TREBALL DE FINAL DE MÀSTER (Àrea Medicina)

Anàlisi de les dades del Dia Mundial de les Malalties Minoritàries a Twitter

Autor: Joaquim de Dalmases Juanet

Consultors: Laia Subirats Maté
Elisenda Bonet Carne

Juliol 2020

VISIÓ GENERAL DEL TREBALL

Treball

1. Introducció

- OBJECTIUS DEL TREBALL.
- PLANIFICACIÓ DEL TREBALL.
- PRODUCTES OBTINGUTS.

FASE	Inici	Final
Definició i planificació del treball final.	19/02/2020	**01/03/2020**
Estat de l'art o anàlisi del mercat.	02/03/2020	**22/03/2020**
Disseny i implementació del treball.	23/03/2020	**09/06/2020**
Redacció de la memòria.	10/06/2020	**24/06/2020**
Presentació i defensa del projecte.	25/06/2020	**30/06/2020**
Defensa Pública.	01/07/2020	**08/07/2020**

2. Marc de referència

'Estat de l'art'

- COM PODEN AJUDAR LES XARXES SOCIALS? PER QUÈ TWITTER?
- DETECCIÓ DE COMUNITATS / DETECCIÓ DE TEMÀTIQUES.
- ANÀLISI DE CONTINGUTS I ANÀLISI DE SENTIMENT.
- INCIDÈNCIA SOBRE LES MALALTIES MINORITÀRIES.

3. Disseny i implementació de l'anàlisi

- CAPTACIÓ I EMMAGATZEMATGE DE DADES.
- GENERACIÓ DEL DATASET DE DADES PREPROCESSAT.
- ANÀLISI DE LES DADES
  MÈTODES NO SUPERVISATS: TASCA D'AGRUPAMENT.
- MODELITZACIÓ AMB ELS ALGORISMES KMEANS, DBSCAN, JERÀRQUIC AGLOMERATIU.
- AVALUACIÓ DE RESULTATS.

4. Conclusions

OBJECTIUS DEL TREBALL

Tecnològics

1. Capturar dades en temps real de la xarxa social Twitter.
2. Aprenentatge automàtic amb mètodes no supervisats.

Socials

1. Detectar comunitats d'usuaris i temàtiques principals.
2. Contribuir a la lluita contra les Malalties minoritàries.

AVALUAR TECNOLOGIES PER L'ADQUISICIÓ D'INFORMACIÓ VALUOSA DE LA XARXA SOCIAL TWITTER

Detecció de patrons en tuits amb aprenentatge automàtic.

(Anàlisi de text/Anàlisi de sentiment)

Denúncies
Esdeveniments
Problemàtiques
Tendències

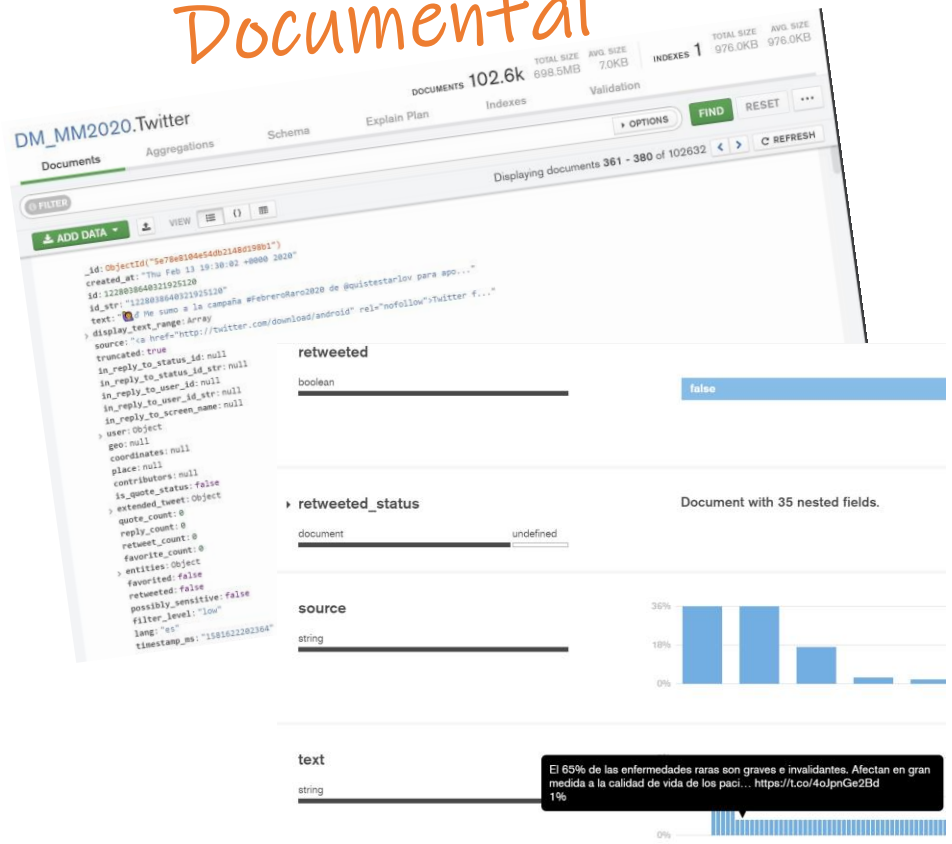
Inclusió en el full de Ruta de les entitats de suport.

Suport a Centres Mèdics

Suport a Pacients

PRODUCTES OBTINGUTS

Base de Dades Documental



102.632 tuits període **13/02/2020 – 30/03/2020**

Objecte original complet - API de Twitter

Dataset de Modelització

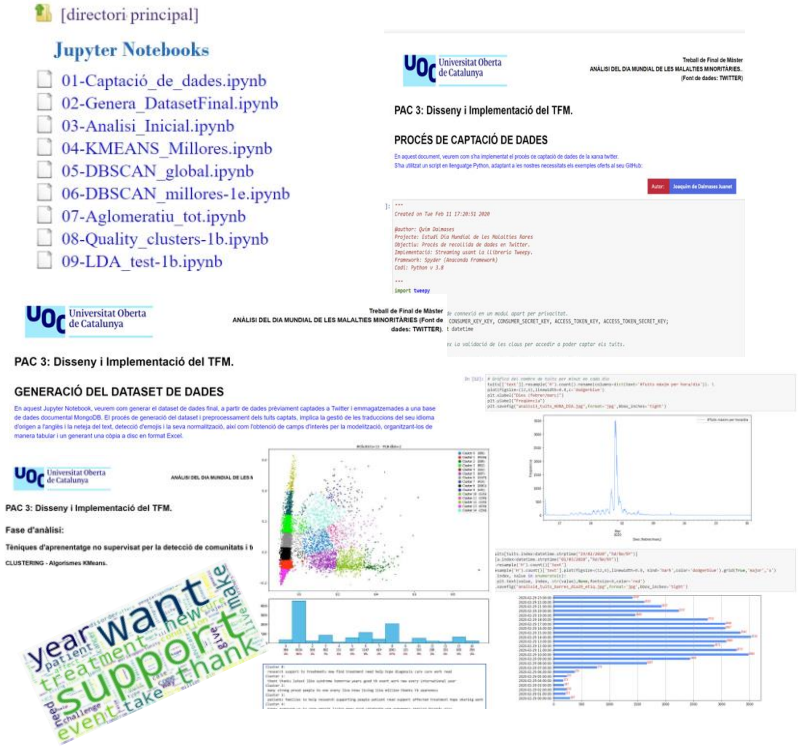
_id	Identificador de tuit únic assignat per la base de dades.
created_at	'Timestamp' (Data/Hora) de creació del tuit.
text_x	Text original del tuit, tal com es va capturar.
text_net	Text format per cadenes de caràcters sense puntuació, 'stopwords', ni dígits numèrics, cadenes amb nombre repetit d'espais 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.

Informació sintetitzada i processada.

1. Text original
2. Text traduït l'anglès, netejat, normalitzat
3. Camps Subjectivitat/Polaritat per anàlisi de sentiment.
4. Disponibilitat d'informació de context: hashtags, mencions a usuari, emojis

Material/Recursos I eines d'anàlisi

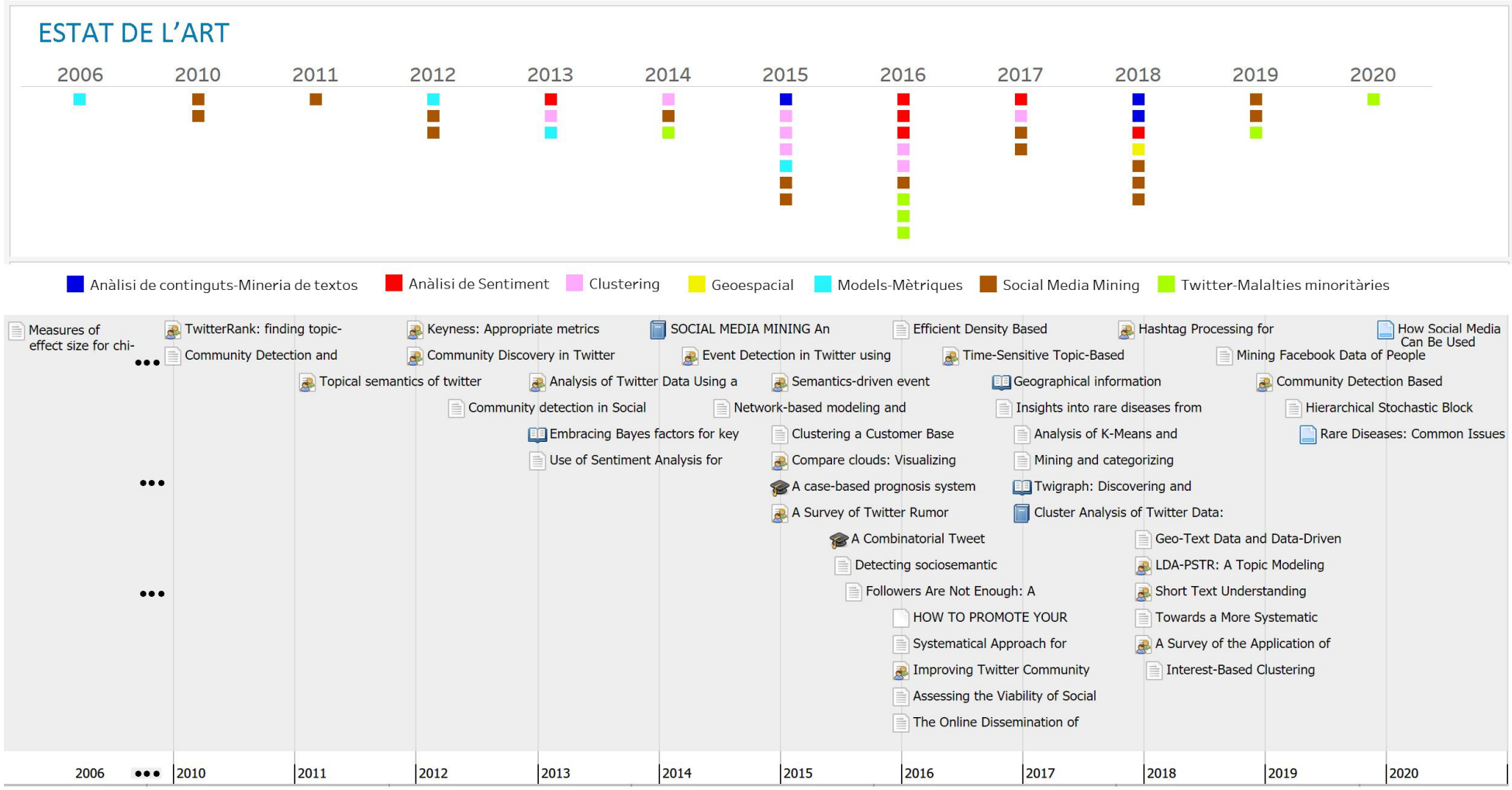
https://github.com/QuimDJ/TFM_DataScience_UOC



GitHub Accessible del projecte:

1. Jupyter Notebooks en format original i html.
2. Jupyter Notebooks addicionals.
3. Exportació en format XLSX de les dades Twitter.
4. Document de memòria TFM.

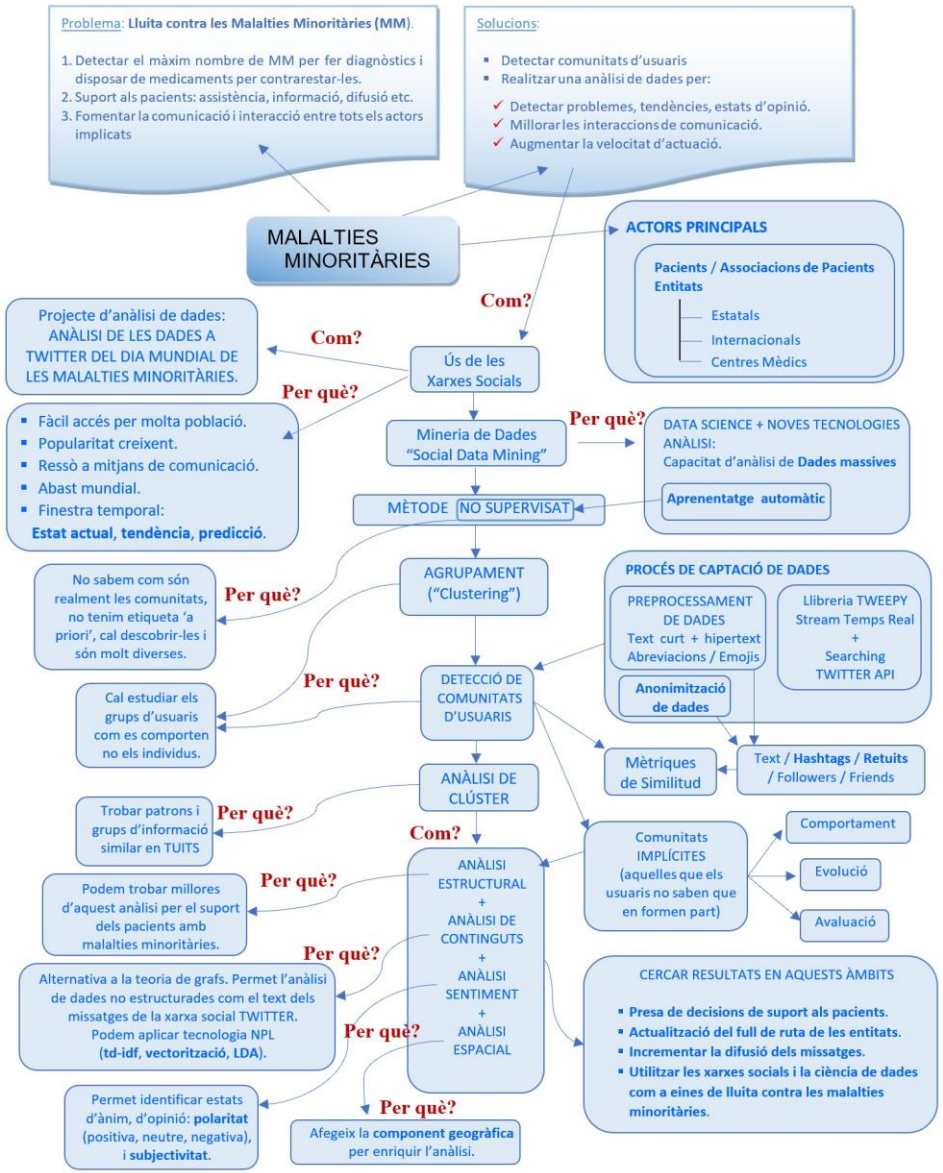
ESTAT DE L'ART



Infografia per data de publicació:

1. Notorietat de treballs a partir de 2012.
2. Producció més significativa de treballs esdevé entre els anys 2015 i 2018.

DISSENY DE L'ANÀLISI

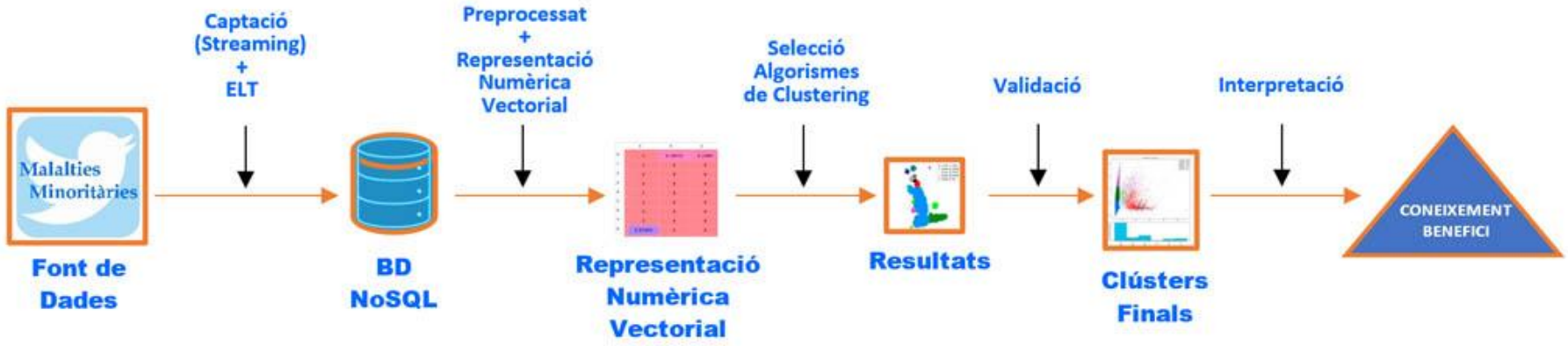


REPTES

1. Gestionar el contingut de l'allau de tuits.
2. Traducció de tuits a l'idioma anglès.
3. Representar numèricament els tuits de manera òptima, per poder conèixer de manera fiable, la seva distància o similitud.
4. Obtenció de grups o clústers de qualitat amb bona separabilitat de temàtiques.

PROCEDIMENT D'ANÀLISI

PIPELINE de l'anàlisi



Font de dades: La xarxa social Twitter. Missatges text de longitud limitada a 288 caràcters. Contingut: text, URL's, hashtags, emojis.

Captació: Flux de dades en temps real, sobre hashtags específics de la temàtica Malalties Minoritàries (MM).

Emmagatzematge en base de dades: NoSQL de tipus documental MongoDB.

Representació numèrica del text dels tuits: Vectorització del text utilitzant una estratègia TF-IDF (importància relativa de les paraules).

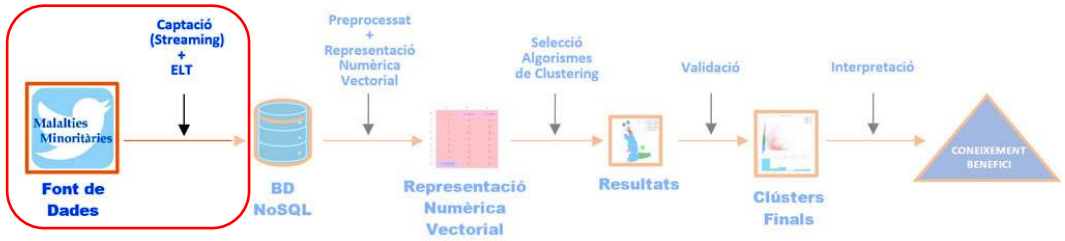
Modelització: Models d'aprenentatge no supervisat. Algorismes KMeans, DBSCAN i jeràrquic aglomeratiu. Agrupem tuits semblants o similars. Funció de similitud: Distància euclidiana sobre matriu TF-IDF, i similitud del cosinus

Validació de resultats: Optimització de paràmetres, per l'obtenció de grups el més compactes i distanciats entre si possibles.

Clústers finals: Obtenció de comunitats on trobem les temàtiques més ben definides.

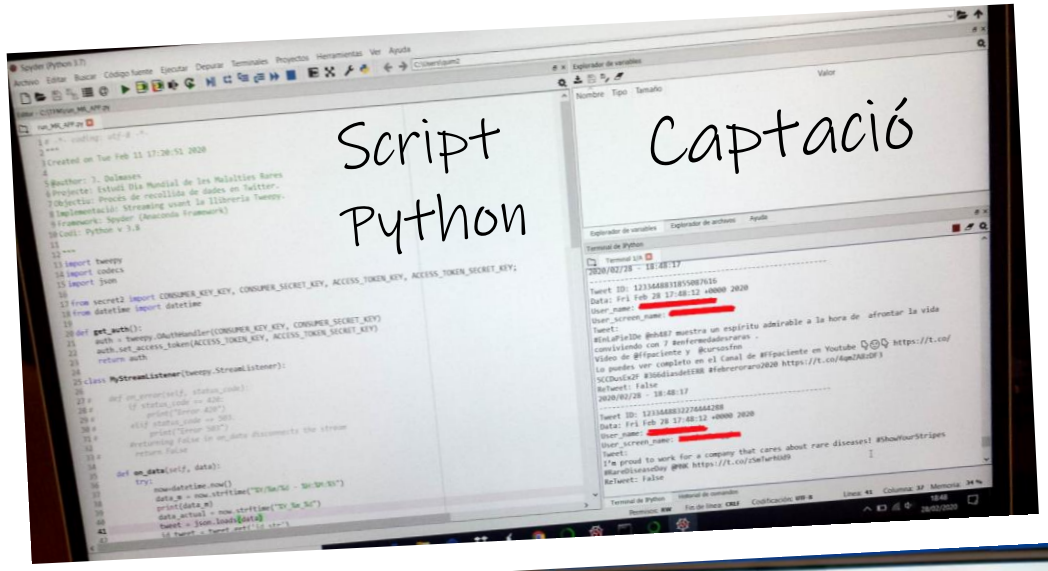
Coneixement i benefici: Detecció de temàtiques importants, opinions positives i negatives (com denúncies per manca de suport, demandes de tractaments, manca de diagnòstic) o difusió d'històries personals.

CAPTACIÓ DE DADES



CAPTACIÓ:

1. Llibreria TWEEPY.
2. Planificació dades d'interès:

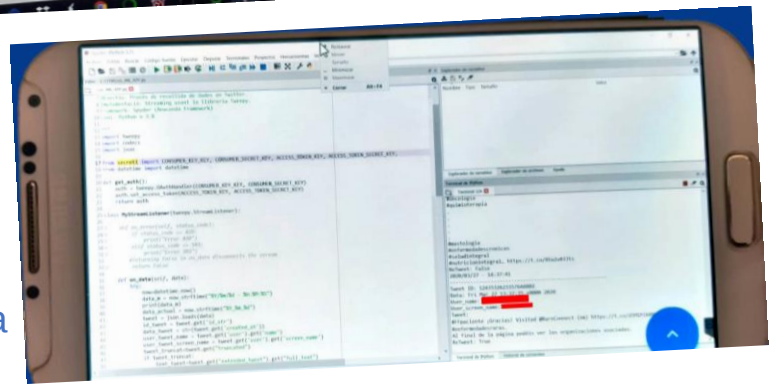


ESTRUCTURAL

1. **Nombre de tweets de cada dia**.
2. **Nombre de tweets per hora**.
3. Nombre d'usuaris amb el nombre de tweets més alt.
4. Si un tweet **és retweet** o no.
5. **Component geogràfica** si existeix. Sinó per country_code/llenguatge.
   a. Camps: geo/place/lang/**country_code**/location/Coordenades
6. Clustering: concentració de tweets.
   a. Red/Orange/Yellow/Green per **intensitat de concentració**
7. Com a **dades descriptives**:
   a. Usuari(name/screen_name)
   b. Data (timestamp)
   c. Text del tweet.

CONTINGUTS

8. Anàlisi del contingut del Tweet -> Sobre el contingut del camp Text del Tweet.
   a. **Estadística descriptiva**: Mitjana/Desv Estàndard/diagrames Box-Plot.
   b. **Anàlisi de sentiment** (llibreria TexBlob)-> Polaritat i Subjectivitat.
   c. **Word cloud**: Anàlisi (visual).



Monitoratge

Seguiment i vigilància constant.

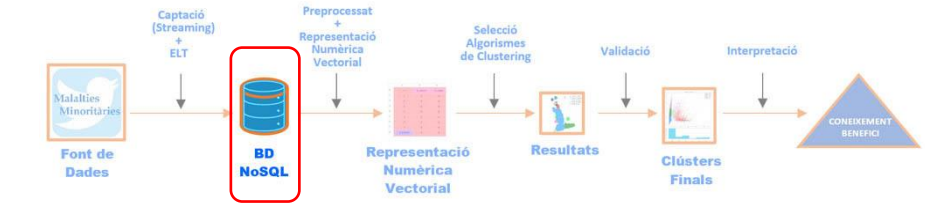
HASHTAGS CAPTURATS

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

Objecte 'Tweet' API Twitter

{"_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"}

EMMAGATZEMAMENT

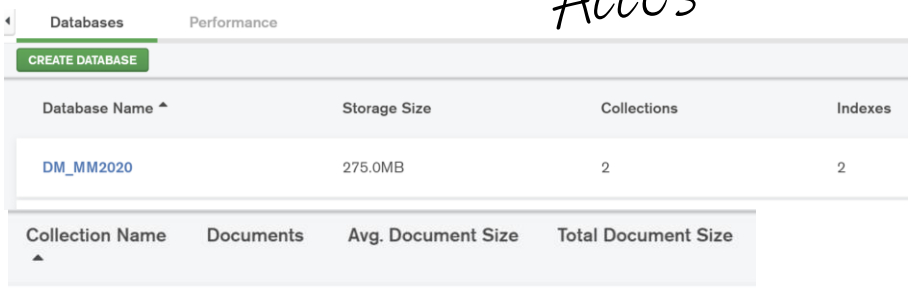


Client BD:
MongoDB Compass
Accés programat
Llibreria Pymongo

Base de dades documental

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

Accés



Importació

1. A fet falta importar els fitxers text resultat de la captura.
2. Millora aplicable

Inserció directa a temps real en la base de dades

Document = Tuit

```
_id: ObjectId("5e78e80f4e54db2148d19749")
created_at: "Thu Feb 13 09:58:19 +0000 2020"
id: 1227894764659499008
id_str: "1227894764659499010"
text: "RT @FEDER_ONG: Las 29 obras ganadoras del concurso fotográfico #EResAr..."
source: "<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter fo..."
truncated: false
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: Object
geo: null
coordinates: null
place: null
contributors: null
› retweeted_status: Object
is_quote_status: false
quote_count: 0
reply_count: 0
retweet_count: 0
favorite_count: 0
› entities: Object
favorited: false
```

Consulta

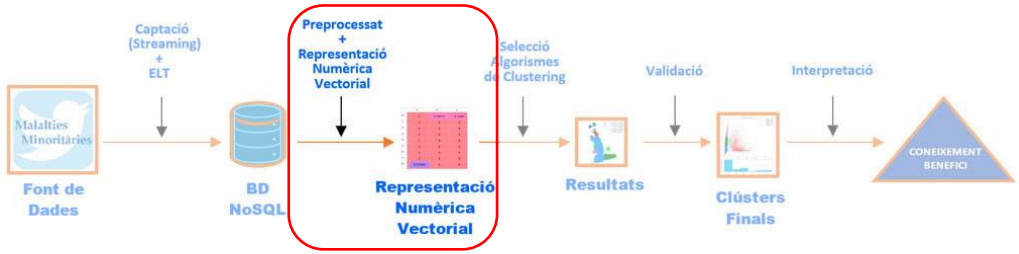
Còpia de seguretat

```
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)
```

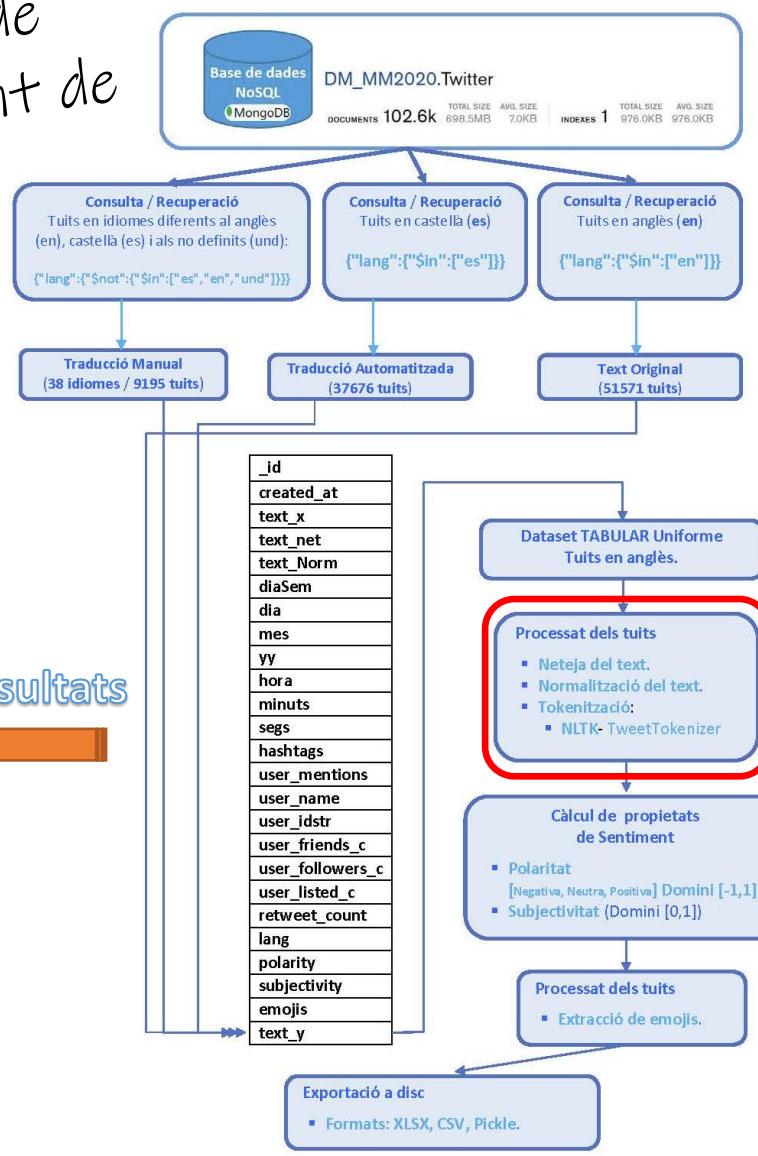
Restaurar a MVirtual

```
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
```

PREPROCESSAMENT DE DADES



Estratègia de Processament de les dades



PREPROCESSAT:

Capturats: 102.682
Descartats: 4.190 **(per falta d'idioma).**
Considerats: 98.442
9195 **d'idiomes diferents del castellà i l'anglès.**
37.676 **en castellà.**
51.571 **en anglès.**

TRADUCCIÓ

(dels tuits a l'idioma anglès)

\+

NETEJAT i DEPURACIÓ

DEL CONTINGUT TEXT de cada tuit.

1. TRADUCCIÓ GRATUITA:
   - IBM Watson
   - Google Translator

NETEJAT DEL TEXT:

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.

Dataset de modelització

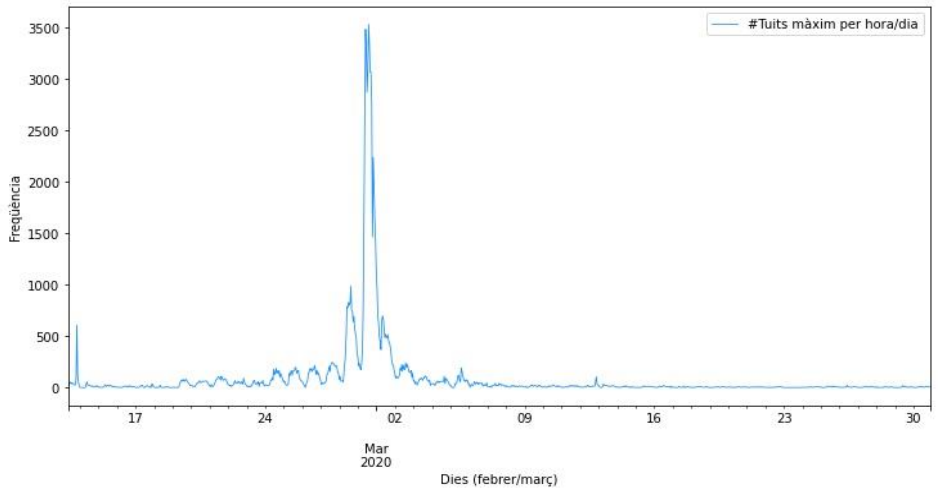
_id	Identificador de tuit únic assignat per la base de dades.
created_at	'Timestamp' (Data/Hora) de creació del tuit.
text_x	Text original del tuit, tal com es va capturar.
text_net	Text format per cadenes de caràcters sense puntuació, 'stopwords', ni dígits numèrics, cadenes amb nombre repetit d'espais 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.

Resultats

EXPLORACIÓ INICIAL I ANÀLISI TEMPORAL

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.

Tuits máx hora/dia



Mitjana de tuits / dia
'dia normal'
durant el període
100 tuits

29 de febrer

Participació més alta.

Màxim nombre de tuits proper 140 tuits/min

Des de 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

Màxim diari: 3530 tuits.

(horari UTC)

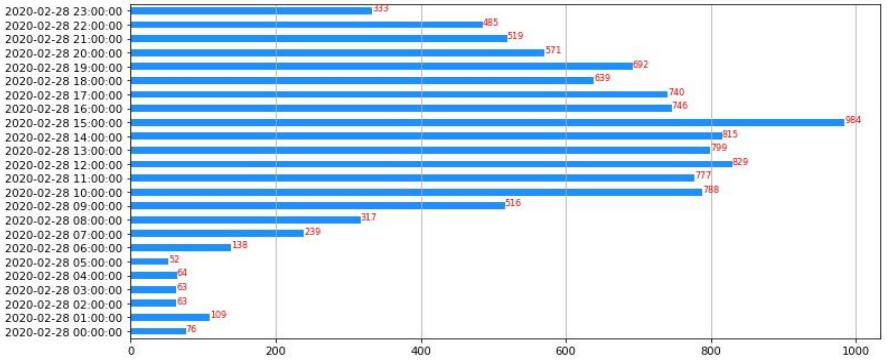
Wordcloud

Sobre el corpus de text de tots els tuits capturats.

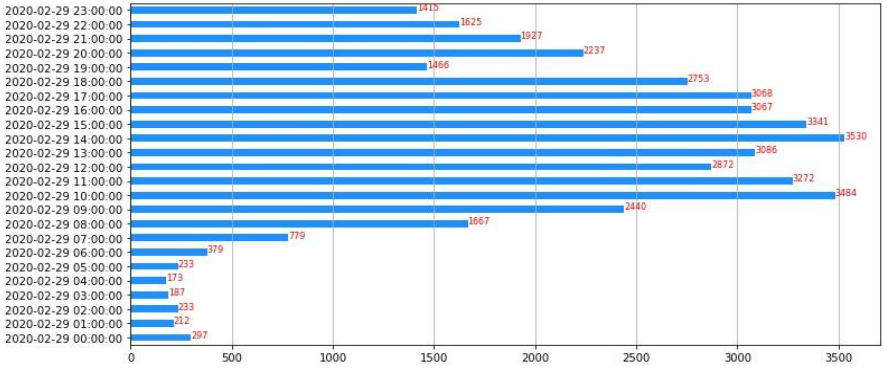
Observem una primera orientació de les temàtiques presents.

Evolució temporal

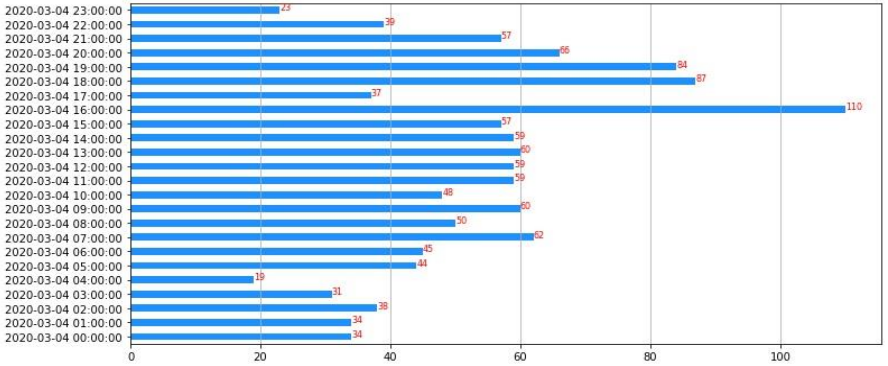
Evolució d'enviament per les dies 28 i 29 de febrer i 4 de març.



Dia 28/02/20

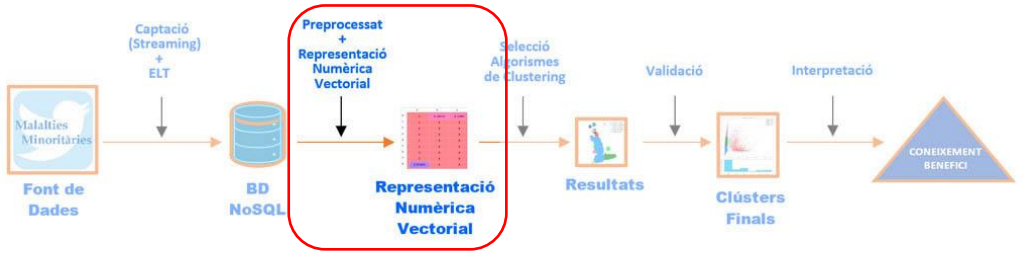


Dia 29/02/20



Dia 04/03/20

REPRESENTACIÓ NUMÈRICA DELS TUITS



Estratègia

REPRESENTACIÓ VECTORIAL
En un espai n-dimensional de cada tuit.

MATRIU TF-IDF
Files: TUITS
Columnes: Diccionari de paraules

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.

$$tf\text{-}idf(t, d) = tf(t, d) * idf(t)$$
$$idf(t) = \log[(1 + n) / (1 + df(t))] + 1$$
(en la implementació de sklearn)

Per què?

1. En aplicar *Tf-idf*, les paraules que es produeixen amb freqüència dins d'un tuit però no freqüentment en la resta de tuits, reben una ponderació més alta, ja que es parteix de la suposició que aquestes paraules són més significatives en relació amb el contingut del tuit.
2. Per aplicar models d'aprenentatge automàtic NO SUPERVISAT: Clustering. necessitem representar numèricament els tuits. VECTORITZACIÓ.
3. Poder AGRUPAR tuits per **estar propers** o per la seva **similitud**.
MÈTRICA: Distància euclidiana o Similitud vectorial (vectors unitaris d'igual direcció).

Exemple

Tuit 0 = 'cada dia menjo pa' Tuit 1 = 'cada nit bec aigua'
Tuit 2 = 'plou cada dia' Tuit 3 = 'ningú sap res'

REPRESENTACIÓ VECTORIAL DELS TUITS
MATRIU TF-IDF

Vectors

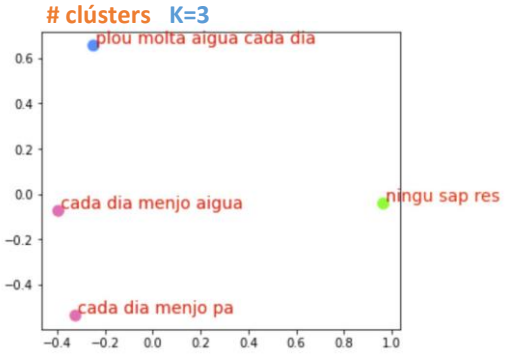
		aigua	bec	cada	dia	menjo	molta	ningu	nit	pa	plou	res	sap
Tuit 0	0	0.000000	0.00000	0.366747	0.453005	0.57458	0.000000	0.00000	0.00000	0.57458	0.000000	0.00000	0.00000
Tuit 1	1	0.453005	0.57458	0.366747	0.000000	0.00000	0.000000	0.00000	0.57458	0.00000	0.000000	0.00000	0.00000
Tuit 2	2	0.412640	0.00000	0.334067	0.412640	0.00000	0.523381	0.00000	0.00000	0.00000	0.523381	0.00000	0.00000
Tuit 3	3	0.000000	0.00000	0.000000	0.000000	0.00000	0.000000	0.57735	0.00000	0.00000	0.000000	0.57735	0.57735

REDUCCIÓ DE DIMENSIONALITAT – PCA – 2 COMPONENTS
REPRESENTACIÓ 2D - X_PCA

```
X_PCA

matrix([[-0.32255927, -0.53726923],
        [-0.39427913, -0.07509857],
        [-0.24749509,  0.65474422],
        [ 0.96433349, -0.04237641]])
```

	x	y	text
0	-0.322559	-0.537269	cada dia menjo pa
1	-0.394279	-0.075099	cada dia menjo aigua
2	-0.247495	0.654744	plou molta aigua cada dia
3	0.964333	-0.042376	ningu sap res

AGRUPAR – APLICANT UN MODEL – KMeans i una mètrica
EN REPRESENTACIÓ 2D



	x	y	text	cluster
0	-0.322559	-0.537269	cada dia menjo pa	2
1	-0.394279	-0.075099	cada dia menjo aigua	2
2	-0.247495	0.654744	plou molta aigua cada dia	0
3	0.964333	-0.042376	ningu sap res	1

CAL DETERMINAR EL Nº DE CLÚSTERS ÒPTIM

APRENENTATGE NO SUPERVISAT - CLUSTERING

Objectiu

AGRUPAR TUITS PER SIMILITUD DEL SEU CONTINGUT TEXT

Grups de tuits molt semblants, tractaran TEMATIQUES similars

Grups de tuits molt semblants, definiran COMUNITATS d'usuaris

Les classes dels grups no es coneixen 'a priori' sinó que han de ser descobertes dins dels tuits.

Selecció d'algorismes

	AVANTATGES	DESAVANTATGES
KMeans **Algorisme clàssic en clustering i considerat recentment el més rellevant dins del conjunt de mètodes no supervisats.**	• Bon rendiment per detectar clústers convexos o de formes esfèriques i de mides similars.	• Necessita el nombre de clústers on agrupar els tuits. • Sensible a valors atípics.
DBSCAN **Agrupa per densitat de veïns. Pot resoldre situacions on KMeans no agrupi bé.**	• Bon rendiment per detectar clústers de formes NO ESFÈRIQUES. • Associa grups a regions amb densitat de veïns (més similar a la funció del cervell humà). • No necessita el nombre de clústers com a paràmetre.	• Funciona millor si la densitat dels grups existents es semblant. • Hem de configurar de manera òptima més paràmetres d'entrada.
Jeràrquic AGLOMERATIU **Aporta una estructura jeràrquica afegida a l'agrupament.**	• Permet una anàlisi visual. • Fàcil aplicar altres mètriques de similitud.	• Cal avaluar diferents tipus d'enllaç o criteris d'agrupament. • Efecte cadena: succeeix quan s'agrupen elements que no s'haurien d'agrupar. • Sensible a valors atípics.

Entre els mètodes d'agrupament no supervisat diferenciem segons el tipus de clúster creat:

AGRUPAMENT ESTRICTE

Un tuit només pot pertànyer a un clúster. (**clústers o grups disjunts**)

KMeans, DBSCAN, AGLOMERATIU

AGRUPAMENT DIFÚS

Un tuit, pot pertànyer a més d'un clúster.

OPTIMITZACIÓ DE PARÀMETRES

Objectiu

OBTENIR L'AGRUPACIÓ ÒPTIMA

1. **Màxima compactació interior en els grups.**
2. **Grups el màxim de separats.**

IDENTIFICACIÓ

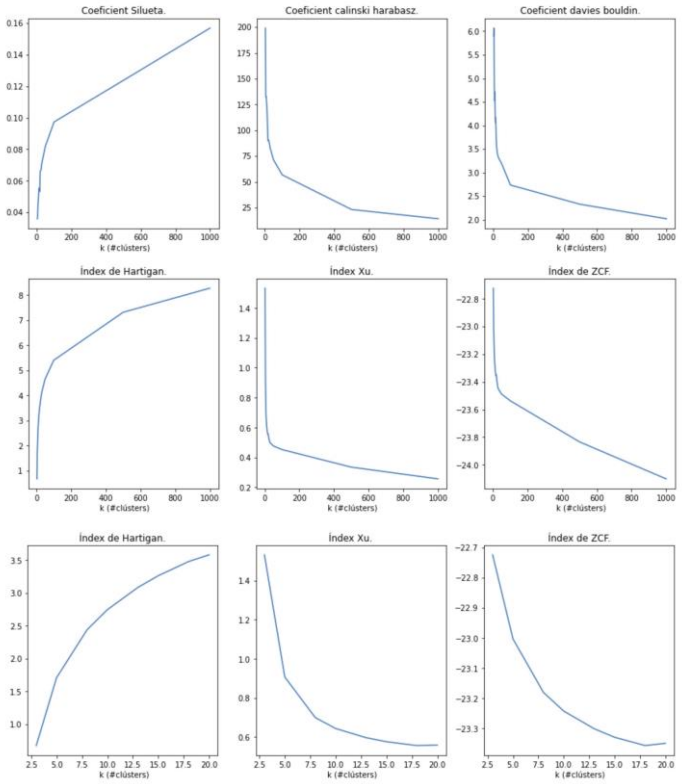
COMUNITATS D'USUARIS

TEMÀTIQUES PRINCIPALS

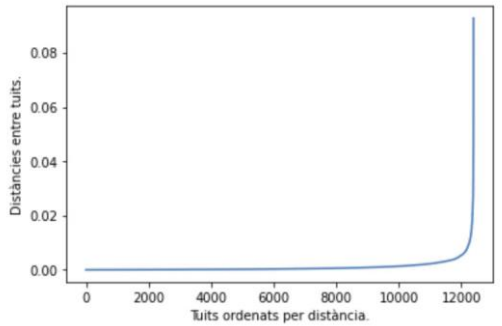
El nombre de clústers ÒPTIM aproximat **K=15**

KMeans

Cerca del **k** (#clústers) **òptim** mitjançant coeficients de qualitat de clúster.

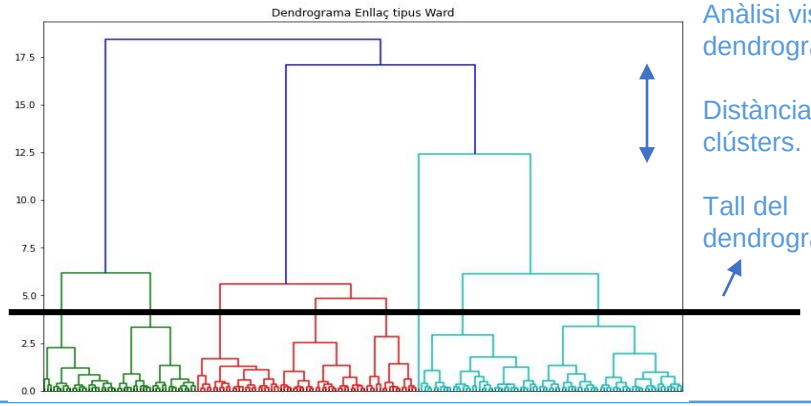


DBSCAN



Cerca de paràmetres òptims: **eps** i **min_samples**.
Estudi de densitat de veïns a partir de l'algorisme K-NN.

JERÀRQUIC AGLOMERATIU



Anàlisi visual per dendrograma:

Distància entre clústers.

Tall del dendrograma.

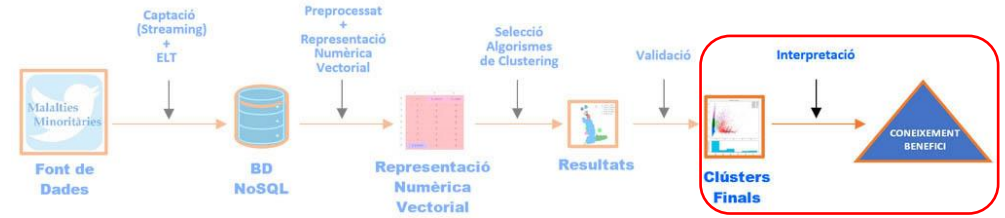
VISUALITZACIÓ DE RESULTATS

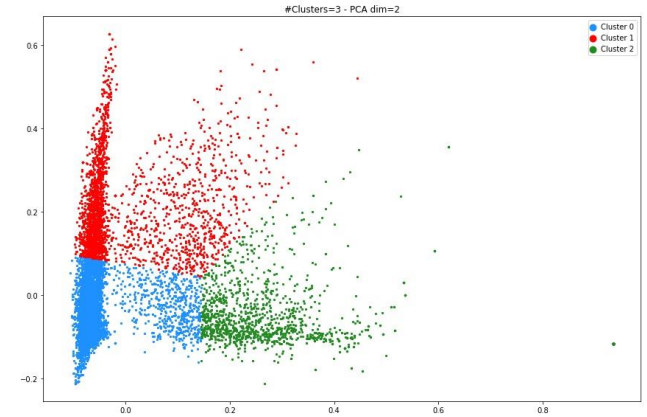
KMeans

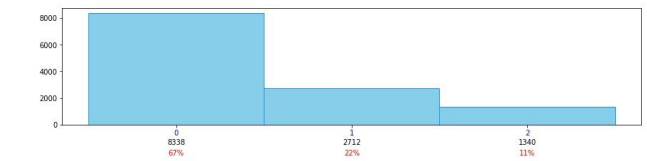
Visualització 2D

Histograma

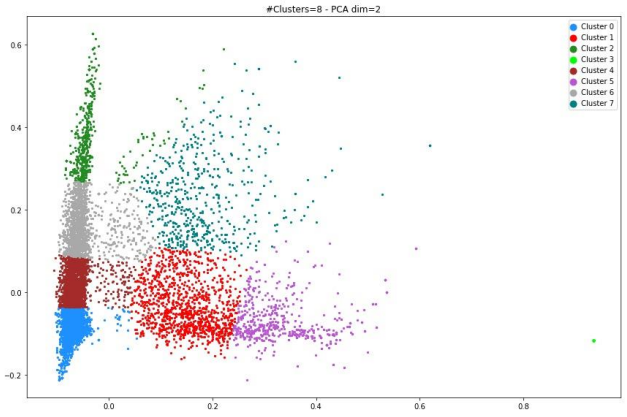
Paraules significatives de clúster

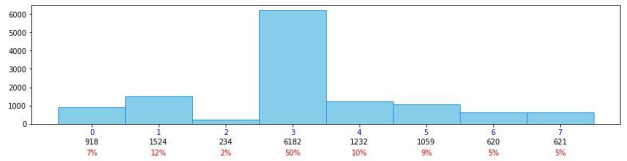




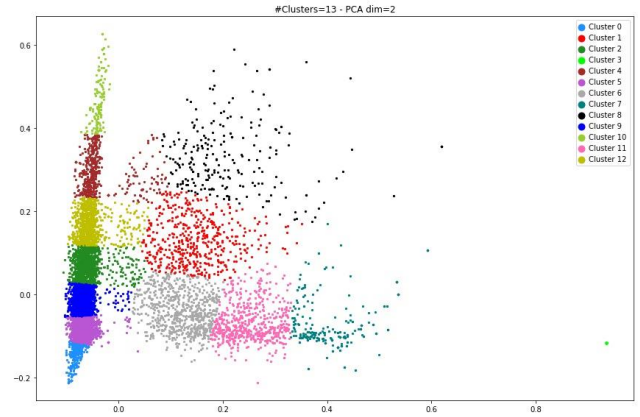


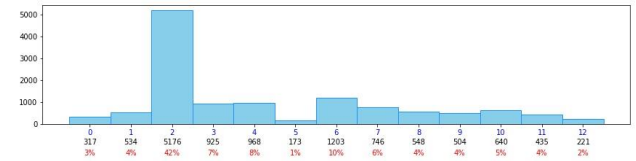
Cluster 0 amb 8338 usuaris.
research thank us patients support one know thanks like many happy tomorrow help life great
Cluster 1 amb 2712 usuaris.
people awareness million raise living to support patients one raising wide help affected around know
Cluster 2 amb 1340 usuaris.
to one celebrated support know every research us want year make work people celebrate many



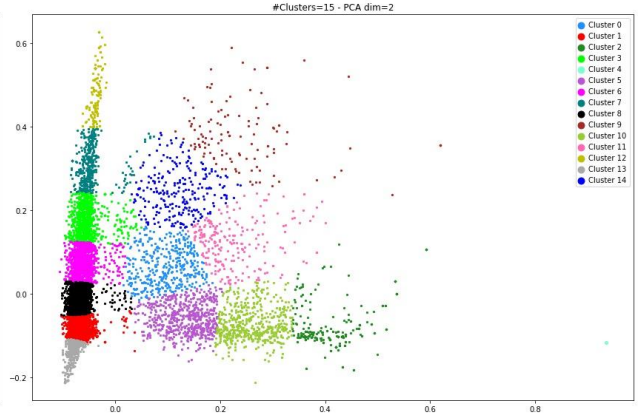


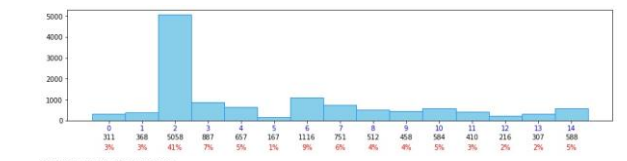
Cluster 0 amb 918 usuaris.
to celebrated want support celebrate us know international every work th make year learn many
Cluster 1 amb 1524 usuaris.
support patients proud families event us to many campaign work thank help official show community
Cluster 2 amb 234 usuaris.
happy everyone us to people also year many friends amazing living celebrate good one feb
Cluster 3 amb 6182 usuaris.
thank know thanks us like tomorrow syndrome great years life latest th many help also
Cluster 4 amb 1232 usuaris.
people million living wide to around affected know live suffer affect many support treatment affects
Cluster 5 amb 1059 usuaris.
awareness raise raising to help people patients impact lives living year please families support us
Cluster 6 amb 620 usuaris.
research support to treatments care medical us life patients find treatment help need new together
Cluster 7 amb 621 usuaris.
one to people year every million know many diagnosed affects syndrome us get hope like





Cluster 0 amb 317 usuaris.
great work see thank event to support team thanks awareness people like family us many
Cluster 1 amb 534 usuaris.
th year to feb every last st old people us tomorrow many awareness years join
Cluster 2 amb 5176 usuaris.
thank thanks latest like many syndrome tomorrow years good work to proud also event health
Cluster 3 amb 925 usuaris.
awareness raise raising to help people patients impact lives living families support please year public
Cluster 4 amb 968 usuaris.
people million living wide to around affected affect suffer live many know one awareness treatment
Cluster 5 amb 173 usuaris.
celebrated to last th tomorrow year awareness every satur know hope campaign want since first
Cluster 6 amb 1203 usuaris.
us know life people to help many thank let like please live get care learn
Cluster 7 amb 746 usuaris.
patients families to research patient treatment read hope supporting help sharing work story support learn
Cluster 8 amb 548 usuaris.
one to people every year know million diagnosed affects many syndrome get living hope like
Cluster 9 amb 504 usuaris.
research support to treatments find new need hope help treatment cure diagnosis care work people
Cluster 10 amb 640 usuaris.
to celebrate make international want give every years work people celebrating many see learn tomorrow
Cluster 11 amb 435 usuaris.
support to show families people please want thank proud suffer help living patients need us
Cluster 12 amb 221 usuaris.
happy us everyone to also year people friends many living good celebrate one awareness amazing

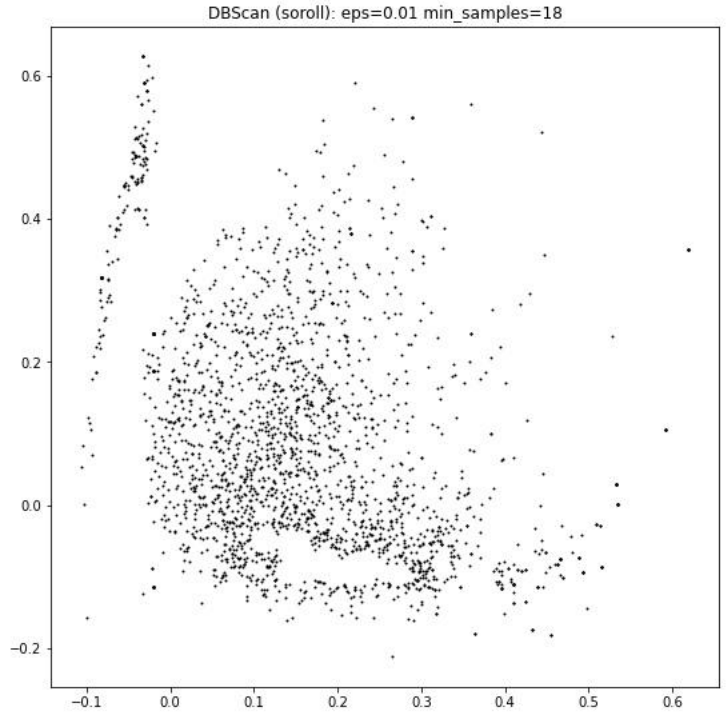


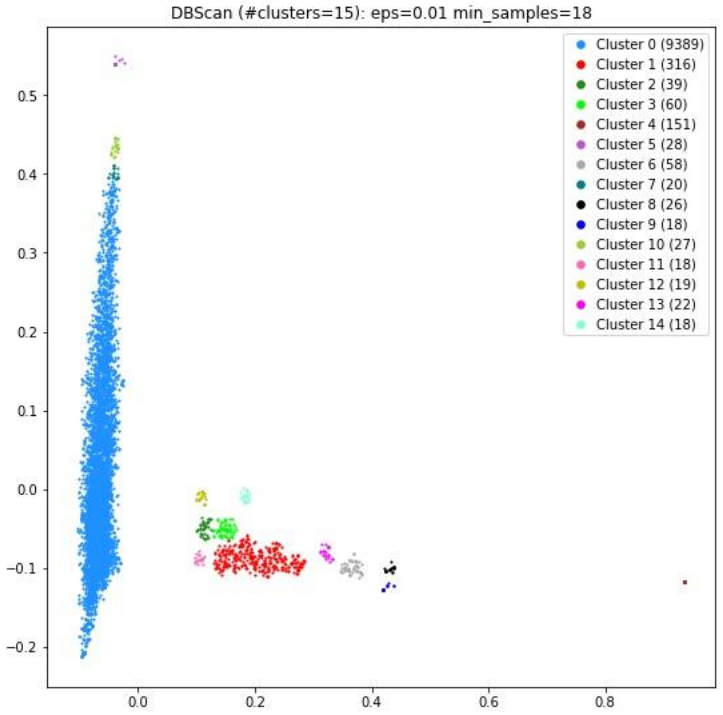


Cluster 0 amb 311 usuaris.
great work see thank event to support team thanks awareness people like family us many
Cluster 1 amb 368 usuaris.
th feb to year every tomorrow people us support one awareness join many international st
Cluster 2 amb 5058 usuaris.
thank thanks latest like many syndrome to years good tomorrow work proud event every new
Cluster 3 amb 887 usuaris.
awareness raise to raising help impact people patients lives year living support public families please
Cluster 4 amb 657 usuaris.
million people living wide around to one live know affected awareness affect suffer support treatment
Cluster 5 amb 167 usuaris.
celebrated to last th tomorrow year every awareness satur hope campaign since want first know
Cluster 6 amb 1116 usuaris.
us know life to help thank let like many join care please get people story
Cluster 7 amb 751 usuaris.
patients families research to supporting help read people treatment hope patient affected support sharing story
Cluster 8 amb 512 usuaris.
one to people every year know diagnosed affects many million get like hope best never
Cluster 9 amb 458 usuaris.
research support to treatments new find help need treatment diagnosis hope work cure people care
Cluster 10 amb 584 usuaris.
to celebrate work years make international want year learn give celebrating tomorrow every many know
Cluster 11 amb 410 usuaris.
support to show families people please thank want proud suffer living patients help need campaign
Cluster 12 amb 216 usuaris.
happy everyone us to year people living many good celebrate one awareness amazing friends also
Cluster 13 amb 307 usuaris.
also to one known year research life us people know many every tomorrow important support
Cluster 14 amb 588 usuaris.
people to affected know many living life live suffer affects affect every point care less

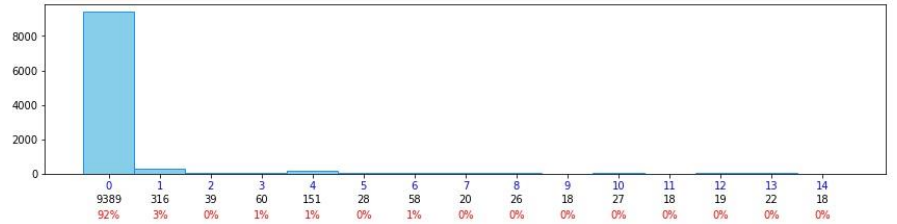
VISUALITZACIÓ DE RESULTATS

Visualització 2D

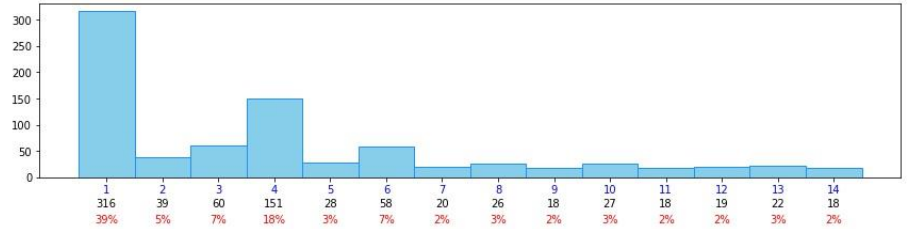




Histogrames



Visualitzant el clúster més gran.



Visualització sense el clúster més gran.

Paraules significatives de clúster

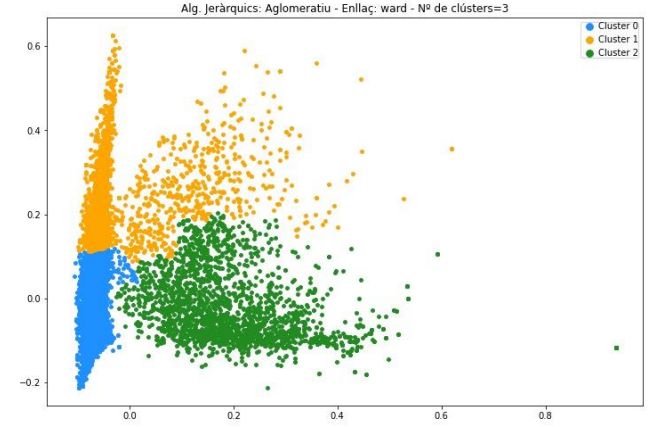
```
Cluster 0: people research support patients one us awareness thank know many help tomorrow like thanks life
Cluster 1: to give celebrated diagnosed make like also want visibility see year international work illness syndrome
Cluster 2: to life treatments time family please us patients still important 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
```

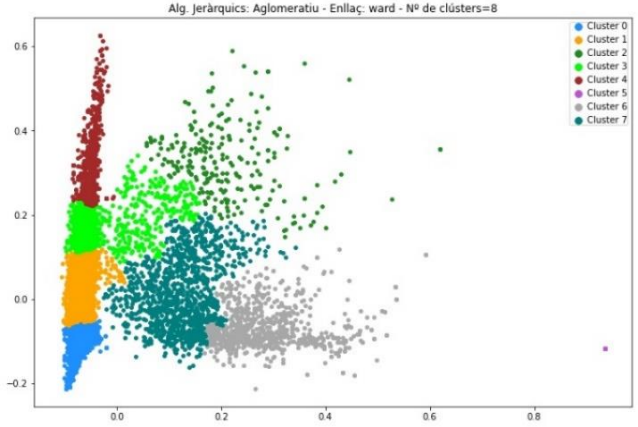
VISUALITZACIÓ DE RESULTATS

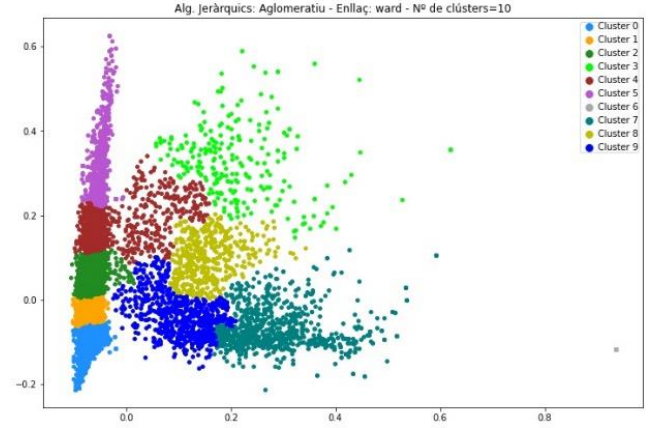
Paraules significatives de clúster

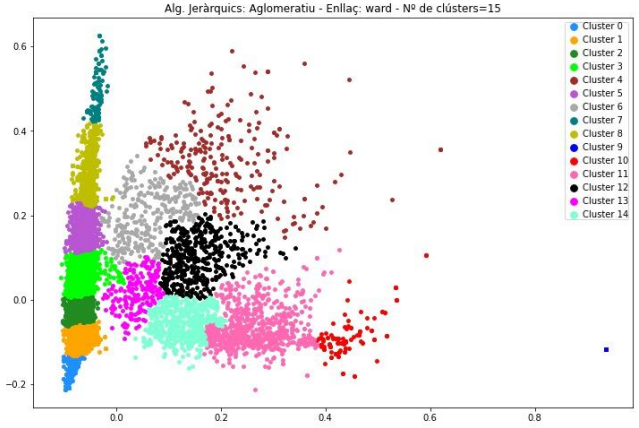
```
Cluster 0: thank latest thanks happy daily much news great sharing good last everyone health work event
Cluster 1: research great like syndrome us good years tomorrow also event work international th love thank
Cluster 2: know patients many one help us research support proud care treatment every 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
```

Visualització 2D

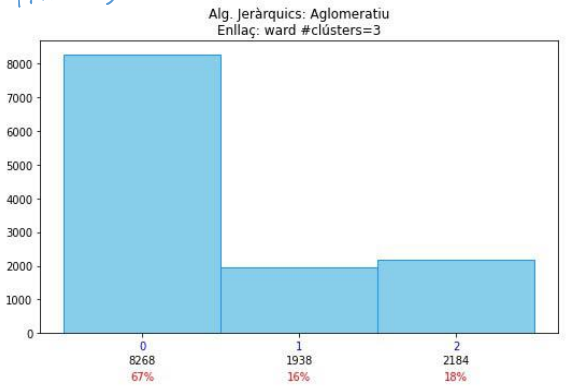


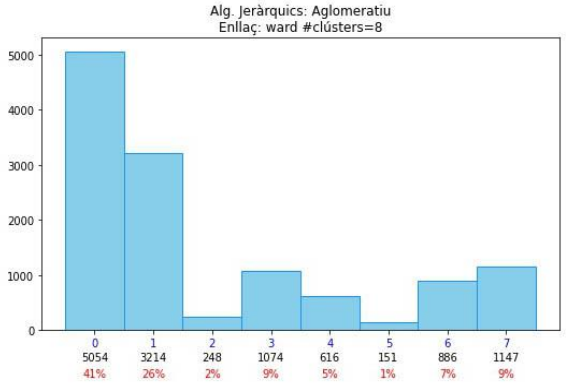


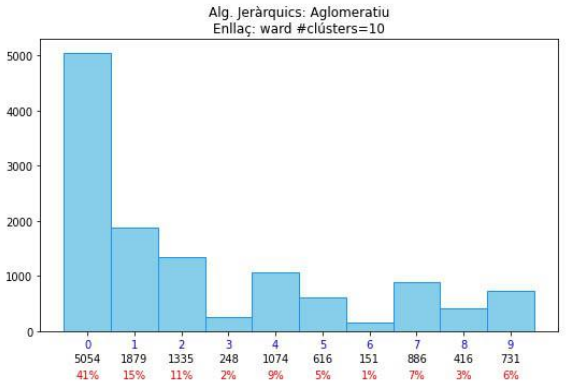


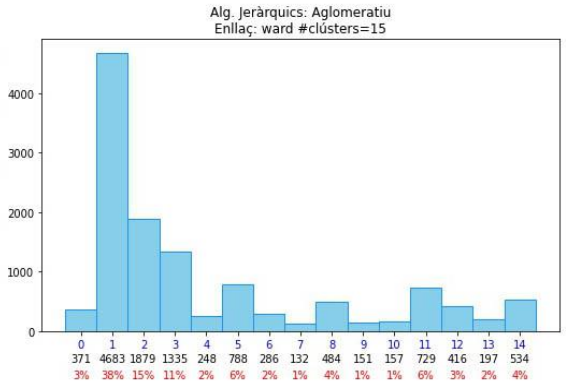


Histograma









INTERPRETACIÓ DE TEMÀTIQUES

Agrupament (KMeans)

```
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
```

Exemple d'Interpretació de Temàtiques

Cluster 0: Suport a la investigació de tractaments nous per afectats que esperen un diagnostic.
Cluster 1: Agraiments per un any més d'èxit en la resposta a un esdeveniment en relació al DMMM.
Cluster 2: Fa ressò en ser conscients de les malalties minoritàries (MM).
Cluster 3: Petició de suport a la investigació de tractaments i suport a les families de pacients.
Cluster 4: Satisfacció de tots de celebrar el dia i prendre consciencia de les MM.
Cluster 5: Sobre treballar per internacionalitzar i aprendre cada any amb la celebració del dia MM.
Cluster 6: Petició de suport pels qui pateixen malalties minoritàries i per les seves famílies.
Cluster 7: Temps de diagnostic un any a pacients i millorar l'esperança per milions de persones.
Cluster 8: Petició d'ajuda, animar a que la gent conegui durant el dia MM.
Cluster 9: Necessitat de campanyes cada any i ser conscients de les MM.
Cluster 10: Reduir/vigilar el nombre de diagnostics als afectats que pateixen transtorns.
Cluster 11: Investigació i la importància de promocionar-ho el dia 29.
Cluster 12: Agraiments envers la celebració d'un event, bona resposta de l'esdeveniment per part de les persones.
Cluster 13: Sensibilització i ser conscients pels que pateixen per obtenir tractaments / milions de persones.
Cluster 14: Petició de campanyes de suport a pacients i families a viure a més del dia MM.

ANÀLISI DE SENTIMENT amb KMeans

Objectiu

Conèixer temàtiques especialitzades.

Opinions a favor

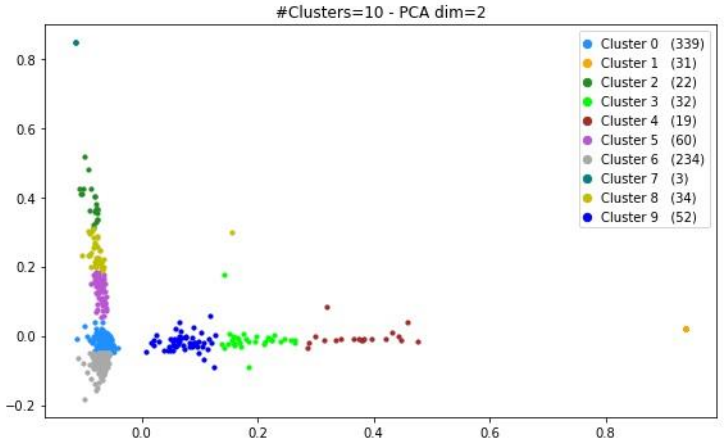
Agraïments, millores, casos d'èxit.

Opinions en negatives

Denúncies, peticions, problemàtiques.

Criteri aplicat

Identificar i extreure informació subjectiva (**valoracions personals**):
Positiva, subjectivitat >0.5 i polaritat>0.5
Negativa subjectivitat>0.5 i polaritat<-0.5
o neutra (polaritat=0) dels recursos de forma quantificable.



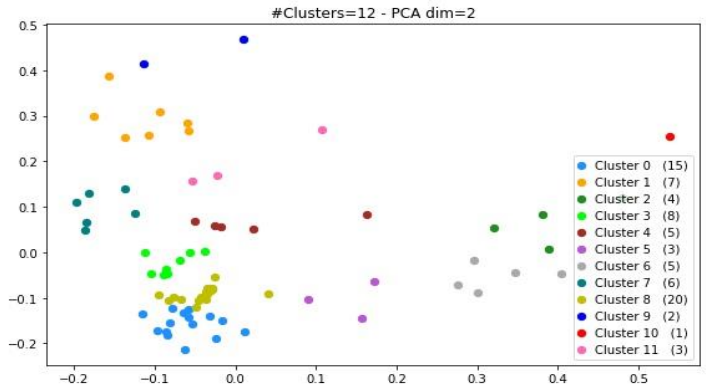
Obtenim les opinions personals:

Positives

Cluster 1:
"Amazing work inspiring awareness families"
Cluster 2:
"... proud suport work many patients..."

Negatives

Cluster 0:
Petició ajuda per migranyes
Cluster 1:
Descripció de la malaltia
Cluster 2:
Manca de tractaments a diagnòstic
Cluster 11:
Patiments insuportables




```
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
```

```
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
```

CONCLUSIONS FINALS

A continuació exposem les **conclusions** més rellevants resultat d'aquest treball de final de màster:

1. En l'àmbit del dia mundial de les malalties minoritàries, s'han capturat, preprocessat, emmagatzemat, modelat i analitzat mitjançant mètodes no supervisats, dades de la xarxa social Twitter, en un període de 47 dies al voltant del dia 29 de febrer de 2020.

2. S'han obtingut un dataset de modelització amb dades processades, una base de dades documental amb totes les dades dels tuits originals, i com resultat d'una tasca d'agrupament dels tuits, s'han obtingut les comunitats d'usuaris i les temàtiques principals.

3. Mitjançant un anàlisi de sentiment, hem extret la informació subjectiva rellevant sobre opinions a favor i en contra de pacients, hospitals, centres mèdics i entitats de suport.

4. S'han capturat patrons de denúncies, peticions de suport, sentiments i sensacions personals dels pacients que poden guiar la presa de decisions futures i l'elaboració de fulls de ruta de les entitats que lluiten contra les malalties minoritàries.

5. S'ha demostrat, que cal més recerca i investigació per reduir el temps en diagnosticar correctament les malalties que pateixen els pacients i disposar de nous tractaments, per donar el suport adequat als pacients i les seves famílies.

6. La xarxa social Twitter és una eina útil per difondre el missatge de cada persona que pateix una malaltia minoritària i les noves tecnologies en matèria d'aprenentatge automàtic no supervisat, una bona forma d'analitzar-lo, per erradicar la invisibilitat i solitud en que actualment els afectats estan immersos.

CONCLUSIONS FINALS

Voler conèixer
les malalties minoritàries
i qui les pateix
és el primer pas
per combatre-les.

Aquest treball de final de màster s'ha basat en la Ciència de Dades per la seva lluita.

Gràcies per la vostra atenció i suport !