

FINAL PROJECT – SEARCH ENGINE

Students:

- | | |
|--|--------------------------|
| • Name: Ingrid Pérez Aguilera | NIA: 205536 |
| Email: ingrid.perez03@estudiant.upf.edu | U-number: U149892 |
| • Name: Clara Reolid Sánchez | NIA: 207531 |
| Email: clara.reolid01@estudiant.upf.edu | U-number: U151201 |

Subject: Information Retrieval and Web Analytics

Group: P102

DATA COLLECTION

The topic chosen for the project has been “Covid-19”. Therefore, in order to scrape Twitter Data relevant to the topic, the following keywords were selected:

- covid
- coronavirus
- pandemic
- virus
- #covid
- #covid19
- #coronavirus

The data collection process took approximately 5 hours.

The statistics of the data are the following:

- **Number of tweets** → 500526
- **Number of original tweets, i.e. tweets that are not retweets of other tweets** → 174423
- **Number of total tweets, by scraping the underlying tweets of those that were retweets** → 189909
- **Number of users with original tweets** → 143752

SEARCH ENGINE

PRE-PROCESSING

Description of the pre-processing strategy

The tweet text pre-processing strategy used is:

- Remove emojis
- Remove “RT” at the beginning of tweets that are not original

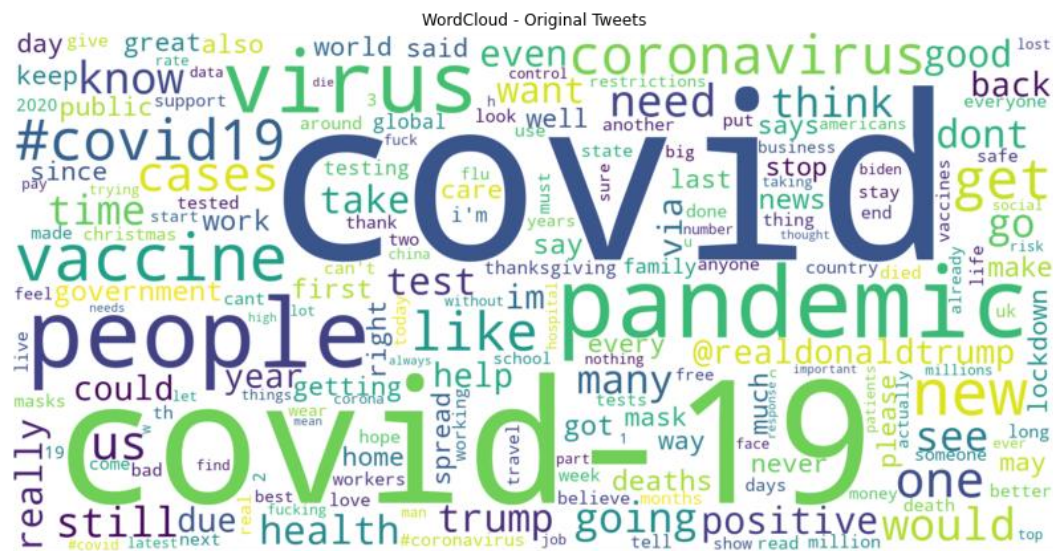
- Remove URLs and Webpage links
- Remove punctuation
- Convert to lowercase
- Remove stopwords
- Perform stemming

To create the search engine, only the text search field of the tweets has been used.

WORDCLOUD

WordCloud generated by the whole corpus (or at least a relevant sample). Only the words considered for the inverted index have to be included.

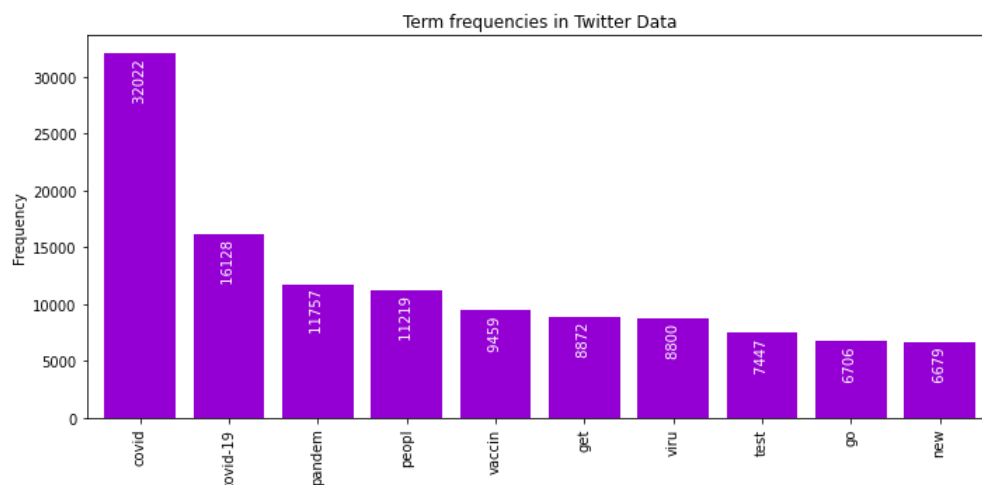
The Wordcloud of the words considered in the inverted index is:



BAR PLOT

Bar plot of the 10 most frequent words (same set of the WordCloud)

Similarly, the corresponding bar plot is:



OUR SCORE – INTERACTION

Description and example to describe your score to rank the documents


The created score used to rank the documents aims to assess the number of interactions the tweet has had. However, when filtering by the original tweets only and limiting the number of interactions to the information available in the scraped data, some of the tweets might be poorly scored. The number of interactions is considered to be the sum of the number of retweets, the number of likes/favourites, the number of replies and the number of quotes.

$$\text{Score} = \text{Retweets} + \text{Likes} + \text{Replies} + \text{Quotes}$$

As the data is scraped from Twitter, original tweets contain the information at the time of creation, which will be always 0 retweets, 0 likes, 0 replies and 0 quotes. Therefore, in order to obtain the updated information, it is necessary to iterate over the not original tweets to update the information about the interaction. However, this approach failed, since none of the original tweets were in the retweeted tweets.

The approach taken then was that for all the tweets that were retweets we would obtain the underlying original tweet and add it to the dataset of original tweets. Therefore, now the data contains tweets whose retweet, like, quote and reply count are non-zero.

For example, for the following tweet:

PPL RLY THINK THE VIRUS WILL MAGICALLY VANISH ON JANUARY 1ST, 2021


The obtained score was:

$$\text{Score} = \text{Retweets} + \text{Likes} + \text{Replies} + \text{Quotes} = 1770 + 39436 + 88 + 446 = 41740$$

SCORE COMPARISON

Screenshot to compare the difference between the classical ranking and the one by your score (screenshot of the output generated by the search-engine)

Searching the query “virus” using tf-idf index produces the following results:

=====

Top 20 results out of 8591 for the seached query:

```

1. Tweet: It's not about a virus User: montyigueldo Date: Tue Nov 24 09:46:08 +0000 2020 Hashtags: [] Likes: 0 Retweets: 0
2. Tweet: Just about a virus User: sunnysinghb82 Date: Mon Nov 23 21:08:28 +0000 2020 Hashtags: [] Likes: 0 Retweets: 0
3. Tweet: This is not about a virus User: YayNarwhals Date: Mon Nov 23 19:51:22 +0000 2020 Hashtags: [] Likes: 0 Retweets: 0
4. Tweet: this IS NOT about virus User: ShadyCrossing Date: Mon Nov 23 18:22:42 +0000 2020 Hashtags: [] Likes: 0 Retweets: 0
5. Tweet: What virus User: RobertTrenn Date: Mon Nov 23 18:21:44 +0000 2020 Hashtags: [] Likes: 0 Retweets: 0
6. Tweet: ❤️it's virUS ❤️ https://t.co/dNpbe9ojSS User: couple_images Date: Thu Nov 12 18:30:15 +0000 2020 Hashtags: [] Likes: 26045 Retweets: 4479
7. Tweet: @PattyArquette She had the virus and has antibodies. She will not spread the virus User: Debshell154 Date: Mon Nov 23 20:47:32 +0000 2020 Hashtags: [] Likes: 0 Retweets: 0
8. Tweet: @GovAndyBeshear "Virus virus virus, fear, death Virus, I'm sure he only one whom can save you morons from this viru... https://t.co/4Ig5lVwvWT User: FreeKentucky202 Date: Mon Nov 23 20:53:21 +0000 2020 Hashtags: [] Likes: 0 Retweets: 0
9. Tweet: Mosquito-Infecting Virus Espirito Santo Virus Inhibits Replication and Spread of Dengue Virus. https://t.co/Xyz4Ke8Frr User: mosquito_papers Date: Sun Nov 22 11:06:32 +0000 2020 Hashtags: [] Likes: 0 Retweets: 0
10. Tweet: @Elex_Michaelson Shut the F up douche. Virus is gonna virus. Nothing you can do to stop a virus. If masks and... https://t.co/ivpQ67xJBG User: SanchoDawg Date: Mon Nov 23 18:28:29 +0000 2020 Hashtags: [] Likes: 0 Retweets: 0

```

11. Tweet: Coronavirus is the common-cold

A..

Cold is not a disease

Virus is not a living organism

Virus has no nuclei

Virus... https://t.co/PcTM35FW6a User: Rent_not_buy Date: Fri Jul 17 20:14:30 +0000 2020 Hashtags: [] Likes: 90 Retweets: 79

```

12. Tweet: 🍷🍷 HAPPY VIRUS 🍷🍷🍷 https://t.co/JjyiUg1tV1 User: velaespa Date: Tue Nov 24 11:12:26 +0000 2020 Hashtags: [] Likes: 0 Retweets: 0
13. Tweet: You mustn't afraid of this virus https://t.co/LxStsEviqJ User: Braddok2070 Date: Tue Nov 24 10:52:55 +0000 2020 Hashtags: [] Likes: 0 Retweets: 0
14. Tweet: Never about the virus User: Caesch_33 Date: Tue Nov 24 10:26:02 +0000 2020 Hashtags: [] Likes: 2 Retweets: 1
15. Tweet: What global virus? https://t.co/GTVPM6wM0 User: RobNicholls15 Date: Tue Nov 24 10:10:31 +0000 2020 Hashtags: [] Likes: 0 Retweets: 0
16. Tweet: happy virus https://t.co/tVLFgjZk8 User: jwoomybo Date: Tue Nov 24 10:06:02 +0000 2020 Hashtags: [] Likes: 0 Retweets: 0
17. Tweet: fck the virus User: for_yjy1101 Date: Tue Nov 24 10:04:53 +0000 2020 Hashtags: [] Likes: 0 Retweets: 0
18. Tweet: Nothing to do with a virus User: nonetaken101 Date: Tue Nov 24 10:03:35 +0000 2020 Hashtags: [] Likes: 0 Retweets: 0
19. Tweet: This has nothing to do with a virus User: nonetaken101 Date: Tue Nov 24 10:01:49 +0000 2020 Hashtags: [] Likes: 0 Retweets: 0
20. Tweet: Let's name the virus THE TRUMP VIRUS - how about that? User: torewalaker Date: Tue Nov 24 09:49:11 +0000 2020 Hashtags: [] Likes: 0 Retweets: 0

```

Whereas, when using our defined score, the results are:

=====

Top 20 results out of 8591 for the seached query:

```

1. Tweet: stardew valley comes with this crazy virus that forces google chrome to open 53 tabs of stardewvalleywiki, 7 of the... https://t.co/pForRRKkA5p User: _itsjackielee Date: Wed Jan 22 05:51:09 +0000 2020 Hashtags: [] Likes: 5255 Retweets: 14150 Quotes: 531 Replies: 270
2. Tweet: The scariest thing about this Corona Virus is that you being careful is not enough, because your survival also depe... https://t.co/dgDStjp4Vp User: Deshysmalls Date: Sun Mar 22 17:35:16 +0000 2020 Hashtags: [] Likes: 3242 Retweets: 18205 Quotes: 548 Replies: 251
3. Tweet: Nancy Pelosi is an 80-year-old virus without a cure. User: MrMichaelBurkes Date: Wed Nov 18 18:01:53 +0000 2020 Hashtags: [] Likes: 37033 Retweets: 5553 Quotes: 665 Replies: 1130
4. Tweet: PPL RLY THINK THE VIRUS WILL MAGICALLY VANISH ON JANUARY 1ST, 2021 🤔🤔🤔 User: quackity4k Date: Sun Oct 11 19:25:32 +0000 2020 Hashtags: [] Likes: 39436 Retweets: 1770 Quotes: 88 Replies: 446
5. Tweet: The history will be a President who thoroughly conceded to a virus that could have been defeated and who refused to... https://t.co/iPUCCwF87M User: davidplouffe Date: Sun Nov 15 20:55:06 +0000 2020 Hashtags: [] Likes: 2639 Retweets: 6587 Quotes: 674 Replies: 436
6. Tweet: Right now, the most potent weapon against this virus is a mask. It is our shared duty to protect our fellow Americans... https://t.co/ZxHbwBYttE User: Transition46 Date: Tue Nov 10 00:41:34 +0000 2020 Hashtags: [] Likes: 2839 Retweets: 4585 Quotes: 543 Replies: 392
7. Tweet: ❤️it's virUS ❤️ https://t.co/dNpbe9ojSS User: couple_images Date: Thu Nov 12 18:30:15 +0000 2020 Hashtags: [] Likes: 26045 Retweets: 4479 Quotes: 828 Replies: 118
8. Tweet: better idea: name the virus after him! User: jimmykimmel Date: Fri Nov 20 17:52:54 +0000 2020 Hashtags: [] Likes: 18639 Retweets: 1088 Quotes: 371 Replies: 2103
9. Tweet: Another childhood friend has died of Covid, and three of my friends are now fighting it. This virus plays no favori... https://t.co/J881uDuEkN User: ConnieSchultz Date: Wed Nov 18 22:24:24 +0000 2020 Hashtags: [] Likes: 1919 Retweets: 1823 Quotes: 100 Replies: 468

```

```

10. Tweet: Me on my way to the link up knowing I'm safe from the virus https://t.co/NeADMxR3A5 User: drew2wavy
Date: Wed Sep 09 09:34:08 +0000 2020 Hashtags: [] Likes: 14718 Retweets: 3465 Quotes: 50 Replies: 16
11. Tweet: pornhub be like "yo phone got a virus" okay and? play the fuckin video User: jeshvs
6:38:49 +0000 2020 Hashtags: [] Likes: 15446 Retweets: 1370 Quotes: 42 Replies: 65
12. Tweet: As the virus overwhelmed us, they died trying to protect us.

They were more than a statistic.

The third in a ser... https://t.co/nKrxdiTz1L User: FacesOfCOVID Date: Wed Oct 28 15:55:00 +0000 2020 Hashtags:
[] Likes: 9359 Retweets: 4956 Quotes: 1575 Replies: 277
13. Tweet: RT IF YOU WANT ME TO BEAT THE FUCK OUT OF THE JABRONI CORONA VIRUS User: the_ironsheik Date: Mon A
pr 13 00:53:41 +0000 2020 Hashtags: [] Likes: 8906 Retweets: 5922 Quotes: 219 Replies: 169
14. Tweet: I might got that tourettes virus User: twomad Date: Fri Nov 20 05:59:09 +0000 2020 Hashtags: [] Lik
es: 11984 Retweets: 144 Quotes: 15 Replies: 130
15. Tweet: 🍀 Happy St. Patrick's Day🍀

DAY 1 Of Quarantine 🦠
Corona Virus Ain't Stop Us 🦠

[1] Full Video Available [1]
-... https://t.co/1EZC7WKhB3 User: onlyfansx9x26x Date: Tue Mar 17 14:11:22 +0000 2020 Hashtags: [] Likes: 9055
Retweets: 3100 Quotes: 38 Replies: 42
16. Tweet: On #COVID-19: If you want to do something useful today, go to Chinatown -- buy a meal, go shopping. The vir
us atta... https://t.co/vSNx1hFUzW User: RonaldKlain Date: Fri Feb 28 14:05:00 +0000 2020 Hashtags: ['COVID-
19'] Likes: 6934 Retweets: 2467 Quotes: 1399 Replies: 1087
17. Tweet: 5/6 It was U.S. that started the argument about the origin of #COVID19. U.S. first used "Chinese virus" &
"Wuhan vi... https://t.co/eHjNOiU51c User: zlj517 Date: Tue Mar 24 15:08:17 +0000 2020 Hashtags: ['COVID19'] Lik
es: 7636 Retweets: 998 Quotes: 167 Replies: 1488

18. Tweet: @realDonaldTrump 9 months on and he is *still* calling it the China Virus 😏 User: murray_nyc Date:
Mon Nov 16 16:49:27 +0000 2020 Hashtags: [] Likes: 7718 Retweets: 105 Quotes: 26 Replies: 255
19. Tweet: Wait a minute ... that means he already had the virus ... and he travelled all over the world last December
... Fol... https://t.co/09co10b6Dg User: pixelatedboat Date: Fri Nov 20 19:39:24 +0000 2020 Hashtags: [] Lik
es: 7057 Retweets: 752 Quotes: 12 Replies: 46
20. Tweet: "PCR test cannot tell you if you have virus or dead chunk of it. No other virus ever on this planet has been
accomp... https://t.co/pRXs4b3N1k User: dockaurG Date: Tue Nov 17 13:29:01 +0000 2020 Hashtags: [] Likes: 4496 Ret
weets: 2546 Quotes: 611 Replies: 151

```

By comparing the results obtained, it is possible to see that with our score we are retrieving the most popular tweets for the query, but not necessarily the match the intention of the user searching.

RESEARCH QUESTION

OUTPUT ANALYSIS

This research question has been answered using only the first 100000 tweets of the original data and then it has been pre-processed by removing the tweets that are not original and getting the underlying tweets of those tweets that were retweets. The duplicates have been removed.

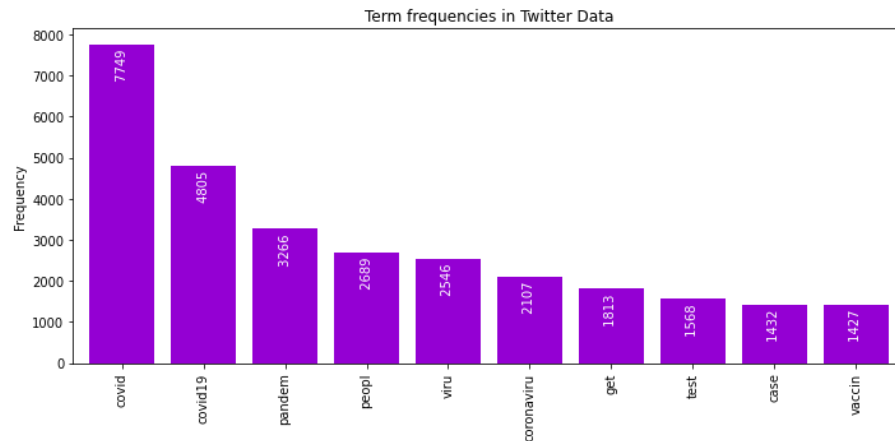
SECTION A

Choose 10 queries to run.

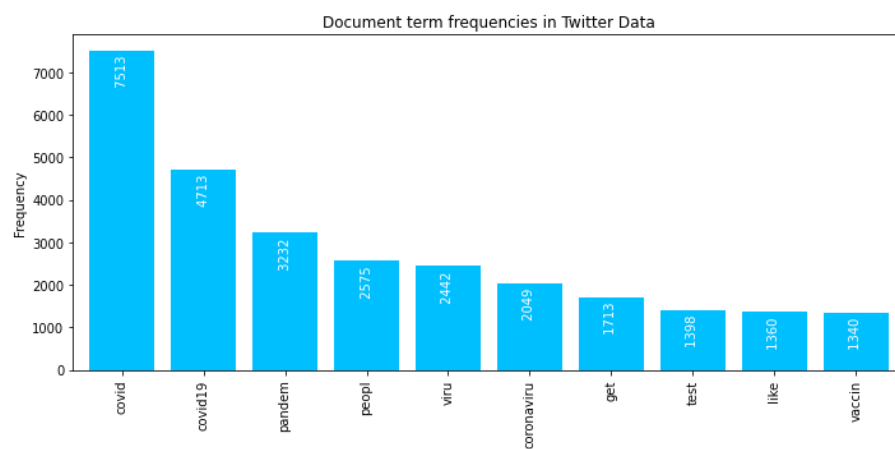
The 10 queries to run have chosen by taking into consideration the terms with most appearances in the dataset, the terms that appear in more documents and the terms that have higher idf.

The following plots summarise such information:

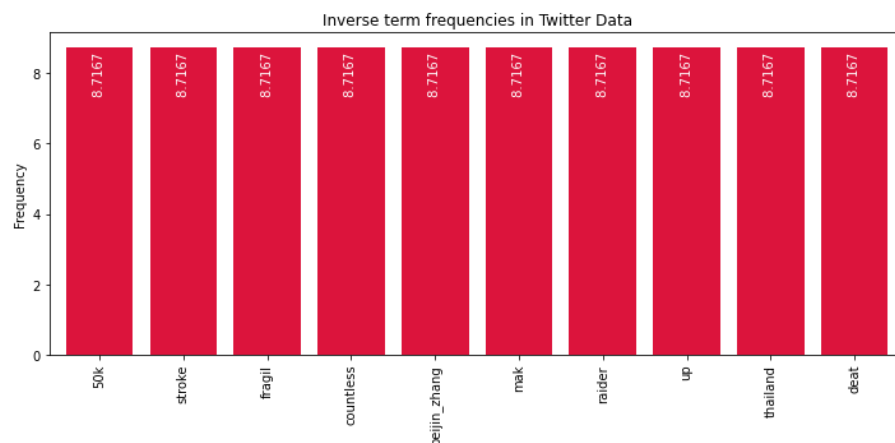
- Words with higher frequency in the documents



- Words with higher document frequency



- Words with higher inverse document frequency that appear in at least 5 documents



Therefore, the 10 queries are:

1. covid
2. pandemic
3. people
4. vaccine

5. virus
6. stroke
7. fragile
8. test
9. countless
10. thailand

These have been chosen as the most popular terms in the dataset and those with highest idf that appear in at least 5 documents.

SECTION B

Return a top-20 list of documents for each of the 10 queries, using tf-idf + cosine similarity methods.

The top-20 list of documents for each query can be found in the file: "top20_tfidf.tsv"

SECTION C

Return a top-20 list of documents for each of the 10 queries, using word2vec + cosine similarity.

The top-20 list of documents for each query can be found in the file: "top20_word2vec.tsv"

SECTION D

Can you imagine a better representation than word2vec? Justify your answer. (HINT - what about Doc2vec? Sentence2vec? Which are the pros and cons?)

The word2vec representation uses a neural network model to learn word associations from a large corpus of text. It represents each distinct word with a vector. Once trained, such a model can detect synonymous words or suggest additional words for a partial sentence. The vectors are chosen carefully such that a simple mathematical function (the cosine similarity between the vectors) indicates the level of semantic similarity between the words represented by those vectors.

The idea of word2vec can be extended to sentences and complete documents. In this case, the intention is to create a numeric representation of a sentence or a document, respectively, regardless of its length. In this case, instead of learning feature representations for words, these are learned for sentences or documents.

The same idea can be extended to sentences and complete documents where instead of learning feature representations for words, you learn it for sentences or documents. sentence2vec can be thought of as a mathematical average of the word vector representations of all the words in the sentence.

Regarding the applications, a word2vec effectively captures semantic relations between words hence can be used to calculate word similarities or fed as features to various NLP tasks such as sentiment analysis etc. However, words can only capture so much, there are times when you need relationships between sentences and documents and not just words. Such models can be used to detect plagiarism or similarity between sentences.

SECTION E

Choose one vector representation, tf-idf or word2vec, and represent the tweets in a two-dimensional scatter plot through the t-sne algorithm. To do so, you may need first to represent the word as a vector, and then the tweet, i.e., resulted as the average value over the words involved. Any other option rather than t-sne may be used, but needs to be justified.

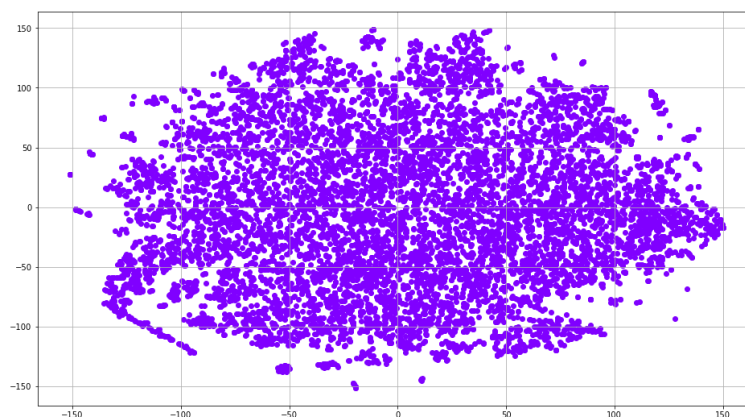
Using the word2vec representation of the words in the tweet, it is possible to approximate a document2vec model, where documents are tweets, by averaging the word2vec representation of the words present in the tweet.

RESEARCH QUESTION 1A

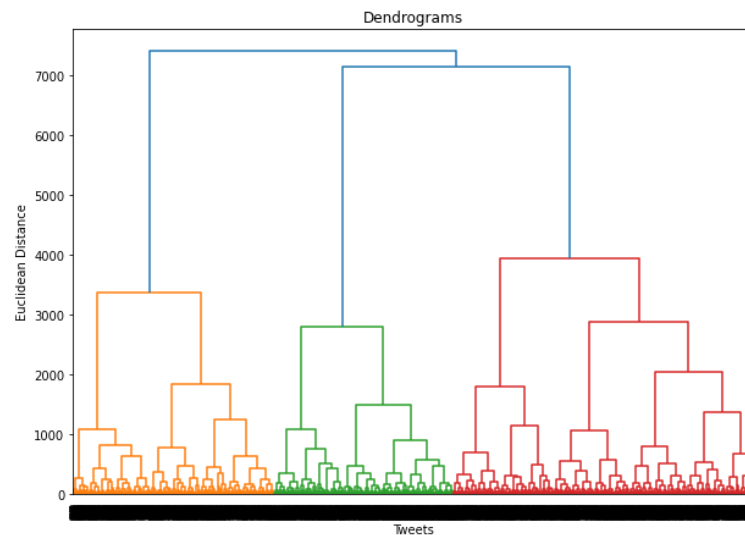
Are you able to detect some subgroups within your tweets representation? Are you able to perform some clustering over the tweets and detect some topics within the conversation? How do you choose the best possible number of clusters?

Due to memory problems, it was not possible to run the hierarchical clustering in all of the 100 k tweets, it has been performed only on a sample of 10000 tweets.

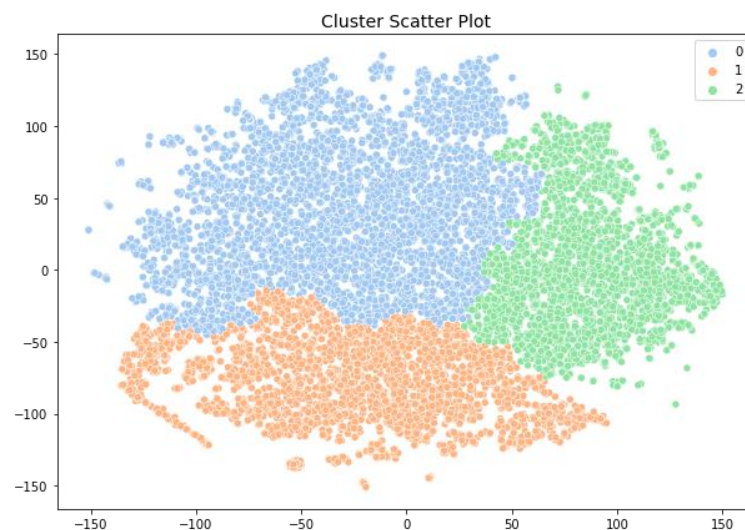
The output of the TSNE algorithm result in the following plot:



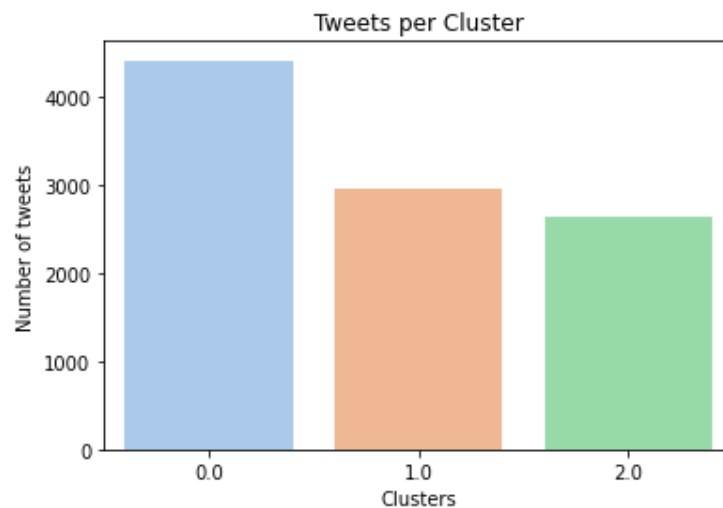
From it, it is possible to see that there is not a clear division of the tweets in different groups. Therefore, in order to perform some clustering and detect some topics within the conversation, we have applied hierarchical clustering. This approach already produces the number of optimal clusters by displaying the dendrogram:



Then, we can apply agglomerative clustering with the number of clusters obtained from the dendrogram, which in this case, it is 3. This results in the following tweet plot with each tweet assigned a colour corresponding to the cluster it belongs to:



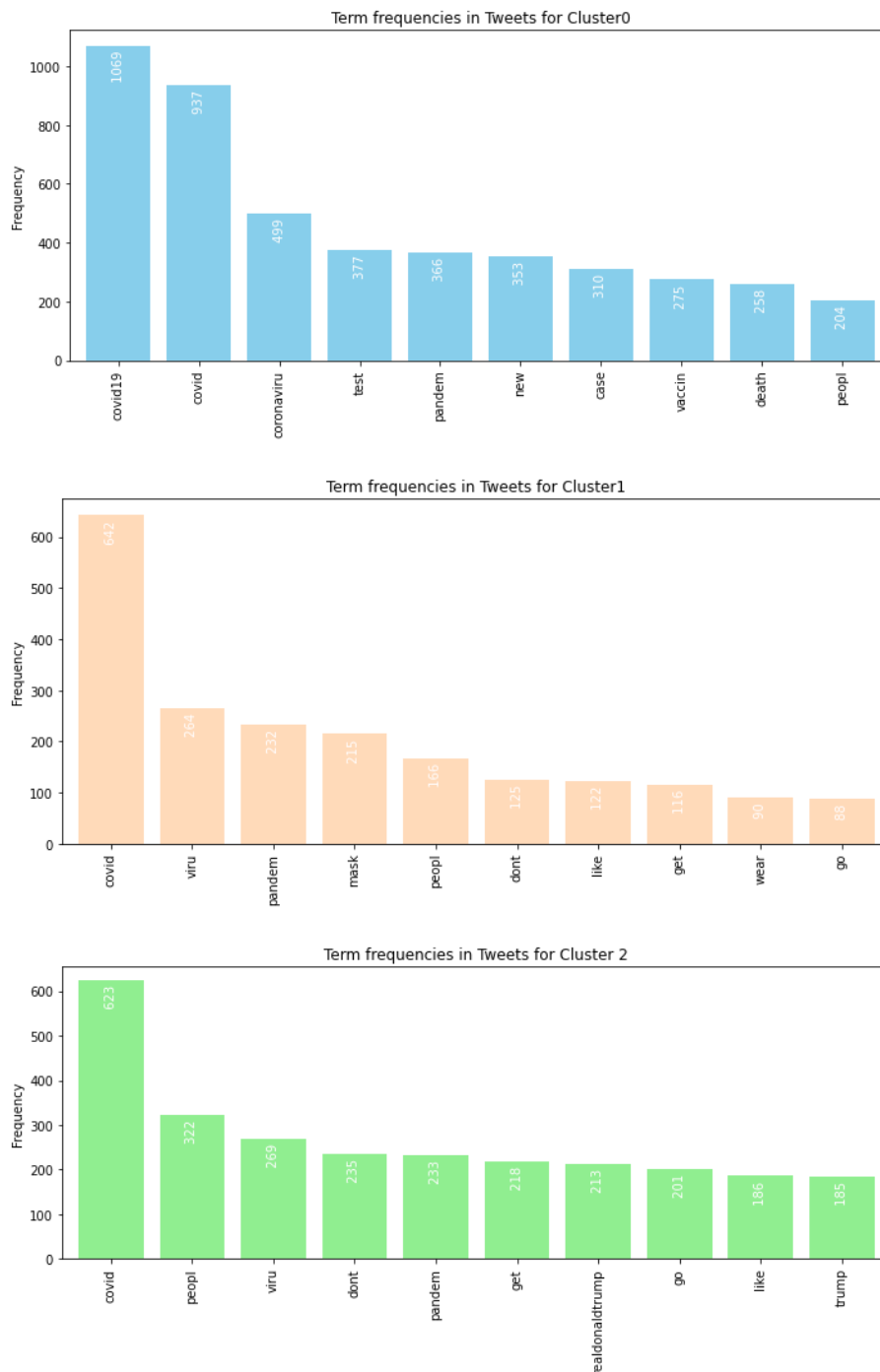
The number of tweets per cluster is:



RESEARCH QUESTION 1B

What are the 5 most relevant keywords in the tweets that are part of each cluster? To what extent these keywords characterize/separate well the clusters?

Since the 5 most relevant keywords in the tweets were not significant to characterise them, the top 10 have been displayed instead. The top 10 words for each cluster are:



Although there are some words that are common in all three clusters, it is possible to identify a common topic for each:

- Cluster 0 → Vaccine and COVID-19 deaths

- Cluster 1 → Mask wearing
- Cluster 2 → Donald Trump and COVID-19.

Therefore, these words do characterise well the clusters.

OUTPUT DIVERSIFICATION

SECTION A

Define the best possible number of clusters and assign the cluster labels to each document.

The number of clusters to be used will be the same ones as in the previous exercise. However, the amount of data used is only 10000 original tweets, since it was computationally unfeasible to classify all the data.

SECTION B

Define a diversity score which the aim is to diversify the final output. This score is assigned to the list of returned documents for the input query.

The approach to assign a diversity score follows the reference: Song, Kai & Tian, Yonghong & Gao, Wen & Huang, Tiejun. (2006). Diversifying the image retrieval results. Proceedings of the 14th Annual ACM International Conference on Multimedia, MM 2006. 707-710. 10.1145/1180639.1180789.

The approach taken is:

1. Compute the matrix M to store the similarity between a pair of tweets defined as the Jaccard similarity on the number of words.

The similarity between document d_j and document d_k is:

$$m_{jk} = \text{sim}(d_j, d_k) = \begin{cases} \frac{|w_j \cap w_k|}{|w_j \cup w_k|} & \text{if } k \neq j \\ 0 & \text{if } k = j \end{cases}$$

where w_j is the words contained in document d_j .

2. Compute the Topic Richness based on the intuition that the higher score an image's neighbour obtains, the higher the image score will be. This is similar to PageRank and, therefore, it can be computed by obtaining the eigenvector with highest eigenvalue of the matrix defined as:

$$A = cM + \frac{1-c}{n}E$$

where c is a dumping factor to prevent that if the matrix M contains zeros, the eigenvector computation does not fail and E is the matrix of ones.

3. The diversity score of a ranking for a given query is defined as the average diversity score of the n documents returned in the ranking. Then, the diversity

score of a document d_j in a set of documents $D = \{d_1, \dots, d_n\}$ annotated with m_j different words is:

$$DS_D(d_j) = \frac{1}{m_j} \sum_{j=1}^{m_j} \frac{1}{N_{w_j}^D}$$

where $N_{w_j}^D$ denotes the number of documents in set D that contain the word w_j .

SECTION C

Now, defining a method to diversify the output through the diversity score defined above, try to generate a still relevant but more diverse final top-k list of documents.

The re-ranking approach taken is:

1. Compute 100 top documents using the traditional ranking method.
2. For each of the top-k documents that want to be retrieved:
 - a. Rank the output of traditional ranking according to Topic Richness.
 - b. Add to the re-ranking results the document with highest Topic Richness.
 - c. Recompute the Topic Richness for the remaining documents in the following way:

$$TR(d_k) = TR(d_k) - M_{jk} \cdot TR(d_j)$$

where d_j is the chosen document and M_{jk} is the entry in the similarity matrix.

RESEARCH QUESTION 2A

Test your new method on some queries, comparing the 2 outputs, before and after the re-ranking and comment the results.

The queries selected are:

- test
- coronavirus
- realdonaldtrump

The output for the top20 ranking of these queries can be found in the file "top20_ranking.tsv".

The output for the top20 reranking of these queries can be found in the file "top20_reranking.tsv".

The ranking results for the query "test" are:

1. Tweet: "As the @WHO's Dr. Tedros said very emphatically at the very outset: "Test, test, test". As a physician, I wish he'... <https://t.co/n3Dn2Nwn1E> User: manigreeva Date: Sun Nov 22 11:12:43 +0000 2020 Hashtags: [] Cluster: 0.0 TopicRichness 9.512778224714583e-05

2. Tweet: Loeffler tests positive for Covid but undergoing further testing <https://t.co/5u1PDNmhne> User: bote930 Date: Sun Nov 22 11:26:46 +0000 2020 Hashtags: [] Cluster: 0.0 TopicRichness 0.0018463648458519586

3. Tweet: @mel_faith1 The thing is is there is NO test for Covid! The test is for a virus only User: susanlee52 Date: Sun Nov 22 11:16:15 +0000 2020 Hashtags: [] Cluster: 0.0 TopicRichness -0.00032220929354842965

4. Tweet: Testing. Testing. Covid/TheDanes: 'They kept it up for a month.1.8% of mask wearers tested positive & 2.1% of the u... <https://t.co/Oc5tGJDuuH> User: Fuerza_Mundial Date: Sun Nov 22 11:04:34 +0000 2020 Hashtags: [] Cluster: 0.0 TopicRichness 0.0007356799139667318

5. Tweet: Loeffler tests positive for Covid but undergoing further testing <https://t.co/c06EbcUbZA> User: DeeFonta Date: Sun Nov 22 10:56:01 +0000 2020 Hashtags: [] Cluster: 0.0 TopicRichness 0.005544894615846776

6. Tweet: @GBPublicHealth Is every patient that takes a Covid test taking a flu test too? Did you stop testing for the flu al... <https://t.co/mTMNKwvEox> User: theresabeverag1 Date: Sun Nov 22 11:17:51 +0000 2020 Hashtags: [] Cluster: 0.0 TopicRichness 0.0001079470604193416

7. Tweet: Georgia Sen. Kelly Loeffler tests positive for Covid but is undergoing further testing. <https://t.co/rYLPJhyaQw> User: MrsVSNCO9 Date: Sun Nov 22 11:09:31 +0000 2020 Hashtags: [] Cluster: 0.0 TopicRichness 0.00021246960902998092

8. Tweet: KARMA.....Georgia Sen. Kelly Loeffler tests positive for Covid but is undergoing further testing - CNN... <https://t.co/JFRCgDaqko> User: m0m19001 Date: Sun Nov 22 11:18:43 +0000 2020 Hashtags: [] Cluster: 0.0 TopicRichness -2.3258388781474113e-05

9. Tweet: @safiume Or better testing: <https://t.co/Iaif2cdF7b> User: nahumshalman Date: Sun Nov 22 11:17:39 +0000 2020 Hashtags: [] Cluster: 0.0 TopicRichness 1.5596213889095625e-05

10. Tweet: #YEDNOnCovid19 have you been tested for #coronavirus ?if not ,visit the #coronavirus center to get tested?... <https://t.co/dt3dDD20xd> User: MutieMumbua Date: Sun Nov 22 11:16:00 +0000 2020 Hashtags: ['YEDNOnCovid19', 'coronavirus'] Cluster: 0.0 TopicRichness -0.0002665741964422599

11. Tweet: ...and testing negative for ethics. User: DanDiego Date: Sun Nov 22 11:10:43 +0000 2020 Hashtags: [] Cluster: 0.0 TopicRichness -0.00047904748160322987

12. Tweet: @DailyMirror Mass testing should not stop at 2 test but should be continuously carried out until we know where the virus is and who has it . User: STSIMON95984452 Date: Sun Nov 22 11:09:40 +0000 2020 Hashtags: [] Cluster: 0.0 TopicRichness -0.00026378831698467195

13. Tweet: @JonLoflin that's the problem. this test cannot tell and as virus is more spread in a population, the % of tests t... <https://t.co/A0Kd0X4xeF> User: boriaguato Date: Sun Nov 22 11:05:43 +0000 2020 Hashtags: [] Cluster: 0.0 TopicRichness -0.00028522254319739924

14. Tweet: Over 170 million coronavirus tests have already been given in the U.S. and it's territories. The tests are available... <https://t.co/mNXppr2AQp> User: suebrown1212 Date: Sun Nov 22 10:58:53 +0000 2020 Hashtags: [] Cluster: 0.0 TopicRichness -0.00046320579387470166

15. Tweet: **Nepal COVID-19 Status**
Tests (PCR): 1,660,075
Tests (RDT): 312,402
Positive: 220,308
Recovered: 199,024
In Isolation... <https://t.co/UZeNF075zy> User: NepalCorona Date: Sun Nov 22 11:19:23 +0000 2020 Hashtags: [] Cluster: 0.0 TopicRichness 0.00018272871298968536

16. Tweet: Kelly Loeffler tests positive for Covid but undergoing further testing - CNNPolitics #Mtnews #Georgia #CDC <https://t.co/BevWrVugTC> User: stewardmagazine Date: Sun Nov 22 11:17:41 +0000 2020 Hashtags: ['Mtnews', 'Georgia', 'CDC'] Cluster: 0.0 TopicRichness -0.0004969932538576234

17. Tweet: "Testing doesn't matter" -my family in reference to COVID testing before any thanksgiving plans.....

Word... whe... <https://t.co/aOnzp7EccS> User: megadillo Date: Sun Nov 22 11:12:42 +0000 2020 Hashtags: [] Cluster: 0.0 TopicRichness 3.294920318003694e-05

18. Tweet: @jksmith34 Test and trace is a joke everywhere and testing someone for a virus that can catch a minute after the te... <https://t.co/iEpxhLsWzb> User: SFotonium Date: Sun Nov 22 11:10:08 +0000 2020 Hashtags: [] Cluster: 0.0 TopicRichness -3.818399839290601e-05

19. Tweet: BOMBHELLLLL! DO NOT GET TESTED FOR COVID. PCR TESTS ARE a TOTAL FRAUD. 100% PROOF | jmviverlivre <https://t.co/o3GLvykjNkr> User: oscar226622 Date: Sun Nov 22 11:09:31 +0000 2020 Hashtags: [] Cluster: 0.0 TopicRichness 0.00026555644196820863

20. Tweet: Most #COVID19 test sites are open everyday, including at the weekend.

If you have any symptoms, book a test.

Rem... <https://t.co/LBotsgxI1b> User: HyndburnCouncil Date: Sun Nov 22 11:08:13 +0000 2020 Hashtags: ['COVID19'] Cluster: 0.0 TopicRichness -5.467215652865927e-05

The reranking results for that same query are:

```

1. Tweet: Loeffler tests positive for Covid but undergoing further testing
https://t.co/c06EbCubZA User: DeeFonta Date: Sun Nov 22 10:56:01 +0000 2020 Hashtags: [] Cluster: 0.0 TopicRichne
ss 0.005544894615846776
2. Tweet: Loeffler tests positive for Covid but undergoing further testing
https://t.co/5uLPDNmhne User: bote930 Date: Sun Nov 22 11:26:46 +0000 2020 Hashtags: [] Cluster: 0.0 TopicRichne
ss 0.001820787038797336
3. Tweet: Testing. Testing. Covid/TheDanes: 'They kept it up for a month.1.8% of mask wearers tested positive & 2.
1% of the u... https://t.co/Oc5tGJDuuH User: Fuerza_Mundial Date: Sun Nov 22 11:04:34 +0000 2020 Hashtags: [] Clu
ster: 0.0 TopicRichness 0.0007304527081733344
4. Tweet: BOMBHELLLLL! DO NOT GET TESTED FOR COVID. PCR TESTS ARE a TOTAL FRAUD. 100% PROOF | jmviverlivre https://t.c
o/3GLvykjNKR User: oscar226622 Date: Sun Nov 22 11:09:31 +0000 2020 Hashtags: [] Cluster: 0.0 TopicRichne
ss 0.00025967328755546683
5. Tweet: Georgia Sen. Kelly Loeffler tests positive for Covid but is undergoing further testing.
https://t.co/rYLP1hyaQw User: MrsVSNCO9 Date: Sun Nov 22 11:09:31 +0000 2020 Hashtags: [] Cluster: 0.0 Top
icRichness 0.00018948810820956137
6. Tweet: Nepal COVID-19 Status
Tests (PCR): 1,660,075
Tests (RDT): 312,402
Positive: 220,308
Recovered: 199,024
In Isolation... https://t.co/UZEnF075zy User: NepalCorona Date: Sun Nov 22 11:19:23 +0000 2020 Hashtags: [] Clu
ster: 0.0 TopicRichness 0.00017581949270011724
7. Tweet: @GBPublicHealth Is every patient that takes a Covid test taking a flu test too? Did you stop testing for the
flu al... https://t.co/mTMNKwEox User: theresabeaverag1 Date: Sun Nov 22 11:17:51 +0000 2020 Hashtags: [] Cluster: 0.
0 TopicRichness 0.00010098637185379432
8. Tweet: "As the @WHO's Dr. Tedros said very emphatically at the very outset: "Test, test, test". As a physician, I wi
sh he'... https://t.co/n3Dn2Nwn1E User: manigreeva Date: Sun Nov 22 11:12:43 +0000 2020 Hashtags: [] Cluster: 0.
0 TopicRichness 9.135135520304554e-05
9. Tweet: "Testing doesn't matter" -my family in reference to COVID testing before any thanksgiving plans.....

Word... whe... https://t.co/aOnzp7EccS User: megadillo Date: Sun Nov 22 11:12:42 +0000 2020 Hashtags: [] Clu
ster: 0.0 TopicRichness 2.6278035694216336e-05
10. Tweet: @safiume Or better testing: https://t.co/Iaif2cdF7b User: nahumshalman Date: Sun Nov 22 11:17:39 +
0000 2020 Hashtags: [] Cluster: 0.0 TopicRichness 8.727732723691942e-06
11. Tweet: @jksmith34 Test and trace is a joke everywhere and testing someone for a virus that can catch a minute after
the te... https://t.co/iEpxhLsWzb User: SFotonium Date: Sun Nov 22 11:10:08 +0000 2020 Hashtags: [] Cluster: 0.
0 TopicRichness -4.189142340017504e-05
12. Tweet: KARMA.....Georgia Sen. Kelly Loeffler tests positive for Covid but is undergoing further testing - CNN... htt
ps://t.co/JFRGdDagko User: m0m19001 Date: Sun Nov 22 11:18:43 +0000 2020 Hashtags: [] Cluster: 0.0 TopicRichne
ss -4.535403672692421e-05
13. Tweet: Most #COVID19 test sites are open everyday, including at the weekend.

If you have any symptoms, book a test.

Rem... https://t.co/LBotsgxI1b User: HyndburnCouncil Date: Sun Nov 22 11:08:13 +0000 2020 Hashtags: ['COVID19'] Clu
ster: 0.0 TopicRichness -5.833236441457625e-05
14. Tweet: @DailyMirror Mass testing should not stop at 2 test but should be continuously carried out until we know wher
e the virus is and who has it . User: STSIMON95984452 Date: Sun Nov 22 11:09:40 +0000 2020 Hashtags: [] Cluster: 0.
0 TopicRichness -0.0002676692835679989
15. Tweet: #YEDNOnCovid19 have you been tested for #coronavirus ?if not ,visit the #coronavirus center to get tested?... h
ttps://t.co/dt3dDD20xd User: MutieMumbua Date: Sun Nov 22 11:16:00 +0000 2020 Hashtags: ['YEDNOnCovid19', 'corona
virus', 'coronavirus'] Cluster: 0.0 TopicRichness -0.00027142026437980003
16. Tweet: @JonLoflin that's the problem. this test cannot tell and as virus is more spread in a population, the % of t
ests t... https://t.co/A0Kd0X4xeF User: boriaguato Date: Sun Nov 22 11:05:43 +0000 2020 Hashtags: [] Cluster: 0.
0 TopicRichness -0.0002885272263087329
17. Tweet: @mel_faith1 The thing is is there is NO test for Covid! The test is for a virus only User: susanlee52
Date: Sun Nov 22 11:16:15 +0000 2020 Hashtags: [] Cluster: 0.0 TopicRichness -0.00033129577427838485

18. Tweet: Over 170 million coronavirus tests have already been given in the U.S. and it's territories. The tests are av
ailabl... https://t.co/mNXppr2AQp User: suebrown1212 Date: Sun Nov 22 10:58:53 +0000 2020 Hashtags: [] Cluster: 0.
0 TopicRichness -0.0004662272775657904
19. Tweet: ...and testing negative for ethics. User: DanDiego Date: Sun Nov 22 11:10:43 +0000 2020 Hashtags:
[] Cluster: 0.0 TopicRichness -0.0004842466563299587
20. Tweet: Kelly Loeffler tests positive for Covid but undergoing further testing - CNNPolitics #Mtnews #Georgia #CDC h
ttps://t.co/BevWrVugTC User: stewardmagazine Date: Sun Nov 22 11:17:41 +0000 2020 Hashtags: ['Mtnews', 'Georgia', 'CD
C'] Cluster: 0.0 TopicRichness -0.0004969932538576234

```

It is possible to see that the results are slightly different, specially when it comes to topic richness since the re-ranked results contain those tweets have cover more topics.

The diversity score for the query is:

QUERY	DIVERSITY
test	12.5794
coronavirus	13.7359
realdonaldtrump	15.0723

RESEARCH QUESTION 2B

What about the coverage? Any difference between the two rankings (with AND without diversity score)?

The spearman correlation is:

QUERY	RANKING
test	0.1398
coronavirus	0.1805
realdonaldtrump	0.0391

The coverage comparison of the two rankings for some queries is:

QUERY	RANKING	RERANKING
test	1	2
coronavirus	1	2
realdonaldtrump	2	3

LINK ANALYSIS

COMMENT

Given the difference of approaches proposed by the teachers, we have implemented first the NDCG approach and, given the bad results, we tried the accuracy approach.

Regarding the NDCG approach, the implementation follows the idea:

1. Generate the retweet graph
2. Select a subset of edges to be the test set
3. Remove the test set edges from the graph
4. Obtain all possible recommendations from the deleted graph for those nodes in the test set
5. Obtain the top 10 recommendations for a subset of nodes that have at least an edge in the test set for each of the proposed algorithms.
6. Compute the NDCG score for each of the test nodes and compute the average NDCG score of the nodes to obtain the NDCG score of the algorithm.

The NDCG scores for the algorithms were:

- Adamic Adar → 0.1166

- Jaccard $\rightarrow 0.1205$
- PageRank $\rightarrow 0.3915$

Therefore, since the results were not as expected, we took the following approach:

1. Generate the retweet graph.
2. Select a subset of edges to be the test set.
3. Remove the test set edges from the graph.
4. Obtain all possible recommendations from the deleted graph for those nodes in the test set.
5. Use the algorithms to predict the probability of the edge for each edge in the test set.
6. Compute the accuracy of the algorithms and the precision and recall.

Note that when computing the PageRank, we only compute it for the top 10 recommendations, otherwise the computation time would be very high. In this way, we assume that if a node is not in the top 10 its PageRank would be so low that its predicted score would be 0.

RESEARCH QUESTION 3A

Which is the best algorithm among the 4 selected in terms of accuracy?

The accuracy for the algorithms is:

- Adamic-Adar $\rightarrow 0.4968$
- Jaccard $\rightarrow 0.4173$
- PageRank $\rightarrow 0.5$
- Alternative Least Squares $\rightarrow 0.5$

The average precision for algorithms is:

- Adamic-Adar $\rightarrow 0.4999$
- Jaccard $\rightarrow 0.4985$
- PageRank $\rightarrow 0.5000$
- Alternative Least Squares $\rightarrow 0.5000$

Therefore, it is possible to state that, in terms of accuracy, PageRank and Alternative Least Squares are the ones that perform best. However, in terms of average precision, PageRank, Adamic Adar and Alternative Least Squares have more or less the same values.

RESEARCH QUESTION 3B

Now, trying to exploit other features, like text from tweets or other users' information try to answer the next question.

Propose a new strategy to predict the links in the test-set. Which is the accuracy of the new algorithm if compared with the previous ones? Explain in detail the strategy you used and prove its effectiveness.

Ideally, we would have liked to predict whether there exists an edge between two users by taking into account the number of common followers and common friends. However, this information is computationally and time expensive to obtain due to the crawling limitations of the Twitter API.

Therefore, we have chosen to predict whether an edge exists by the number of retweets, favourites, quotes and replies of the underlying tweet. This means that, if user A retweets a tweet from user B, the interaction metrics would be collected for the tweet being retweeted.

In order to do so, when getting the interaction metrics for the possible recommendations, we check whether user of n1 or user of n2 have that information and assign it to the link with source n1 and destination n2 if n1 contains the information or otherwise.

Therefore, we now have a dataset with the following information:

- Source → User Screen Name
- Destination → User Screen Name of the user whose tweet was retweeted
- Retweets → Number of retweets of the tweet of the user in destination
- Favourites → Number of favourites of the tweet of the user in destination
- Quotes → Number of quotes of the tweet of the user in destination
- Likes → Number of favourites of the tweet of the user in destination
- Edge → Indicates whether the edge is in the graph or it is a possible recommendation.

Then, the dataset containing the information has been split with 80% of data for training and 20 % of data for testing.

The model used to classify whether the edge was present or not is an SVM classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall.

A support-vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class, since in general the larger the margin, the lower the generalization error of the classifier.

Hence, when using this approach, the accuracy in the test set is:

$$accuracy = 0.9629$$

and the average precision score is:

$$avg.precision = 0.9597$$