Joan Medina Martí
Guillem Morgó Homs

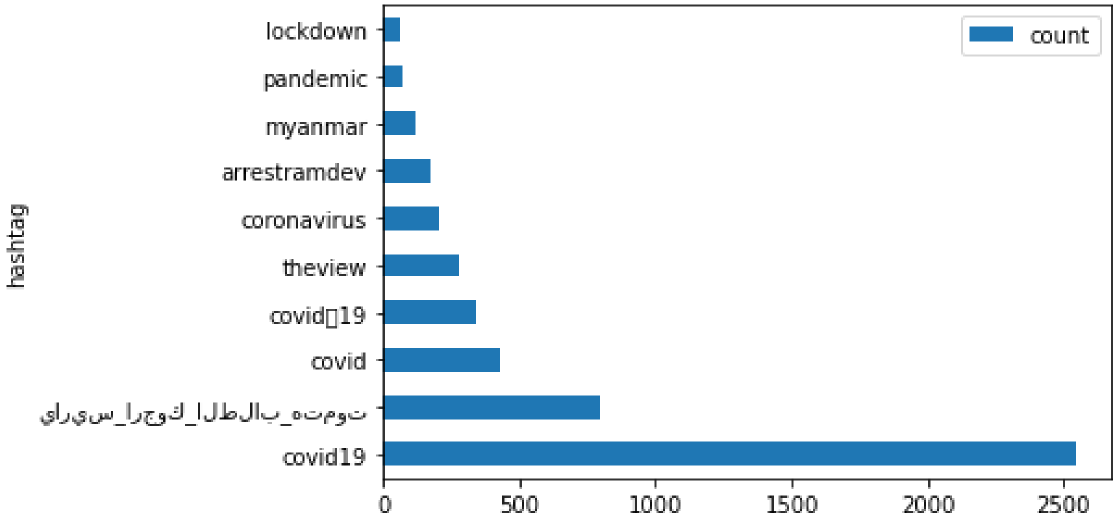
SECTION 1: DATA COLLECTION

Data information

The total number of tweets we obtained directly from the scrapper is more than 50 thousands of tweets, from which we had:

- Number of retweets: 45734
- Number of original tweets: 23727
- Number of unique users: 52502

It has also been analysed the hashtag frequencies and the image below represent a bar plot showing it:



We can see that the majority of them are hashtags related with the coronavirus such us #covid19, #coronavirus, #pandemic and so on, which means that the scrapper has done a good performance.

Keywords

For the data collection, the array of words that were passed to the scrapper was the following one:

["#covid19", "#Covid19", "#COVID19", "#covid", "#coronavirus", "#pcr", "#Covid-19", "#quarantine", "confinement", "Covid19", "coronavirus", "COVID19", "PCR", "Covid-19", "#Covid-19", "Coronavirus", "Antibody", "Social distancing", "epidemic", "pandemic", "Asymptomatic", "vaccine"]

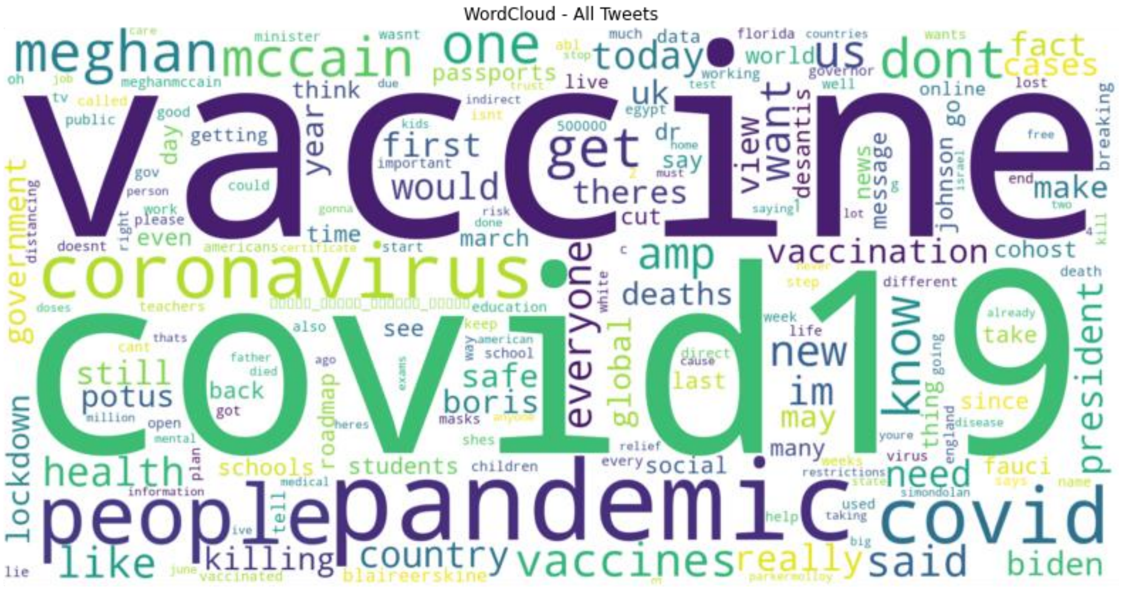
We tried to use all the words related to this topic so the scrapper was able to collect as many tweets as it can.

Time for the data collection

The algorithm for the data collection took a lot of time to do it, so we decided to use the 50,000 tweets we had so far that lasted around 1 hour to be collected

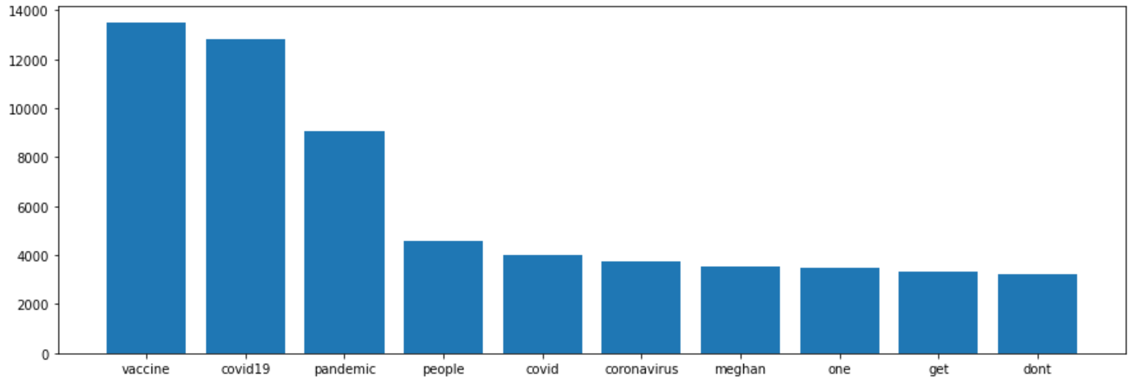
SECTION 2: SEARCH ENGINE

WordCloud & Bar plot



In this WordCloud we can see the most common words which are Covid19 or vaccine which is logic given that we are scrapping tweets related to this topic. In a lower order we can also see less related words which are frequent within the tweets such as ‘Meghan’, ‘people’, ‘government’, etc.

In this bar plot below we can see more precisely the frequency of the most used words:



It is curious to see that it is being more talked about the vaccine rather than covid itself.

Joan Medina Martí
Guillem Morgó Homs

Pre-processing

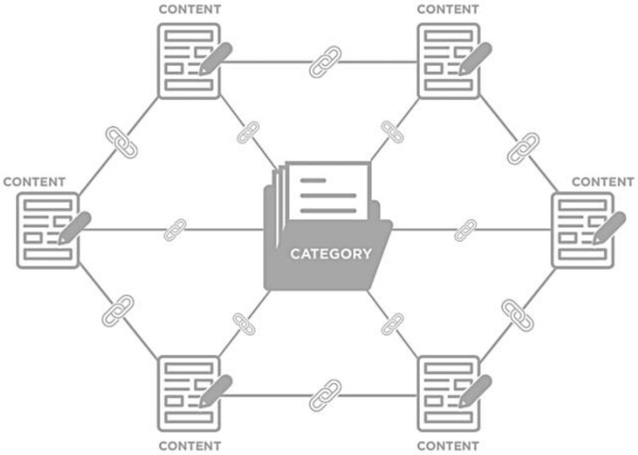
To be able to perform all the computations for the search engine, first of all one need to do some data manipulation such as cleaning the data in order to normalize it. All the functions implemented to make it possible are the following ones:

- Remove links
- Remove hashtags
- Convert text to lowercase
- Remove text marks: such as interrogative sign, comas, dots, etc.
- Remove accents
- Remove emojis
- Split text and numbers into different words
- Remove alone numbers
- Remove multiple whitespaces
- Remove punctuation marks
- Remove stopwords: most frequent words which are non-informative such as articles
- Stemming: Convert all remaining words into their root word: playing, players → play

Cluster-based search engine

Our developed ranking consists on making a tf-idf score but instead of returning the top k elements we only return the first one (text column) and we define it as a new query. We have developed a new function in order to make a kind of cluster (similar tweets) to the initial one, returning the top 19 nearest ones. Thus, we have developed another score ranking different from the initial one (tf-idf) and having as an output 20 similar tweets that match with the query.
The screenshots showing the comparison of the two score rankings is shown in the code, printing as an output the returned documents of each query.

We developed a cluster-based search engine which return the top 20 most common tweets. It consists of making a tf-idf score but instead of returning the top k elements, it only returns the first one which is the tweet with the highest similarity with the query. Then we introduce this tweet as a new query, so the function returns the most similar tweets. Hence, the final 20 tweets are composed of the most similar tweet of the query and the most 19 similar tweets with that first tweet. The first tweet represents the category that the other tweets have to talk about.

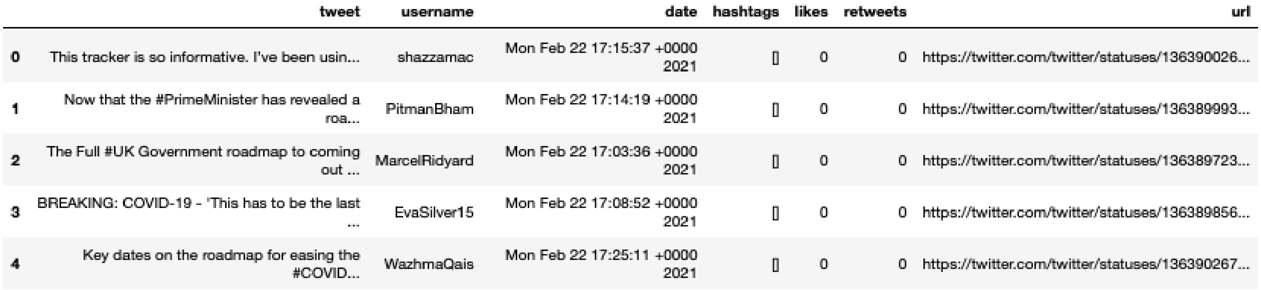


Joan Medina Martí
Guillem Morgó Homs

Comparison of both rankings

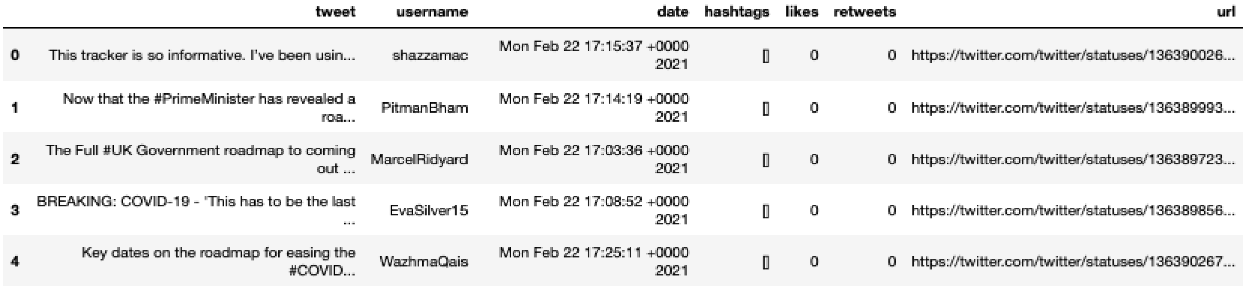
The query used to make the comparison between the tweets returned by both of the rankings is: *lockdown*

- 5 most relevant tweets using TF-IDF and cosine similarity



	tweet	username	date	hashtags	likes	retweets	url
0	This tracker is so informative. I've been usin...	shazzamac	Mon Feb 22 17:15:37 +0000 2021	[]	0	0	https://twitter.com/twitter/statuses/136390026...
1	Now that the #PrimeMinister has revealed a roa...	PitmanBham	Mon Feb 22 17:14:19 +0000 2021	[]	0	0	https://twitter.com/twitter/statuses/136389993...
2	The Full #UK Government roadmap to coming out ...	MarcelRidyard	Mon Feb 22 17:03:36 +0000 2021	[]	0	0	https://twitter.com/twitter/statuses/136389723...
3	BREAKING: COVID-19 - 'This has to be the last ...	EvaSilver15	Mon Feb 22 17:08:52 +0000 2021	[]	0	0	https://twitter.com/twitter/statuses/136389856...
4	Key dates on the roadmap for easing the #COVID...	WazhmaQais	Mon Feb 22 17:25:11 +0000 2021	[]	0	0	https://twitter.com/twitter/statuses/136390267...

- 5 most relevant tweets using cluster-based ranking



	tweet	username	date	hashtags	likes	retweets	url
0	This tracker is so informative. I've been usin...	shazzamac	Mon Feb 22 17:15:37 +0000 2021	[]	0	0	https://twitter.com/twitter/statuses/136390026...
1	Now that the #PrimeMinister has revealed a roa...	PitmanBham	Mon Feb 22 17:14:19 +0000 2021	[]	0	0	https://twitter.com/twitter/statuses/136389993...
2	The Full #UK Government roadmap to coming out ...	MarcelRidyard	Mon Feb 22 17:03:36 +0000 2021	[]	0	0	https://twitter.com/twitter/statuses/136389723...
3	BREAKING: COVID-19 - 'This has to be the last ...	EvaSilver15	Mon Feb 22 17:08:52 +0000 2021	[]	0	0	https://twitter.com/twitter/statuses/136389856...
4	Key dates on the roadmap for easing the #COVID...	WazhmaQais	Mon Feb 22 17:25:11 +0000 2021	[]	0	0	https://twitter.com/twitter/statuses/136390267...

The retrieved tweets seem to be the same which means that it is a good search engine, but it could also mean that we have few tweets.
The main similarity between the ranking is that they will always have the same most relevant tweet which will be the first one because it is chosen using the same technique.

SECTION 3: RESEARCH QUESTIONS

RESEARCH QUESTION 1

The queries that have been used are the following:

```
1. trump has covid
2. impact of coronavirus
3. how to face off covid
4. lockdowns does not work
5. how good is the vaccine?
6. vaccine of Pfizer
7. vaccine 2021
8. no more masks
9. I think I have covid
10. what are covid symptoms
```

The results of the two search engines algorithm have been stored in the folder 'results' with the names of RQ1b and RQ1c respectively.

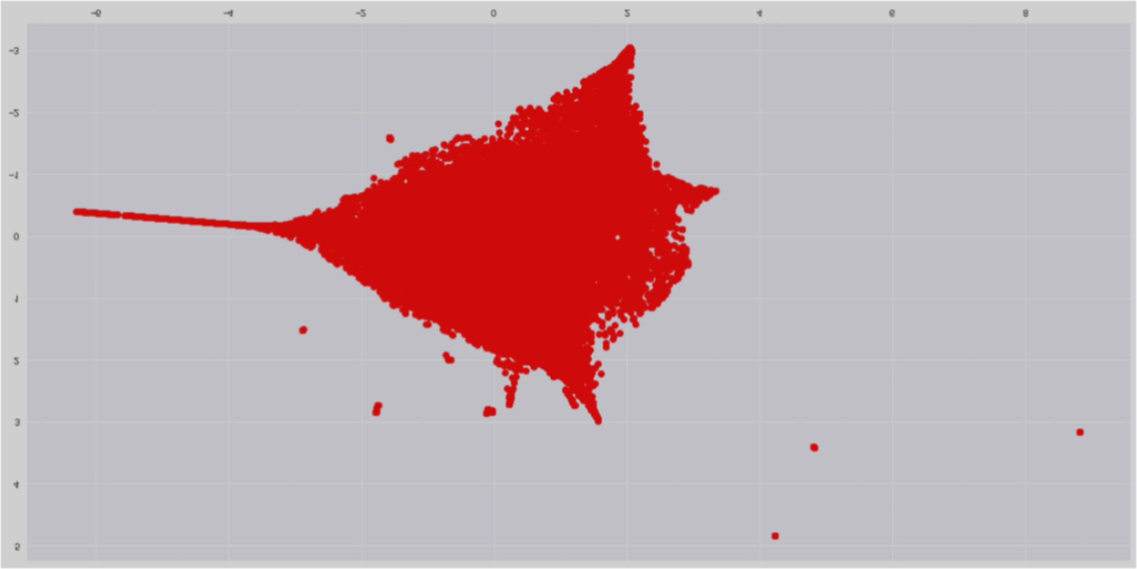
Joan Medina Martí
Guillem Morgó Homs

Better representation than word2vec

If we compare the performance between word2vec and sentence2vec, we can say word2vec is more specific which doesn't mean that is has a better accuracy. But it is true that when one persona search for a tweet using a very weird word, word2vec will work better than sentence2vec.

In this case I think that due to al tweets talk about the same topic (coronavirus), all the main words of the queries will be very similar such as coronavirus, vaccine and so on, which means that all the tweets will be highly related with those words. So, a sentence2vec search engine will be better in this case.

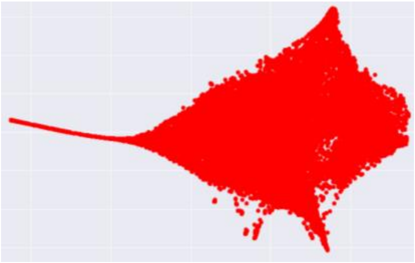
TSNE plot

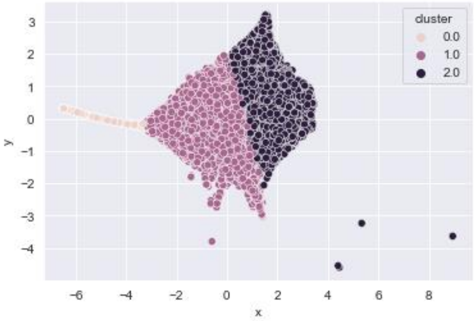


RQ 1A

We can clearly see that, in fact, there's only one cluster. But, because we want to detect some topics' clusters, we will decide the number of clusters manually.

This plot remembers to a kind of manta ray and the subgroups have been done following the anatomy of this animal: upper body, lower body and the queue.





Joan Medina Martí
Guillem Morgó Homs

RQ 1B

Cluster 1		Cluster 2		Cluster 3	
kifj	15	pandem	2564	vaccin	5370
sahilkhanup	14	covid19	1848	get	1492
cp	9	new	601	covid	859
rahulkohli	9	us	598	covid19	824
stevebi	8	peopl	581	peopl	750

These keywords don't seem to separate the clusters because we can see that some words are repeated within the clusters but a reasonable answer to this problem is that we have recorded to few tweets.

RESEARCH QUESTION 2

The score formula we used to evaluate how high is the diversity of the retrieved tweets, is the following:

$$top = size(retrieved\, tweets)$$

$$best = \frac{\#clusters * \sqrt{\frac{top}{\#clusters}}}{top}$$

The best variable is the highest diversty of retireved tweets. Example:
Given 10 tweets belonging to 3 clusters the formula gives the following result

$$best = \frac{3 * \sqrt{\frac{10}{3}}}{10}$$

Then we compute the score of the real data which is the sum of the score of each cluster:

$$score_i = \sqrt{\#tweets_i}$$

$$score = \sum_i score_i$$

The final score is the combination of the real score and the best possible score:

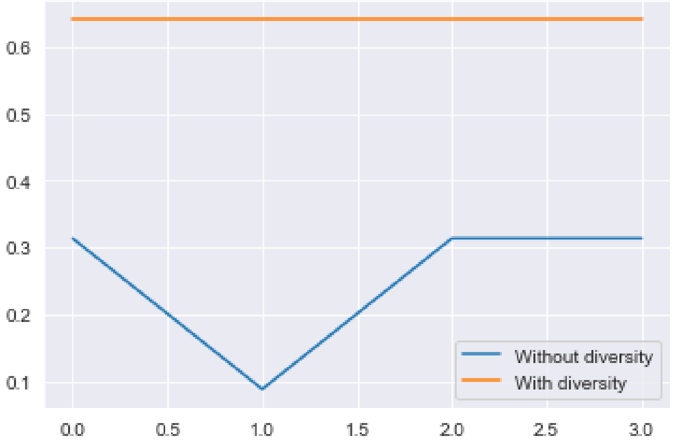
$$(score/best)\%1$$

If the real score is equal to the best score, the result will be 1, which is the maximum score.

Joan Medina Martí
Guillem Morgó Homs

RQ2A - RQ2B

The score plot and their respective queries we used to test our score of diversity are the following:



```
Query: Pfizer
Query: vaccine
Query: Biden
Query: lockdowns
```

The orange line represents the score of the tweets after diversifying them. The blue line represents the tweets score without diversifying them.

Cluster WITHOUT diversifying	Cluster WITH diversifying
0	0
1	0
1	0
2	1
1	1
1	1
1	2
1	2
2	2
1	2

RESEARCH QUESTION 3

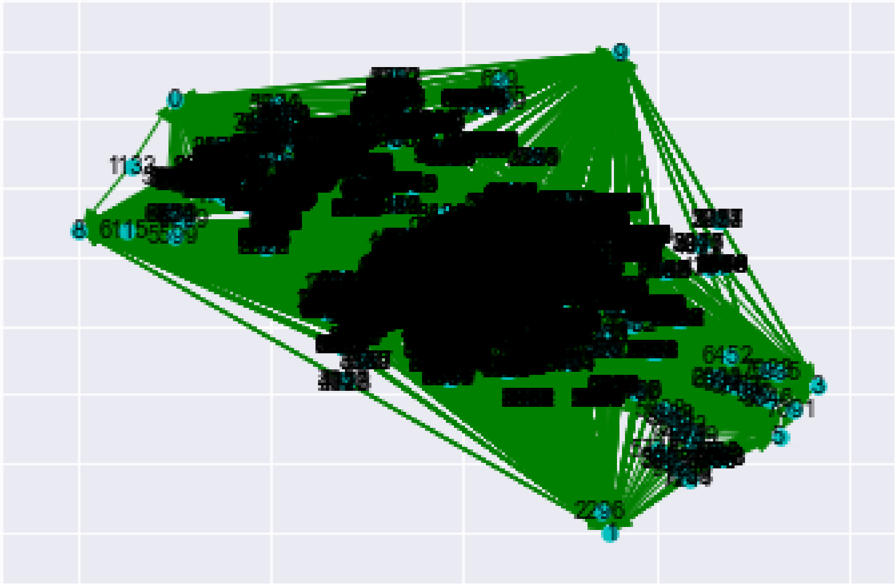
The initial graph consisted of:
```
Nodes: 11074
Edges: 7678
```

After the train test/split the graph belonging to the training set consisted of:
```
Nodes: 9539
Edges: 5831
```

The best algorithm from all that have been analyzed is that one named *Alternating Least Squares* with an efficiency of 97%.

Joan Medina Martí
Guillem Morgó Homs

The following plot is the recommendation graph:


9
0
1132
8
6115
5589
6452
3
5
2296
1