DATA COLLECTION

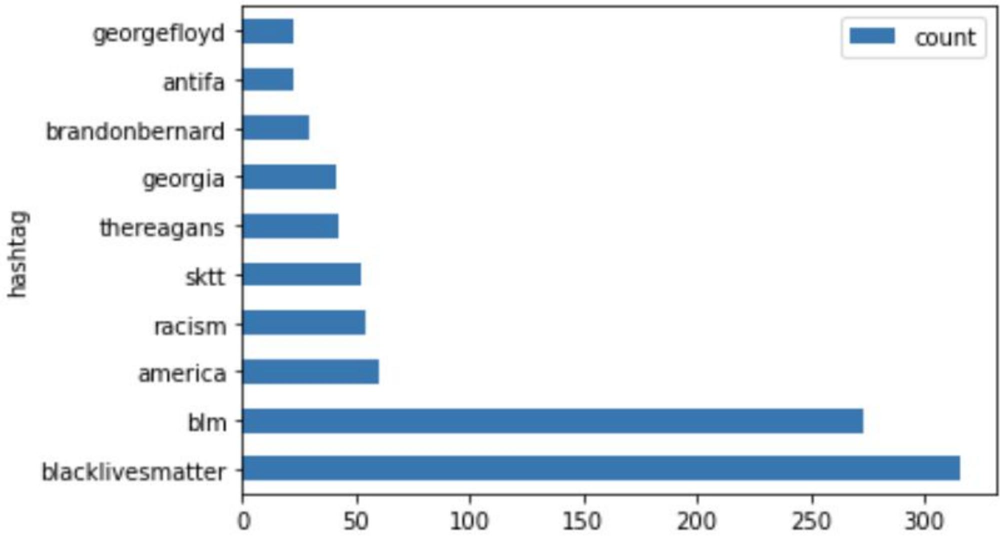
Data information

The total number of tweets we obtained directly from the scrapper is more than 100 thousands tweets, from which we had:

- Number of retweets: 60522
- Number of original tweets: 44735
- Number of unique users: 23539

However, this number is significantly reduced after removing the retweeted ones and the duplications. After those two steps, we have a total of 11546 tweets.

Moreover, we have also analyzed the hashtags that appear more frequently in our dataset and here we have a barplot representing this information.



Keywords for scrapper

During the collection, the query passed to the scrapper was:

['BlackLivesMatter OR AllLivesMatter OR BlueLivesMatter OR
GeorgeFloyd OR BLM OR racism America’]

This way, all tweets containing any of these keywords would be retrieved and sent in the json file.

Time spent in collecting data

Due to the “rate limit reached” problem with the scrapping we lasted around 2 hours to get all the tweets.

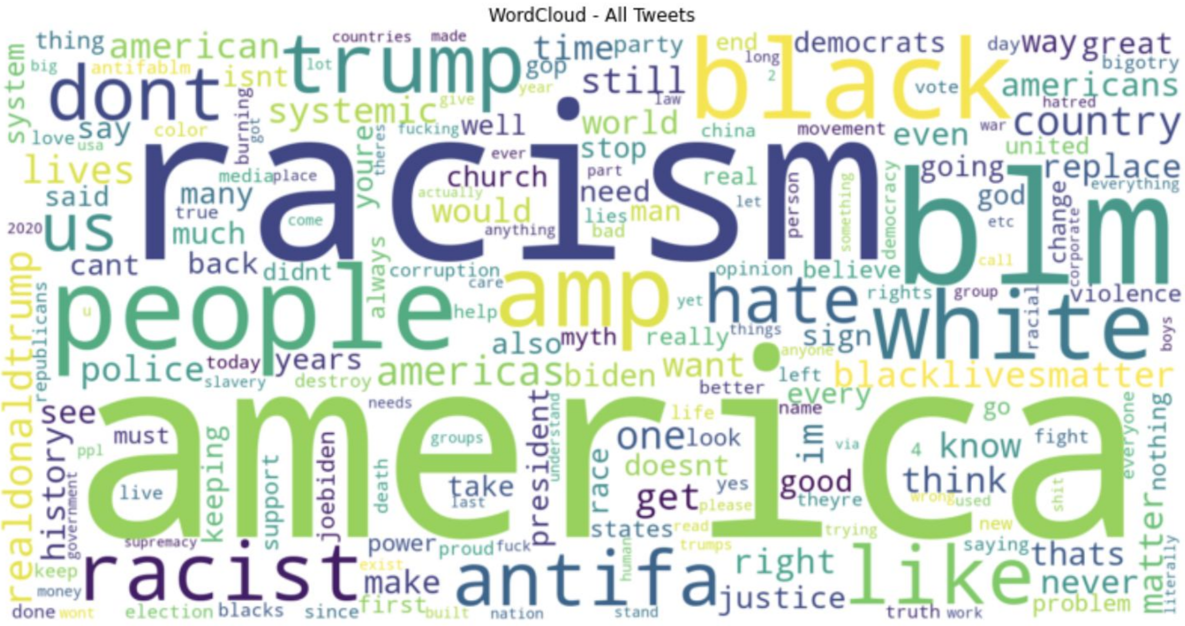
SEARCH ENGINE

Description of the pre-processing strategy

In order to clean the text before processing them and to normalize them, we followed this steps:

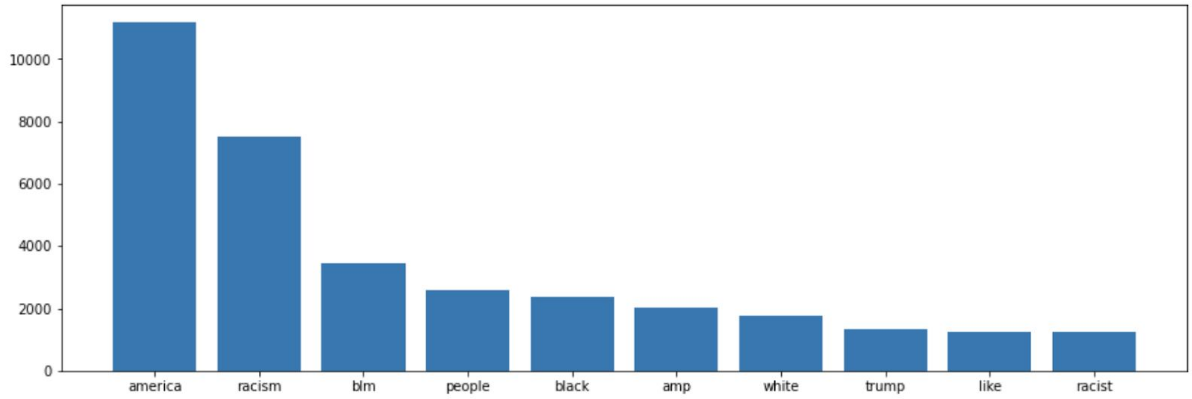
- Transform all words to lowercase.
- Remove accents.
- Remove emojis.
- Replacement of apostrophe (it's --> its).
- Replacement of symbols (*, ?, ...) by space.
- Remove stop words.
  These are words really relevant in the whole collection, so they are not informative of a concrete document. This means that we should remove them before indexing the documents as they don't have to be taken into account in the similarity.
- Stemming
  We don't want to consider as different all words that are from the same family. Therefore, we keep the root and remove the different endings.

WordCloud



We can see that the words that appear the most are mainly america and racism. Also we have a lot of tweets including the words: trump, blm, antifa, racist and black.

Barplot of 10 most frequent words



Using the barplot we can see the same most frequent words that we could extract from the previous wordcloud. However, in this case we can see clearer the order the follow in frequency, as in the previous wordcloud we get an intuition but apart from "america " and "racism" we could not give an accurate order of most frequent words.

Our score

This score has to involve the popularity of a tweet, so we use the tf-idf as the first score, and then we add score to it depending on how many retweets and favorites have. First of all we get all the retweet_count of the docs of the query, and we calculate the mean, it will be used later. Then we get the list of all the retweet_count of these docs, and we distribute the range between 0 to 0.5 to every count. For example:

If we obtain that the retweet_counts are [11, 8, 5, 0, 0, 0] we will divide the range [0, 0.5] in 4 parts [11, 8, 5, 0] because all the docs that have retweet_count 0, will add 0 to the score.

So, 11 will have 0.5, 8 → 0.33, 5 → 0.166 and 0 → 0.

Then we multiply this number obtained by the mean of retweet_counts and then we add it to each tf-idf score.

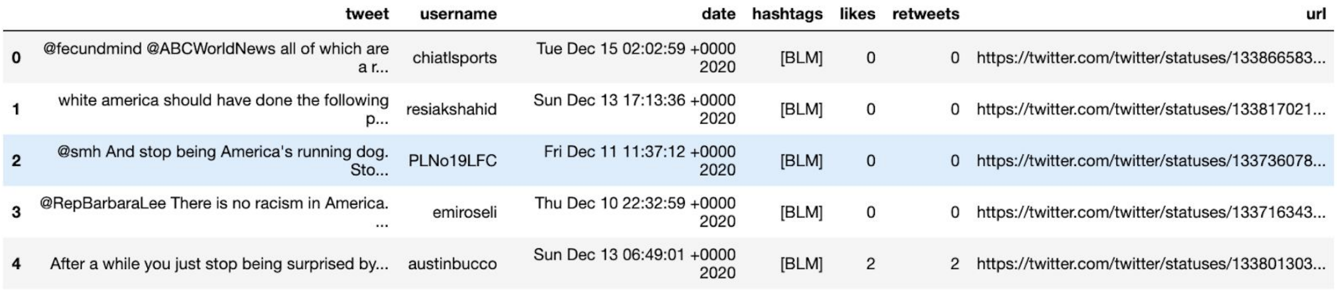
We do the same process but now with the favorite_count. And also add it to the score resulted by the previous process. And finally we obtain a new score related to the popularity of a tweet.

Having this the tweets that are retweeted and have been added to favorite most, are more relevant. We have considered that a retweet and a favorite tweet has the same weight, if we talk about relevance.

<u>Screenshot comparing both rankings</u>

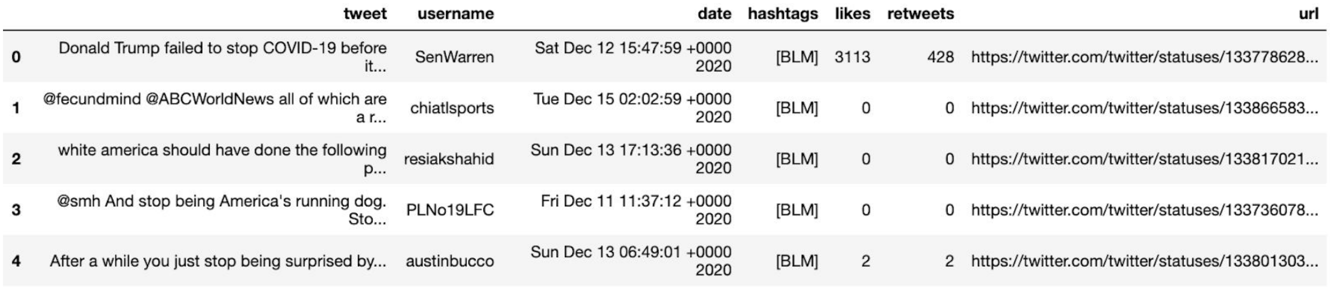
For the query = ‘stop racism’ we obtain:

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



	tweet	username	date	hashtags	likes	retweets	url
0	@fecundmind @ABCWorldNews all of which are a r...	chiatlsports	Tue Dec 15 02:02:59 +0000 2020	[BLM]	0	0	https://twitter.com/twitter/statuses/133866583...
1	white america should have done the following p...	resiakshahid	Sun Dec 13 17:13:36 +0000 2020	[BLM]	0	0	https://twitter.com/twitter/statuses/133817021...
2	@smh And stop being America's running dog. Sto...	PLNo19LFC	Fri Dec 11 11:37:12 +0000 2020	[BLM]	0	0	https://twitter.com/twitter/statuses/133736078...
3	@RepBarbaraLee There is no racism in America. ...	emiroseli	Thu Dec 10 22:32:59 +0000 2020	[BLM]	0	0	https://twitter.com/twitter/statuses/133716343...
4	After a while you just stop being surprised by...	austinbucco	Sun Dec 13 06:49:01 +0000 2020	[BLM]	2	2	https://twitter.com/twitter/statuses/133801303...

- 5 most relevant tweets using our score



	tweet	username	date	hashtags	likes	retweets	url
0	Donald Trump failed to stop COVID-19 before it...	SenWarren	Sat Dec 12 15:47:59 +0000 2020	[BLM]	3113	428	https://twitter.com/twitter/statuses/133778628...
1	@fecundmind @ABCWorldNews all of which are a r...	chiatlsports	Tue Dec 15 02:02:59 +0000 2020	[BLM]	0	0	https://twitter.com/twitter/statuses/133866583...
2	white america should have done the following p...	resiakshahid	Sun Dec 13 17:13:36 +0000 2020	[BLM]	0	0	https://twitter.com/twitter/statuses/133817021...
3	@smh And stop being America's running dog. Sto...	PLNo19LFC	Fri Dec 11 11:37:12 +0000 2020	[BLM]	0	0	https://twitter.com/twitter/statuses/133736078...
4	After a while you just stop being surprised by...	austinbucco	Sun Dec 13 06:49:01 +0000 2020	[BLM]	2	2	https://twitter.com/twitter/statuses/133801303...

We can see that using our score, the most popular tweet is the first retrieved.

Research Question 1

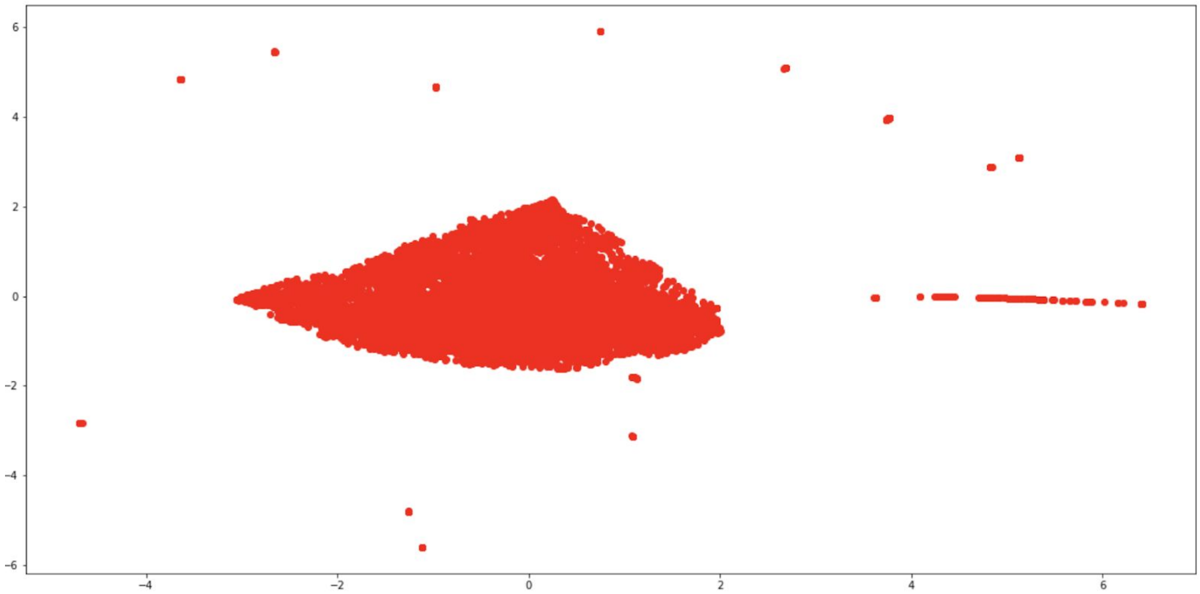
<u>List of 10 selected queries</u>

1. Stop racism
2. Racism is everywhere
3. Blm president biden
4. Racism in usa
5. I can’t breath
6. Rest in peace
7. Blm George Floyd
8. Justice for George Floyd
9. Justice system
10. All lives matter

<u>Answer to Question d → Can you imagine a better representation than word2vec?</u>

As we know Word2Vec algorithm builds distributed semantic representation of words.With Doc2vec or sentence2vec, the same approach is defined but instead of learning feature representation of words, you lear documents or sentences.As we are dealing with tweets, maybe using word2vec is not the best options, as word within the tweets have some relationships, and with word2vec you can not capture that.

Output given by t-sne



Answer to question RQ-1A

In the above plot we can see a main center cluster and many other small clusters surrounding it. A simple clustering on t-sne representation can be done, however we should consider that tsne does not preserve distances nor density. It only preserves nearest-neighbours.Therefore if we apply any classical clustering algorithm such us k-means clustering or any other, we can reach misleading results, ie. These clusters may be artificials clusters created by t-sne.

The best numbers of clusters can be done manually, we can set this value to 2, as is the number of big clusters in the image

Answer to question RQ 1B - a table for each cluster may help.

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?

Five more relevant words for cluster 0:

‘america’	7080
‘racism’	4995
‘’	3637
‘america’	3220
‘blm’	2278
‘peopl’	2142
‘black’	2106
‘racism’	1807
‘ amp’	1739
‘white’	1672

Five more relevant words for cluster 1:

‘replac’	495
‘america’	466
‘ ’	406
‘black’	323
‘live’	323
‘matter’	301
‘church’	252
‘ amp’	248
‘sign’	247
‘racist’	246

Research Question 2

Describe the diversity score and the post-processing strategy

This score is assigned to the list of returned documents for the input query;
As we have previously assigned a cluster to each of the tweets in the previous exercise, we can use this label in order to define a diversify score to each output, to do so, this score will be defined as the even distribution of unique clusters among all the top-k items of the query.i.e If we define 5 clusters among all the documents,an output of 10 documents, where there are a pair of each unique cluster, its output score will be 1.The re-ranking is done by sampling the top document for each cluster, therefore, in the top-5 tweets we will have one for each of the 5 clusters. The top-10 will have 2 documents for each of the clusters

List of chosen queries.
We select the same queries as before:

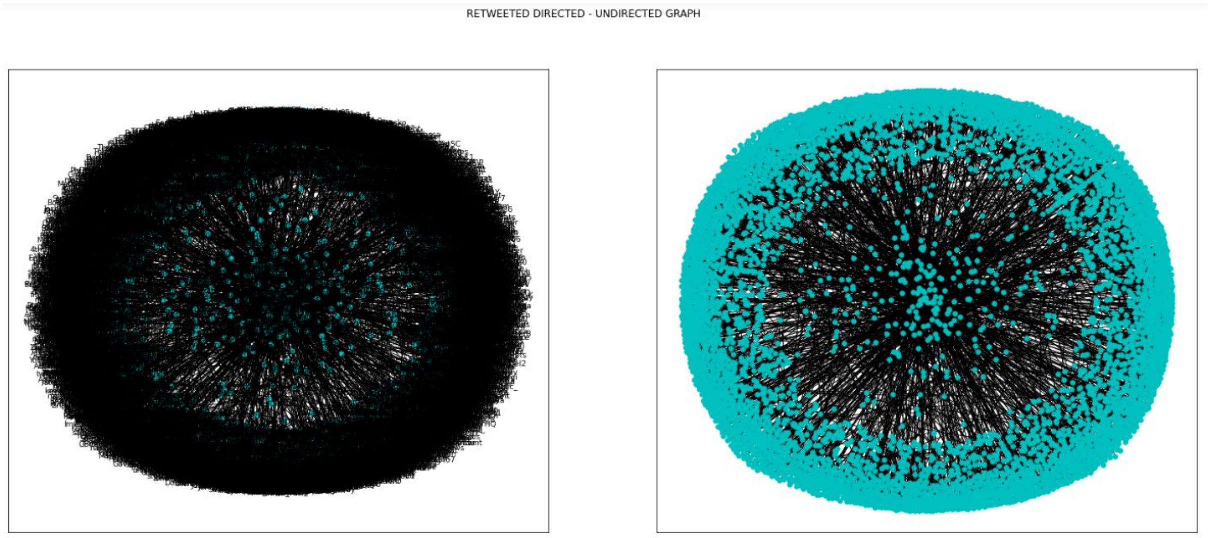
1. Stop racism
2. Racism is everywhere
3. Blm president biden
4. Racism in usa
5. I can't breath
6. Rest in peace
7. Blm George Floyd
8. Justice for George Floyd
9. Justice system
10. All lives matter

Answer to RQ 2A - plots and tables may help
What about the coverage? Any difference between the two rankings By the two rankings?
The coverage differs a lot from the two rankings, as our diversity score is mainly based on the coverage.

Research Question 3

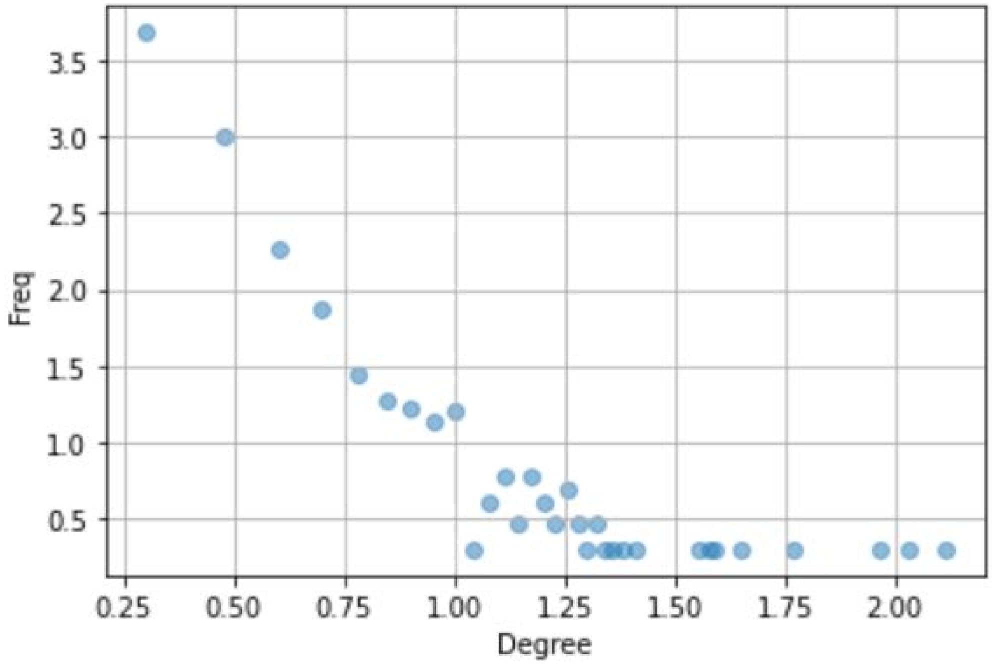
Summary statistics of the retweet graph, train and test.

Here we can see a plot of the graph, the first graph is directed with labels and the second one is the same but undirected and without labels.



RETWEETED DIRECTED - UNDIRECTED GRAPH

The next image shows the distribution of the degrees of the graphs, scaled with logs.



We can see that there are a lot of nodes with a small degree and only a few nodes with large degrees.

The graph has 6201 nodes connected by 4658 edges.

The test set is obtained by the nodes involved in the 20% of the edges of the graph obtained randomly, the next step is removing from the graph those nodes and the resulting graph has 5270 nodes connected by 3400 edges. From there we obtained the nodes at distance 2 and these are the potential recommendations.

Answer to RQ 3A - tables and details about trained models may help

The evaluation of the models are done using this formula:

$$metric = \frac{\#correct}{sizeofU}$$

Models	Metric
Personalized PageRank	0.9757
AdamicAdar	-
ALS	0.9782
Node2vec	-