer treatment know awareness help
Cluster 5:
 research support to care treatments patients medical life us find help treatment families new work
Cluster 6:
 to celebrated celebrate every us make work want year many international learn th people see
Cluster 7:
 us tomorrow great th new health thank patients to help work see join patient year
Cluster 8:
 awareness raise raising to help patients people impact lives living year families join support please
Cluster 9:
 support campaign official to event show families people patients please proud thank want help living
```

Figura Nº 46: Resultats de l'execució de l'algorisme KMeans per K=10.




```
Cluster 0:
 one to people year every know diagnosed million syndrome affects hope get best life us
Cluster 1:
 many strong proud people to one every living live like know million thanks go us
Cluster 2:
 awareness raise raising to help people impact patients lives families year living please support public
Cluster 3:
 celebrated to th last tomorrow year every awareness campaign satur know want hope years since
Cluster 4:
 research support to treatments need new help find hope read treatment patients cure care families
Cluster 5:
 thank thanks latest like syndrome tomorrow years good work th every year event also new
Cluster 6:
 patients families to patient treatment read hope sharing story research supporting work support learn help
Cluster 7:
 support to show people families please want thank suffer patients proud need living help us
Cluster 8:
 great work see thank event to support team thanks awareness people like family us much
Cluster 9:
 us know life to people help thank like please let join get tomorrow learn care
Cluster 10:
 happy us everyone to year also people friends many living celebrate good one awareness amazing
Cluster 11:
 to celebrate make work international years every give year want people th celebrating see learn
Cluster 12:
 people million living wide around to affected suffer affect live one know awareness affects treatment
```

Figura Nº 47: Resultats de l'execució de l'algorisme KMeans per K=13.




```
Cluster 0:
 research support to treatments new find treatment need help hope diagnosis cure care work read
Cluster 1:
 thank thanks latest like syndrome tomorrow years good th event work new every international year
Cluster 2:
 many strong proud people to one every live know living like million thanks th awareness
Cluster 3:
 patients families to help research supporting people patient read support affected treatment hope sharing work
Cluster 4:
 happy everyone us to year people living many good celebrate one awareness amazing friends also
Cluster 5:
 to celebrate work make years give international want th year celebrating every learn see tomorrow
Cluster 6:
 awareness raise raising to help impact people lives patients year public please support families living
Cluster 7:
 one to people year every diagnosed know affects million syndrome get hope best like us
Cluster 8:
 us know life to help thank let join please like care tomorrow get people learn
Cluster 9:
 celebrated to last th tomorrow year every awareness satur hope know campaign since want years
Cluster 10:
 people affected to living know suffer affects disorder every affect less life care diagnosed number
Cluster 11:
 also to one known life research year us many people know every support tomorrow important
Cluster 12:
 great work see thank event to support team thanks awareness people like much us family
Cluster 13:
 million people living wide around to one know live awareness affect affected suffer support treatment
Cluster 14:
 support to show families people please want thank proud suffer patients living need campaign help
```

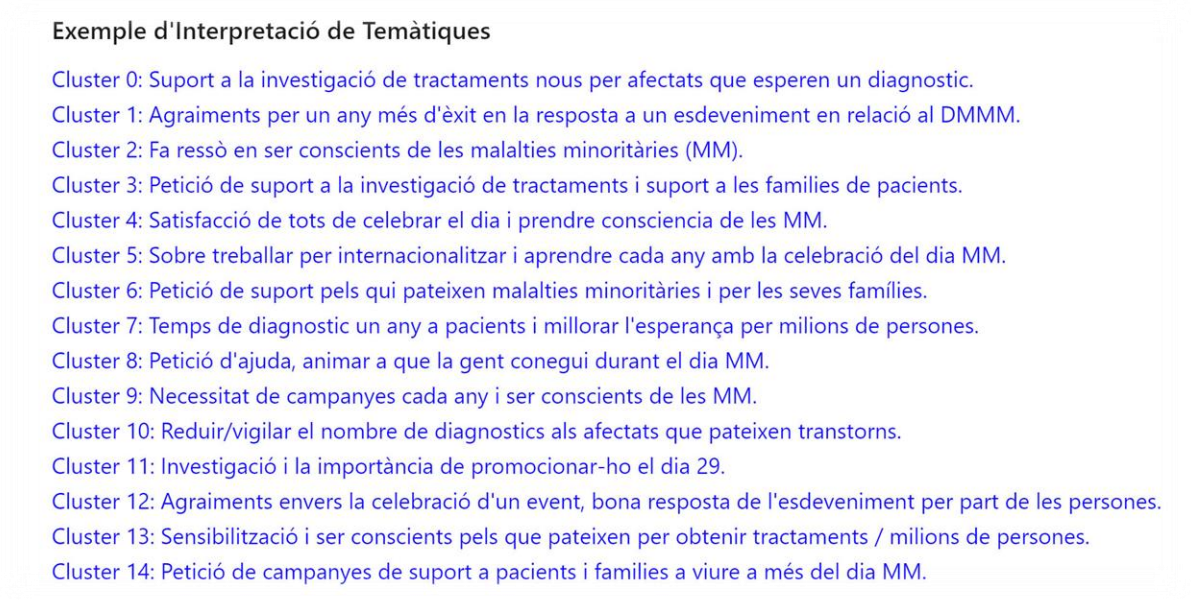
Figura Nº 48: Resultats de l'execució de l'algorisme KMeans per K=15.

L'algorisme KMeans, sabem que té bon rendiment per detectar clústers convexos, isotròpics amb formes esfèriques, i que siguin de mides similars. A més a més també coneixem que pot ser força sensible a l'existència de valors extrems o atípics, (Amat, 2017). Per k=3 no obtenim una bona separació del possible conjunt de clústers existent, però si podem extreure una bona informació dels temes més rellevants de manera concisa, perquè tot i que els clústers tenen molta barreja de temàtiques ens mostren les paraules més freqüents en un nombre reduït de grups. Per tant sobre aquesta base anirem obrint el ventall de variacions. Recerca, celebració i consciència al suport i ajuda de pacients són les tres temàtiques que podríem extreure com a fonamentals. Si analitzem els resultats per k=6 clústers, podem observar que els temes particionats en k=3 es mantenen i detectem en el clúster 4 una nova temàtica pels tractaments relacionats amb la recerca i el suport en aquesta àrea. Per k=8 apareixen noves formes de suport i ajuda a les famílies i pacients, mitjançant campanyes oficials, mentre seguim mantenint els temes anteriors. La temàtica de celebració fa incidència a ser internacional i a ser anual en referència a la celebració del Dia Mundial de les Malalties Minoritàries (endavant DMMM). Per k=10 es mantenen les temàtiques i s'especialitzen alguns temes, com per exemple la recerca (via genètica) i sobre la diagnosi sobre síndromes. També es fa més èmfasi a l'èxit de celebració de campanyes i esdeveniments que es fan. Queda constància de què part del suport als pacients i famílies és per realitzar diagnòstics i trobar nous tractaments. A partir de k=13 adquirim la capacitat d'especificar conceptes, perquè sobre els temes base que queden definits en la partició per k=10, obtenim matisos i les paraules contingudes als clústers en s donen més informació. Per exemple, per k=13 en el clúster 4 obtenim la següent sèrie: 'research suport to treatments need new help find treatment patients...' que és una temàtica que s'intuia en les primeres particions (k=6 o k=8), però aquí estan compactes en el mateix clústers i donen informació específica sobre la temàtica de recerca. Per k=15, en el clúster 0, és el mateix, però inclou informació sobre la recerca per poder diagnosticar. Aspecte que queda més definit en el clúster 10 on apareixen paraules com 'affects' i 'disorder'. Per tant podem comprovar que sobre un volum massiu de tuits podem particionar les temàtiques més importants i anar aclarint aspectes específics o afinant els detalls.

Per contra la dispersió existent en la matriu de vectorització fa que no detectem grups satèl·lit dels grups importants.

També observem que per la naturalesa de l'algorisme, tendeix a tessel·lar l'espai de modelització, i es fan divisions d'una temàtica única quan no hauria de passar. L'exemple el tenim en la regió allargada de la part esquerra que es va fraccionant de manera sistemàtica quan augmentem el nombre de clústers.

A manera d'exemple d'anàlisi i benefici del clustering aplicat, es mostra la següent llista de temàtiques extretes del clustering amb k=15.



Un altre aspecte a destacar és el fet que als clústers identificats i analitzats, no hem obtingut valoracions negatives. És notori que predomina la satisfacció per poder celebrar el DMMM, i contribuir a ajudar als afectats i les seves famílies, i celebrar el fet de seguir-ho fent any rere any. Per analitzar aquest fet, s'ha aplicat una anàlisi de sentiment al dataset de modelització.

3.5.2.1 Anàlisi de sentiment amb KMeans.

Aprofitant que es disposa de la informació de la subjectivitat i la polaritat mitjanes de tots els tuits de cada usuari, s'ha dividit el dataset inicial en dos grups. Un primer grup amb subjectivitat >0.5 i polaritat>0.5 i un segon amb subjectivitat>0.5 i polaritat<-0.5.

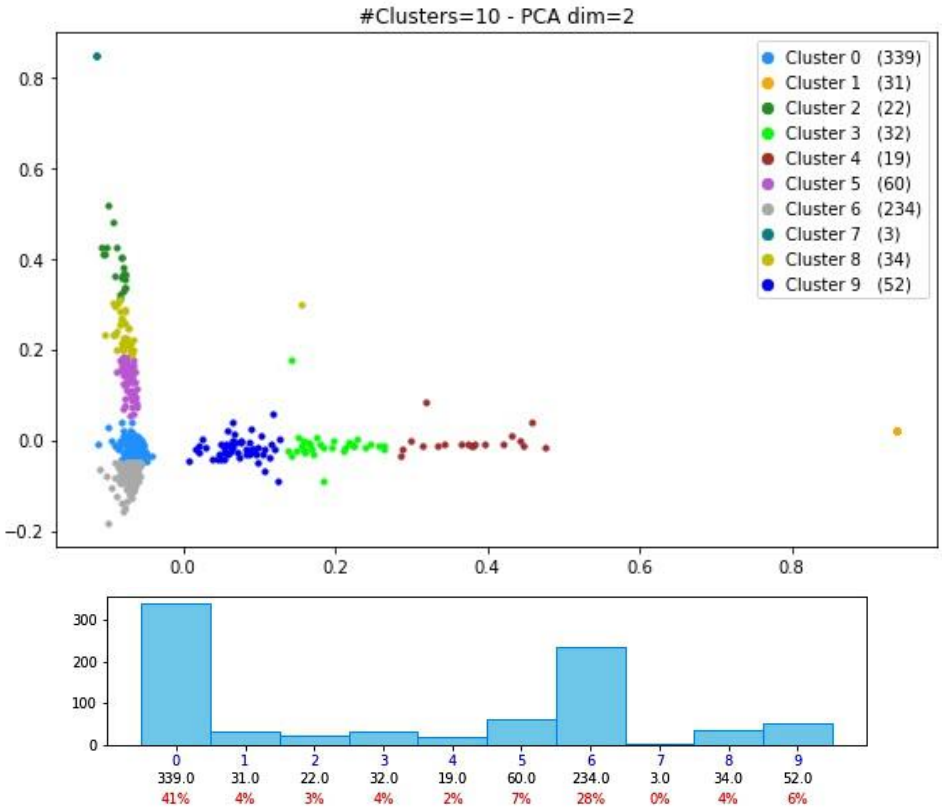
A continuació es mostra en la figura 49 els resultats obtinguts per k=10.

Es pot comprovar com els missatges positius són força diferents pel fet d'haver aplicat un factor de subjectivitat als missatges, per sobre del 0.5 quan el rang de valors possible està entre 0 i 1. Trobem valoracions sobre aspectes de les temàtiques detectades, com per exemple en el clúster 1: 'amazing work inspiring awareness families' o en el clúster 6: 'proud support work many patients strong part community to us join million every team families'.

Els clústers obtinguts estan força agrupats i poden ser detectats sense problema per l'algorisme.

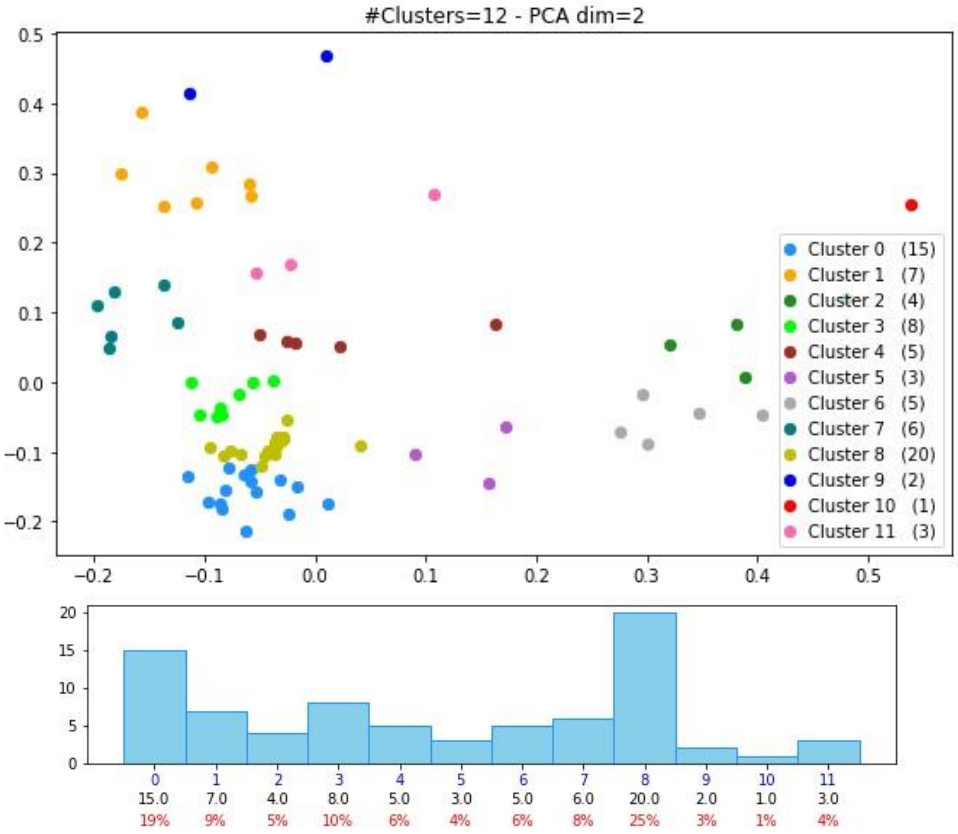
Pel cas d'opinions negatives o en contra que ha motivat l'anàlisi de sentiments, obtenim molta informació nova. Les temàtiques detectades es concentren bàsicament en peticions d'ajuda, descripcions de malalties i descripcions de casos personals. En concret sobre símptomes de la malaltia, diagnòstics erronis, situacions de frustració i solitud o angoixa que deriven en demanar més encert en els diagnòstics. Exemples d'això són la petició d'ajuda per migranyes en el clúster 0, descripció de la malaltia en el clúster 1, denuncia per manca de tractaments a un diagnòstic en el clúster 2, o per un diagnòstic erroni per ús del

medicament ‘cbd’ per tractar un problema com si fos trastorn neurològic (esquizofrènia) en el clúster 4, o estrès, por i patiments insuportables en el clúster 11.




```
Cluster 0:
 happy beautiful to latest important awareness nice excellent people living one support interesting year best
Cluster 1:
 amazing work inspiring awareness families daughter friend to support like around things resilient meet raising
Cluster 2:
 happy national friends celebrating affected everyone us to forms forward fortunate food fort force forces
Cluster 3:
 wonderful thank latest diario el to work de helping sharing support great spend us news
Cluster 4:
 uncommon to pathologies spain cancers people estimated treatment lupus wide th million research beauty thank
Cluster 5:
 great work see support to awareness people know opportunity team research time come job interview
Cluster 6:
 proud support work many patients strong part community to us join million every team families
Cluster 7:
 love someone abigail champ sending would full need paints rock star community adorable bringing tell
Cluster 8:
 awesome event dr great sure way make perfect like could thanks feb welcome us forward
Cluster 9:
 good morning to keep fight always see fighting work job struggle know spread time learn
```

Figura Nº 49: Execució de l’algorisme KMeans per l’anàlisi de sentiment de tuits a favor.




```
Cluster 0:
 illness invisible horrible painful let suffer aura migraine please terrible change research weird fight stop
Cluster 1:
 cold always reason feet hands urticaria allergy allergic count um milling think they summer oh
Cluster 2:
 happened years almost three thing worst writing suffered awful hysterical struggling alone diagnosed badly treated
Cluster 3:
 sick pope common support life wrestlers forget fam showing joins patients chosen million exigimos concience
Cluster 4:
 devastating conditions psp awareness please support share fondness convivir magnitude people raise cbd neurological misdiagnosis
Cluster 5:
 bad tts bunny derived solution understands god spends hands one friend forget foundation four fondness
Cluster 6:
 one to frustrating suffers adrenoleucodystrophy hijo sick called talked talk press continue voice society give
Cluster 7:
 condition know improve dreadful needs care comprehensive honor might things sick business look patient miserable
Cluster 8:
 suffers everyone makes to played house keep fundamental jets corrupt what sorry violent dissease stupid
Cluster 9:
 even we show sickly invisibles sometimes use explain say resource alagille difficult raras want must
Cluster 10:
 thank using crazy peace go history bring sharing siderosis superficial write thorough others battled experience
Cluster 11:
 fucking estres morning fault stress buzz tortures also legitimately overcome scares core fear really terrifying
```

Figura Nº 50: Execució de l'algorisme KMeans per l'anàlisi de sentiment de tuits en contra.

En aquest últim cas el vocabulari i el nombre d'instàncies no era molt gran i tot i això l'aplicació de l'algorisme ens ha proporcionat els aspectes principals. De nou observem que l'obtenció d'una matriu de vectorització equilibrada és part fonamental en els resultats finals de l'agrupament.

3.5.3 Modelització amb l'algorisme DBSCAN.

DBSCAN és un algorisme d'agrupament no supervisat, que ens aporta noves característiques respecte de KMeans. El seu mode de funcionament està basat en agrupar les instàncies tal com ho realitzaria el nostre cervell. S'identifica

regions amb alta densitat d'instàncies, respecte de les que tenen baixa densitat (que es consideren com un tipus de soroll). Amb aquesta tècnica, DBSCAN és capaç de detectar clústers amb formes allargades o formes no esfèriques que poden ser de diferents mides i que l'algorisme KMeans detecta erròniament. Perquè una observació formi part d'un clúster, hi ha d'haver un mínim d'observacions veïnes dins d'un radi de proximitat especificat. A més a més l'algorisme considera que els clústers estan separats per regions buides o amb poques observacions (Amat, 2017) .

Seguint el mateix patró de procediment, apliquem l'algorisme DBSCAN al dataset de modelització per veure si és capaç de millorar els resultats de KMeans.

A continuació es mostra el procediment seguit per avaluar el dataset de modelització millorat amb l'algorisme DBSCAN

3.5.3.1 Optimització del paràmetre 'eps'.

Amb el mateix procediment descrit en la primera temptativa amb DBSCAN, s'ha calculat i representat les distàncies entre tuits 2 a 2, per identificar el valor òptim de 'eps'. A partir d'aquest valor del paràmetre, empíricament s'ha triat el valor de 'min_samples'. En la figura 51, la gràfica un valor òptim del paràmetre 'eps' quan es produeix un canvi pronunciat. Per visualitzar el resultat apliquem reducció de dimensionalitat per PCA i treballar amb la matriu dispersa utilitzem la llibreria delayedsparse:

```
# Visualitzacio amb PCA
Xz = matriu_tfidf
pca=delayedsparse.PCA(n_components=2)
X_PCA = pca.fit(Xz).transform(Xz)
```

```
neigh = NearestNeighbors(n_neighbors=2)
X=X_PCA
nbrs = neigh.fit(X)
distances, indices = nbrs.kneighbors(X)
```

```
distances = np.sort(distances, axis=0)
distances = distances[:,1]
plt.plot(distances)
plt.xlabel("Tuits ordenats per distància.")
plt.ylabel("Distàncies entre tuits.")
```

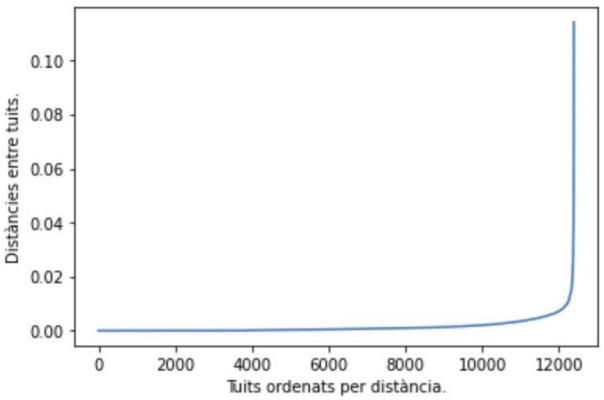


Figura Nº 51: Càlcul del valor òptim del paràmetre eps.

3.5.3.2 Execució i visualització del model DBSCAN.

El valor del paràmetre eps es troba al voltant de 0.01. Mostrem les proves empíriques efectuades per trobar una bona configuració de paràmetres eps i min_samples. Primer s'ha cercat un radi eps molt restrictiu per saber les densitats màximes on es troben i s'ha usat un nombre de veïns mínim:

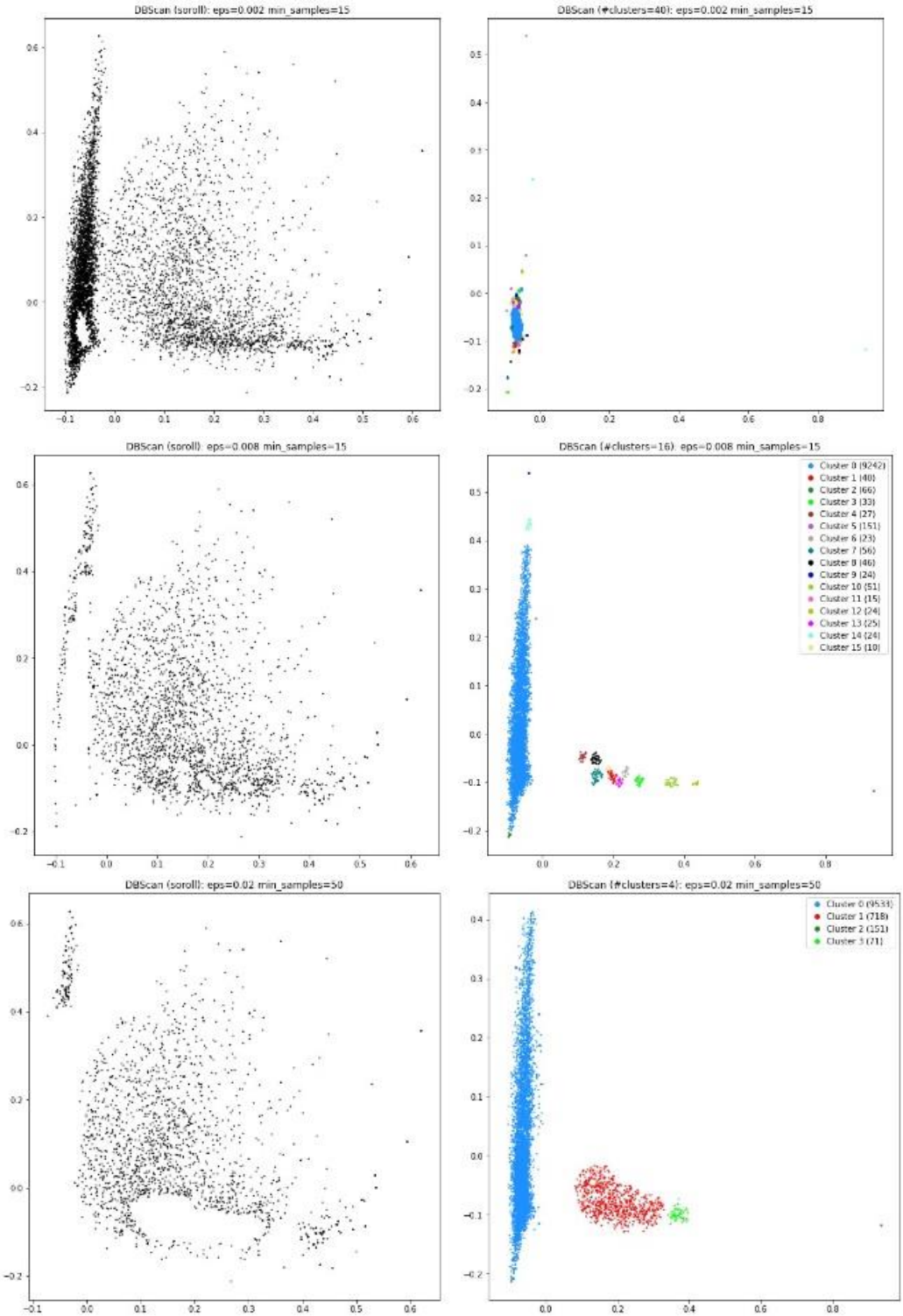
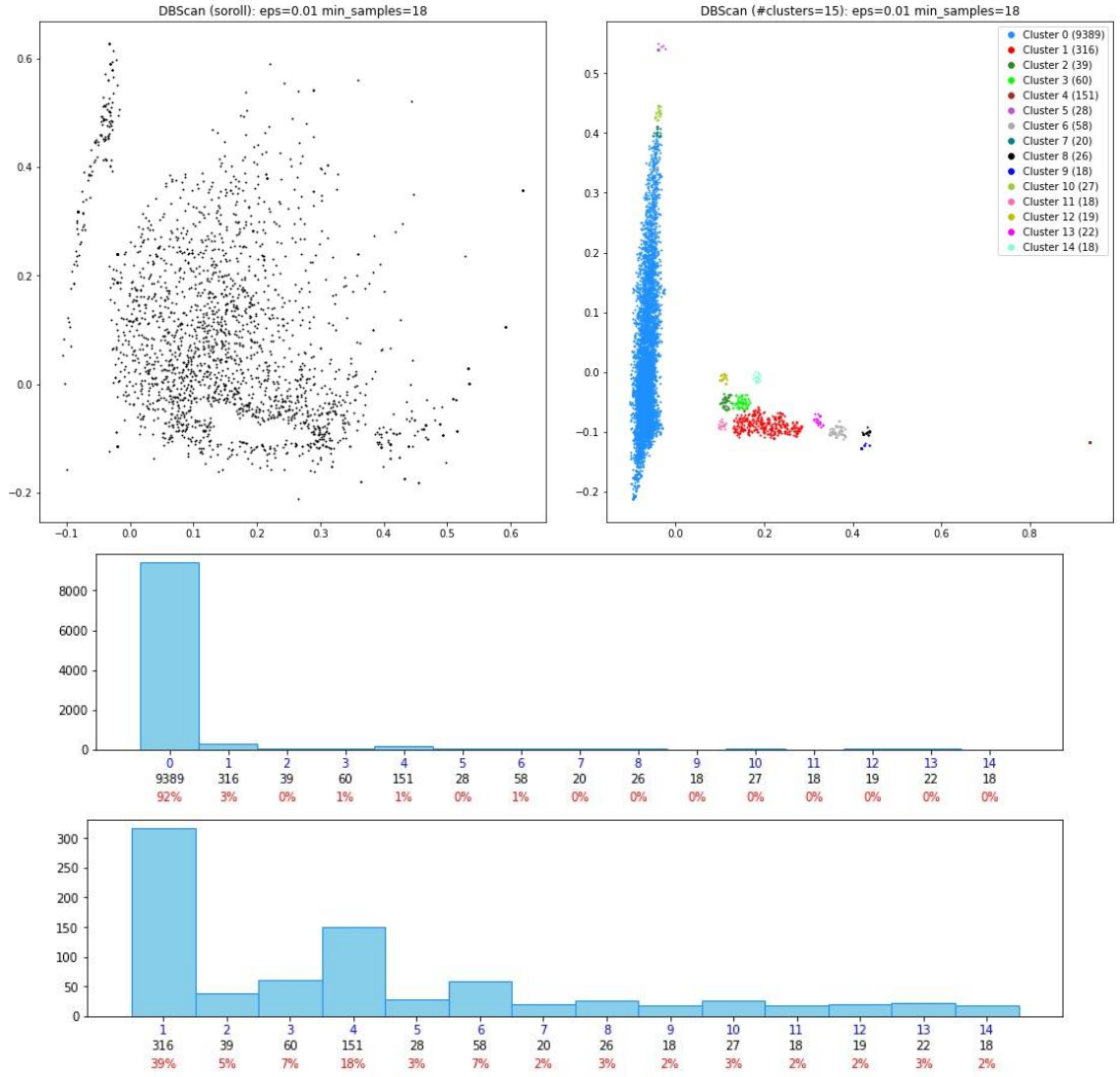


Figura Nº 52: Comportament de l'algorisme DBSCAN per valors del paràmetre eps.

S’ha detectat en el clúster 0 utilitzant un radi eps molt petit i amb molta densitat de tuits. Tot i que a la figura 52 s’han mesurat amb min_samples=50 arriba a ser de min_samples=300 en la regió inferior. La problemàtica que cal resoldre en utilitzar DBSCAN, és que si tenim densitats molt dispars, ens costa trobar una configuració que detecti tots els clústers. En la figura 53 es mostra una configuració de paràmetres compensada per trobar les temàtiques més significatives.




```
Cluster 0: people research support patients one us awareness thank know many help tomorrow like thanks life
Cluster 1: to give celebrated diagnosed make like also want visibility see year international work illness syndrome
Cluster 2: to life treatments time family please us patients still important also th like help year
Cluster 3: to every one family many year syndrome cancer also us learn like condition help years
Cluster 4: to yester hope helping help heard hear health hard happy group great good going go
Cluster 5: people million wide living know learn raise awareness around many live impact lives help global
Cluster 6: to stripes celebrating via remember done say best different national two look follow kids makes
Cluster 7: people million support international living wide awareness around raise chronic live join one new affected
Cluster 8: to celebrate want read hope new story family please learn give hard happy group great
Cluster 9: celebrated to event years syndrome yester going help heard hear health hard happy group great
Cluster 10: people one million wide awareness suffer living raise around help live support different show together
Cluster 11: to cancer could even better syndrome diagnosis thanks make like thank something helping news us
Cluster 12: to like awareness old life year first strong children every fighting know often many long
Cluster 13: to together work important research celebrate years let make share every th learn little need
Cluster 14: to one many year research know support fight genetic st diagnosed syndrome see family let
```

Figura Nº 53: Agrupament per DBSCAN del dataset de modelització millorat.

Els resultats obtinguts són similars als vistos en l'agrupació de KMeans. Les temàtiques importants se centren en els temes de: la conscienciació social, la petició de recerca i investigació de nous tractaments, les denuncies i promoció de casos, l'agraïment per la celebració d'esdeveniments que promouen i difonen els problemes de les malalties minoritàries (i en concret els relacionats amb el DMMM). També obtenim informació nova com:

1. Concepte de visibilitat (clúster 1).
2. Més mencions al càncer (clústers 3 i 11).
3. Treball en conjunt i compartir-ho (clúster 13).
4. Lluita en silenci o llargues lluites contra les malalties (clúster 12).
5. Temps de resposta en tractaments (clúster 2).

En la figura 53 es podem observar dos histogrames, el primer inclou el clúster 0 i el segon l'exclou per observar amb més detall les proporcions de la resta de clústers.

Novament, perdem algunes temàtiques en concentrar-nos en les més significatives i si es vol més detall hauríem de fer l'agrupament en regions específiques o per filtratge específic del vocabulari. A mesura que anem fent més restrictiu el nombre de veïns sobre el radi definit en 0.015, podem trobar configuracions compactes de clústers que són Quant a resultats de temàtiques comparables als anteriors, com és el cas del mostrat en la figura 54.

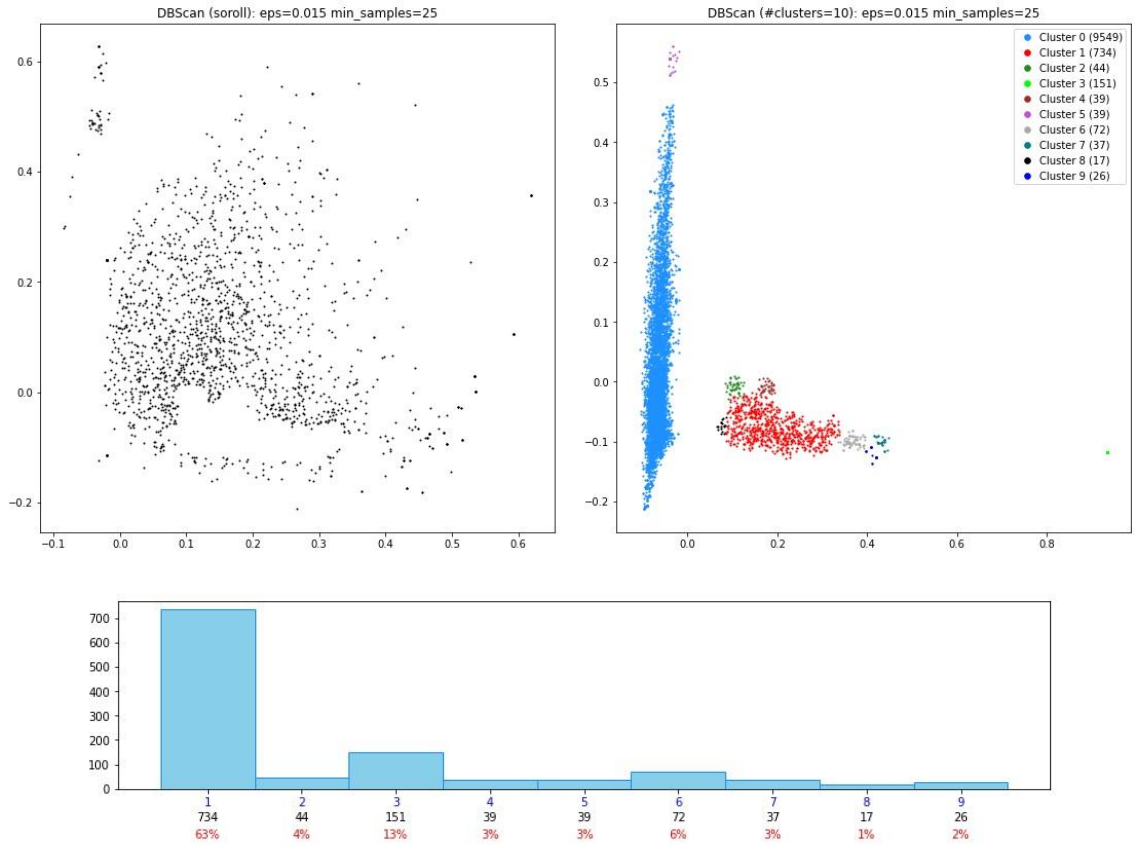
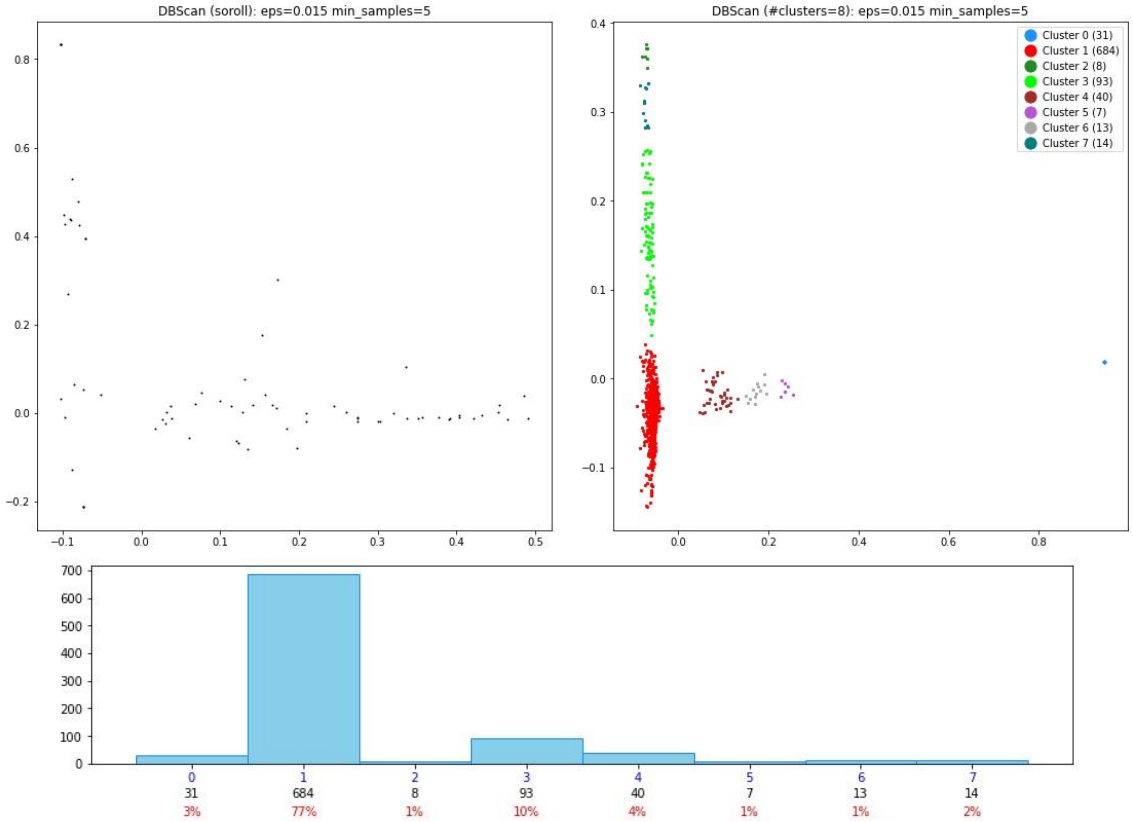


Figura Nº 54: Agrupament per DBSCAN amb paràmetres eps=0.015 i min_samples=25.

3.5.3.3 Anàlisi de sentiment amb DBSCAN.

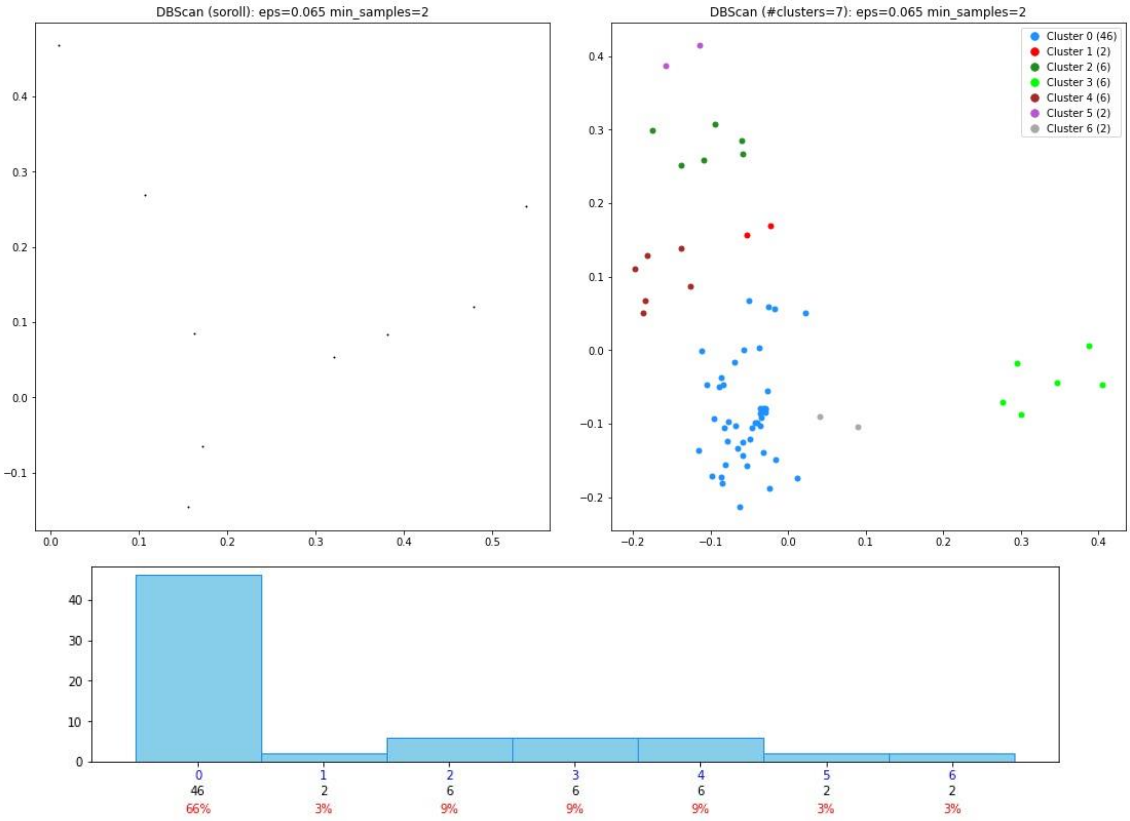
Finalment presentem també els resultats pel cas d'anàlisi de sentiment, però usant l'algorisme DBSCAN.




```
Cluster 0: happy áras follow forefront forces force for football food following followers folk forget focused focus
Cluster 1: great to good love amazing beautiful thank uncommon people work awareness wonderful latest see important
Cluster 2: proud stand pride anna anything something strong affected trialcard million many group research support football
Cluster 3: proud work to support awareness us join patients community families wide help every treatments raise
Cluster 4: happy fri see research life know to us makes fellow type year check hard never
Cluster 5: happy celebrate couple overjoyed documenting biogen neuromyotonia probably maha everybody dia experience take glad able
Cluster 6: happy fellow bold princesses ending remission baby genetically mutated sleepy af entities equalcare leapyear deserving
Cluster 7: proud support patients dyskinesieciliaireprimitive labmates strong part students russell living many mateo helps cares
           company
```

Figura Nº 55: Agrupament per DBSCAN pel cas particular de tuits positius o favorables.

Dels resultats de la figura 55, podem observar en la gràfica de la dreta, que els clústers obtinguts són més compactes i clars. Determinen un conjunt de temàtiques molt similar a les obtingudes amb KMeans amb un menor nombre de clústers. En la gràfica de l'esquerra el nombre de punts classificats com a soroll també és mínim i a més delimita els clústers detectats.




```
Cluster 0: cold fucking years thank devastating people hands bad happened worst illness others psp conditions know
Cluster 1: nurses highlight genetics role sick fundamental diagnosis thinking knows tell trach ventilator asleep daughter needs
Cluster 2: sick common support awareness life care comprehensive showing fam joins society voice improve million concience
Cluster 3: suffers called fundamental makes year old remember neurofibromatosis genetic hijo adrenoleucodystrophy one to thousand minori
Cluster 4: please raise stop awareness hope share sick research weird invisible always devastating causes excruciating pain
Cluster 5: pope sick wrestlers forget patients chosen find finding flare fondness force foundation four france friend
Cluster 6: god spends understands bad allah itis jamal sad brother worship hate makes one friend force
```

Figura Nº 56: Agrupament per DBSCAN pel cas particular de tuits negatius o en contra.

Per l'agrupament d'opinions negatives o en contra no obtenim bons resultats i és necessari cercar per separat en els clústers perquè les densitats són molt diferents entre el grup majoritari i la resta de clústers. Com a exemple oferim els clústers resultants per eps=0.0016 i min_samples=3 a la figura 57.

```
Cluster 0: crap wrong house played total aura come dies health migraine jets corrupt violent horrible obviously
Cluster 1: breakdown amino prevents toxic acid phe extremely known brain fukuoka asia pacific japan individually meetup
Cluster 2: socio problem sanitary serious frightening fought found foundation four framework france francis free freely friend
Cluster 3: das hoxe nursities lunch ill ftd four framework france francis free freely friend friends frightening
Cluster 4: unknown hurtful intention insecure deadly afro pif detected circumstances juncture reminding say expensive per may
Cluster 5: fundamental makes appear physiotherapy suffers levels shocked cases high approximately three two okay happen afraid
Cluster 6: half painful buzz tortures fault morning stress fucking sunflower serve newly aims nonprofit established student
Cluster 7: stupid dissease pee chronic possible alcohol drunk without spherocytosis spleen hereditary anymore doctrine philosophies piles
Cluster 8: count allergic allergy wishing jeremy urticaria tough teens unbelievable dia simple buzzing worship jamal allah
```

Figura Nº 57: Agrupament per DBSCAN pel cas de tuits negatius de detall.

En aquest darrer exemple tenim resultats més semblants als obtinguts anteriorment pel cas del dataset sense filtrar per valors de subjectivitat i polaritat. Obtenim temàtiques més específiques amb més detall en les paraules.

En resum DBSCAN és força útil quan els clústers no segueixen formes convexes o esfèriques i tenen formes allargades com en el nostre cas, però la diferencia de densitat en la multitud de grups fa que ens sigui més útil en l'estudi del detall malalties cròniques, problemes ('toxic àcid clúster 1) o 'alcohol drunk' (clúster 7).

3.5.4 Modelització amb l'algorisme jeràrquic de tipus aglomeratiu.

El tercer tipus d'algorisme que s'ha analitzat és l'algorisme jeràrquic de tipus aglomeratiu (en anglès agglomerative hierarchical clustering o 'down up'). El funcionament de l'algorisme és molt senzill:

1. El procés s'inicia considerant cadascuna de les observacions com un clúster individual (fulles de l'estructura jeràrquica final).
2. Iterativament se segueixen els passos següents fins que totes les observacions pertanyen a un únic clúster:
    a. Es calcula la distància entre cada possible parell dels n clústers. Aquí és necessari definir una mesura de similitud definint una distància i un enllaç.
    b. Es fusionen els parells de clústers més similars.
3. Segons objectiu decidir on tallar l'estructura d'arbre generada.

Per visualitzar l'arbre resultant, s'han utilitzat gràfics de tipus dendrograma, que són diagrames que mostren les agrupacions successives que es van generant en un algoritme jeràrquic.

Per construir un dendrograma aglomeratiu haurem inicialment d'establir amb quina mètrica de distància desitjarem treballar i quin criteri d'enllaç de grups o segments utilitzarem. El següent pas serà considerar cada instància del joc de dades com un grup o segment en si mateix i a partir d'aquí començarem a calcular distàncies entre grups.

Per tant és bàsic en aquest tipus d'algorismes definir de quina manera es defineix la similitud de dos tuits un cop representats numèricament en l'espai n-dimensional utilitzant la matriu de vectorització per tf-idf.

La prova, consisteix a l'aplicació de l'algorisme aglomeratiu utilitzant distància euclidiana i criteris d'enllaç de tipus **ward** (minimització de la variància), **simple**, **complet**, **mitjana** (en anglès average) i **centroide**. També s'ha experimentat amb altres mètriques com la similitud del cosinus però els resultats no han variat en excés.

Per tal de tenir clara la similitud aplicada en cada cas s'ha inclòs una breu descripció i unes característiques de partida extretes de l'estudi de (Amat, 2017):

Ward : S'agrupen els 2 clústers amb menor increment de la variància total intra-cluster.
Simple : Es fusionen aquells clústers que al calcular la distància entre tots els elements d'un clúster i de l'altre contenen la distància mínima.
Complet : Igual que en el cas anterior però considerant la màxima distància.
Mitjana : Es fusionen aquells clústers que calculant la mitjana de les distàncies entre els elements de cada clústers, aquesta és mínima.

Centroide : Es fusionen aquells clústers que tenen la distància entre els seus centroides mínima.

Els enllaços de tipus ward, complet i average coneixem que generen dendrogrames més compensats, però la seva idoneïtat depèn de cada cas d'aplicació. L'enllaç simple és el menys conservador i el que es veu més afectat per l'existència de valors atípics i pateix l'efecte cadena, que succeeix quan es fusionen grups que no cal agrupar. Permet obtenir clústers amb formes allargades o no el·líptiques. L'enllaç complet no està afectat per l'efecte cadena però és sensible a l'existència de valors atípics. L'enllaç mitjana, la majoria de vegades no millora els resultats dels enllaços simple i complet però és una solució de compromís entre els dos.

3.5.4.1 Execució i visualització del model aglomeratiu jeràrquic.

En aquest apartat, es mostren els resultats obtinguts al construir i executar el model aglomeratiu jeràrquic usant com a distància la distància euclidiana i després la similitud del cosinus i com a tipus d'enllaç i per aquest ordre els tipus ward, simple, complet, mitjana i centroide. Per cada cas es mostra el dendrograma obtingut, la representació del clustering en 2D, l'histograma associat i el contingut de cada clúster pels casos de nombre de clústers 3,8,10 i 15 de manera que es pot apreciar l'evolució en l'agrupació. En la figura

```python
def tokens_mes_propers(vector_clusters, vectorizer, mat_vect, topk=10):
    paraules = vectorizer.get_feature_names()
    relevant_labels = set(vector_clusters)
    for this_label in relevant_labels:
        matching_rows = np.where(vector_clusters == this_label)[0]
        coeff_sums = np.sum(mat_vect[matching_rows], axis=0).A1
        sorted_coeff_idxs = np.argsort(coeff_sums)[::-1]
        print('Cluster {}: '.format(this_label), end='')
        for idx in sorted_coeff_idxs[:topk]:
            print('{} '.format(paraules[idx]), end='')
        print()
```

```python
# Visualització de tota la jerarquia de clústers
# usant un enllaç de tipus ward (Minimització variança):
lnk="Ward"
plt.figure(figsize=(12,8))
time_start = time.time()
Z1=hc.ward(X_PCA)
dn=dendrogram(Z1,
              truncate_mode='lastp', # Mostra només els últims p clústers units.
              p=200, # valor de p
              leaf_rotation=90., # rotació d'etiquetes en l'eix de les abcises.
              leaf_font_size=None, # mida de la font en les etiquetes de l'eix de les abcises.
              no_labels=True,
              show_contracted=True # val True quan s'aplica el 'truncate_mode'.
             )
temps=(time.time()-time_start)/60
plt.title("Dendrograma Enllaç tipus {}".format(lnk))
plt.savefig("dn_ward.jpg",format='jpg',bbox_inches='tight')
plt.show()
print("#Visualització dendrograma (PCA):",np.shape(X_PCA),"\n Durada: ",int(temps) if temps>0 else 0,"minut/s ", \
      int((temps-int(temps))*60),"segons.")
```

Figura Nº 58: Scripts Python: per el dendrograma i la visualització dels continguts dels clústers.

La funció *tokens_mes_propers* és equivalent a la utilitzada amb l'algorisme DBSCAN i mostra els mots més freqüents de cada clúster. El script de la part inferior de la figura 58 representa gràficament el dendrograma per cada enllaç.

3.5.4.2 Enllaços per distància euclidiana.

En aquest apartat es mostren els resultats de l'algorisme jeràrquic aglomeratiu, calculats per una mètrica de distància euclidiana.

3.5.4.3.1 Enllaç de tipus ward.

En la figura 59, es pot observar el dendrograma global i en la part inferior els diferents talls del dendrograma per valors de k=3,8,10 i 15 clústers.

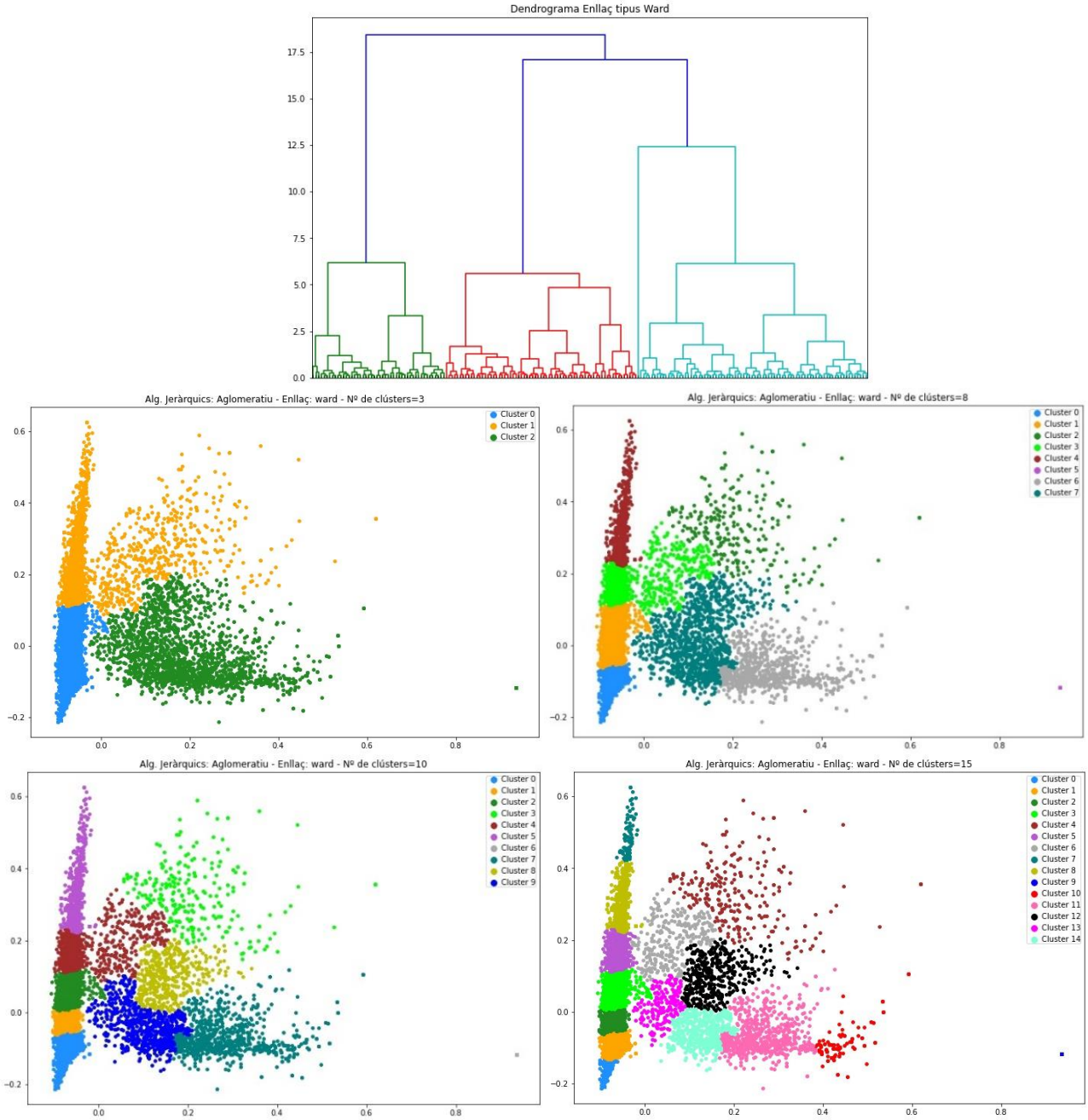
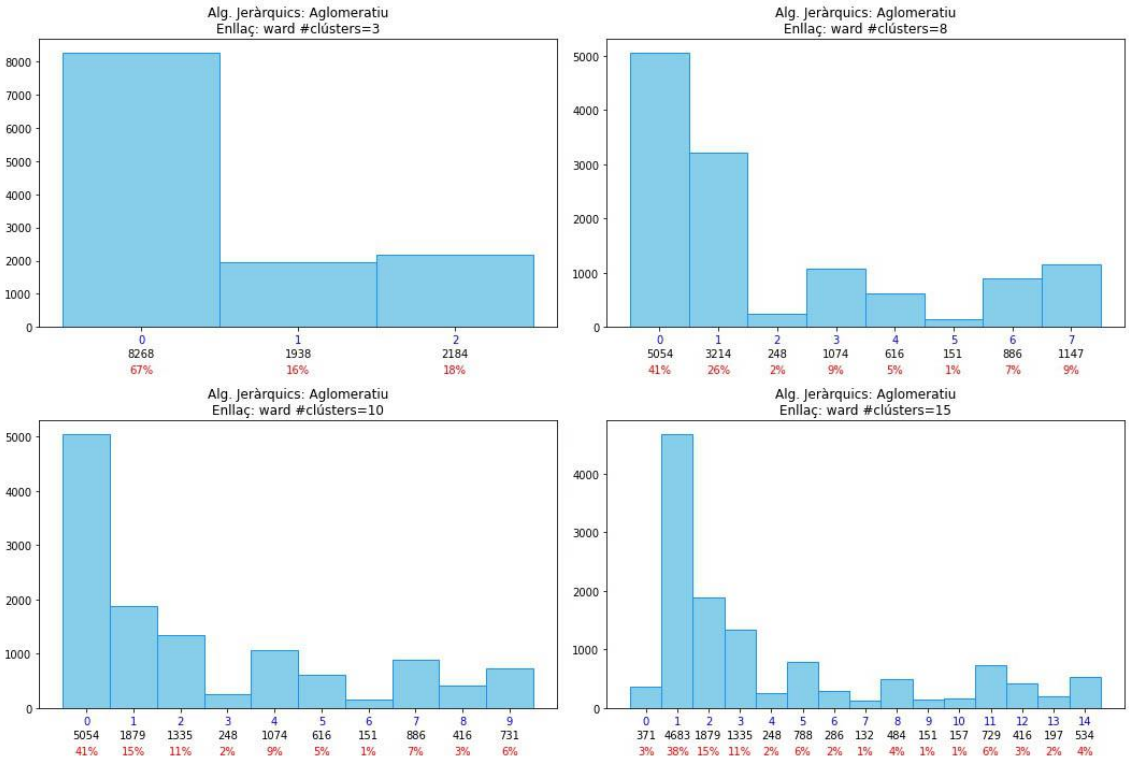


Figura Nº 59: Dendrograma i scatterplot 2D del model jeràrquic aglomeratiu per k=3,8,10,15.




```
num_clusters=15
tall=hc.fcluster(Z1, num_clusters, criterion='maxclust')-1
tokens_mes_propers(tall, tfidf_vect, matriu_tfidf, 15)
```

```
Cluster 0: thank latest thanks happy daily much news great sharing good last everyone health work event
Cluster 1: research great like syndrome us good years tomorrow also event work international th love thank
Cluster 2: know patients many one help us research support proud care treatment every families learn year
Cluster 3: awareness support one patients help people raise living research families us raising many know every
Cluster 4: to people million living awareness wide around raise one learn know raising st families affect
Cluster 5: people awareness raise million living patients many affected support one know help life lives live
Cluster 6: to awareness people living million raise support patients help many wide affected live us around
Cluster 7: people million living wide one around awareness support raise know live suffer affected help many
Cluster 8: people million living awareness raise one affected wide around support help know live patients families
Cluster 9: to yester hope helping help heard hear health hard happy group great good going go
Cluster 10: to celebrated celebrate international one want support read tomorrow awareness little research us know event
Cluster 11: to celebrated one support year work research us want know every th make also many
Cluster 12: to awareness raise people support patients impact know help lives living families learn million one
Cluster 13: to us thank awareness patients share research story support see get help event work one
Cluster 14: to us thank syndrome research like one patients every year many also good years time
```

Figura Nº 60: Histogrames associats al model jeràrquic aglomeratiu i contingut dels clústers detectats.

Els resultats ens permeten obtenir les temàtiques principals ja obtingudes amb els algorismes de clustering KMeans i DBSCAN. Les temàtiques comprenen agraïments i satisfacció per celebrar esdeveniments que donen suport a les malalties minoritàries amb caràcter internacional, la necessitat de recerca, la consciència del patiment de pacients i famílies, la necessitat de recerca i tractaments. En els resultats observem, com es fa explícita la naturalesa de l'algorisme jeràrquic aglomeratiu i podem observar el refinament de les temàtiques en els mateixos clústers, que comparteixen d'inici paraules, i després focalitzen cap a una direcció el contingut de la temàtica. Com a exemple d'aquest fet, els clústers 13 i 14, on s'identifica agraïment 'to us thank' de manera conjunta i es focalitza en cada cas, en la consciència i ajuda de pacients en compartir i donar suport al clúster 13 ('awareness patients share') i sobre la recerca de síndromes en pacients al clúster 14 ('syndrome research like one patients').
També en aquest cas, es detecta un clúster que concentra el nombre més gran d'instàncies amb molta diferència de la resta. La part de tuits que no pot

classificar l'algorisme, queda inclosa en aquest clúster, com passava amb KMeans i DBSCAN.

3.5.4.3.2 Enllaç de tipus simple.

En la figura 61, es pot observar el dendrograma global i en la part inferior els diferents talls del dendrograma per valors de k=3,8,10 i 15 clústers.

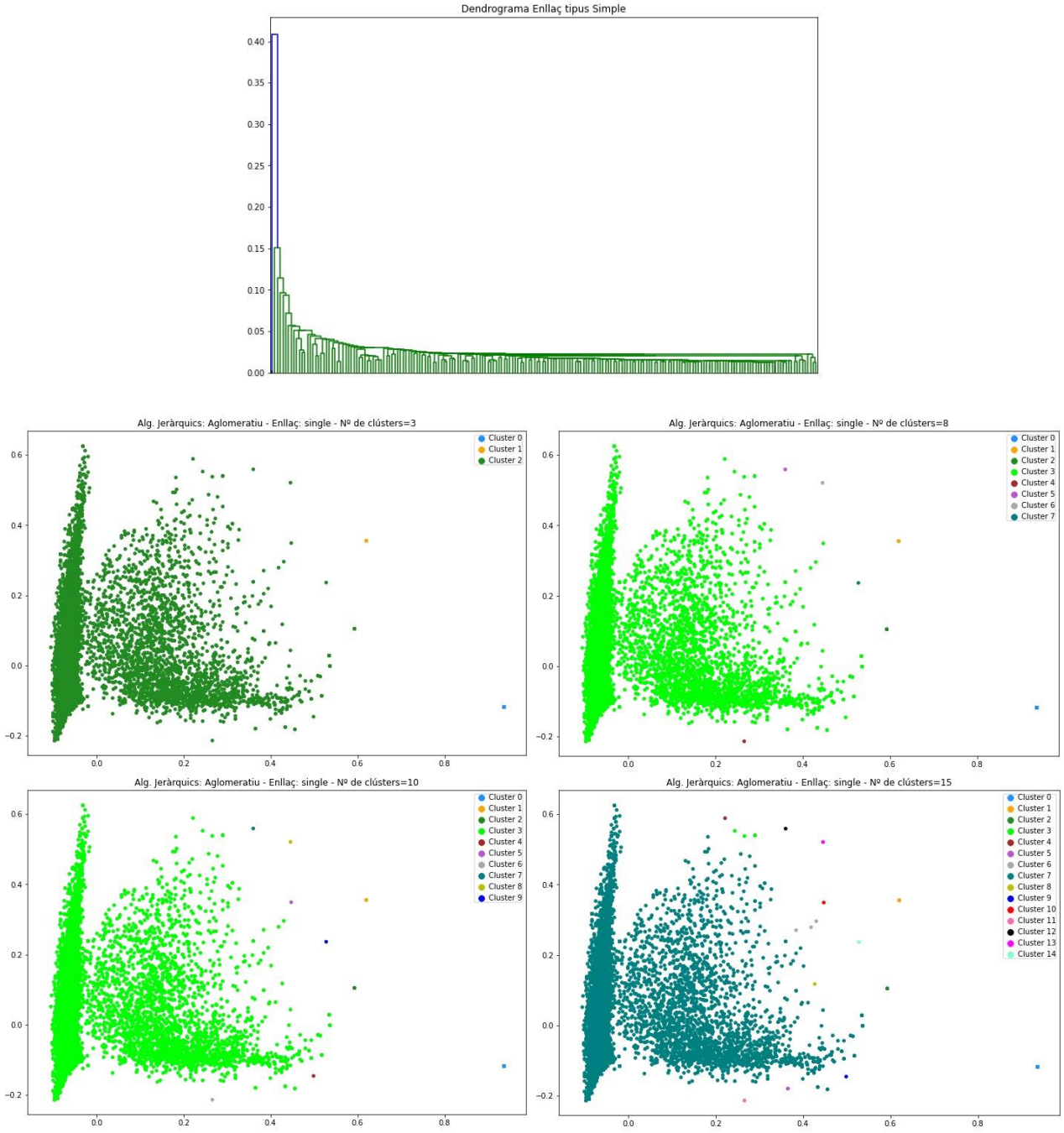
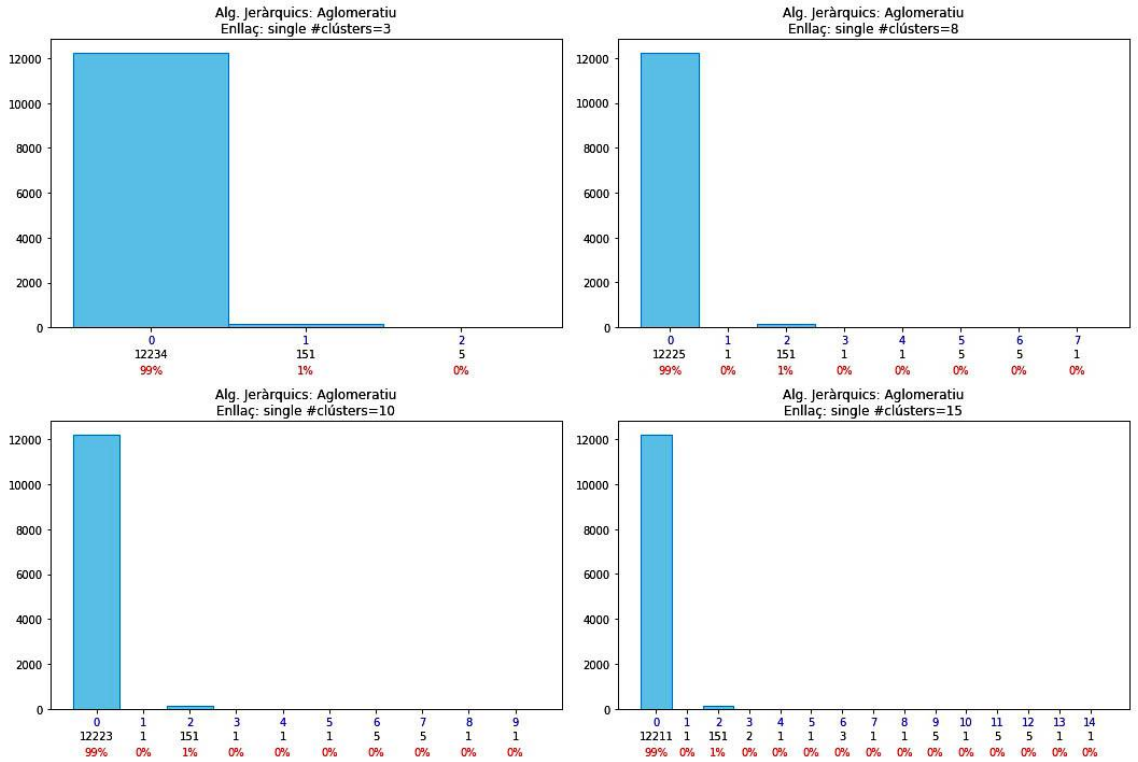


Figura Nº 61: Dendrograma i scatterplot 2D del model jeràrquic aglomeratiu per k=3,8,10,15.




```
num_clusters=15
tall=hc.fcluster(Z2, num_clusters, criterion='maxclust')-1
tokens_mes_propers(tall, tfidf_vect, matriu_tfidf, 15)
```

```
Cluster 0: to yester hope helping help heard hear health hard happy group great good going go
Cluster 1: people to yester honor help heard hear health hard happy group great good going go
Cluster 2: awareness to yester going honor helping help heard hear health hard happy group great good
Cluster 3: living million people wide to around families one awareness affects great honor helping help heard
Cluster 4: around learn raise living million one awareness people to hear health hard good happy heard
Cluster 5: latest to yester go helping help heard hear health hard happy group great good going
Cluster 6: people to live many know go help heard hear health hard happy group great good
Cluster 7: to people awareness support one research patients us know million thank many help living raise
Cluster 8: know awareness to yester going helping help heard hear health hard happy group great good
Cluster 9: to want work thank go helping help heard hear health hard happy group great good
Cluster 10: one people to yester helping help heard hear health hard happy group great good going
Cluster 11: latest thanks to yester going helping help heard hear health hard happy group great good
Cluster 12: living million people to go help heard hear health hard happy group great good going
Cluster 13: million people to going helping help heard hear health hard happy group great good yester
Cluster 14: million to yester go helping help heard hear health hard happy group great good going
```

Figura Nº 62: Histogrames associats al model jeràrquic aglomeratiu i contingut dels clústers detectats.

L'enllaç de tipus simple, no proporciona una bona diferenciació de clústers. El fet que hi hagi moltes paraules que representen a diverses temàtiques i que són molt majoritàries respecte d'altres, fa que altres temes no els detectem per quèdar relegats per els primers. A més a més, es produeix l'efecte cadena i clústers que haurien d'estar diferenciats queden fusionats.

En aquest exemple, es fa encara més evident, que el clúster majoritari (cluster 0) inclou molts tuits que són agrupables amb identitat i similitud pròpia i que l'algorisme amb aquest criteri de distància no discrimina.

3.5.4.3.3 Enllaç de tipus complet.

En la figura 63, es pot observar el dendrograma global i en la part inferior els diferents talls del dendrograma per valors de k=3,8,10 i 15 clústers.

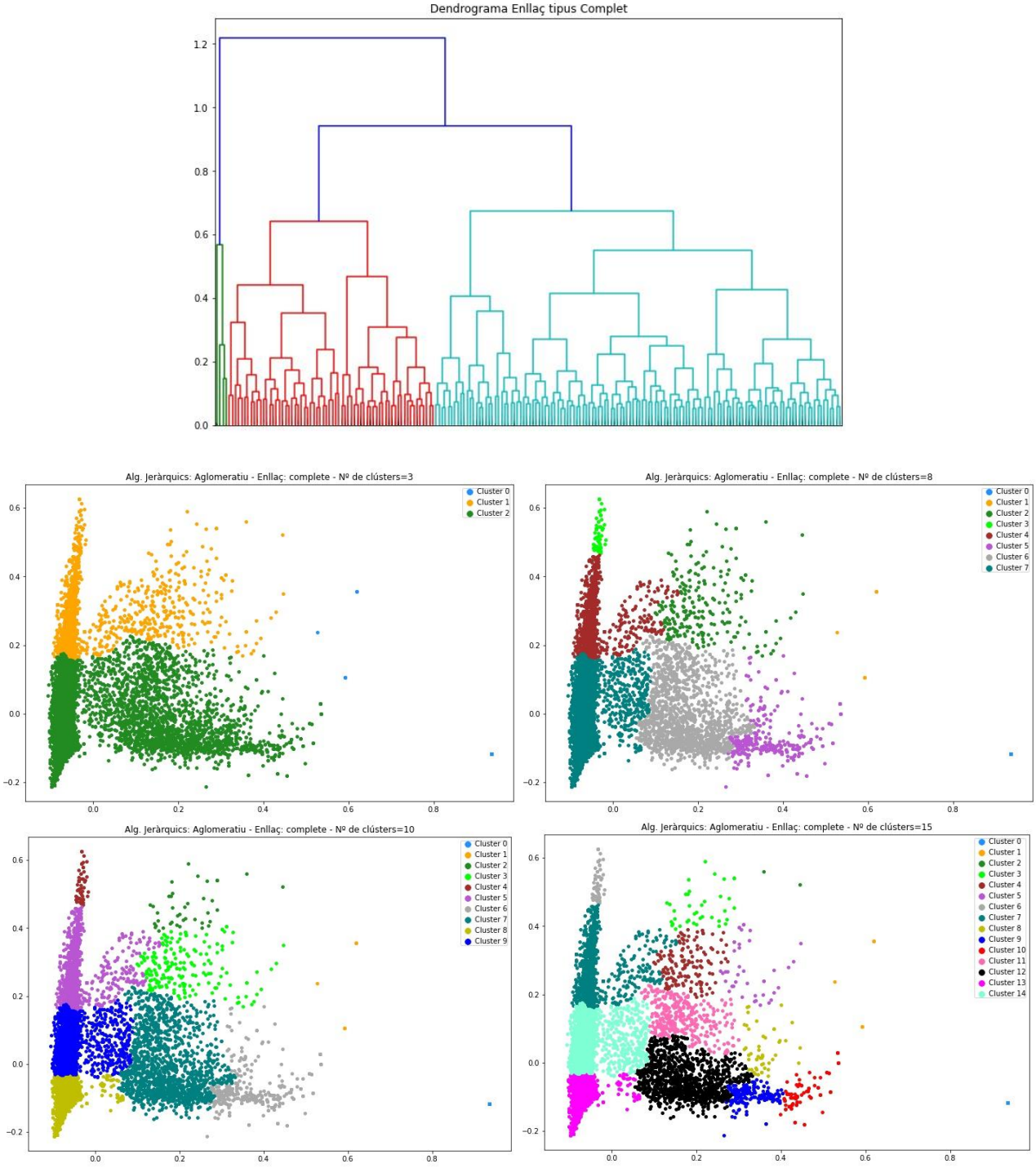
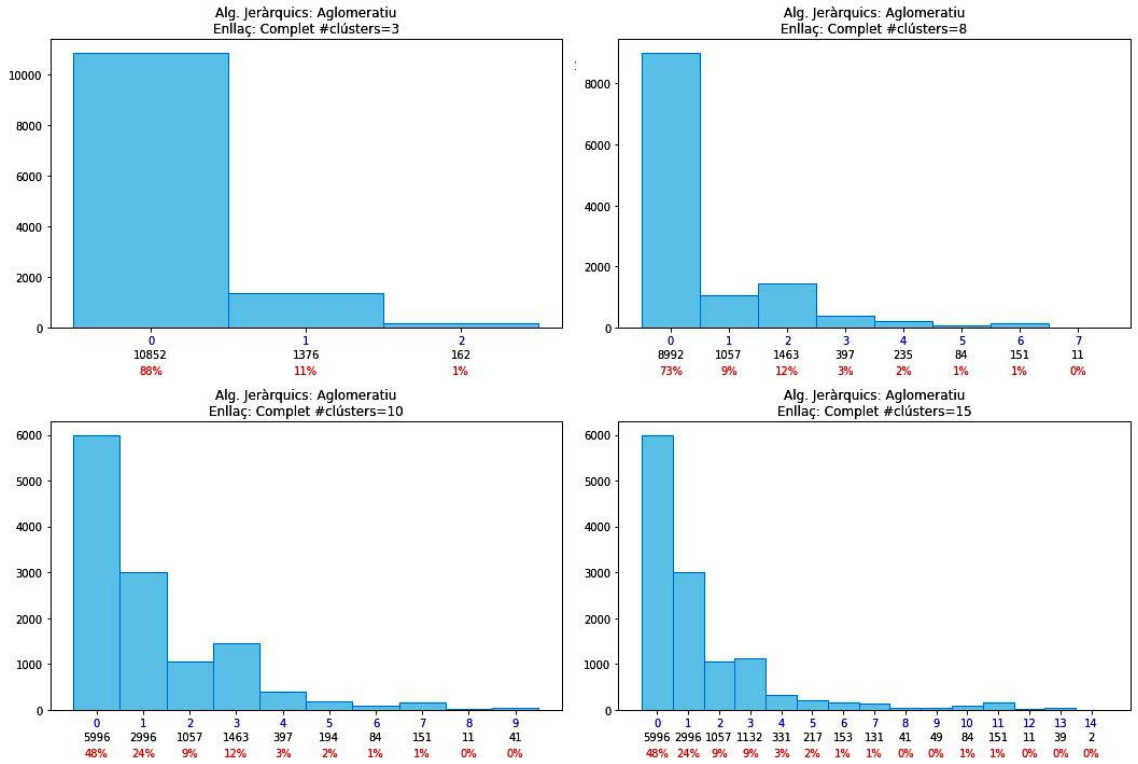


Figura Nº 63: Dendrograma i scatterplot 2D del model jeràrquic aglomeratiu per k=3,8,10,15.




```
num_clusters=15
tall=hc.fcluster(Z3, num_clusters, criterion='maxclust')-1
tokens_mes_propers(tall, tfidf_vect, matriu_tfidf, 15)
```

```
Cluster 0: to yester hope helping help heard hear health hard happy group great good going go
Cluster 1: to awareness people million going helping help heard hear health hard happy group great good
Cluster 2: million people to living go help heard hear health hard happy group great good going
Cluster 3: million people living to around wide awareness raise suffer one learn live help raising know
Cluster 4: people to million awareness living raise wide around st raising affected one year support learn
Cluster 5: to people million affect know suffer awareness one raise living celebrate suffering live affects around
Cluster 6: people million living wide around awareness one raise support know live families proud show th
Cluster 7: people million awareness living raise one wide affected support around patients help know many live
Cluster 8: to one awareness support know living raise affected raising show want international th proud work
Cluster 9: to celebrated work celebrating research celebrate important us year th stripes years via also little
Cluster 10: to celebrated celebrate international one want support read tomorrow research us know event see thanks
Cluster 11: to awareness people raise support patients impact million families lives living raising help many know
Cluster 12: to one support us year every research know awareness like learn many also make patients
Cluster 13: thank research thanks happy us great latest like tomorrow syndrome years work life th good
Cluster 14: support awareness one patients people help many us know research families raise treatment every learn
```

Figura Nº 64: Histogrames associats al model jeràrquic aglomeratiu i contingut dels clústers detectats.

El criteri d'enllaç complet, també ens proporciona una divisió de clústers com era el cas de l'enllaç ward. Podem obtenir les temàtiques globals però la qualitat dels resultats és inferior. Molts tuits mal agrupats tenen afectació, incloent-hi paraules que després quan cerquem les paraules més importants de cada clúster ens fan perdre capacitat per diferenciar les temàtiques. Aquest efecte es produeix de manera similar en els següents tipus d'enllaç mitjana (average) i centroide, tal com es pot observar en el conjunt de figures 65, 66, 67 i 68.

En general els resultats obtinguts són inferiors a l'algorisme KMeans i DBSCAN. Les característiques pròpies del dataset on de partida hi ha molta diversitat implica no cometre errors en agrupacions intermèdies, per no acumular-los en el resultat final.

Un altre aspecte es què en general, l'algorisme aglomeratiu no ens ha permès detectar els clústers allargats i ens ha fraccionat o tessel·lat clústers sense necessitat.

3.5.4.3.4 Enllaç de tipus mitjana.

En la figura 65, es pot observar el dendrograma global i en la part inferior els diferents talls del dendrograma per valors de k=3,8,10 i 15 clústers.

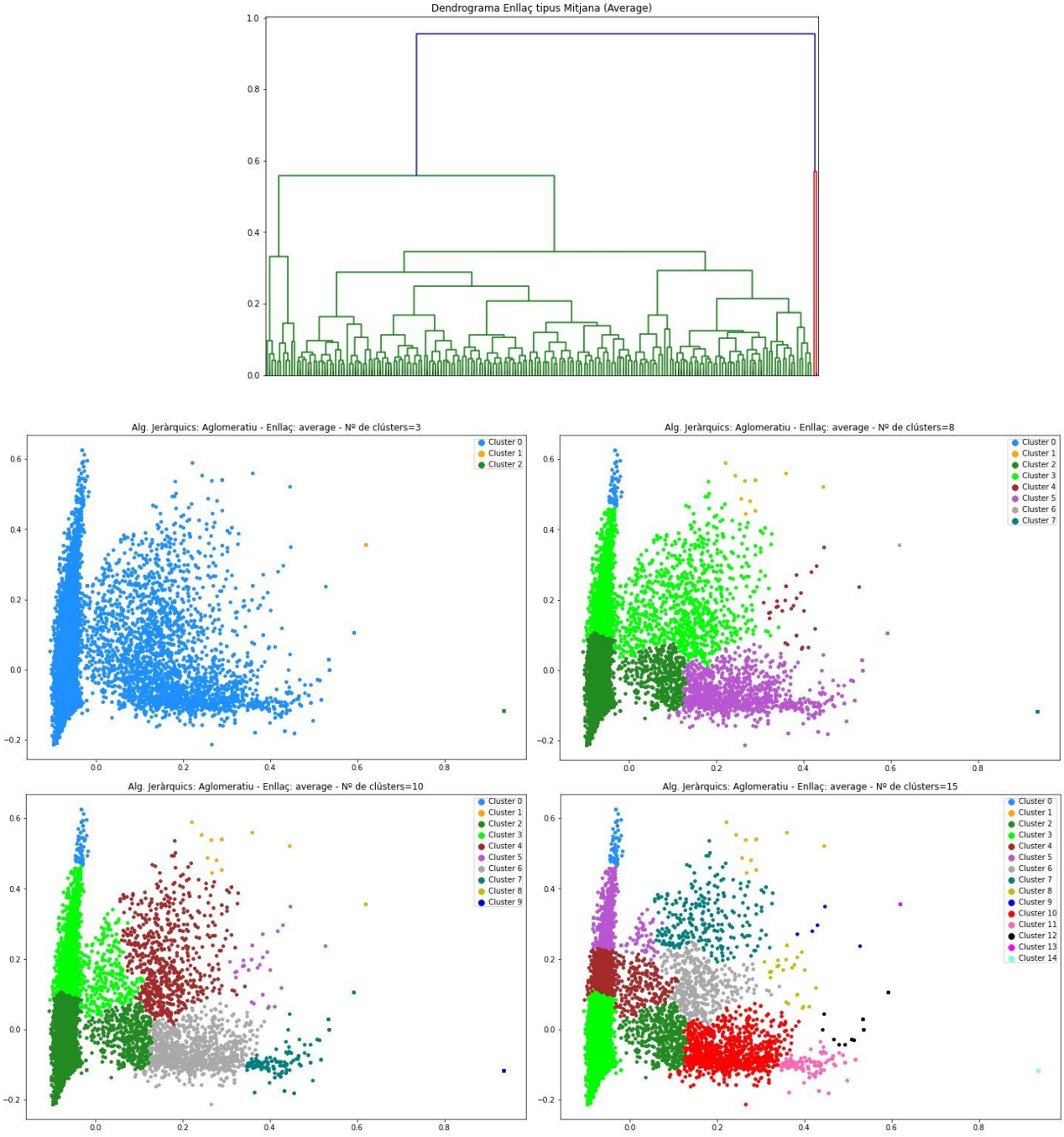
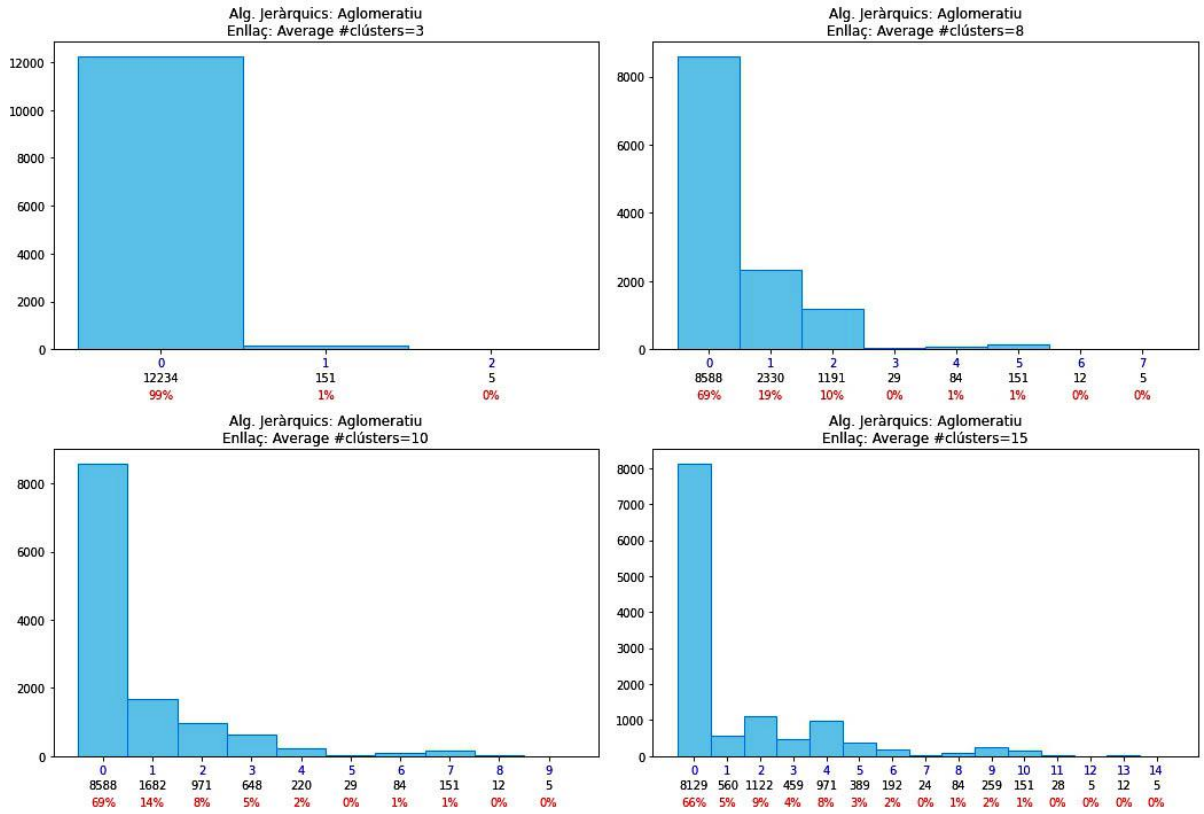


Figura Nº 65: Dendrograma i scatterplot 2D del model jeràrquic aglomeratiu per k=3,8,10,15.




```
num_clusters=15
tall=hc.fcluster(Z4, num_clusters, criterion='maxclust')-1
tokens_mes_propers(tall, tfidf_vect, matriu_tfidf, 15)
```

```
Cluster 0: people million living wide around awareness one raise support know live families proud show th
Cluster 1: million people living to wide around one live awareness learn raise help families genetic full
Cluster 2: to us research thank awareness support patients life help story one like share also time
Cluster 3: research thank us patients support one know thanks like happy tomorrow many help great life
Cluster 4: people awareness raise million patients living support many one affected know help us tomorrow families
Cluster 5: people million living awareness raise one wide affected around support help know live patients suffer
Cluster 6: to awareness people raise support patients impact lives living million help know learn many raising
Cluster 7: people to million living awareness wide around raise one learn know raising affected support live
Cluster 8: to people awareness affect know one million raising th living raise fighting without common special
Cluster 9: to people million live many know one forward friends help heard hear health fighting find
Cluster 10: to one support celebrated us every year research work want many learn make know also
Cluster 11: to celebrated celebrate international want stripes celebrating read tomorrow via little research us remember event
Cluster 12: to one support awareness know patients wide affected learn families many hard going health hear
Cluster 13: people to yester honor help heard hear health hard happy group great good going go
Cluster 14: to yester hope helping help heard hear health hard happy group great good going go
```

Figura Nº 66: Histogrames associats al model jeràrquic aglomeratiu i contingut dels clústers detectats.

3.5.4.3.5 Enllaç de tipus centroide.

En la figura 67, es pot observar el dendrograma global i en la part inferior els diferents talls del dendrograma per valors de k=3,8,10 i 15 clústers.

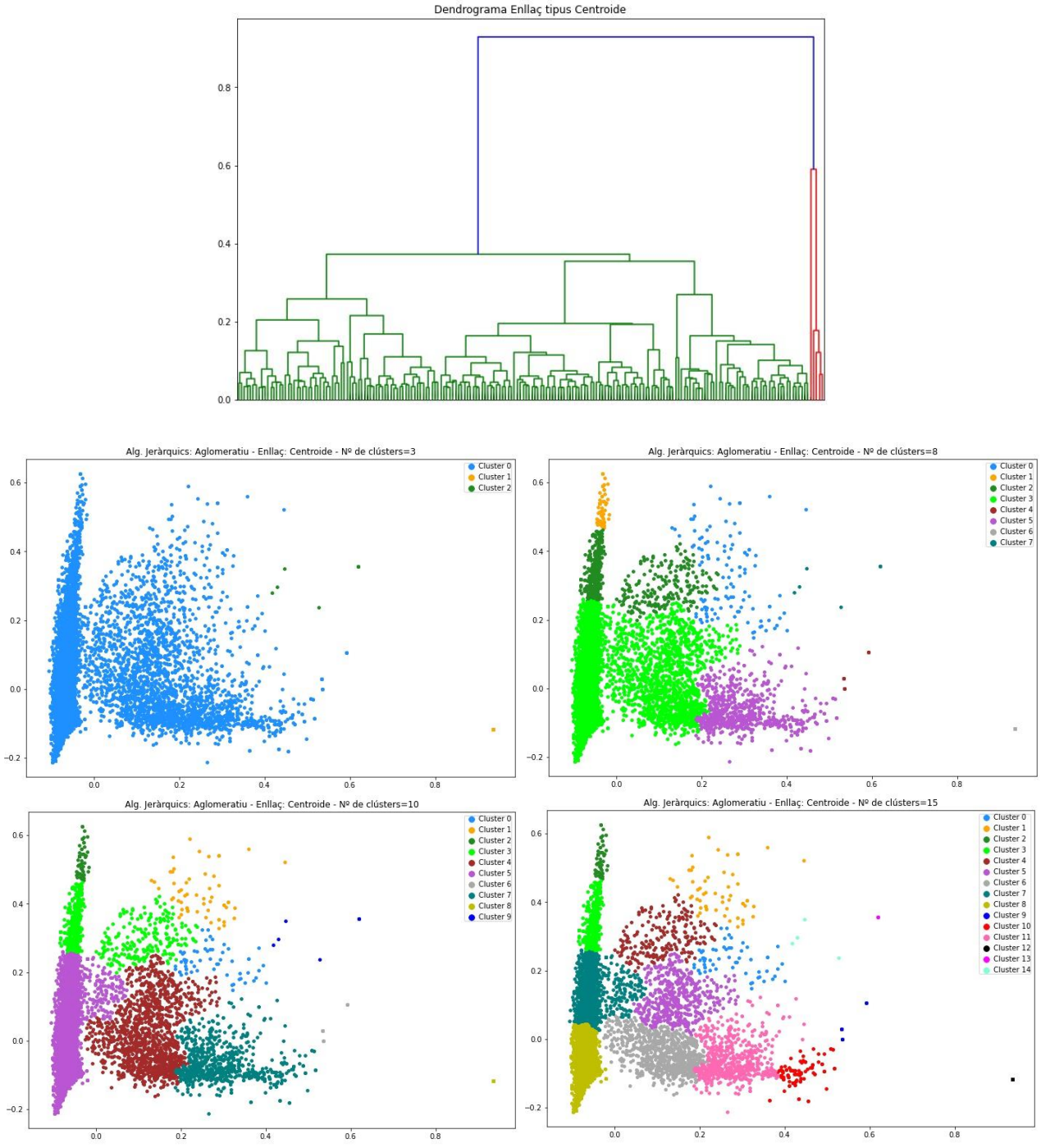
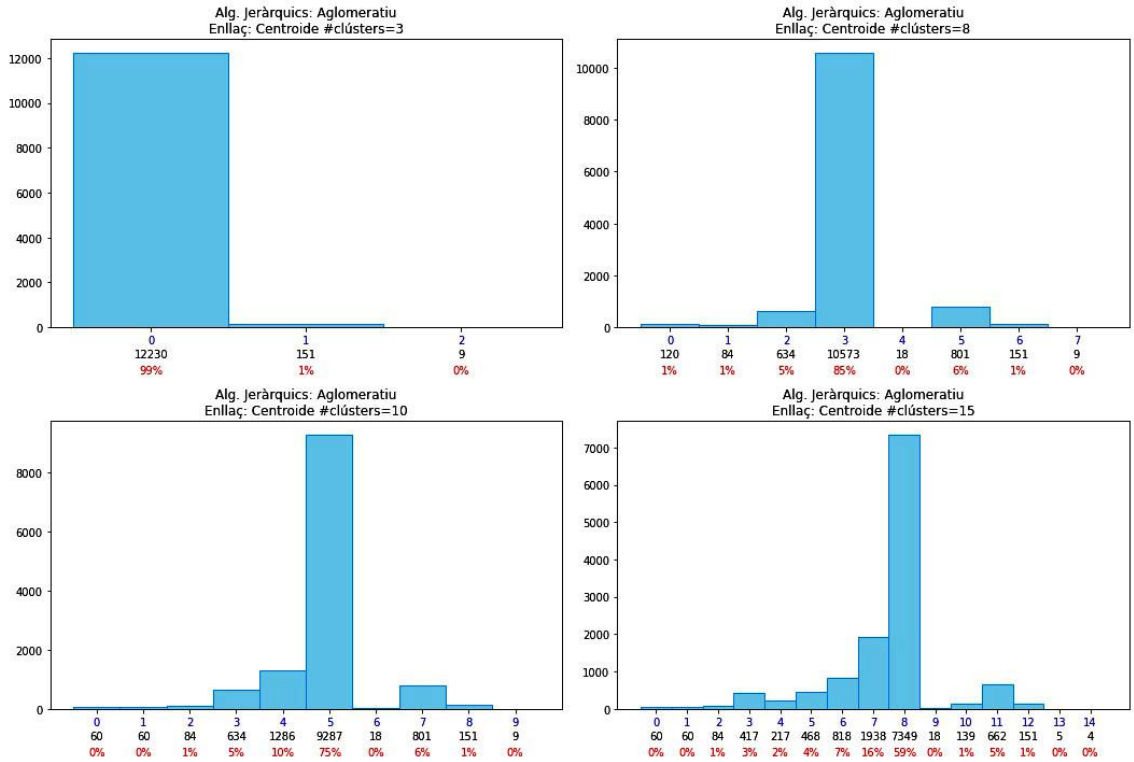


Figura Nº 67: Dendrograma i scatterplot 2D del model jeràrquic aglomeratiu per k=3,8,10,15.




```
num_clusters=15
tall=hc.fcluster(Z6, num_clusters, criterion='maxclust')-1
tokens_mes_propers(tall, tfidf_vect, matriu_tfidf, 15)
```

```
Cluster 0: to people million awareness affect living know one suffer wide raise help every patients around
Cluster 1: million people to living around awareness wide raise one suffer families st learn suffering year
Cluster 2: people million living wide around awareness one raise support know live families proud show th
Cluster 3: people million living one wide awareness affected around raise support help know suffer live patients
Cluster 4: people million to living awareness wide raise around support affected learn live help raising patients
Cluster 5: to awareness people raise patients support impact living lives help million know raising families learn
Cluster 6: to us thank research like patients one syndrome every share support awareness story also year
Cluster 7: people awareness support raise one living patients million help many know affected raising us families
Cluster 8: research thank us patients thanks happy know like tomorrow great many life work support th
Cluster 9: to one support awareness yester good helping help heard hear health hard happy group great
Cluster 10: to celebrated celebrate international want read tomorrow little research us know event see thanks th
Cluster 11: to celebrated one support work us year know research want awareness th every make many
Cluster 12: to yester hope helping help heard hear health hard happy group great good going go
Cluster 13: people to yester honor help heard hear health hard happy group great good going go
Cluster 14: to people million many know one forward great helping help heard hear fighting health find
```

Figura Nº 68: Histogrames associats al model jeràrquic aglomeratiu i contingut dels clústers detectats.

3.5.4.3 Ús de la mètrica similitud de cosinus.

En aquest apartat es mostren els resultats calculats usant una mètrica de similitud de cosinus, i s'ha tractat de comprovar si es milloraven els resultats. No s'ha observat millores considerables. Per la matriu tf-idf, dos vectors normalitzats es consideren similars, si tenen la mateixa direcció, és a dir si la similitud del cosinus es zero. La màxima diferència la trobem quan són ortogonals.

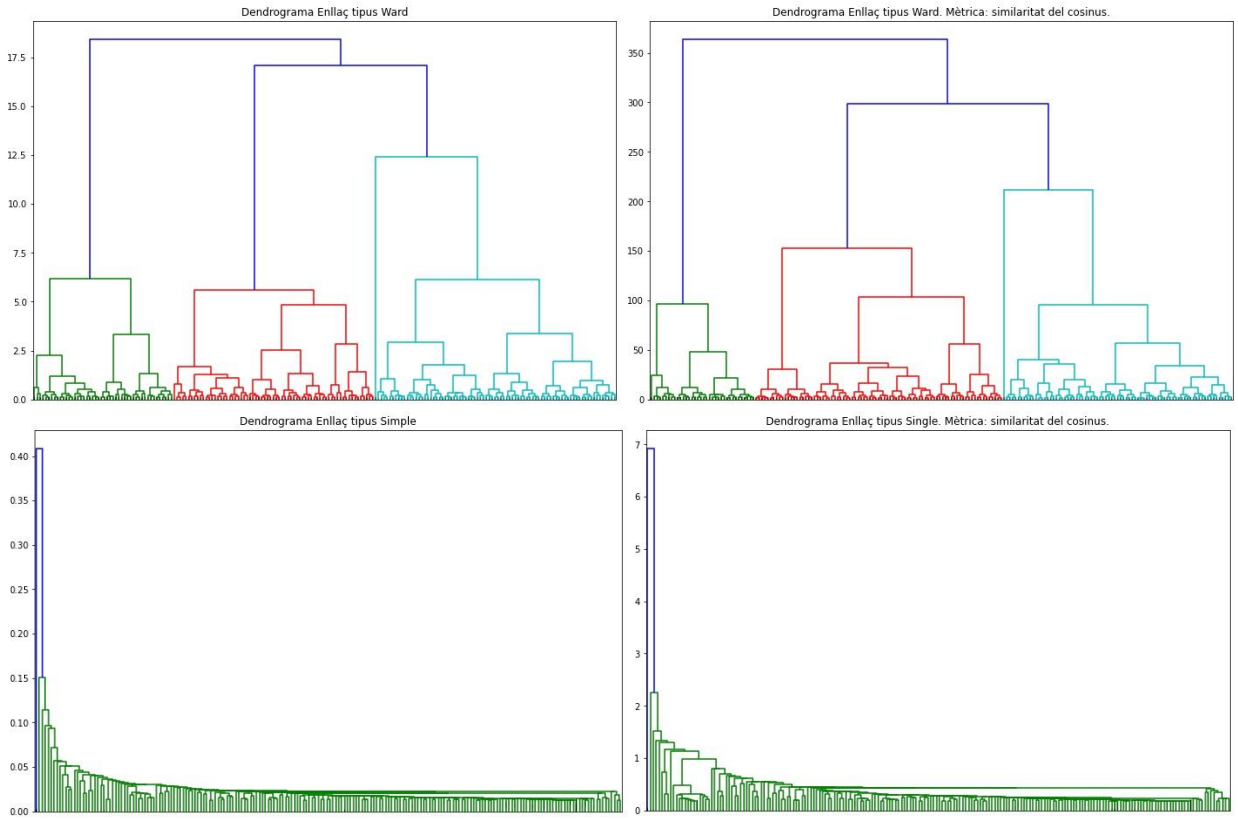
Utilitzant altres mètriques de distància

```
# Similaritat del cosinus.
similaritat = 1 - cosine_similarity(matriu_tfidf)
similaritat.shape
# Reducció de la dimensionalitat amb PCA:
#pca = PCA(n_components=2)
Xz = similaritat
n_comp=2
print("\nCreant el model i matriu PCA - Nº Components=",n_comp)
pca=delayedsparse.PCA(n_components=n_comp)
X_PCA_cosine = pca.fit(Xz).transform(Xz)
print("Dimensions de les dades reduïdes amb PCA:", np.shape(X_PCA_cosine))
```

Figura Nº 69: Càlcul de la matriu de vectors de similitud per cosinus.

3.5.4.3.1 Comparativa de dendrogrames.

En la figura 70, es pot observar per tots els casos, el dendrograma calculat sobre la matriu tf-idf a l'esquerra, i el dendrograma sobre la matriu de similitud a la dreta. S'ha decidit mostrar els dendrogrames pel seu anàlisi visual, sense fer més comprovacions. En cas d'especial interès, es pot comprovar la qualitat de cada dendrograma representant les distàncies entre clústers. Una manera pràctica de comprovar-ho, consisteix a calcular el coeficient de correlació entre les distàncies 'cophenetic' del dendrograma (altura dels nodes) i la matriu de distàncies original.



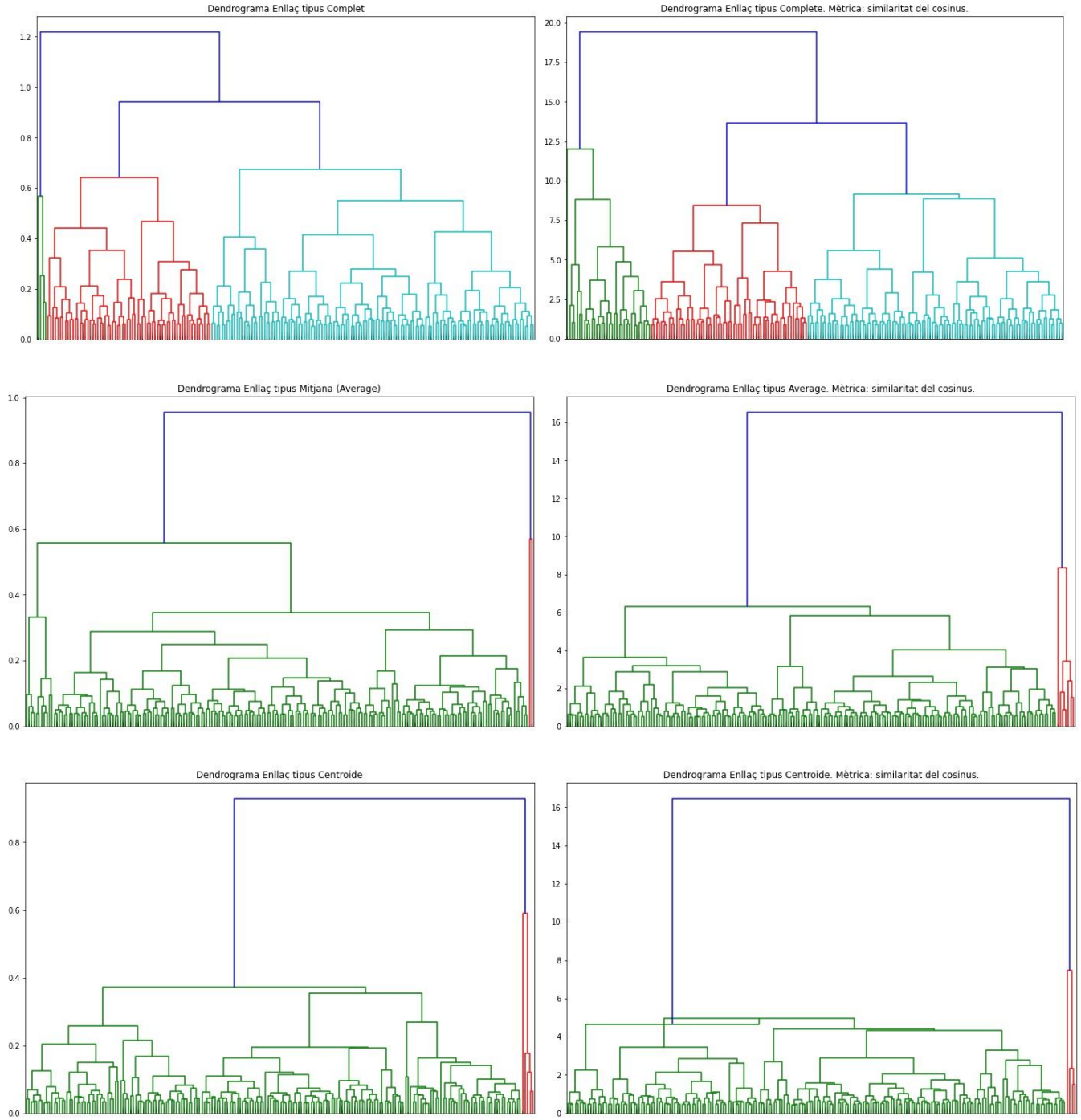


Figura Nº 70: Comparativa dels dendrogrames amb mètrica euclidiana i similitud del cosinus.

3.5.4.3.2 Comparativa pels casos ward i complet

Per finalitzar la comparació de les dues mètriques, en la figura 71, es mostra la gràfica d'agrupament pels diferents talls k=3,8,10 i 15 del dendrograma i es comparen amb els representats pel cas de distància euclidiana. Aquesta comparativa només es presenta pels casos d'enllaç ward figura 71, i pel cas d'enllaç complet figura 72.

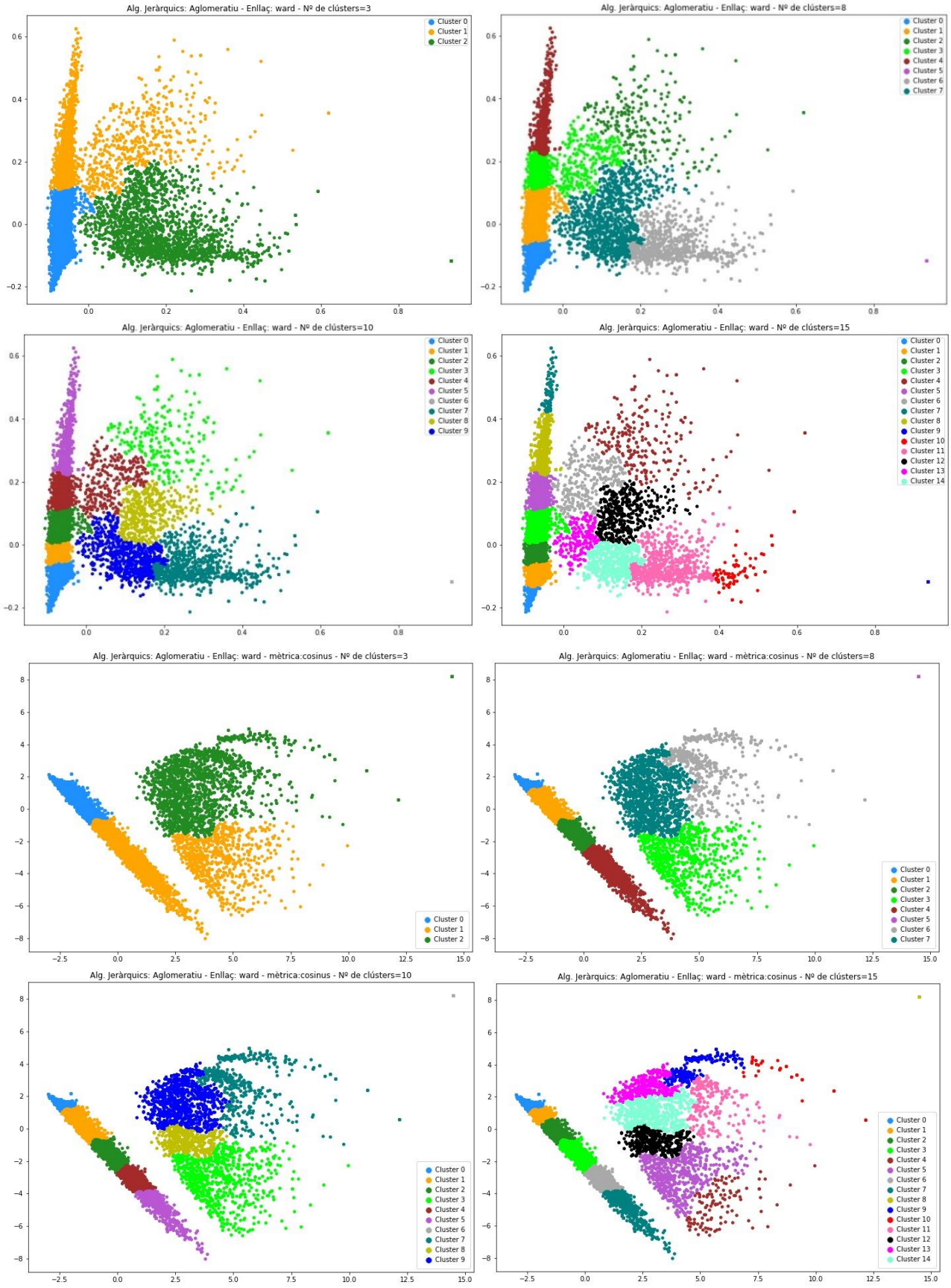


Figura Nº 71: Comparativa entre clústers amb mètrica euclidiana i similitud del cosinus (enllaç ward).

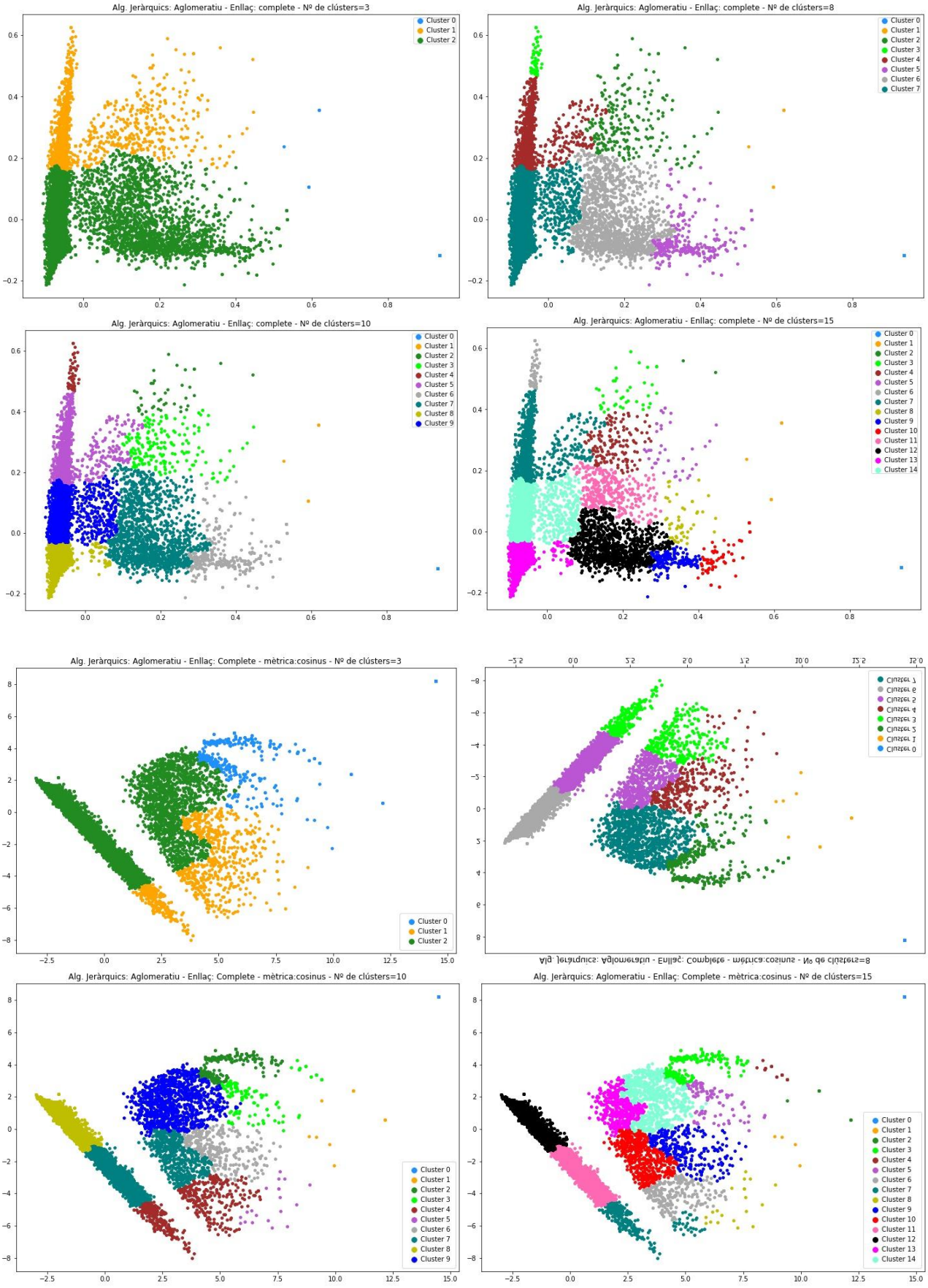


Figura Nº 72: Comparativa de clústers amb mètrica euclidiana i similitud del cosinus (enllaç complet).

Pel darrer cas podem observar més diferències però al final, pel nombre de clústers seleccionat (tall del dendrograma), no s'observen diferències significatives com s'ha comentat anteriorment. No podem generalitzar aquest fet, perquè dependrà de cada cas i domini de dades. Deduïm que com treballem amb tuits de longitud molt semblant, la variabilitat de longitud entre documents és baixa, i no afecta la mètrica. Per documents text en general on les seves longituds són més variables, és recomana fer servir la similitud del cosinus.

3.6 Detecció de temàtiques per mètode Latent Dirichlet Allocation.

Per tal de disposar d'un càlcul de referència alternatiu per les temàtiques detectades en aquesta anàlisi, s'exposa breument un cas senzill d'aplicació del mètode Latent Dirichlet Allocation (LDA) sobre el dataset de modelització millorat.

LDA és un model probabilístic generatiu, basat en la idea que cada document pot descriure's mitjançant una distribució de temes i cada tema pot descriure mitjançant una distribució de paraules.

Utilitzem la llibreria gensim per calcular un diccionari de paraules a partir del text de tots els tuits utilitzant la funció **Dictionary** del mòdul **corpora**. En aquest format el text es pot passar a la funció **doc2bow** que crea el comptatge de paraules i després inicialitzar el model **Lda**, tal com es pot observar en el primer bloc de la figura 73. Un cop creat el model, podem indicar el nombre de temes que volem que el model detecti. La sortida del model determina un conjunt de temes, formats per un grup de paraules ordenades per un coeficient de significació, que utilitzem per determinar cada temàtica. Si simplifiquem la sortida podem visualitzar la llista referent equivalent als clústers que s'han analitzat anteriorment en els tres tipus d'algorismes d'aprenentatge no supervisat, com es pot veure a la figura 74.

```
doc_clean=tuits.text
doc_clean=list(doc_clean.str.split())

dictionary = corpora.Dictionary(doc_clean)
doc_term_matrix = [dictionary.doc2bow(doc) for doc in doc_clean]
doc_term_matrix

Lda = gensim.models.ldamodel.LdaModel
```

```
# executem LDA per 15 temes
ldamodel = Lda(doc_term_matrix, num_topics=15, id2word =dictionary, passes=50)

print(ldamodel.print_topics())
d15=pd.DataFrame(ldamodel.print_topics(),columns=["topic","text"])
d15
```

```
[(0, '0.017*"wish" + 0.017*"create" + 0.016*"right" + 0.015*"national" + 0.012*"ings" + 0.012*"matter" + 0.012*"received" + 0.011*"man" + 0.010*"oh" + 0.010*"kids"'),
(1, '0.051*"happy" + 0.026*"opportunity" + 0.024*"full" + 0.019*"might" + 0.018*"ill" + 0.018*"research" + 0.018*"care" + 0.018*"life" + 0.017*"geer" + 0.017*"opportunit
ies"'), (2, '0.044*"ank" + 0.021*"great" + 0.020*"much" + 0.020*"ll" + 0.017*"video" + 0.016*"amazing" + 0.014*"good" + 0.013*"work" + 0.012*"via" + 0.011*"done"'), (3,
'0.036*"s" + 0.014*"morrow" + 0.012*"celebrated" + 0.012*"international" + 0.010*"la" + 0.009*"patient" + 0.009*"research" + 0.008*"year" + 0.008*"event" + 0.008*"satu
r"'), (4, '0.029*"care" + 0.026*"heal" + 0.022*"campaign" + 0.018*"access" + 0.017*"social" + 0.016*"diagnosis" + 0.015*"s" + 0.015*"treatment" + 0.015*"event" + 0.013
*"official"'), (5, '0.036*"love" + 0.024*"little" + 0.018*"sick" + 0.017*"want" + 0.015*"bit" + 0.014*"op" + 0.012*"de" + 0.010*"happen" + 0.010*"put" + 0.009*"send"'),
(6, '0.026*"know" + 0.023*"one" + 0.016*"s" + 0.015*"people" + 0.014*"like" + 0.012*"also" + 0.012*"life" + 0.011*"many" + 0.011*"even" + 0.010*"need"'), (7, '0.024*"syn
ome" + 0.018*"disorder" + 0.017*"diagnosed" + 0.016*"cancer" + 0.015*"called" + 0.013*"condition" + 0.013*"genetic" + 0.012*"type" + 0.011*"caes" + 0.010*"like"'), (8,
'0.048*"anks" + 0.046*"late" + 0.035*"daily" + 0.024*"coronavir" + 0.019*"news" + 0.010*"unfortunately" + 0.010*"intereing" + 0.009*"oand" + 0.008*"defined" + 0.008*"hou
rs"'), (9, '0.013*"new" + 0.012*"time" + 0.012*"find" + 0.012*"looking" + 0.011*"s" + 0.010*"better" + 0.010*"link" + 0.010*"research" + 0.010*"synome" + 0.009*"spai
n"'), (10, '0.045*"read" + 0.030*"ory" + 0.027*"sharing" + 0.020*"beautiful" + 0.020*"become" + 0.019*"families" + 0.019*"experience" + 0.018*"supporting" + 0.017*"hope"
+ 0.015*"ar"'), (11, '0.028*"support" + 0.022*"help" + 0.020*"awareness" + 0.016*"ank" + 0.012*"please" + 0.011*"share" + 0.010*"get" + 0.010*"work" + 0.009*"raise" + 0.
009*"patients"'), (12, '0.052*"s" + 0.047*"people" + 0.031*"awareness" + 0.027*"million" + 0.022*"living" + 0.019*"patients" + 0.018*"raise" + 0.013*"support" + 0.012*"w
ide" + 0.012*"many"'), (13, '0.015*"child" + 0.011*"school" + 0.010*"lack" + 0.010*"europe" + 0.010*"old" + 0.010*"son" + 0.008*"life" + 0.008*"mean" + 0.008*"eye" + 0.0
08*"born"'), (14, '0.018*"remember" + 0.017*"forget" + 0.013*"big" + 0.012*"make" + 0.012*"s" + 0.012*"good" + 0.010*"ay" + 0.010*"re" + 0.010*"play" + 0.009*"march"')]
```

Figura Nº 73: Implementació de la detecció de temàtiques pel model Latent Dirichlet Allocation.

```
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer

import gensim
from gensim import corpora
```

```
import re
def temes(d,prob=False):
    clusters=[]
    l=[]
    for i in range(0,d.shape[0]):
        if prob:
                print("Cluster {}: {}".format(i,d.text[i]))
                clusters.append(i)
                l.append(d.text[i])
        else:
            clusters.append(i)
            paraules=[k.replace('"','') for k in re.findall(r'\".*?\"',d.text[i])]
            l.append(paraules)
            print("Cluster {}: {}".format(i,paraules))

    r=pd.DataFrame({'clusters':clusters,'paraules':l})
    return r
```

```
temes(d15)
```

```
Cluster 0: ['wish', 'create', 'right', 'national', 'ings', 'matter', 'received', 'man', 'oh', 'kids']
Cluster 1: ['happy', 'opportunity', 'full', 'might', 'ill', 'research', 'care', 'life', 'geer', 'opportunities']
Cluster 2: ['ank', 'great', 'much', 'll', 'video', 'amazing', 'good', 'work', 'via', 'done']
Cluster 3: ['s', 'morrow', 'celebrated', 'international', 'la', 'patient', 'research', 'year', 'event', 'satur']
Cluster 4: ['care', 'heal', 'campaign', 'access', 'social', 'diagnosis', 's', 'treatment', 'event', 'official']
Cluster 5: ['love', 'little', 'sick', 'want', 'bit', 'op', 'de', 'happen', 'put', 'send']
Cluster 6: ['know', 'one', 's', 'people', 'like', 'also', 'life', 'many', 'even', 'need']
Cluster 7: ['synome', 'disorder', 'diagnosed', 'cancer', 'called', 'condition', 'genetic', 'type', 'caes', 'like']
Cluster 8: ['anks', 'late', 'daily', 'coronavir', 'news', 'unfortunately', 'intereing', 'oand', 'defined', 'hours']
Cluster 9: ['new', 'time', 'find', 'looking', 's', 'better', 'link', 'research', 'synome', 'spain']
Cluster 10: ['read', 'ory', 'sharing', 'beautiful', 'become', 'families', 'experience', 'supporting', 'hope', 'ar']
Cluster 11: ['support', 'help', 'awareness', 'ank', 'please', 'share', 'get', 'work', 'raise', 'patients']
Cluster 12: ['s', 'people', 'awareness', 'million', 'living', 'patients', 'raise', 'support', 'wide', 'many']
Cluster 13: ['child', 'school', 'lack', 'europe', 'old', 'son', 'life', 'mean', 'eye', 'born']
Cluster 14: ['remember', 'forget', 'big', 'make', 's', 'good', 'ay', 're', 'play', 'march']
```

Figura Nº 74: Implementació de la detecció de temàtiques pel model Latent Dirichlet Allocation

3.7 Avaluació de resultats en els algorismes d'agrupament.

Com s'ha comentat la dificultat d'avaluar el resultat del procés d'agrupament, resideix en què no es disposa de la solució òptima. Per tant cal utilitzar criteris que ens certifiquin no la solució òptima sinó una bona qualitat dels grups essent així probable que estiguem molt a prop de la solució òptima. Aquest procés es coneix com a validació de l'agrupament (en anglès cluster validation). A (Palacio-Niño & Berzal, 2019) es proposen diverses consideracions per la validació dels grups obtinguts, dels quals destaquem els utilitzats pel cas d'agrupament de documents text:

1. Cerciorar-se de què el dataset estudiat sigui agrupable.
   Per aquest punt, s'ha visualitzat sempre els resultats aplicant la reducció de dimensionalitat per PCA i s'ha representat en 2 dimensions.
2. Determinar el correcte nombre de clústers.
   S'han aplicat índexs per mesurar la qualitat interior dels clústers (figures 31 i 40) per l'aplicació de l'algorisme de KMeans, i s'han cercats valors òptims pels paràmetres de l'algorisme DBSCAN (figura 51), amb aquestes referències s'ha utilitzat l'algorisme jeràrquic aglomeratiu amb varietat d'enllaços. S'han obtingut els agrupaments de tuits, partint de paràmetres òptims. Valorant els resultats dels tres tipus d'algorismes, el nombre de clústers òptim el podem situar en 15.
3. Comprovar la qualitat dels clústers o grups.
   Implícitament en els processos d'optimització del punt anterior s'han considerat les característiques de compactació interna o cohesió interna de cada clúster conjuntament amb la característica de separació màxima entre clústers.

4. Comparació de resultats per determinar el millor.
   S'han elaborat visualitzacions per diversos valors òptims possibles pels paràmetres de cada algorisme, en el cas de KMeans per diferents valors de k, per l'algorisme DBSCAN diferents valors dels paràmetres eps i min_samples i per l'algorisme jeràrquic aglomeratiu diferents valors del nombre de clústers, en base a talls sobre el seu dendrograma, per diferents tipus de mètrica i enllaç de distància.

El criteri d'avaluació final que s'ha seguit, ha estat la validació de la separabilitat de temàtiques en els grups obtinguts. Per fer-ho s'han considerat les paraules més properes als centroides dels grups trobats per tal d'obtenir les paraules més significatives de cada tuit, determinant les temàtiques dominants en cada clúster.

Sota aquesta base operativa, l'agrupament, ha estat orientat ha obtenir el conjunt de temàtiques més significatives del dataset de modelització millorat. format per un conjunt de tuits, on l'usuari és únic i el contingut text conté la suma de tots els seus tuits.

Implícitament en el procés d'agrupament, s'han calculat també les comunitats d'usuaris existents. Però, d'acord amb el criteri inicial d'aquest estudi, de mantenir la privacitat dels usuaris i preservar totes les seves dades personals, no s'ha fet més referència al graf subjacent.

Per avaluar els resultats s'han de tenir en compte les dificultats detectades en l'exploració inicial, i les heretades de la font de dades. Els documents o tuits són textos molt curts, molts cops gairebé sense text, el que fa que al vectoritzar disposem d'una matriu molt dispersa on el nombre de paraules representatives o importants puguin ser molt repetitives en moltes temàtiques. Les millores introduïdes després de la primera temptativa han millorat els resultats finals.

Per KMeans s'han obtingut divisions de grups incorrectes, i no s'han obtingut divisions ideals, però a la pràctica s'ha assolit l'objectiu global amb un bon rendiment final en l'obtenció de temes significatius. En canvi, amb l'algorisme DBSCAN, si s'ha obtingut l'agrupament correcte en els casos on els clústers que l'algorisme KMeans no ha agrupat bé, però el fet que la densitat entre clústers fos molt diferent, ha penalitzat la detecció de temàtiques al 100%. Per aconseguir-ho s'ha comprovat que és necessari aplicar l'algorisme configurant els paràmetres eps i min_samples en cada cas. Tot i això s'ha trobat però una bona configuració per la detecció de temàtiques importants.

En el cas de l'algorisme aglomeratiu hem obtingut resultats satisfactoris pel cas d'enllaç ward o minimització de la variància. Però en conjunt la separabilitat de temàtiques ha estat més costosa. Els clústers detectats però, en molts casos han estat de força més detall que amb els algorismes KMeans i DBSCAN.

Com a avaluació final dels resultats, i de l'experiència de l'estudi dels tres tipus d'algorismes, seria molt útil utilitzar una estratègia conjunta basada en l'aplicació

de l’algorisme KMeans per la identificació inicial dels clústers i la utilització de l’algorisme de DBSCAN per la detecció d’aquells clústers resultat de la teselació incorrecta de KMeans. La utilitat dels algorismes jeràrquics resideix en la facilitat d’aplicar altres tipus de mètriques de distància diferents a la distància euclidiana, com són les distàncies de mahalanobis, jaccard o similitud del cosinus. Per tant poden aportar informació molt útil. En l’estudi s’ha presentat l’anàlisi usant la similitud del cosinus amb resultats semblants als de la distància euclidiana.

Finalment com a element de comparació afegit, s’han calculat les temàtiques del dataset de modelització amb millores mitjançant el mètode Latent Dirichlet Allocation, i hem obtingut resultats similars als trobats en l’estudi, confirmant el bon signe dels resultats finals obtinguts.

4. Conclusions.

En aquest treball, hem afrontat el repte d'analitzar un volum gran de dades en format de tuits de la xarxa social Twitter. La forma de fer-ho, ha estat analitzant el seu contingut text. Altres opcions possibles són l'ús d'informació del perfil dels usuaris, o informació de context com hahstags, URL's, nombre de seguidors o relacions. En canvi, l'estudi s'ha centrat a fer l'anàlisi en la temàtica de les malalties minoritàries sobre la base del contingut text capturat, utilitzant tècniques de processament de llenguatge natural i models d'aprenentatge automàtic no supervisat, complementats amb una anàlisi de sentiments sobre el text.

La captura de dades en temps real mitjançant 'streaming' i l'ús d'una base de dades documental a funcionat força bé. Com a possible millora, es pot implementar l'emmagatzemament en la base de dades en temps real.

Els punts crítics de l'estudi han estat, gestionar l'allau de tuits del dia 29 de febrer, la traducció de tuits a l'idioma anglès, la definició d'un corpus de dades òptim i disposar d'una matriu de vectorització sense soroll, el menys dispersa possible. Ha estat necessari establir millores sobre el dataset inicial, per l'obtenció de grups o clústers de qualitat on disposar d'una bona separabilitat de temes existents als tuits capturats.

Els resultats obtinguts han estat satisfactoris quant als objectius fixats d'obtenir les temàtiques més importants o significatives, i afinar el coneixement sobre opinions a favor i en contra durant el període d'estudi, en els hashtags monitorats.

Aspectes d'interès o d'on podem extreure coneixement:

- El pic màxim de tuits va ser de 140 tuits per minut, durant el dia 29 de febrer al voltant de les 16:00h, i el registre màxim per hora de 3530, aquest mateix dia.
- En mitjana les franges horàries de 08:00h a 09:00h al matí i 16:00h a 17:00h són les de més audiència per executar difusió de missatges.
- La visualització del wordcloud sobre el text del dataset inicial de modelització ha mostrat molt bé, les paraules més usades i ens ha proporcionat una primera impressió de les temàtiques existents.
- Per afinar aquesta primera impressió inicial els models d'aprenentatge no supervisat també assolint els objectius globals, KMeans detectant el conjunt de temàtiques globals, i DBSCAN complementant l'algorisme KMeans determinant realment la forma dels grups existents i els seus centroides. L'algorisme jeràrquic aglomeratiu ha funcionat bé en l'anàlisi de detall i no tant en la detecció inicial de grups.

Les claus importants en l'anàlisi efectuada resideixen en obtenir una matriu de vectorització tf-idf amb un vocabulari ben definit. Cal reduir l'existència de mots amb sintaxi correcta però sense significat, i determinar una funció de similitud òptima. En aquest treball s'han utilitzat mètrica euclidiana i similitud del cosinus

però queda per futurs treballs l'ús d'índexs que combinin diferents característiques de les relacions entre usuaris i d'informació de context. També és interessant en futurs treballs l'aplicació de tècniques semi supervisades on els grups o clústers es calculen amb restriccions sobre els documents.

Per finalitzar aquest apartat de conclusions finals i el treball de final de màster, fem una valoració dels resultats obtinguts en vers el benefici social respecte la seva aplicabilitat en la lluita contra les malalties minoritàries.

L'anàlisi presentat, ens ha permès capturar patrons de denúncies, peticions de suport, sensacions i sentiments personals dels pacients, molt útils per la presa de decisions. Sobre com actuar en el futur o posar èmfasi en l'assistència a pacients o en la modificació del full de ruta de les entitats que donen suport i lluiten contra les malalties minoritàries. D'igual manera disposem del coneixement de quan i com incrementar la difusió dels temes d'interès.

Com a idea i proposta final, donar suport a l'incipient canvi d'hàbit de l'ús més freqüent de les xarxes socials per la difusió de campanyes i esdeveniments i habilitar una segona direcció per la captació de patrons existents. De la mateixa manera que en una conversa és important parlar però encara més escoltar, escoltar les xarxes socials i desenvolupar projectes que es concentrin en detectar les necessitats dels usuaris. En aquest estudi s'ha comprovat què cal més recerca, investigació, diagnòstics en temps inferiors a un any, nous tractaments i conèixer com se senten i evolucionen en el temps els pacients i famílies que pateixen malalties minoritàries en solitud.

5. Glossari.

Agrupament (en anglès clustering)
Procediment d'associació d'elements individuals en conjunts d'elements utilitzant algun tipus de similitud quantitativa o qualitativa. Específicament en aprenentatge automàtic no supervisat s'apliquen diferents tipus d'algorismes que consideren diferents formes d'associació. La màxima característica en aquests casos és que no es disposa a priori cap etiquetat inicial dels individus. També es pot usar per trobar valors atípics.

Agrupament jeràrquic
Un mètode d'agrupament jeràrquic pot ser aglomerat o divisiu, segons si la descomposició jeràrquica es forma de manera descendent (fusió)(en anglès top down) o de dalt a baix (divisió) (en anglès bottom up).

Agrupament jeràrquic aglomerat (en anglès agglomerative hierarchical clustering)
Utilitza una estratègia de baix a dalt. Comença normalment deixant que cada objecte formi el seu propi clúster i fusioni iterativament els grups en grupss més grans i grans, fins que tots els objectes es troben en un sol grup o es satisfan certes condicions de terminació. Per al pas de fusió, troba els dos grups més propers (segons alguna mesura de similitud) i els combina per formar un sol grup.

Agrupament jeràrquic divisiu (en anglès divisive hierarchical clustering)
Utilitza una estratègia de dalt a baix. Comença posant tots els objectes en un grup (en anglès clúster), que és l'arrel de la jerarquia. Després divideix el grup arrel en diversos subgrups més petits i, de manera recursiva, els divideix en altres. El procés de particions continua fins que cada grup al nivell més baix és prou coherent (contenen un sol objecte o els objectes d'un grup són prou semblants entre si).

Anàlisi de Sentiment (en anglès Sentiment Analysis)
L'anàlisi de sentiment es refereix a l'ús del processament de llenguatge natural, anàlisi de text i lingüística computacional per identificar i extreure informació subjectiva, positiva, negativa o neutra dels recursos de forma quantificable.

Aprenentatge automàtic no supervisat
Explorar grans quantitats de dades per extreure informació significativa sobre l'estructura del seu grup considerant les similituds i diferències entre les entitats de dades (en anglès unsupervised learning).

BJSON
Versió comprimida del format JSON.

Bossa de paraules (en anglès 'Bag of Words')
Conjunt de paraules úniques d'un text. S'associa a disposar del compteig de la freqüència d'aparició de cada paraula en un conjunt de textos o corpus. Molts cops la bossa de paraules estarà optimitzada i només contindrà les paraules importants o significatives del text.

Centroide

En un context de mètodes d'agrupament no supervisats, el centroide de cada grup, es aquell punt que el representa i es caracteritza per què la suma de totes les distàncies amb els elements del grup és mínima.

Comunitat social
Una xarxa social és una estructura social formada per individus o bé entitats considerats 'nodes' que estan units per un o més tipus d'interdependència com ara amistat, parentesc, interessos comuns, intercanvis financers, relacions sexuals, creences, coneixements o prestigi.

Comunitat virtual
Les comunitats virtuals són espais de comunicació, intercanvi i treball col·laboratiu amb l'objectiu de potenciar el treball en xarxa, compartir coneixements i experiències, generar debats i elaborar documents, així com fer valdre el treball que s'està fent en benestar social al món local.

Corpus (en anglès el plural és corpora)
En el context de l'àrea de processament de llenguatge natural, descriu una col·lecció de textos. En aquest document fa referència al conjunt de continguts text dels tuits capturats, que han estat o no agrupats per autor.

CRISP-DM (Cross-Industry Standard Process for Data Mining)
Metodologia provada a la indústria per l'aplicació guiada de projectes de mineria de dades.

Densitat
En un context de mètodes d'agrupament no supervisats, la densitat d'un objecte ***k*** es pot mesurar pel nombre d'objectes propers a ***k***.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
És un algorisme d'agrupament per densitat. Té per objectiu trobar objectes centrals, és a dir, objectes que tenen veïnats densos. Connecta objectes centrals i els seus veïnats per formar regions denses com a grups o clústers.

Dendrograma (en anglès dendrogram)
Diagrama de dades en forma d'arbre que organitza les dades en subcategories que es van dividint en unes altres fins a arribar al nivell de detall desitjat.

ELT (en anglès acrònim de Extract Load Transform)
Tècnica que fa referència a com obtenim les dades d'una font i les emmagatzemem en un magatzem de dades. ELT primer extreu les dades, després les carrega en el magatzem i finalment les transforma. Aquesta tècnica no perd informació però fa despesa d'espai. En contraposició a ETL, que primer transforma les dades que extreu i després les càrrega en el magatzem de dades, estalviant espai però amb possible pèrdua d'informació original.

Emoticona (originari del japonès emoji= e [**dibuix**] **+ moji** [**caràcter**]**)**
Conjunt de símbols utilitzats en missatges electrònics que fan la funció d'ideogrames, de manera que mitjançant un dibuix es pot expressar sentiments si són cares de persones o conceptes si són icones gràfiques.

Fluxe de dades (en anglès streaming data)
Els fluxes de dades són dades en temps real o quasi real. Són sistemes capacitats per rebre dades d'un emissor i retransmetre-les a un receptor de manera asíncrona, permetent que emissor i receptor no perdin la comunicació.

Ground truth
En aprenentatge automàtic, fa referència a la precisió de la classificació del conjunt d'entrenament per a tècniques d'aprenentatge supervisat.

Hashtag
Etiqueta utilitzada en els missatges electrònics de la xarxa Twitter per denotar una temàtica o concepte. Es composa del caràcter '#' seguit d'una seqüència alfanumèrica.

Histograma
Representació gràfica d'una variable en forma de barres, on la superfície de cada barra és proporcional a la freqüència dels valors representats.

Isotròpic
Característica dels objectes físics amb propietats idèntiques en totes direccions.

JSON (acrònim de JavaScript Object Notation)
Format de texte senzill per l'intercanvi d'informació.

K-Means
És un algoritme de classificació no supervisat o de segmentació de tipus partional, és a dir, que genera una estructura plana. L'algoritme està basat en la partició d'un conjunt de n observacions en grups de manera que cada observació pertany al grup amb menor distància al seu centroide o punt central en cada instant de l'execució.

Latent Dirichlet Allocation (LDA)
És model probabilístic generatiu per a col·leccions de dades discretes com els corpus de text. LDA és un model bayesià jeràrquic de tres nivells, en el qual cada tema d'una col·lecció es modelitza com a barreja finita sobre un conjunt de temes subjacents. Al seu torn, cada tema es modelitza com una barreja infinita sobre un conjunt subjacent de probabilitats de temes. En el context del model de text, les probabilitats del tema proporcionen una representació explícita d'un document.

Lematització (en anglès lemmatization)
Tècnica usada per obtenir la representació en forma de mot de totes les seves formes flexionades (plural, femení, conjugat). Pot ser morfològica i sintàctica on es té en compte el context on apareix la paraula. S'usa conjuntament amb la tècnica de paraules derivades ('stemming'), que cerca el mateix objectiu.

Mineria de xarxes socials (**en anglès Social Data Mining**)
Integra redes sociales, análisis de redes sociales y minería de datos para proporcionar una plataforma conveniente y coherente para el estudio. Abarca las

herramientas para representar formalmente, medir, modelar y extraer patrones significativos de datos de redes sociales a gran escala.

NLP (Natural Language Processing)
És un subcamp d'informàtica que se centra a permetre als ordinadors d'entendre el llenguatge d'una manera "natural", com fan els humans, com entendre el sentiment del text, el reconeixement de la parla i generar respostes a preguntes.

NLTK (Natural Language Toolkit)
Mòdul o llibreria de programació de codi obert en llenguatge Python, que permet utilitzar algunes de les funcions més bàsiques de NLP, així com models pretractats per a diferents tasques.

NoSQL
Tipus de base de dades alternativa als tradicionals sistemes relacionals, aportant millores per la gestió de dades massives i el seu ús en sistemes distribuïts propis de les aplicacions web. Es caracteritza per la no imposició d'esquema i per la seva escalabilitat. Conceptualment compromet temporalment la seva consistència en vers la seva disponibilitat i capacitat tolerant a falles.

No Supervisat (en anglès unsupervised)
En aprenentatge automàtic, concepte associat per contraposició als mètodes supervisats on existeix un valor solució a un de cercat. Quan no existeix el valor correcte i no podem comprovar el grau de qualitat del valor cercat, es classifiquen com no supervisats. En aquest document els algorismes d'agrupament són no supervisats.

Particionament estricte
Entenem particionament com un sinònim de agrupament. I per estricte quan els grups obtinguts són disjunts.

Particionament difús
Respecte el particionament estricte, en aquest cas, els grups obtinguts si poden compartir elements, o tècnicament permet solapaments.

PCA (acrònim de Principal Component Analysis)
Tècnica per aconseguir reduir el nombre de variables d'un conjunt de dades multivariat, identificant un nombre de variables menor amb una pèrdua d'informació mínima.

Polaritat
En aquest document, terme aplicat a l'àmbit de l'anàlisi de sentiment. La polaritat descriu el grau d'emotivitat d'un text, classificant-lo amb un valor continu en un rang de valors entre -1 i 1, on podem identificar emotivitat negativa amb el valor -1, neutra amb valor 0 i positiva amb valor 1.

Subjectivitat
En aquest document, terme aplicat a l'àmbit de l'anàlisi de sentiment. La subjectivitat descriu el grau d'opinió o expressió d'un sentiment per part de l'autor d'un text (càrrega subjectiva), classificant-lo amb un valor continu en un rang de

valors entre 0 i 1, on podem identificar absència de subjectivitat amb el valor 0, i màxima subjectivitat amb valor 1.

Tf-idf (Term frequency - inverse data frequency)
És una estratègia per assenyalar la importància relativa de les paraules. La freqüència inversa de dades determina el pes de les paraules rares a tots els documents.

Token
Anomenem token a cadascun dels elements obtinguts en un procés de 'tokenització'.

Tokenització (en anglès Tokenitzation)
A la mineria de text o a la cerca de text complet, el procés d'identificació d'unitats significatives dins de cadenes, ja sigui en els límits de paraules, morfemes o tiges, de manera que es poden agrupar els testimonis relacionats, p. ex. tot i que "San Francisco" són dues paraules, es pot tractar com un sol token.

Tokenization is a step which splits longer strings of text into smaller pieces, or tokens. Larger chunks of text can be tokenized into sentences, sentences can be tokenized into words, etc. Further processing is generally performed after a piece of text has been appropriately tokenized. Tokenization is also referred to as text segmentation or lexical analysis. Sometimes *segmentation* is used to refer to the breakdown of a large chunk of text into pieces larger than words (e.g. paragraphs or sentences), while *tokenization* is reserved for the breakdown process which results exclusively in words.

Twitter
És un servei de microblog que permet als seus usuaris enviar i llegir missatges de text d'una llargada màxima de 280 caràcters anomenats tuits o piulades.

Vectorització
Tècnica emprada en aprenentatge automàtica amb l'objectiu de represesentar numèricament les dades. En aquest document, el text dels tuits s'han representat numèricament amb vectors (tf-idf) per poder avaluar la seva similitud i aplicar els models d'agrupament.

Word cloud
Un núvol de paraules (també anomenat núvol d'etiquetes o llista ponderada) és una representació visual de les dades de text. Les paraules solen ser paraules simples i la importància de cadascuna es mostra amb la mida o el color de la lletra.

6. Bibliografia.

Abascal-Mena, R., Lema, R., & Sèdes, F. (2015). Detecting sociosemantic communities by applying social network analysis in tuits. Social Network Analysis and Mining, 5(1), 38. https://doi.org/10.1007/s13278-015-0280-2

Akay, A., Dragomir, A., & Erlandsson, B.-E. (2015). Network-Based Modeling and Intelligent Data Mining of Social Media for Improving Care. IEEE Journal of Biomedical and Health Informatics, 19(1), 210-218. https://doi.org/10.1109/JBHI.2014.2336251

Alnajran, N., Crockett, K., McLean, D., & Latham, A. (2017). Cluster Analysis of Twitter Data: A Review of Algorithms (J. VanDenHerik, A. P. Rocha, & J. Filipe, Eds.). Scitepress. https://doi.org/10.5220/0006202802390249

Amat, J. (2017). Clustering y heatmaps: Aprendizaje no supervisado [Ciéncia de datos]. Clustering y heatmaps: aprendizaje no supervisado. https://www.cienciadedatos.net/documentos/37_clustering_y_heatmaps

Antelmi, A. (2018). Towards a More Systematic Analysis of Twitter Data: A Framework for the Analysis of Twitter Communities. 303-314.

Arora, P., Deepali, & Varshney, S. (2016). Analysis of K-Means and K-Medoids Algorithm For Big Data. Procedia Computer Science, 78, 507-512. https://doi.org/10.1016/j.procs.2016.02.095

Artificial Intelligence Research Institute (IIIA-CSIC), Campus de la UAB, E-08193 Bellaterra, Spain, Gromann, D., Declerck, T., & DFKI GmbH, Saarbrücken, Germany. (2017). Hashtag Processing for Enhanced Clustering of Tuits. RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning, 277-283. https://doi.org/10.26615/978-954-452-049-6_038

Baralis, E., Cerquitelli, T., Chiusano, S., Grimaudo, L., & Xiao, X. (2013). Analysis of Twitter Data Using a Multiple-level Clustering Strategy. En A. Cuzzocrea & S. Maabout (Eds.), Model and Data Engineering (pp. 13-24). Springer. https://doi.org/10.1007/978-3-642-41366-7_2

Bengfort, B., Bilbro, R., & Ojeda, T. (2018). Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning (Edición: 1). O'Reilly Media;

Center for Drug Evaluation and Research. (2019). *Rare Diseases: Common Issues in Drug Development Guidance for Industry*. U.S. Food and Drug Administration; FDA.http://www.fda.gov/regulatory-information/search-fda-guidance-documents/rare-diseases-common-issues-drug-development-guidance-industry-0

Darmon, D., Omodei, E., & Garland, J. (2015). Followers Are Not Enough: A Multifaceted Approach to Community Detection in Online Social Networks. PLOS ONE, 10(8), e0134860. https://doi.org/10.1371/journal.pone.0134860

Davies, W. (2016). Insights into rare diseases from social media surveys. Orphanet Journal of Rare Diseases, 11(1), 151. https://doi.org/10.1186/s13023-016-0532-x

De Boom, C., Van Canneyt, S., & Dhoedt, B. (2015). Semantics-driven event clustering in Twitter feeds. Proceedings of the 5th Workshop on Making Sense of Microposts, 1395, 2-9. http://hdl.handle.net/1854/LU-6887623

Diakopoulos, N., Elgesem, D., Salway, A., Zhang, A., & Hofl, K. (2015). Compare clouds: Visualizing text corpora to compare media frames. In Proc. of IUI Workshop on Visual Text Analytics.

Dutta, S., Das, A. K., Bhattacharya, A., Dutta, G., Parikh, K. K., Das, A., & Ganguly, D. (2019). Community Detection Based Tweet Summarization. A A. Abraham, P. Dutta, J. K. Mandal, A. Bhattacharya, & S. Dutta (Eds.), Emerging Technologies in Data Mining and Information Security (pp. 797-808). Springer. https://doi.org/10.1007/978-981-13-1498-8_70

EURORDIS. (2015). Pàgina web oficial Rare Diseases Europe. https://www.eurordis.org/es

Fani, H., Zarrinkalam, F., Bagheri, E., & Du, W. (2016). Time-Sensitive Topic-Based Communities on Twitter. Proceedings of the 29th Canadian Conference on Artificial Intelligence on Advances in Artificial Intelligence - Volume 9673, 192–204. https://doi.org/10.1007/978-3-319-34111-8_25

FECAMM. (2020). Pàgina oficial Federació Catalana de Malalties Minoritàries. http://www.fcmpf.entitatsbcn.net/

FEDER Catalunya. (2020). Pàgina web oficial. https://enfermedades-raras.org/index.php/feder-cataluna

FEDER. (2020). Pàgina web oficial Federación Española de las Enfermedades Raras. https://enfermedades-raras.org/

Friedemann, V. (2015). Clustering a Customer Base Using Twitter Data. 5.

Gabrielatos, C., & Marchi, A. (2012). Keyness: Appropriate metrics and practical issues. Paper presented at Corpus-assisted Discourse Studies International Conference, Italy.

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online. Journal of Medical Internet Research, 15(11), e239. https://doi.org/10.2196/jmir.2721

Han, Y. (2018). A Survey of the Application of Social Media Mining. https://doi.org/10.2991/iwmecs-18.2018.110

Hand, R. K., Kenne, D., Wolfram, T. M., Abram, J. K., & Fleming, M. (2016). Assessing the Viability of Social Media for Disseminating Evidence-Based Nutrition Practice Guideline Through Content Analysis of Twitter Messages and Health Professional Interviews: An Observational Study. Journal of Medical Internet Research, 18(11), e295. https://doi.org/10.2196/jmir.5811

Hawker, M. D. (2010). Developer's Guide to Social Programming: Building Social Context Using Facebook, Google Friend Connect, and the Twitter API, The (Edición: 1). AddisonWesley Professional; Social Media Mining.

Hu, Y. (2018). Geo-Text Data and Data-Driven Geospatial Semantics. ArXiv. https://doi.org/10.1111/gec3.12404

Infografia CRISP-DM. (2020). Recuperat de https://www.sv-europe.com/crisp-dm-infographic/

Johnston, J. E., Berry, K. J., & Mielke, P. W. (2006). Measures of effect size for chi-squared and likelihood-ratio goodness-of-fit tests. Perceptual and Motor Skills, 103(2), 412-414. https://doi.org/10.2466/pms.103.2.412-414

Kaur, N. (2015). A Combinatorial Tweet Clustering Methodology Utilizing Inter and Intra Cosine Similarity [Thesis, Faculty of Graduate Studies and Research, University of Regina]. https://ourspace.uregina.ca/handle/10294/6549

Lam, A. J. (2016). Improving Twitter Community Detection through Contextual Sentiment Analysis. ACL. https://doi.org/10.18653/v1/P16-3005

Ifrim, G., Shi, B., & Brigadir, I. (2014, 4 d'abril). Event Detection in Twitter using Aggressive Filtering and Hierarchical Tweet Clustering. Second Workshop on Social News on the Web (SNOW), Seoul, Korea, 8 abril 2014. https://researchrepository.ucd.ie/handle/10197/7546

Merinopoulou, E., & Cox, A. (2020). How Social Media Can Be Used to Understand What Matters to People with Rare Diseases. Evidera. Recuperat de https://www.evidera.com/how-social-media-can-be-used-to-understand-what-matters-to-people-with-rare-diseases/

Paez, M. S., Amini, A. A., & Lin, L. (2019). Hierarchical Stochastic Block Model for Community Detection in Multiplex Networks. arXiv:1904.05330 [cs, stat]. http://arxiv.org/abs/1904.05330

Palacio-Niño, J.-O., & Berzal, F. (2019). Evaluation Metrics for Unsupervised Learning Algorithms. arXiv:1905.05667 [cs, stat]. http://arxiv.org/abs/1905.05667

Palomino, M., Taylor, T., Göker, A., Isaacs, J., & Warber, S. (2016). The Online Dissemination of Nature–Health Concepts: Lessons from Sentiment Analysis of Social Media Relating to "Nature-Deficit Disorder". International Journal of Environmental Research and Public Health, 13(1). https://doi.org/10.3390/ijerph13010142

Papadopoulos, S., Kompatsiaris, Y., Vakali, A., & Spyridonos, P. (2012). Community detection in Social Media. Performance and application considerations. Data Mining and Knowledge Discovery, 24(3), 515-554. https://doi.org/10.1007/s10618-011-0224-z

Anumol, B., & Pattani, R. V. (2016). Efficient Density Based Clustering of Tuits and Sentimental Analysis Based on Segmentation. 3(3), 5.

Pew Research Center. (2011). Peer-to-Peer Health Care. Recuperat de: http://www.pewinternet.org/2011/02/28/peer-to-peer-health-care-2/. Accedit en març de 2020.

RareConnect. (s. f.). Recuperat març de 2020, de https://www.rareconnect.org/es/

Rare disease day. Recuperat febrer de 2020, de https://www.rarediseaseday.org/article/about-rare-disease-day

Roig, J. G., Roma, J. C., Alfonso, J. M., & Quiles, R. C. (2017). Minería de datos: Modelos y algoritmos (Edición: 1). Editorial UOC, S.L.

Rooney, E. (2016). HOW TO PROMOTE YOUR RARE DISEASE STORY THROUGH SOCIAL MEDIA. Recuperat març 20202 https://globalgenes.org/wp-content/uploads/2019/01/How-to-Promote-Your-Rare-Disease-Through-Social-Media.pdf

Serrano, E., Iglesias, C. A., & Garijo, M. (2015). A Survey of Twitter Rumor Spreading Simulations. En M. Núñez, N. T. Nguyen, D. Camacho, & B. Trawiński (Eds.), Computational Collective Intelligence (pp. 113-122). Springer International Publishing. https://doi.org/10.1007/978-3-319-24069-5_11

Shi, Q., Wang, Y., Sun, J., & Fu, A. (2018). Short Text Understanding Based on Conceptual and Semantic Enrichment. En G. Gan, B. Li, X. Li, & S. Wang (Eds.), Advanced Data Mining and Applications (pp. 329-338). Springer International Publishing. https://doi.org/10.1007/978-3-030-05090-0_28

Social media data collection tools—Social media data collection tools. (s. f.). Recuperado febrero de 2020, de http://socialmediadata.wikidot.com/

Subirats, L., Reguera, N., Bañón, A. M., Gómez-Zúñiga, B., Minguillón, J., & Armayones, M. (2018). Mining Facebook Data of People with Rare Diseases: A Content-Based and Temporal Analysis. International Journal of Environmental Research and Public Health, 15(9), 1877. https://doi.org/10.3390/ijerph15091877

Sundararaman, D., & Srinivasan, S. (2017). Twigraph: Discovering and Visualizing Influential Words Between Twitter Profiles. En G. L. Ciampaglia, A. Mashhadi, & T. Yasseri (Eds.), Social Informatics (Vol. 10540, pp. 329-346). Springer International Publishing. https://doi.org/10.1007/978-3-319-67256-4_26

Tang, L., & Liu, H. (2010). Community Detection and Mining in Social Media. Synthesis Lectures on Data Mining and Knowledge Discovery, 2(1), 1-137. https://doi.org/10.2200/S00298ED1V01Y201009DMK003

Tai, C.-H., Tan, Z.-H., & Chang, Y.-S. (2016). Systematical Approach for Detecting the Intention and Intensity of Feelings on Social Network. IEEE Journal of Biomedical and Health Informatics. https://doi.org/10.1109/JBHI.2016.2535721

Tweepy/tweepy. (2020). [Python]. Tweepy. https://github.com/tweepy/tweepy (Original work published 2009)

Twitter Developers. (s. f.). Recuperat febrer de 2020, de https://developer.twitter.com/en/apps

Vathi, E., Siolas, G., & Stafylopatis, A. (2017). Mining and categorizing interesting topics in Twitter communities. Journal of Intelligent & Fuzzy Systems, 32(2), 1265-1275. https://doi.org/10.3233/JIFS-169125

Welch, M. J., Schonfeld, U., He, D., & Cho, J. (2011). Topical semantics of twitter links. Proceedings of the fourth ACM international conference on Web search and data mining, 327–336. https://doi.org/10.1145/1935826.1935882

Weng, J., Lim, E.-P., Jiang, J., & He, Q. (2010). TwitterRank: Finding topic-sensitive influential twitterers. Proceedings of the Third ACM International Conference on Web Search and Data Mining - WSDM '10, 261. https://doi.org/10.1145/1718487.1718520

Wilson, A. C. (2013). Embracing Bayes factors for key item analysis in corpus linguistics. En M. Bieswanger & A. Koll-Stobbe (Eds.), New approaches to the study of linguistic variability (pp. 3-11). Peter Lang. https://eprints.lancs.ac.uk/id/eprint/51045/

Zafarani, R., Abbasi, M.A. i Liu H. (2014, juliol). Social Data Mining An Introduction. Cambridge University Press. www.cambridge.org/9781107018853

Zhang, Y., Wu, Y., & Yang, Q. (2012). Community Discovery in Twitter Based on User Interests. Journal of Computational Information Systems. 8 (3), 991-1000

Zhou, K., & Yang, Q. (2018). LDA-PSTR: A Topic Modeling Method for Short Text. ADMA. https://doi.org/10.1007/978-3-030-05090-0_29

7. Annexos.

7.1 Script codi principal del procés de captació de tuits.

En la figura 75, es pot revisar el script principal del procés de captació. Valida les claus de l'aplicació de captació definida en la web de Twitter per poder operar i capturar dades de manera controlada. Per fer-ho activa un objecte de la classe MyStreamListener que escolta indefinidament l'arribada dels tuits en temps real.

```
import tweepy
import codecs
import json
# Situem les claus de connexió en un modul apart per privacitat.
from secret1 import CONSUMER_KEY_KEY, CONSUMER_SECRET_KEY, ACCESS_TOKEN_KEY, ACCESS_TOKEN_SECRET_KEY;
from datetime import datetime

# Funció que definiex la validació de les claus per accedir a poder captar els tuits.
def get_auth():
    auth = tweepy.OAuthHandler(CONSUMER_KEY_KEY, CONSUMER_SECRET_KEY)
    auth.set_access_token(ACCESS_TOKEN_KEY, ACCESS_TOKEN_SECRET_KEY)
    return auth

# Codi principal de l'script

print(" ----- Twitter Procés Captació ----- ")

# Obtenim tuits usant tweepy
auth = get_auth()  # Recuperem un objecte 'auth' usant la funció 'get_auth':
api = tweepy.API(auth)  # Construim un objecte d'API.

# Connecta i escolta el flux de dades o 'stream'
myStreamListener = MyStreamListener()
myStream = tweepy.Stream(auth=api.auth, listener=myStreamListener)

print(">> Capturant tweets sobre:\n#DiaMundialEnfermedadesRaras\n#DíaMundialEnfermedadesRaras\n \
#SomosFEDER\n#EnfermedadesRaras\n#DMEnfermedadesRaras2020\n#DM2020")

myStream.filter(track=['#DiaMundialEnfermedadesRaras,#DíaMundialEnfermedadesRaras,#SomosFEDER, \
#EnfermedadesRaras,#DMEnfermedadesRaras2020,#DM2020'])
```

Figura Nº 75: Codi principal executat pel procés de captació.

7.2 Script d'implementació de la *class MyStreamListener*.

La implementació de la classe MyStreamListener, és la personalització per aquest estudi de la forma en què es capturen els tuits proporcionat per Twitter. En la figura 76, es pot revisar el script amb el codi en llenguatge Python documentat. En el codi es capturen les dades completes dels tuits en format JSON i s'emmagatzemen en format text al mateix temps que es monitoritzen per pantalla per pemetre el seguiment de tot el procés.

```python
# Classe que implementa el flus o stream on es reben els tuits per després ser emmagatzemats.
class MyStreamListener(tweepy.StreamListener):
    # Event que desencadena la captació i emmagatzematge en el mateix directori de l'script.
    def on_data(self, data):
        try:
            # Obtenció de la data en que es rep el tuit.
            now=datetime.now()
            # Definició del format de visualització d'aquesta data.
            data_m = now.strftime("%Y/%m/%d - %H:%M:%S")
            print(data_m)
            data_actual = now.strftime("%Y_%m_%d")
            # obtenció de l'estructura en json de dades del tuit.
            tweet = json.loads(data)
            # Extracció de cada camp d'interés del tuit, però per visualitzar
            # per pantalla i monitoritzar el procés.
            #ID únic del tuit
            id_tweet = tweet.get('id_str')
            # data de creació pròpia del tuit.
            data_tweet = str(tweet.get('created_at'))
            # Nom d'usuari
            user_tweet_name = tweet.get('user').get('name')
            # Àlies de l'usuari
            user_tweet_screen_name = tweet.get('user').get('screen_name')
            # Inspecció del camp 'truncated' per saber si la longitud es de 140 o de 280 caràcters
            # i obtenir el contingut del camp 'text' o del camp 'extended_tweet.full_text'.
            tweet_truncat=tweet.get("truncated")
            if tweet_truncat:
                text_tweet=tweet.get("extended_tweet").get("full_text")
            else:
                text_tweet = tweet.get('text')
            # Inspecció del tuit per saber si és un retuit.
            retweet="False"
            if tweet.get("retweeted_status") is not None:
                retweet = "True"
                if tweet.get("retweeted_status").get("truncated"):
                    text_tweet=tweet.get("retweeted_status").get("extended_tweet").get("full_text")
                else:
                    text_tweet = tweet.get("retweeted_status").get('text')

            if tweet.get("retweet_count")>0 or tweet.get("is_quote_status"):
                retweet = "True"
            else:
                retweet = "False"
            else:
                retweet = "False"
            # Impressió per pantalla de les dades descriptives del tuit
            # per la seva monitorització.
            print("-------------------------------------------------------")
            regTweet=""
            regTweet = "Tweet ID: " + id_tweet
            regTweet = regTweet + "\nData: " + data_tweet
            regTweet = regTweet + "\nUser_name: " + user_tweet_name
            regTweet = regTweet + "\nUser_screen_name: " + user_tweet_screen_name
            regTweet = regTweet + "\nTweet:\n" + text_tweet
            regTweet = regTweet + "\nReTweet: " + retweet

            print(regTweet)
            # Emmagatzematze a disc en el fitxer de notació: "TWEETS_RAW_(data).txt"
            # Exemple: TWEETS_RAW_2020_02_13.txt
            with codecs.open("TWEETS_RAW_"+ data_actual +".txt", "a", encoding="utf-8") as myfile:
                #json.dump(data, myfile, ensure_ascii=False)
                myfile.write(data)
                myfile.close()
            # Emmagatzematze a disc d'una selecció de dades, en el fitxer de notació: "TWEETS_(data).txt"
            # Exemple: TWEETS_2020_02_13.txt
            with codecs.open("TWEETS_"+ data_actual +".txt", "a", encoding="utf-8") as myfile1:
                registre='{"data": "' + data_tweet + '",\n'
                registre = registre + '"name": "' + user_tweet_name + '",\n'
                registre = registre + '"screen_name": "' + user_tweet_screen_name + '",\n'
                registre = registre + '"text": "' + text_tweet + '",\n'
                registre = registre + '"ReTweet: "' + retweet + '"},\n'
                myfile1.write(registre)
                myfile1.close()

        except Exception as e:
            print(data_m)
            print("ERROR: {}".format(e))
        finally:
            return True  # Resta l'script en execució permanent.
```

Figura Nº 76: Script Python de la classe ***MyStreamListener*** del procés de captació.

7.3 Còpia de seguretat i restauració de la base de dades documental.

En la figura 77, es mostra com s'ha realitzat i el resultat de la còpia de seguretat de la base de dades documental on s'han emmagatzemat tots els tuits capturats en temps real.

```
mongodump --host=localhost --port=27017 --collection=Twitter --db=DM_MM2020

writing DM_MM2020.Twitter to
[############............]  DM_MM2020.Twitter  55010/102632  (53.6%)
[########################]  DM_MM2020.Twitter  102632/102632  (100.0%)
done dumping DM_MM2020.Twitter (102632 documents)
```

Figura Nº 77: Còpia de seguretat de la base de dades documental MongoDB.

En la figura 78, es mostra com el procés de restauració de la base de dades en altre servidor diferent per accedir a les dades capturades.

```
mongorestore --host=localhost --port=7017

using default 'dump' directory
preparing collections to restore from
reading metadata for DM_MM2020.Twitter from dump\DM_MM2020\Twitter.metadata.json
restoring DM_MM2020.Twitter from dump\DM_MM2020\Twitter.bson
[##......................]  DM_MM2020.Twitter  64.9MB/698MB  (9.3%)
[#####...................]  DM_MM2020.Twitter  166MB/698MB  (23.7%)
[#########...............]  DM_MM2020.Twitter  263MB/698MB  (37.6%)
[############............]  DM_MM2020.Twitter  364MB/698MB  (52.2%)
[################........]  DM_MM2020.Twitter  466MB/698MB  (66.8%)
[####################.....]  DM_MM2020.Twitter  573MB/698MB  (82.0%)
[########################.]  DM_MM2020.Twitter  675MB/698MB  (96.6%)
[########################]  DM_MM2020.Twitter  698MB/698MB  (100.0%)
no indexes to restore
finished restoring DM_MM2020.Twitter (102632 documents)
done
```

Figura Nº 78: Restauració de la còpia de seguretat en altre servidor de dades.