



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

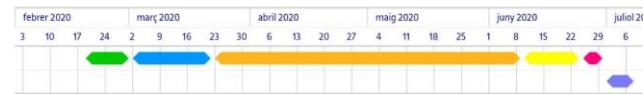
- OBJECTIUS DEL TREBALL.
- PLANIFICACIÓ DEL TREBALL.
- PRODUCTES OBTINGUTS.

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



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

4. Conclusions

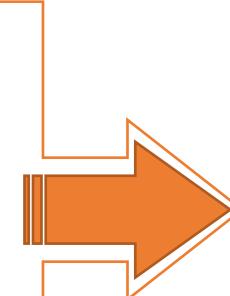
- CAPTACIÓ I EMMAGAZEMATGE DE DADES.
- GENERACIÓ DEL DATASET DE DADES PREPROCESSAT.
- ANÀLISI DE LES DADES
MÉTODES NO SUPERVISATS: TASCA D'AGRUPAMENT.
- MODELITZACIÓ AMB ELS ALGORISMES KMEANS, DBSCAN, JERÀRQUIC AGLOMERATIU.
- AVALUACIÓ DE RESULTATS.



OBJECTIUS DEL TREBALL

Tecnològics

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



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

Detecció de patrons en tuits amb aprenentatge automàtic.

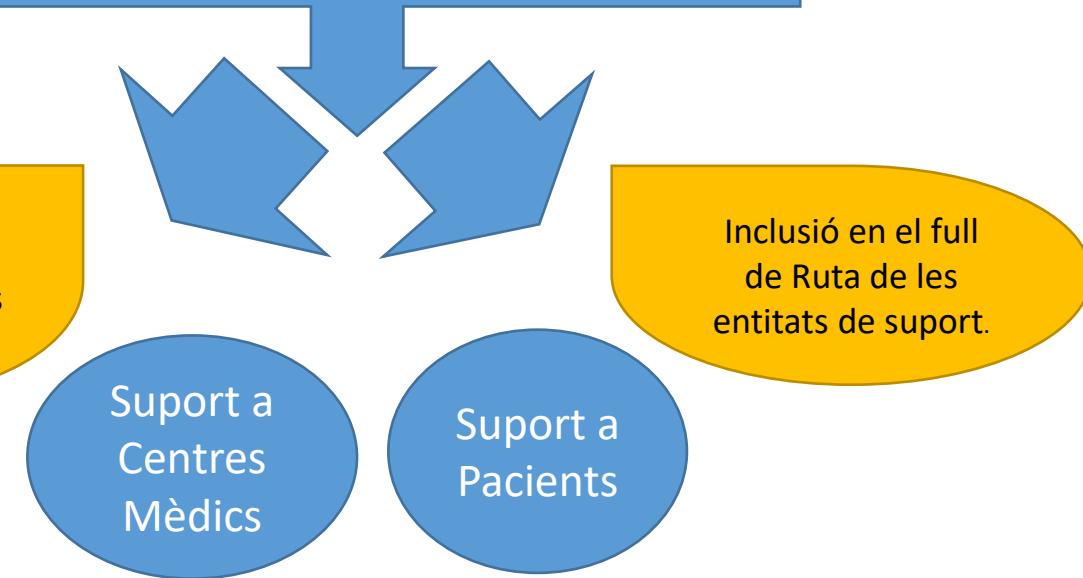
(Anàlisi de text/Anàlisi de sentiment)

Socials

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

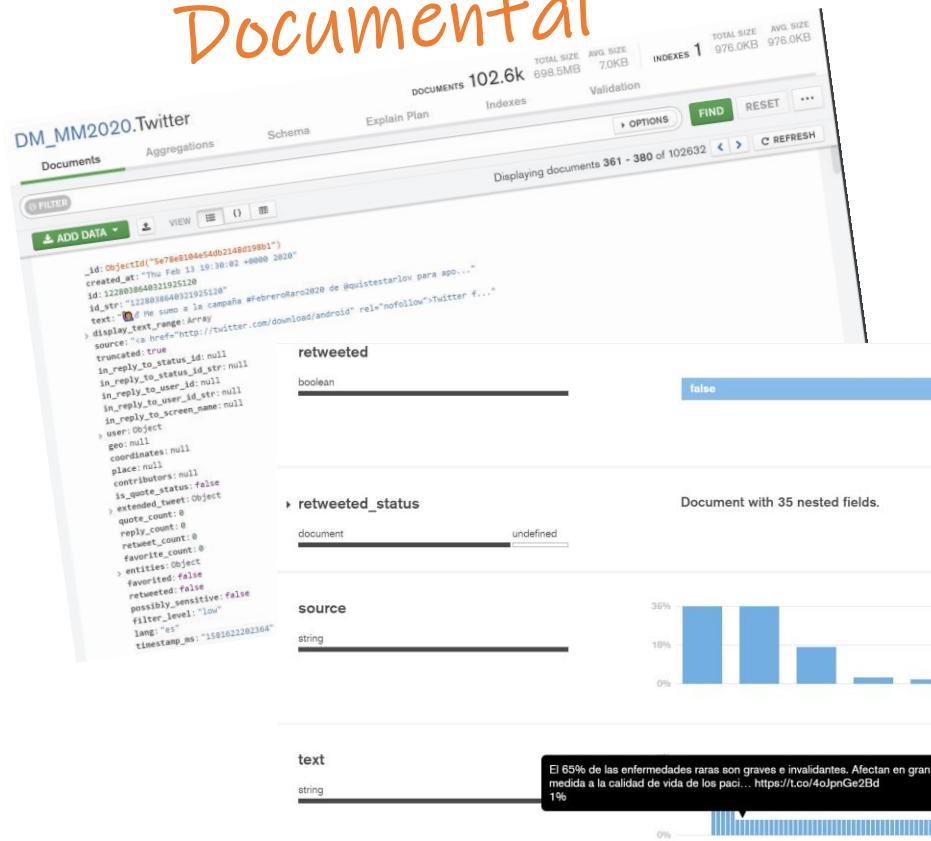


Denúncies
Esdeveniments
Problemàtiques
Tendències





Base de Dades Documental



102.632 tuits période 13/02/2020 – 30/03/2020

Objecte original complet - API de Twitter

Dataset de Modelització

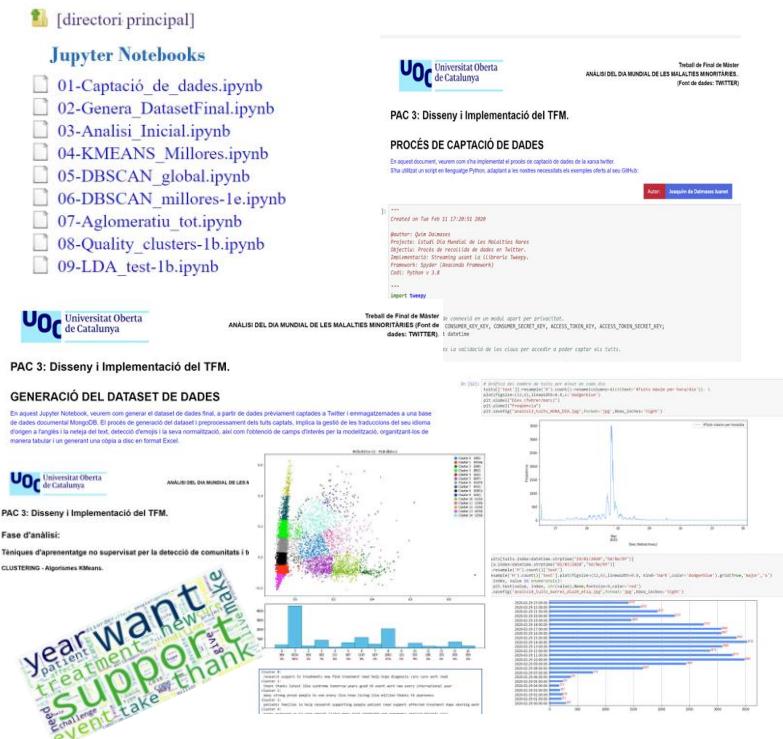
_id	Identificador de tuit únic assignat per la base de dades.
created_at	'Timestamp' (Data/Hora) de creació del tuit.
text_x	Text original del tuit, tal com es va capturar.
text_net	Text format per cadenes de caràcters sense puntuació, 'stopwords', ni dígits numèrics, cadenes amb nombre repetit d'espais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

Material/Recursos I eines d'anàlisi

https://github.com/QuimDJ/TFM_DataScience_UOC

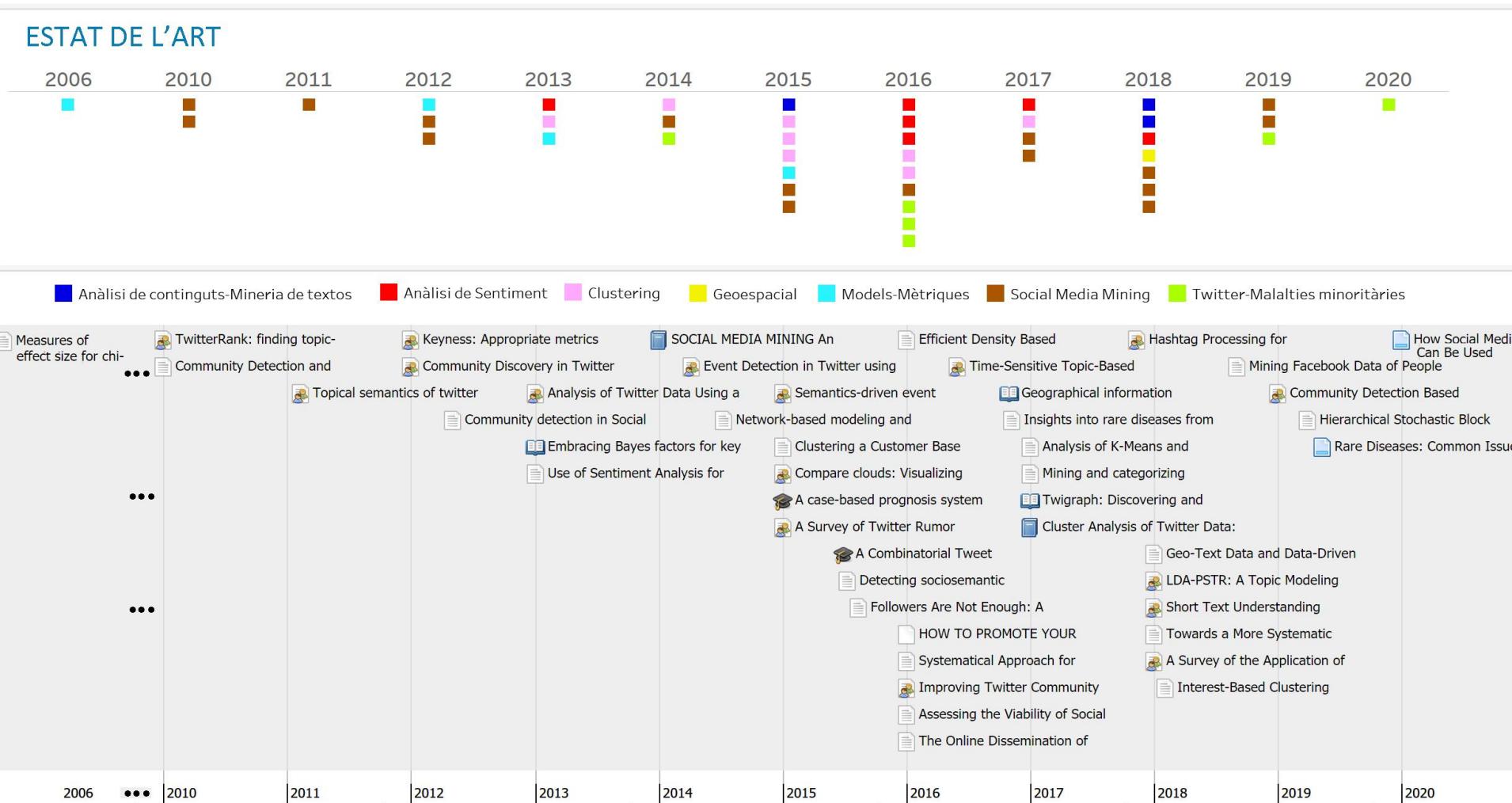


GitHub Accessible del projecte:

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.

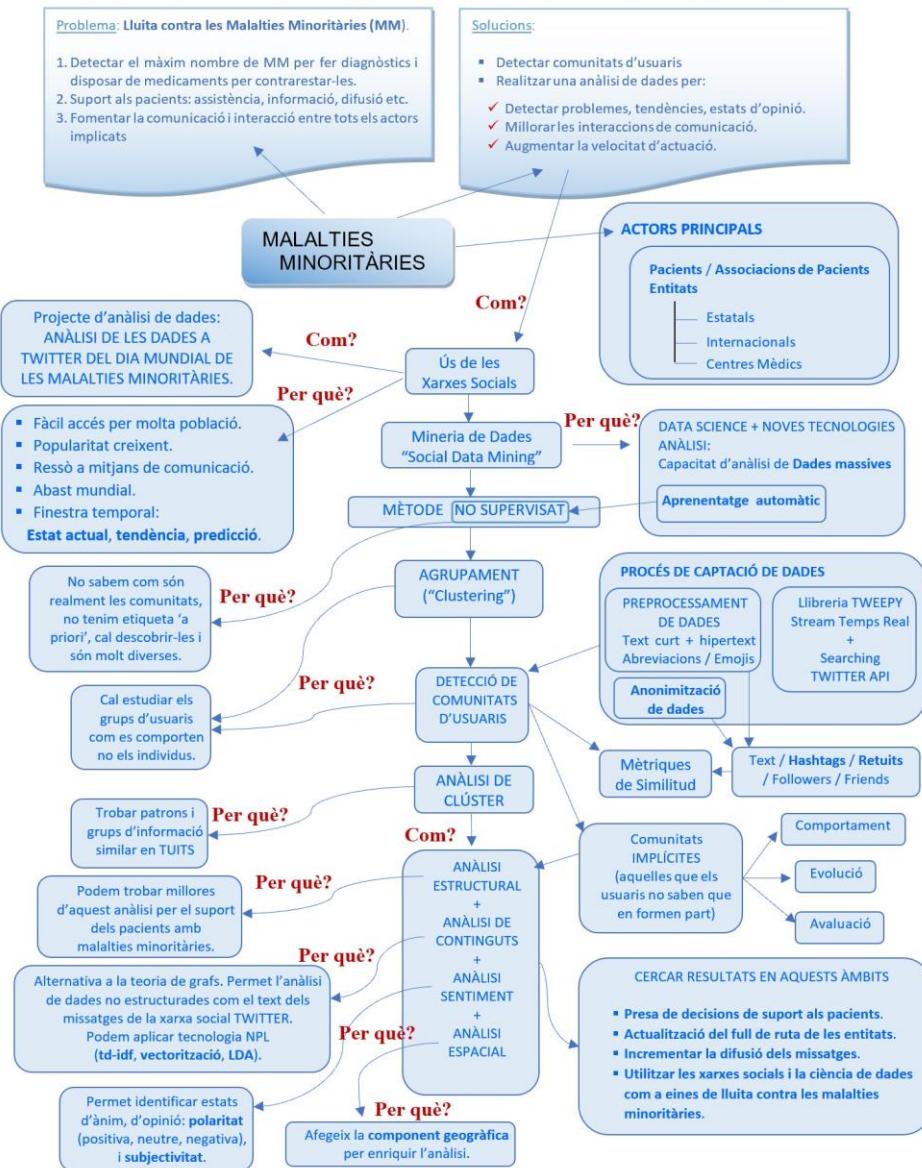


ESTAT DE L'ART



Infografia per data de publicació:

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



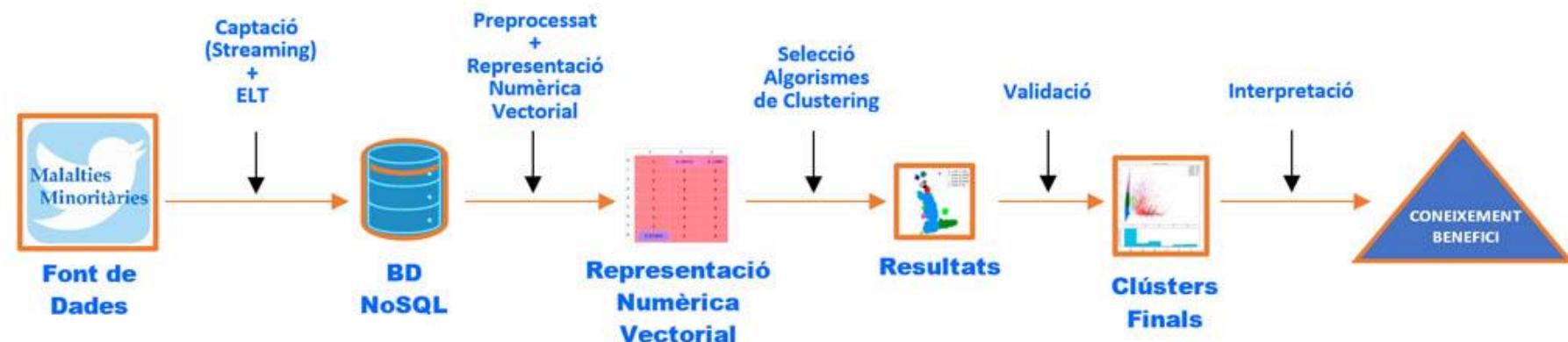
REPTES

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.

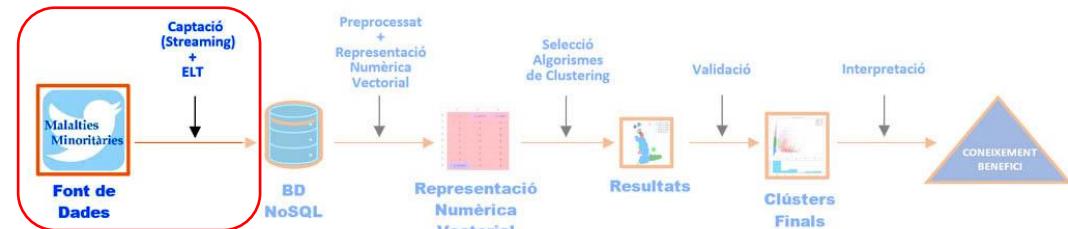
Representació numèrica del text dels tuits: Vectorització del text utilitzant una estratègia TF-IDF (importància relativa de les paraules).

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:

1. Llibreria TWEETY.
 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/llengua/te.
 - 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.

 - Estadística descriptiva: Mitjana/Desv Estàndard/diagrames Box-Plot.
 - Anàlisi de sentiment (llibreria TexBlob)-> Polaritat i Subjectivitat.
 - Word cloud: Anàlisi (visual).

The screenshot shows a Windows desktop environment with several open windows. In the foreground, a terminal window displays the output of a Python script named `run_ML_APR_09.py`. The script uses the tweepy library to interact with the Twitter API. It prints out tweets from a specific user, showing their ID, created_at, and text. The text of the tweets discusses the World Health Organization's Day of the Health Worker and mentions the 'show your stripes' campaign. In the background, there is a file explorer window showing a folder structure and a taskbar with various icons.

```
1 #!/usr/bin/python
2
3 # Created on Tue Feb 11 17:28:51 2020
4
5 #Author: T. Delmas
6 #Project: Estudi Dia Mundial de les Malalties Rares
7 #Objective: Procés de recollida de dades en Twitter.
8 #Implementation: Utilitzant llibreria tweepy
9 #Frameworks: Spyder (Anaconda Framework)
10 #Code: Python v. 3.8
11
12 ...
13
14 import tweepy
15 import codecs
16 import json
17
18 from secret import CONSUMER_KEY, CONSUMER_SECRET_KEY, ACCESS_TOKEN_KEY, ACCESS_TOKEN_SECRET_KEY
19 from datetime import datetime
20
21 def get_auth():
22     auth = tweepy.OAuthHandler(CONSUMER_KEY, CONSUMER_SECRET_KEY)
23     auth.set_access_token(ACCESS_TOKEN_KEY, ACCESS_TOKEN_SECRET_KEY)
24
25     return auth
26
27 class MyStreamListener(tweepy.StreamListener):
28
29     def on_connect(self, status_code):
30         if status_code == 200:
31             print("Connected")
32         else:
33             print(status_code)
34
35     def on_error(self, status):
36         if status['code'] == 420:
37             print("disconnecting")
38             return False
39
40     def on_data(self, data):
41
42         now=datetime.now()
43         date_m = now.strftime("%Y-%m-%d %H:%M:%S")
44         print(date_m)
45         data["date"] = now.strftime("%Y-%m-%d")
46         tweet = json.loads(data)
47         id_tweet = tweet["id"]
48
49         if tweet["text"].startswith("@"):
50
51             if tweet["text"].startswith("@WHOHQ"):
52                 print(tweet["text"])
53
54                 if tweet["text"].startswith("@WHOHQ") and tweet["text"].endswith("showyourstripes"):
55                     print(tweet["text"])
56
57                     if tweet["text"].endswith("showyourstripes") and tweet["text"].startswith("@WHOHQ"):
58                         print(tweet["text"])
59
60                         if tweet["text"].endswith("showyourstripes") and tweet["text"].startswith("@WHOHQ"):
61                             print(tweet["text"])
62
63                             if tweet["text"].endswith("showyourstripes") and tweet["text"].startswith("@WHOHQ"):
64                                 print(tweet["text"])
65
66                                 if tweet["text"].endswith("showyourstripes") and tweet["text"].startswith("@WHOHQ"):
67                                     print(tweet["text"])
68
69                                     if tweet["text"].endswith("showyourstripes") and tweet["text"].startswith("@WHOHQ"):
70                                         print(tweet["text"])
71
72                                         if tweet["text"].endswith("showyourstripes") and tweet["text"].startswith("@WHOHQ"):
73                                             print(tweet["text"])
74
75                                             if tweet["text"].endswith("showyourstripes") and tweet["text"].startswith("@WHOHQ"):
76                                                 print(tweet["text"])
77
78                                                 if tweet["text"].endswith("showyourstripes") and tweet["text"].startswith("@WHOHQ"):
79                                                     print(tweet["text"])
80
81                                                     if tweet["text"].endswith("showyourstripes") and tweet["text"].startswith("@WHOHQ"):
82                                                         print(tweet["text"])
83
84                                                         if tweet["text"].endswith("showyourstripes") and tweet["text"].startswith("@WHOHQ"):
85                                                             print(tweet["text"])
86
87                                                             if tweet["text"].endswith("showyourstripes") and tweet["text"].startswith("@WHOHQ"):
88                                                               print(tweet["text"])
89
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
279
280
281
282
283
284
285
286
287
288
289
289
290
291
292
293
294
295
296
297
298
299
299
300
301
302
303
304
305
306
307
308
309
309
310
311
312
313
314
315
316
317
318
319
319
320
321
322
323
324
325
326
327
328
329
329
330
331
332
333
334
335
336
337
338
339
339
340
341
342
343
344
345
346
347
348
349
349
350
351
352
353
354
355
356
357
358
359
359
360
361
362
363
364
365
366
367
368
369
369
370
371
372
373
374
375
376
377
378
379
379
380
381
382
383
384
385
386
387
388
389
389
390
391
392
393
394
395
396
397
398
399
399
400
401
402
403
404
405
406
407
408
409
409
410
411
412
413
414
415
416
417
417
418
419
420
421
422
423
424
425
426
427
428
429
429
430
431
432
433
434
435
436
437
438
439
439
440
441
442
443
444
445
446
447
448
449
449
450
451
452
453
454
455
456
457
458
459
459
460
461
462
463
464
465
466
467
468
469
469
470
471
472
473
474
475
476
477
478
479
479
480
481
482
483
484
485
486
487
488
489
489
490
491
492
493
494
495
496
497
498
499
499
500
501
502
503
504
505
506
507
508
509
509
510
511
512
513
514
515
516
517
518
519
519
520
521
522
523
524
525
526
527
528
529
529
530
531
532
533
534
535
536
537
538
539
539
540
541
542
543
544
545
546
547
548
549
549
550
551
552
553
554
555
556
557
558
559
559
560
561
562
563
564
565
566
567
568
569
569
570
571
572
573
574
575
576
577
578
579
579
580
581
582
583
584
585
586
587
588
589
589
590
591
592
593
594
595
596
597
598
599
599
600
601
602
603
604
605
606
607
608
609
609
610
611
612
613
614
615
616
617
618
619
619
620
621
622
623
624
625
626
627
628
629
629
630
631
632
633
634
635
636
637
638
639
639
640
641
642
643
644
645
646
647
648
649
649
650
651
652
653
654
655
656
657
658
659
659
660
661
662
663
664
665
666
667
668
669
669
670
671
672
673
674
675
676
677
678
679
679
680
681
682
683
684
685
686
687
688
689
689
690
691
692
693
694
695
696
697
698
699
699
700
701
702
703
704
705
706
707
708
709
709
710
711
712
713
714
715
716
717
718
719
719
720
721
722
723
724
725
726
727
728
729
729
730
731
732
733
734
735
736
737
738
739
739
740
741
742
743
744
745
746
747
748
749
749
750
751
752
753
754
755
756
757
758
759
759
760
761
762
763
764
765
766
767
768
769
769
770
771
772
773
774
775
776
777
778
779
779
780
781
782
783
784
785
786
787
788
789
789
790
791
792
793
794
795
796
797
798
799
799
800
801
802
803
804
805
806
807
808
809
809
810
811
812
813
814
815
816
817
818
819
819
820
821
822
823
824
825
826
827
828
829
829
830
831
832
833
834
835
836
837
838
839
839
840
841
842
843
844
845
846
847
848
849
849
850
851
852
853
854
855
856
857
858
859
859
860
861
862
863
864
865
866
867
868
869
869
870
871
872
873
874
875
876
877
878
879
879
880
881
882
883
884
885
886
887
888
889
889
890
891
892
893
894
895
896
897
898
899
899
900
901
902
903
904
905
906
907
908
909
909
910
911
912
913
914
915
916
917
918
919
919
920
921
922
923
924
925
926
927
928
929
929
930
931
932
933
934
935
936
937
938
939
939
940
941
942
943
944
945
946
947
948
949
949
950
951
952
953
954
955
956
957
958
959
959
960
961
962
963
964
965
966
967
968
969
969
970
971
972
973
974
975
976
977
978
979
979
980
981
982
983
984
985
986
987
988
989
989
990
991
992
993
994
995
996
997
998
999
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1089
1090
1091
1092
1093
1094
1095
1096
1097
1097
1098
1099
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1189
1190
1191
1192
1193
1194
1195
1196
1196
1197
1198
1199
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1289
1290
1291
1292
1293
1294
1295
1296
1297
1297
1298
1299
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1389
1390
1391
1392
1393
1394
1395
1396
1397
1397
1398
1399
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1489
1490
1491
1492
1493
1494
1495
1496
1497
1497
1498
1499
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1589
1590
1591
1592
1593
1594
1595
1596
1597
1597
1598
1599
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1689
1690
1691
1692
1693
1694
1695
1696
1697
1697
1698
1699
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1789
1790
1791
1792
1793
1794
1795
1796
1797
1797
1798
1799
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1889
1890
1891
1892
1893
1894
1895
1896
1897
1897
1898
1899
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1989
1990
1991
1992
1993
1994
1995
1996
1997
1997
1998
1999
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2089
2090
2091
2092
2093
2094
2095
2096
2097
2097
2098
2099
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2189
2190
2191
2192
2193
2194
2195
2196
2197
2197
2198
2199
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2289
2290
2291
2292
2293
2294
2295
2296
2297
2297
2298
2299
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
```

Monitoratge

Seguiment i vigilància constant.

HASHTAGS CAPTURATS

- #DiaMundialEnfermedadesRaras
- #RareDiseaseDay
- #SomosFEDER
- #EnfermedadesRaras
- #DME enfermedadesRaras2020
- #DM2020

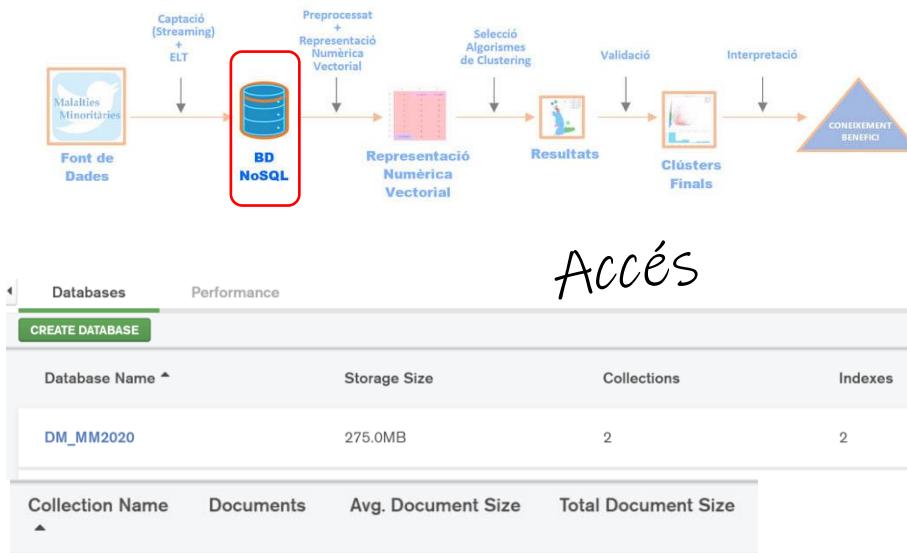
Objecte
'Tweet'
API
Twitter

[{"id": "identifier únic de la base de dades", "created_at": "Wed Feb 12 09:41:33 +0000 2020", "id": "1227528156287991910", "id_str": "1227528156287991910", "text": "Lucha contra las enfermedades raras: https://t.co/Ye13Azqjln 29 - 02 - 2020 - Día mundial de las enfermedades raras 2020 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": "ODI_UOC_24", "screen_name": "odijloc", "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": "FSFBFA", "profile_background_image": null, "profile_background_image_url_https": "", "profile_background_tile": false, "profile_link_color": "#1DA1F2", "profile_sidebar_border_color": "CODEED", "profile_sidebar_fill_color": "#DEE#6", "profile_text_color": "#333333", "profile_use_background_image": true, "profile_image_url": "https://pbs.twimg.com/profile_images/1209460351185694720/u1u3w9M5_normal.jpg", "profile_image_url_https": "https://pbs.twimg.com/profile_images/1209460351185694720/u1u3w9M5_normal.jpg", "default_profile": true, "default_profile_image": false, "follow_request_sent": null, "followed_by": 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/Ye13Azqjln 29 - 02 - 2020 - Día mundial de las enfermedades raras #diamundialenfermedaderarasLas ER o poco frecuentes son aquellas que tienen una baja prevalencia en la población003n: nmenos de 5 de cada 10000 habitantes***.", "display_text_range": [0, 280], "entities": {"hashtags": [{"text": "#diamundialenfermedaderaras", "indices": [122, 150]}], "urls": [{"url": "https://t.co/Ye13Azqjln"}, {"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": [{"url": [{"url": "https://t.co/Ye13Azqjln"}, {"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/1227528156287991910", "display_url": "twitter.com/i/web/status/1u2026", "indices": [117, 140]}], "user_mentions": [], "symbols": []}}, "retweeted": false, "possibly_sensitive": false, "filter_level": "low", "lang": "es", "timestamp_ms": "1581500493486"}]

Objecte 'Tweet' API Twitter



EMMAGATZEMAMENT



Document = Tuit

<code>_id: ObjectId("5e78e80f4e54db2148d19749")</code>
<code>created_at: "Thu Feb 13 09:58:19 +0000 2020"</code>
<code>id: 1227894764659499008</code>
<code>id_str: "1227894764659499010"</code>
<code>text: "RT @FEDER_ONG: Las 29 obras ganadoras del concurso fotográfico #EResAr..."</code>
<code>source: "Twitter fo..."</code>
<code>truncated: false</code>
<code>in_reply_to_status_id: null</code>
<code>in_reply_to_status_id_str: null</code>
<code>in_reply_to_user_id: null</code>
<code>in_reply_to_user_id_str: null</code>
<code>in_reply_to_screen_name: null</code>
<code>> user: Object</code>
<code>geo: null</code>
<code>coordinates: null</code>
<code>place: null</code>
<code>contributors: null</code>
<code>> retweeted_status: Object</code>
<code>is_quote_status: false</code>
<code>quote_count: 0</code>
<code>reply_count: 0</code>
<code>retweet_count: 0</code>
<code>favorite_count: 0</code>
<code>> entities: Object</code>
<code>favorited: false</code>

DM_MM2020.Twitter

Consult

Client BD:
MongoDB Compass
Accés programat
Llibreria Pymongo

Importació

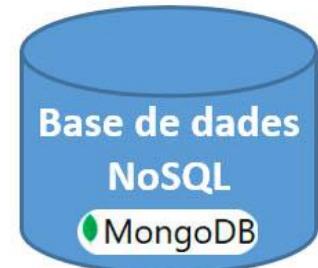
1. A fet falta importar els fitxers text resultat de la captura.
2. Millora aplicable

Inserció directa
a temps real en la
base de dades

DOCUMENTS	102.6k	TOTAL SIZE	698.5MB	AVG. SIZE	7.0KB	INDEXES	1	TOTAL SIZE	976.0KB	AVG. SIZE	976.0KB
Displaying documents 1 - 20 of 525											

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



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)
```

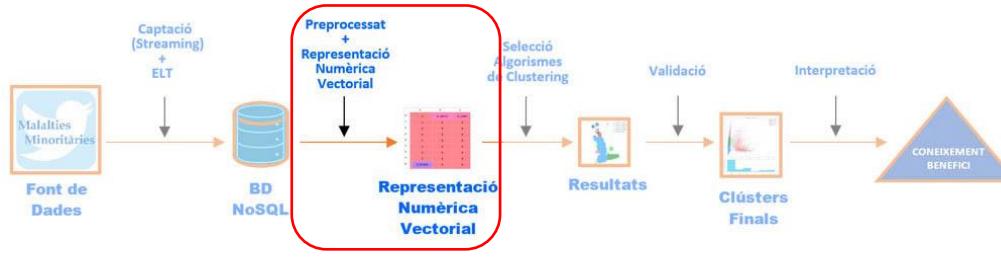
mongorestore --host=localhost --port=7017

```
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
```

Restaurar
a MVirtua



PREPROCESSAMENT DE DADES



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.

PREPROCESSAT:

TRADUCCIÓ

(dels tuits a l'idioma anglès)



NETEJAT i DEPURACIÓ
DEL CONTINGUT TEXT de cada tuit.

1. TRADUCCIÓ GRATUITA:

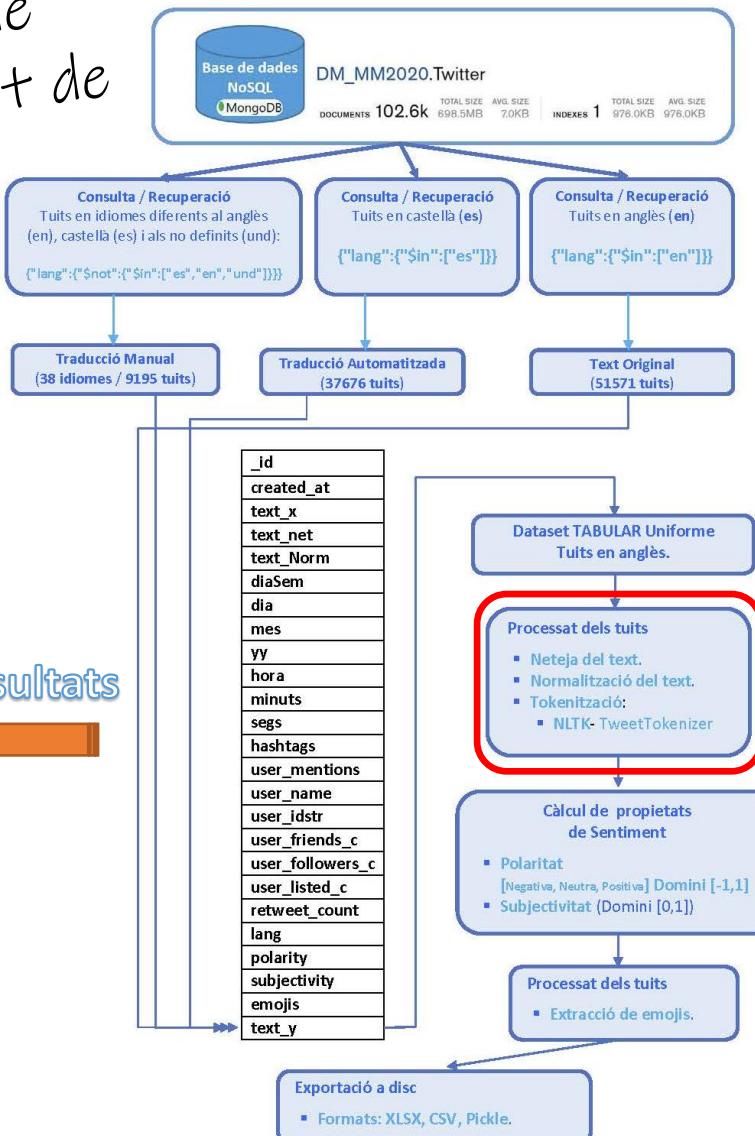
IBM Watson
Google Translator

NETEJAT DEL TEXT:

- Eliminar lesUrls.
- Substituir seqüències de diversos caràcters en blanc per un de sol.
- Eliminar els signes de puntuació i caràcters especials com retorns de carro, alimentació de línia, tabuladors, etc.
- Eliminar els díigits numèrics.
- Eliminar les 'stop words' (paraules molt freqüents però poc significatives).
- Eliminar els caràcters @ i # en mencions d'usuari i hashtags. En aquest cas podem escollir també eliminar-los del text.
- 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 permessa.
- Extracció i eliminació d'emojis.

Dataset de modelització

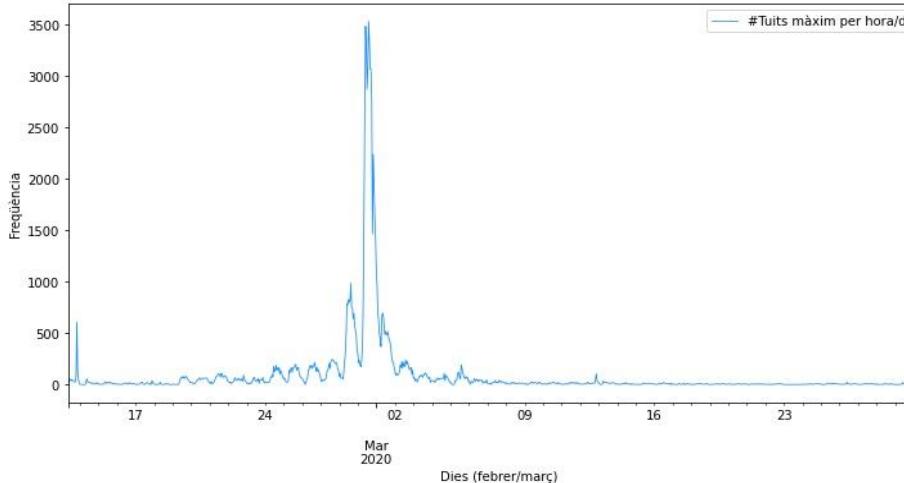
Estratègia de processament de les dades





L'exploració inicial s'ha dut a terme sobre un *dataframe* o taula de dades, indexat per la data de cada tuí, 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 màx horal/dia



Wordcloud

Sobre el corpus de text de tots els tuits capturats



Observem una primera orientació de les temàtiques presents.

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

29 de febre

Participació mes alta

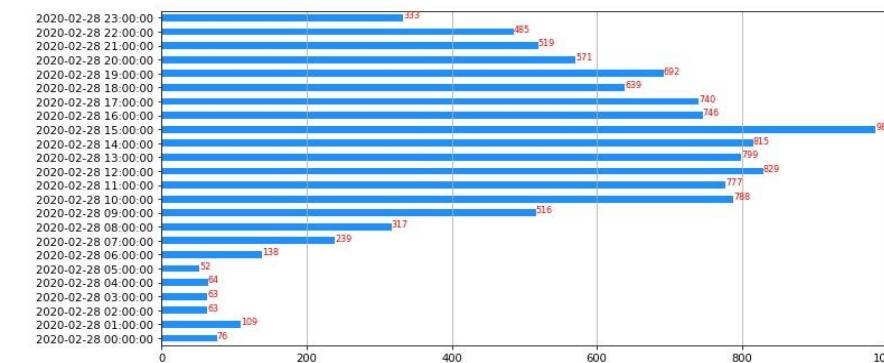
Màxim nombre de tuits proper 140 tuits/mir

Des de 09:00h i dura
quasi bé tot el dia.

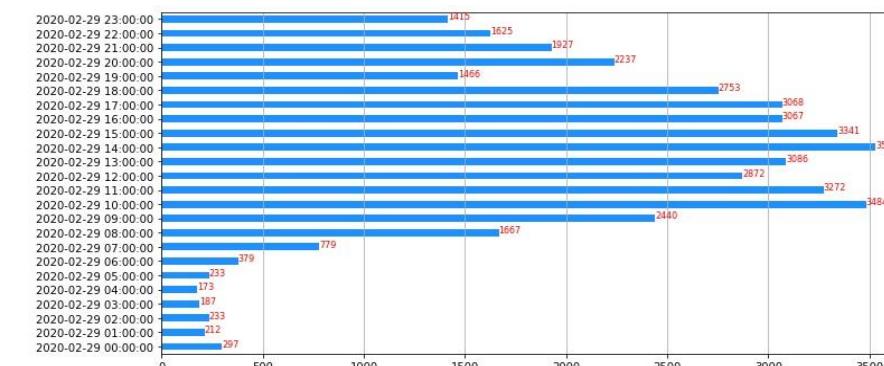
El pic màxim de tot el dia es produeix entre les 14:00h i 15:00h

Màxim diari: 3530 tuits.

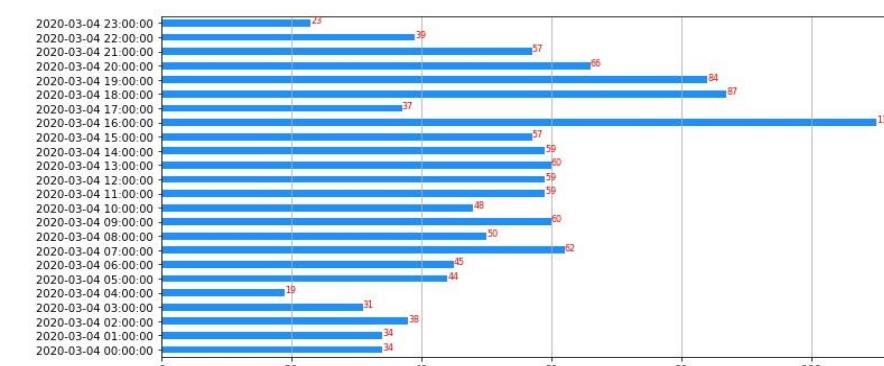
Evolució temporal



Dja 28/02/20



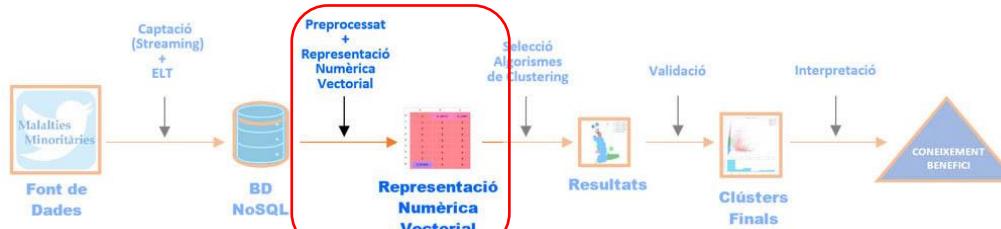
Dia 29/02/20



Dia 04/03/20



REPRESENTACIÓ NUMÈRICA DELS TUITS



Estratègia

REPRESENTACIÓ VECTORIAL

En un espai n-dimensional de cada tuit.

MATRIU TF-IDF

Files: TUTS

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.

$$\text{tf-idf}(t, d) = \text{tf}(t, d) * \text{idf}(t)$$

$$\text{idf}(t) = \log \left(\frac{1 + n}{1 + \text{df}(t)} \right) + 1$$

(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'

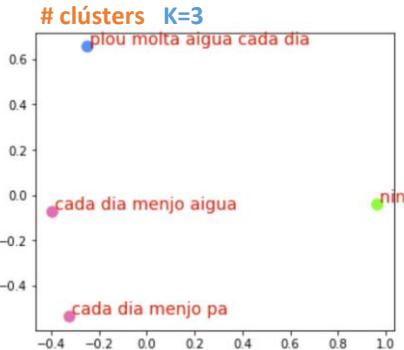
REPRESENTACIÓ VECTORIAL DELS TUITS MATRÍU TF-IDF

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

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

X_PCA	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	ningú sap res

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

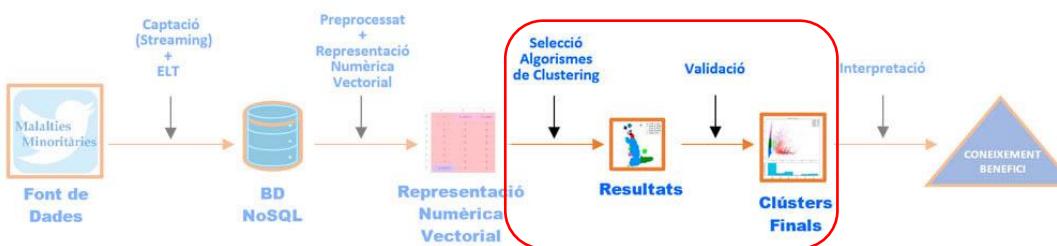


	x	y	text	cluster
0	-0.322559	-0.537269	cada dia menjo pa	2
1	-0.394279	-0.075099	cada dia menjo aigua	2
2	-0.247495	0.654744	plou molta aigua cada dia	0
3	0.964333	-0.042376	ningú sap res	1

CAL DETERMINAR EL Nº
DE CLÚSTERS ÒPTIM



APRENENTATGE NO SUPERVISAT - CLUSTERING



Selecció
d'algorismes

KMeans

Algorisme clàssic en clustering i considerat recentment el més rellevant dins del conjunt de mètodes no supervisats.

AVANTATGES

- Bon rendiment per detectar clústers convexos o de formes esfèriques i de mides similars.

DESAVANTATGES

- 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.

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.

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

AGRUPAMENT ESTRÍCTE

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

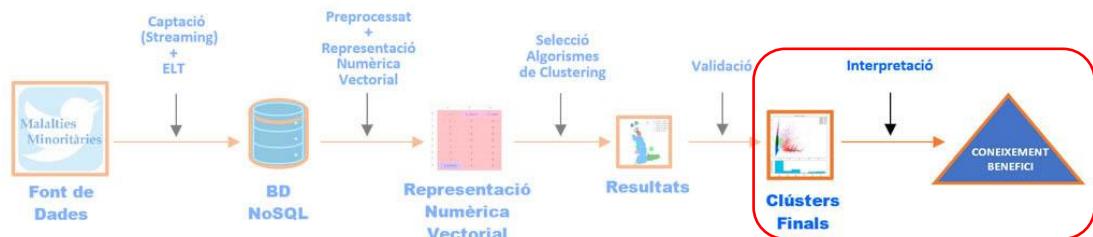
KMeans, DBSCAN, AGLOMERATIU

AGRUPAMENT DIFÚS

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



OPTIMITZACIÓ DE PARÀMETRES

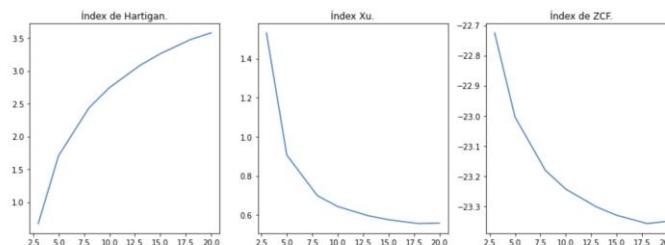
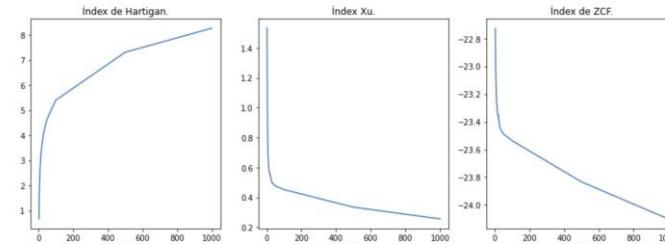
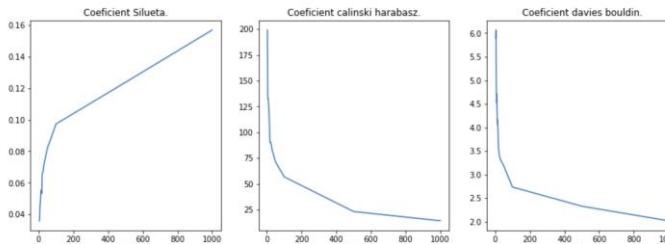


Objectiu

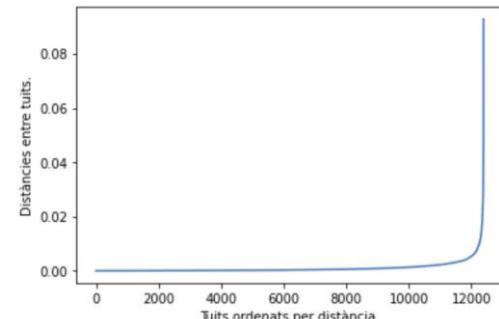
OBTENIR L'AGRUPACIÓ ÒPTIMA

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

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

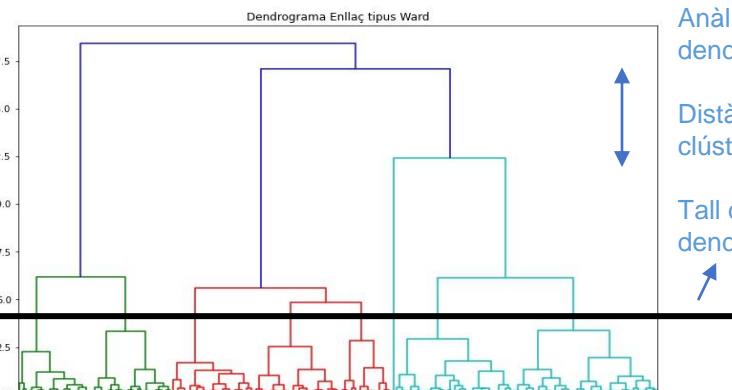


DBSCAN



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.

El nombre de clústers
ÒPTIM aproximat
K=15



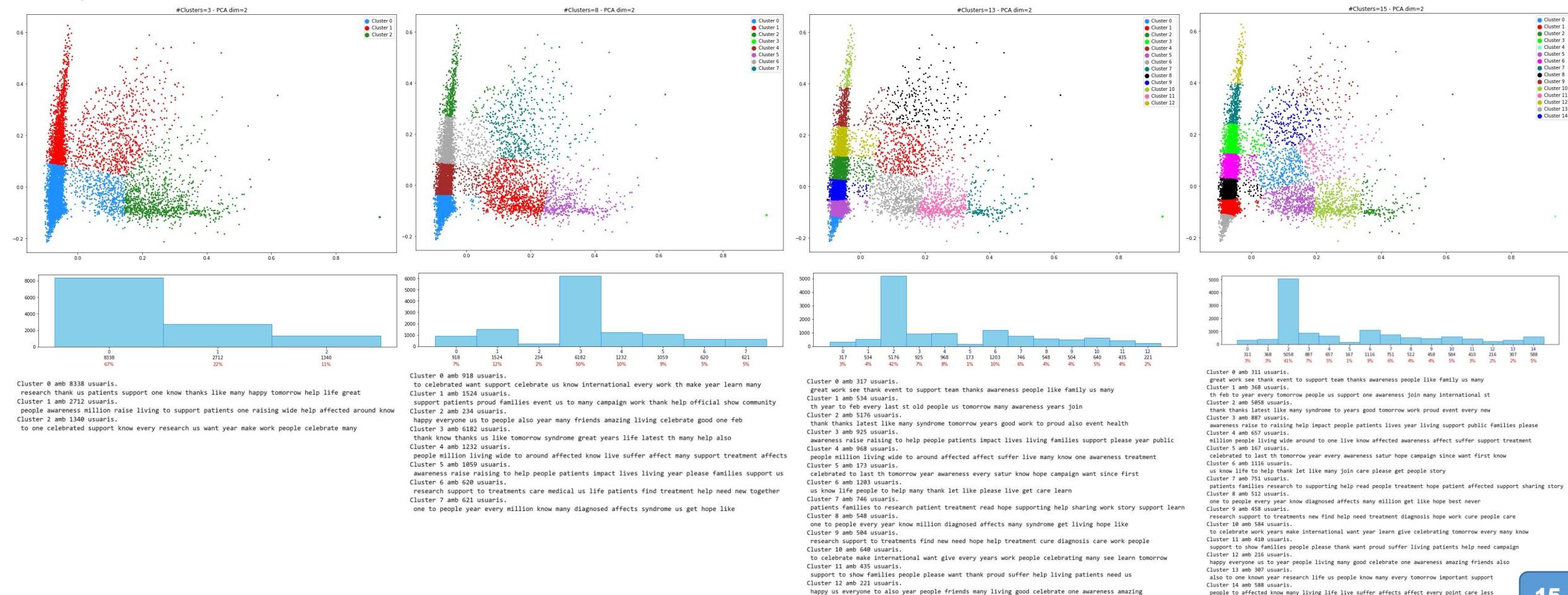
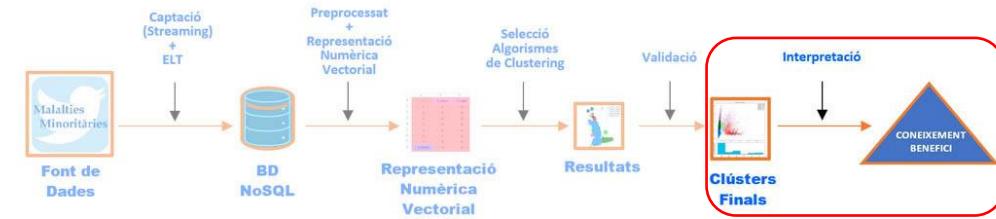
VISUALITZACIÓ DE RESULTATS

KMeans

Visualització 2D

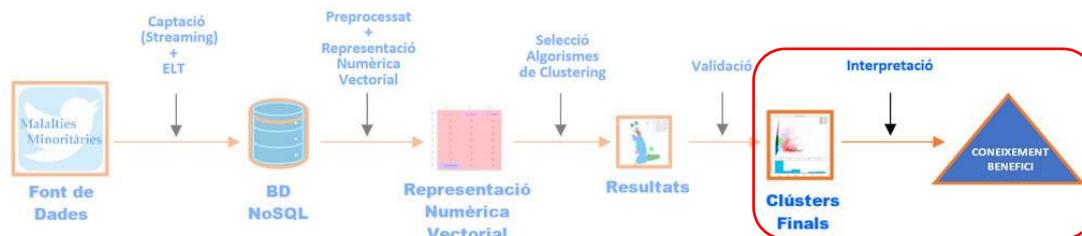
Histograma

Paraules significatives de clúster



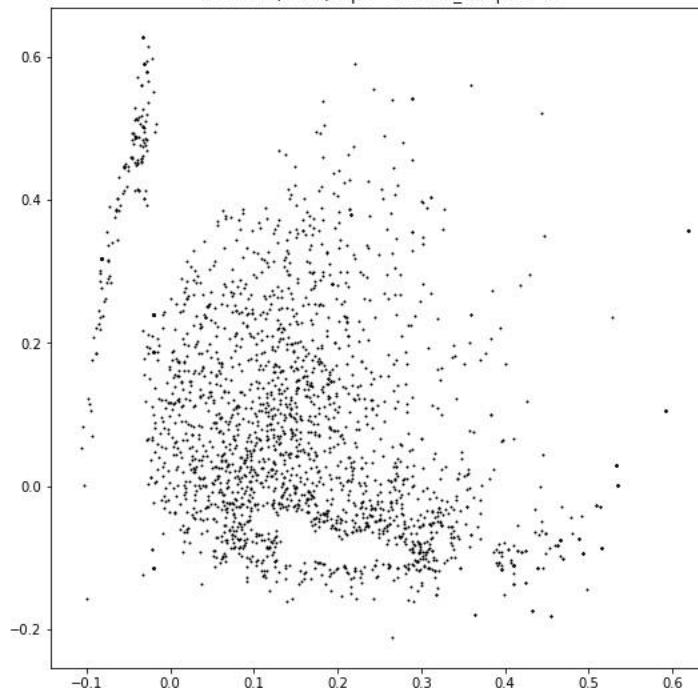


VISUALITZACIÓ DE RESULTATS

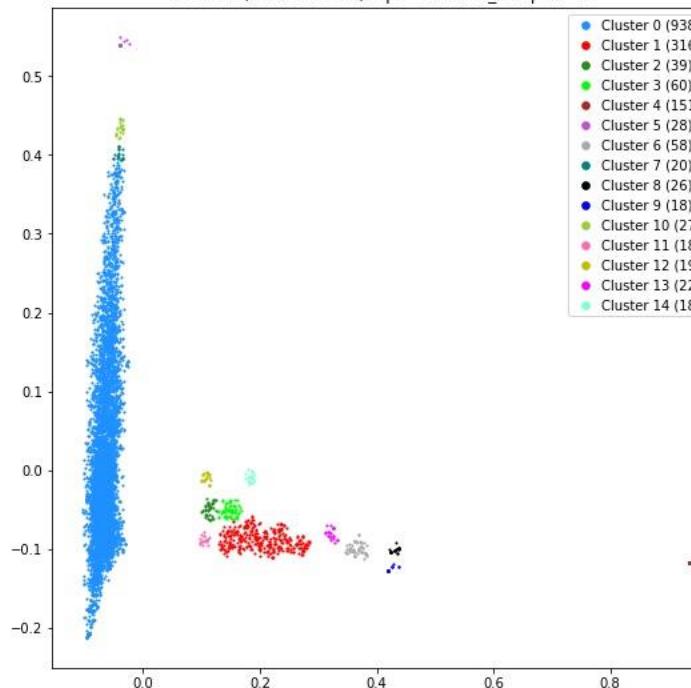


Visualització 2D

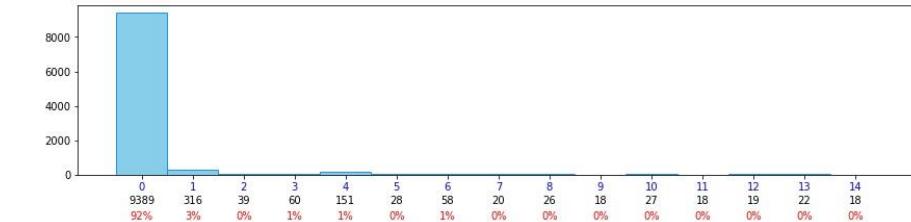
DBScan (soroll): $\text{eps}=0.01 \text{ min_samples}=18$



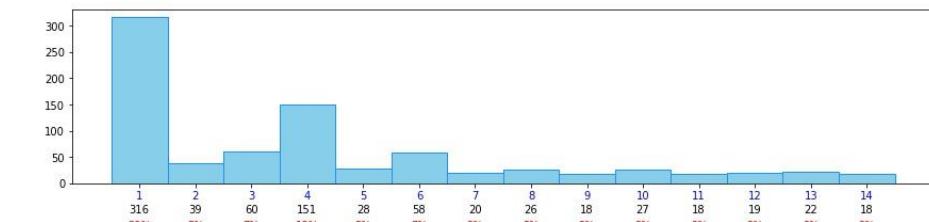
DBScan (#clusters=15): $\text{eps}=0.01 \text{ min_samples}=18$



Histogrammes



Visualitzant el clúster més gran.



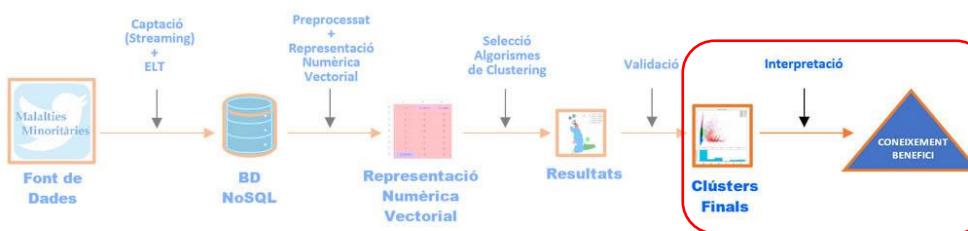
Visualització sense el clúster més gran.

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

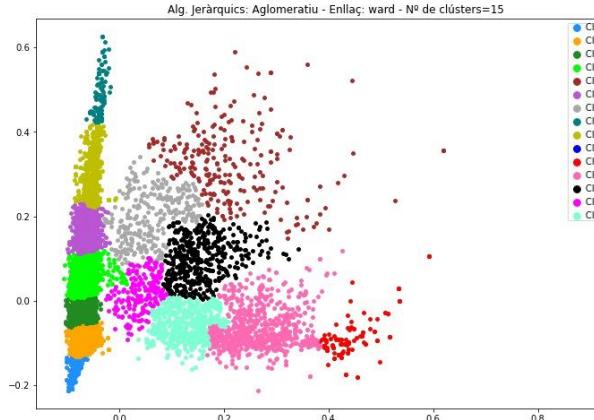
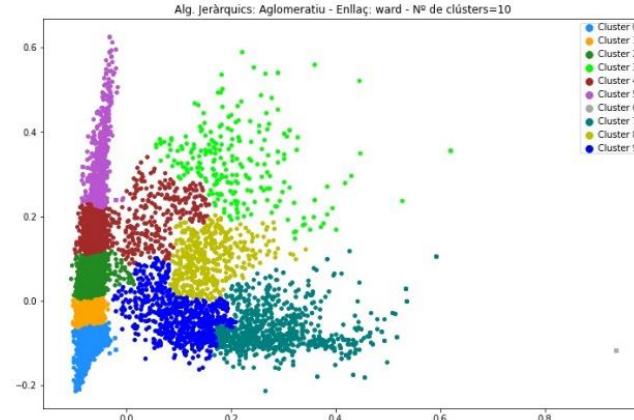
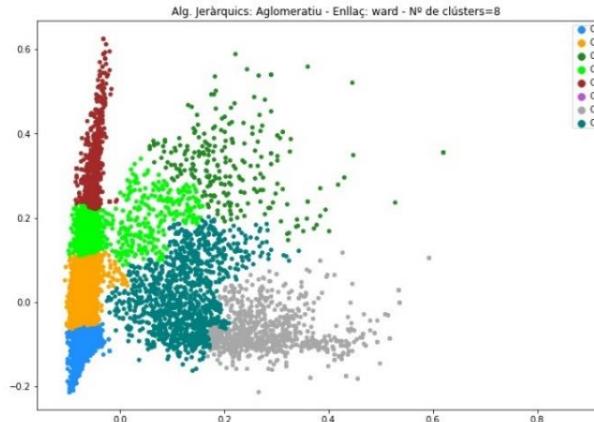
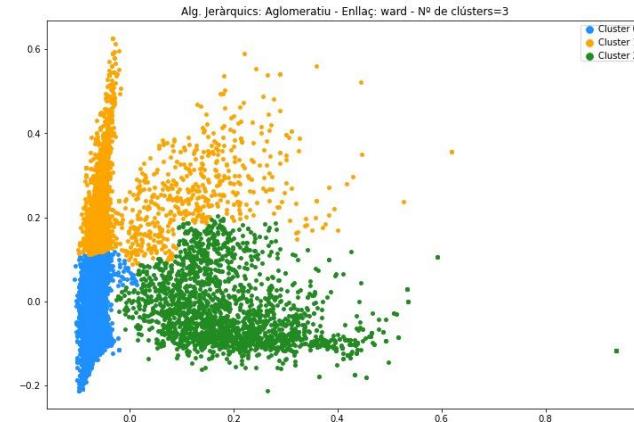


VISUALITZACIÓ DE RESULTATS



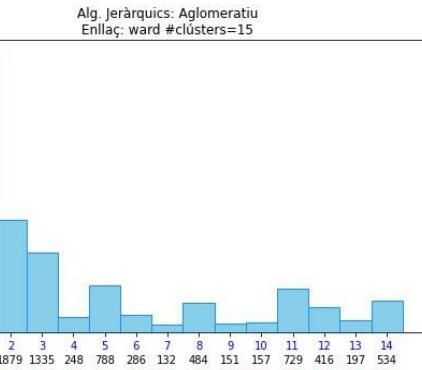
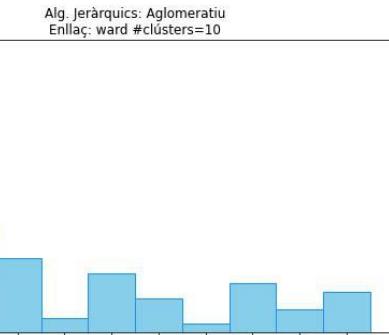
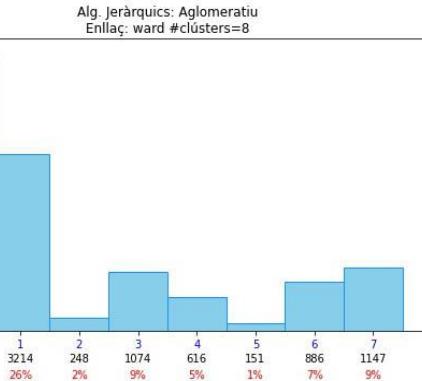
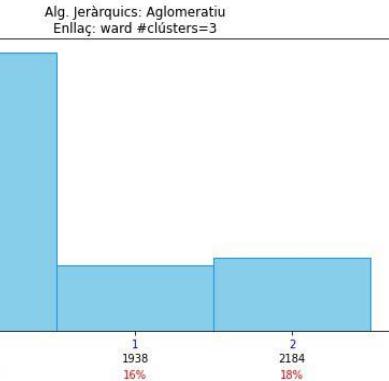
Paraulas
significatives
de clúster

Visualització 2D



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

Histograma





INTERPRETACIÓ DE L'AGRUPAMENT

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 famílies de pacients.

Cluster 4: Satisfacció de tots de celebrar el dia i prendre consciència 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 personnes.

Cluster 14: Petició de campanyes de suport a pacients i famílies a viure a més del dia MM.



Objectiu

ANÀLISI DE SENTIMENT amb KMeans

Conèixer temàtiques especialitzades.

Opinions a favor

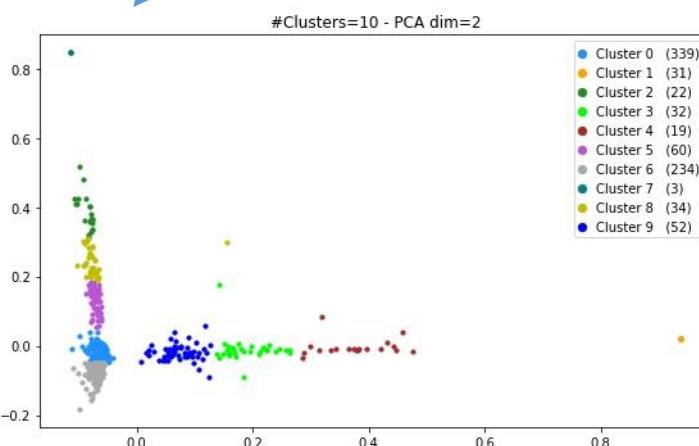
Agraïments, millores, casos d'èxit.

Opinions en negatives

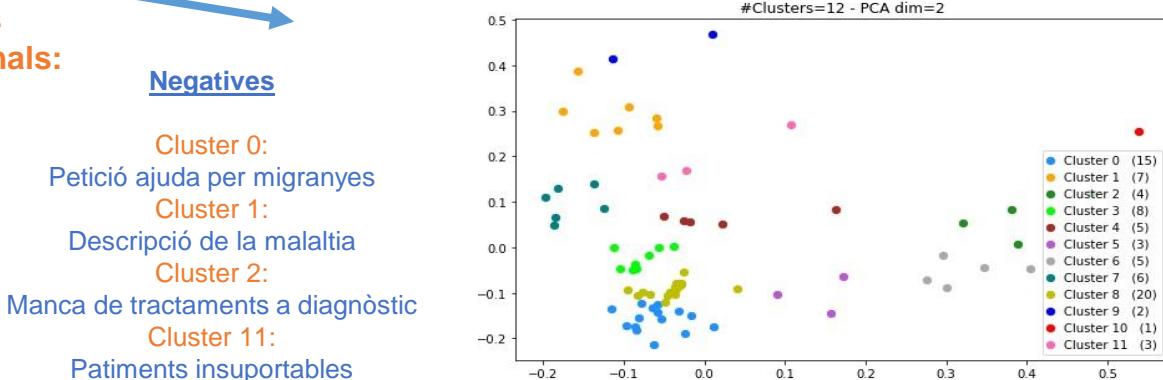
Denúncies, peticions, problemàtiques.

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.



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:
 Cluster 2:
 Cluster 3:
 Cluster 4:
 Cluster 5:
 Cluster 6:
 Cluster 7:
 Cluster 8:
 Cluster 9:
 Cluster 10:
 Cluster 11:



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



Gràcies per la vostra atenció i suport !