Èlia Armengol Escolà
205744

Report & Repository

Section 1 - Data collection

In the first part of our project we had to collect data. We decided that our topic was #BlackLivesMatter so we had to collect tweets related to that subject.

The first step in collecting data was creating a function called MyStreamListener that was able to “listen” to the tweets about a certain topic and save them in a folder of our choosing.

With this function, we were able to collect tweets about our topic with the following keywords: "#BLM", "#BlackLivesMatter", "#BreonnaTaylor", "#GeorgeFloyd", "black lives", "black" and "black people".

The waiting time for the function is 3 minutes and 37 seconds and is able to produce a dataset with all the tweets, 10.000 approximately.

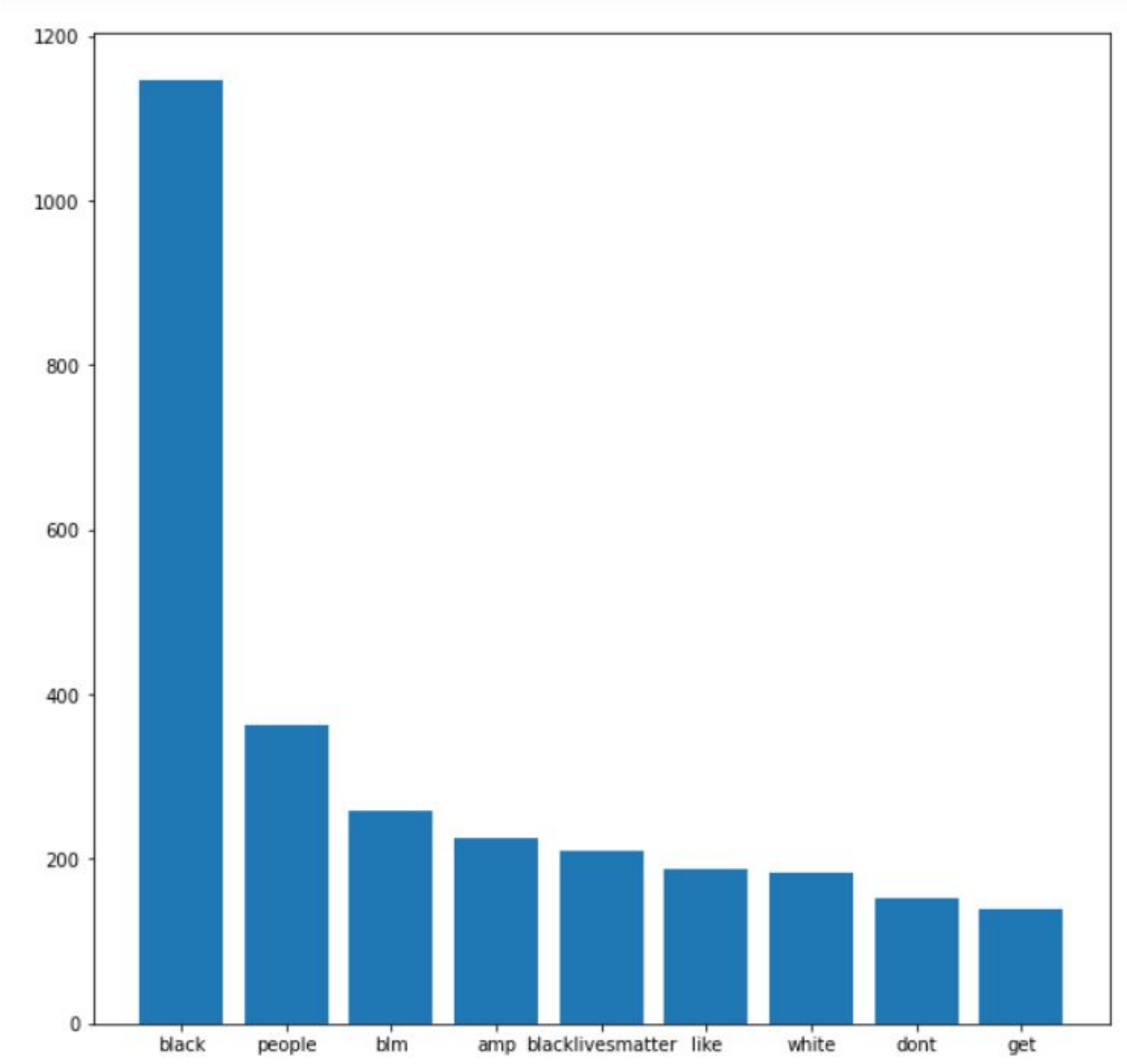
Section 2 - Search Engine

Next, we had to build the search engine, to do so we created a dataset with the tweets that interested us and transform the actual text into a list called lines.
With that list we were able to apply some normalization of the text such as remove punctuation, accents, multiple white spaces, as well as make it all lower case. This is very useful when dealing with text data.

We also created the inverted index and with it we were able to generate a WordCloud of the terms in our tweets:

As we can see here the most important words are: black, people and blm.

We also were able to create a bar plot of the 10 most frequent words as to represent the most important words in a different manner; As you can see the most important words are the same as the ones from the word cloud.



Finally, with the inverted index, we were interested in doing a ranking of our tweets.

First, we performed the TF-IDF ranking and our score which can be seen in the code.