Capstone Project - CIND820

Initial Results

Identifying Trends and Sentiment in Twitter Data

Name: Pallavi Thirunavukarasu

Student Number: 501114919

Supervisor: Dr Ceni Babaoglu

Date: March 07, 2022

Table of Contents

[Initial Results 3](#_Toc97583259)

[Data description 3](#_Toc97583260)

[Overview of the Dataset 4](#_Toc97583261)

[Missing Values 4](#_Toc97583262)

[Variables 5](#_Toc97583263)

[Data Visualization and Summarization 8](#_Toc97583264)

[Data Cleaning and Processing 13](#_Toc97583265)

[Model and Testing 15](#_Toc97583266)

# Initial Results

The dataset to be used for obtaining the initial results is the dataset available at - <https://www.kaggle.com/gpreda/covid19-tweets>. These tweets were collected using Twitter API and a Python script. A query for this high-frequency hashtag (#covid19) was run on a daily basis for a certain time period, to collect a larger number of tweet samples.

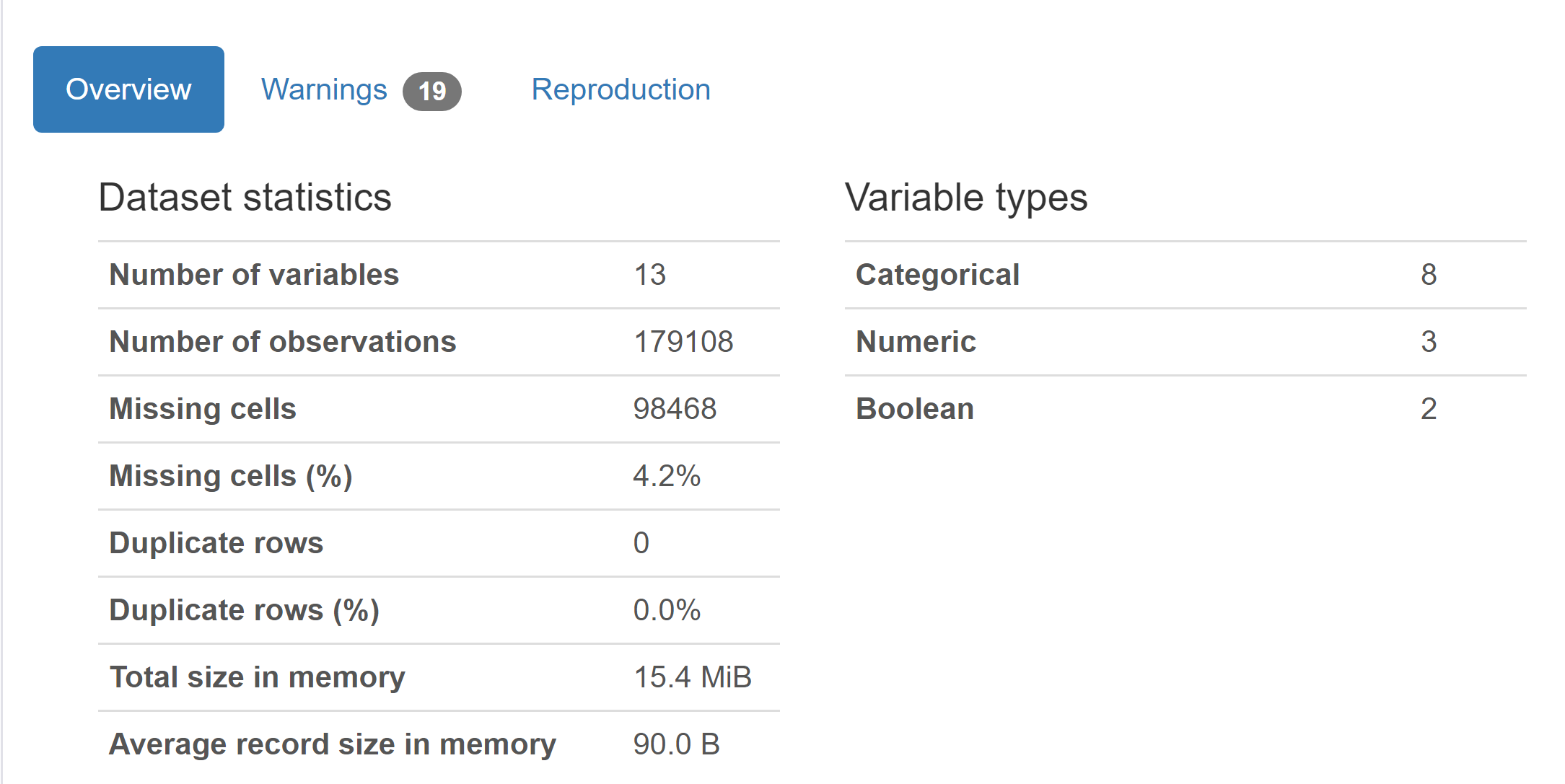
This dataset consists of tweets in English covering dates from 2020-02-29 to 2020-07-24. The tweets were downloaded from no specific region and are global. Retweets are not included in this dataset.

## Data description

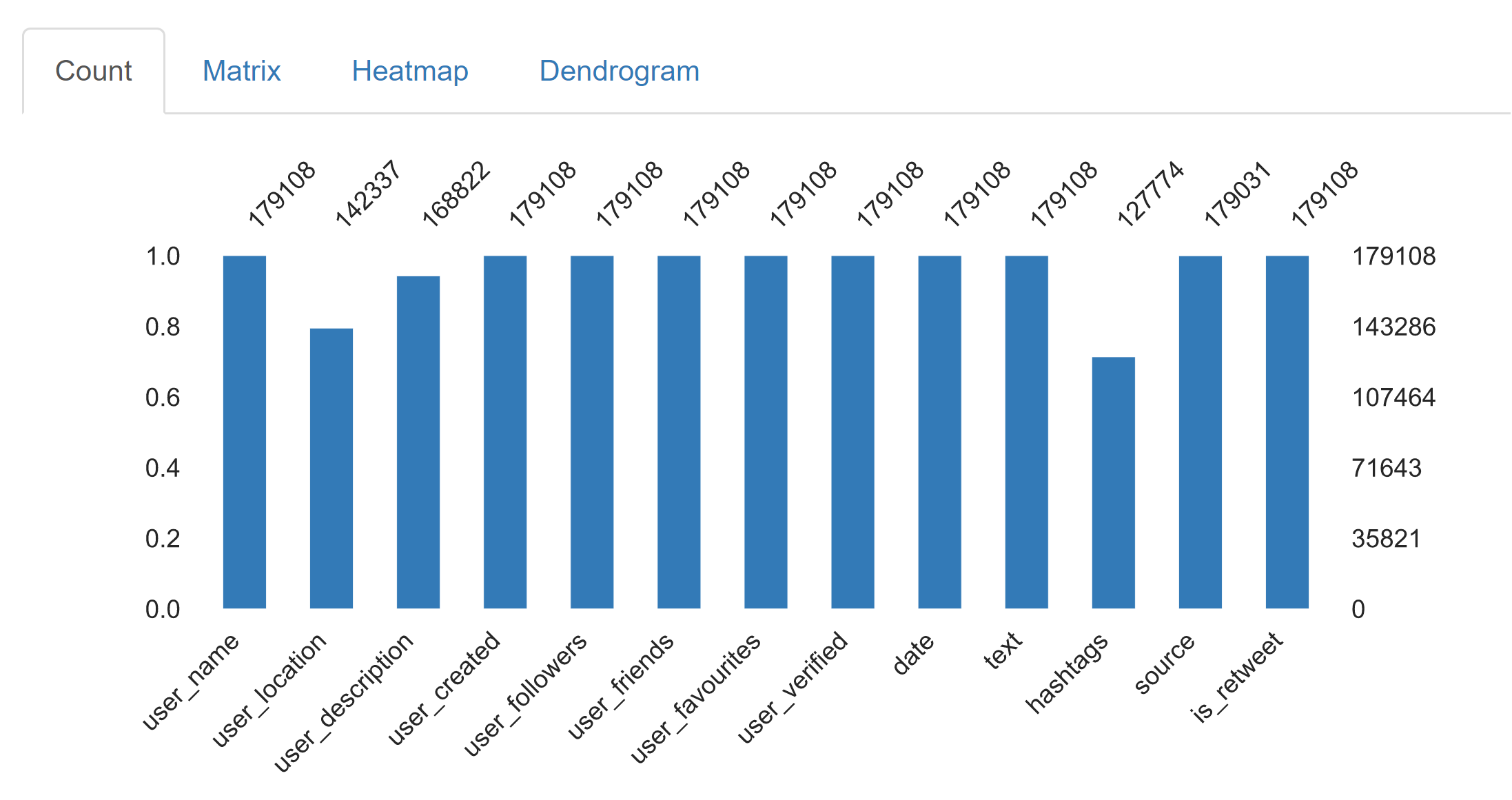
There are 13 columns in this dataset:

1. user\_name – Name of the user on twitter
2. user\_location - Location of the user
3. user\_description – Description of the user on twitter
4. user\_created – When the user was created
5. user\_followers – Number of followers of this user
6. user\_friends – Number of friends of this user
7. user\_favourites – Number of favorites of this user
8. user\_verified – Is the user verified
9. date – Tweet date
10. text -Text of the tweet
11. hashtags – List of hash tags
12. source – Source of the tweet
13. is\_retweet – Tweet is a retweet

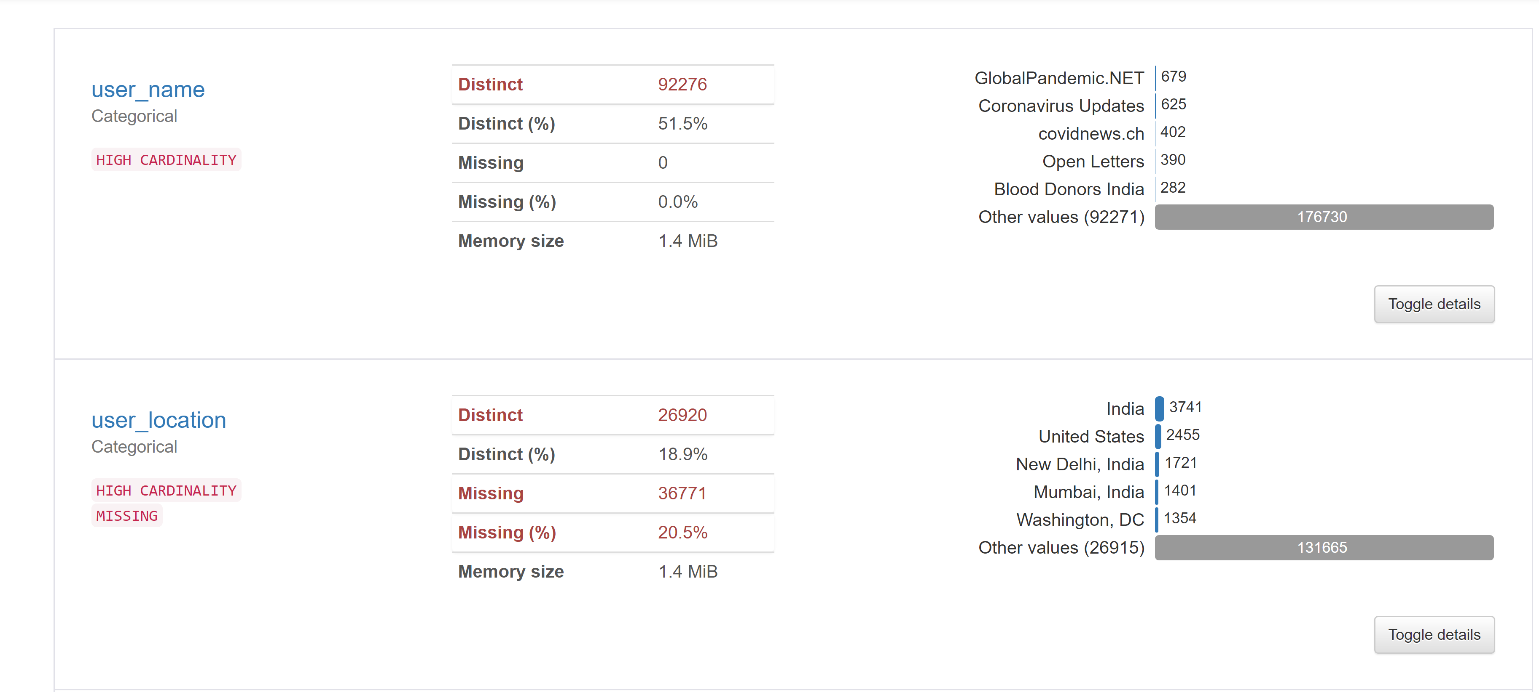
### Overview of the Dataset

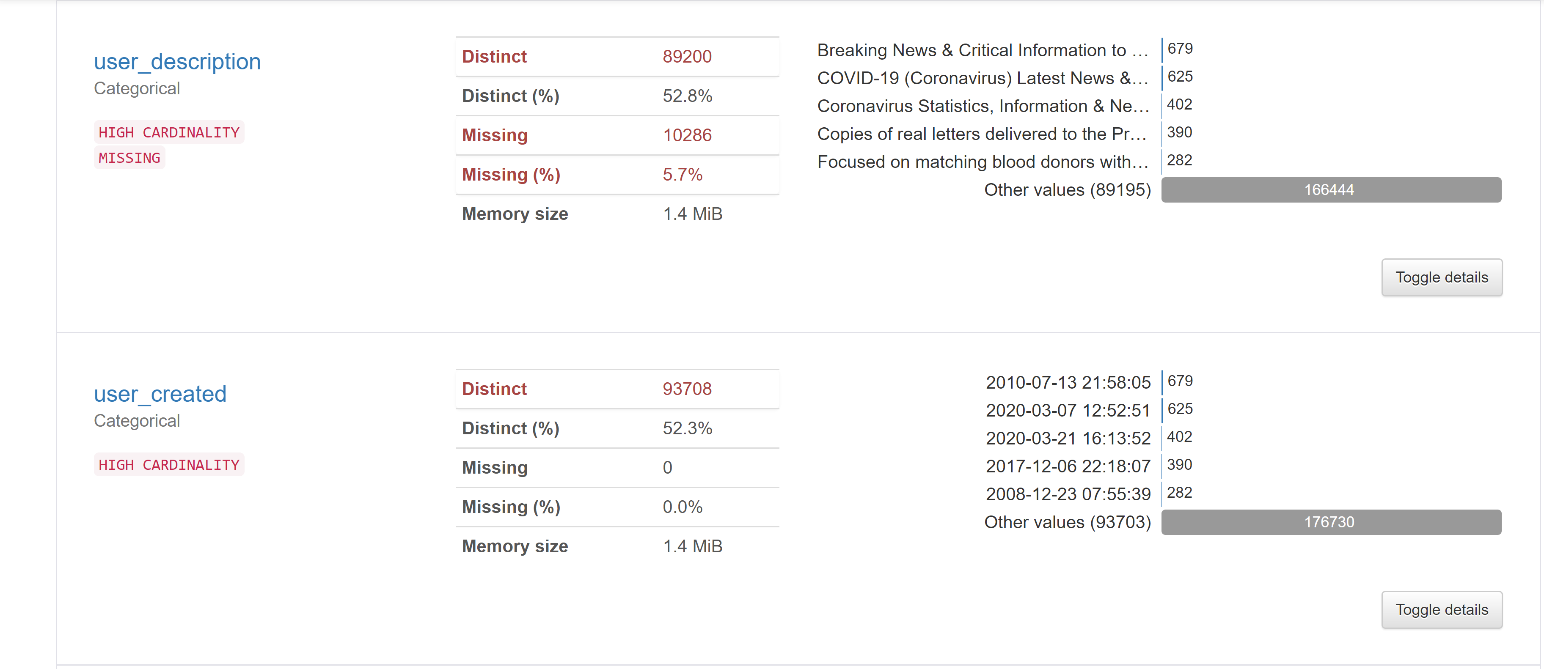


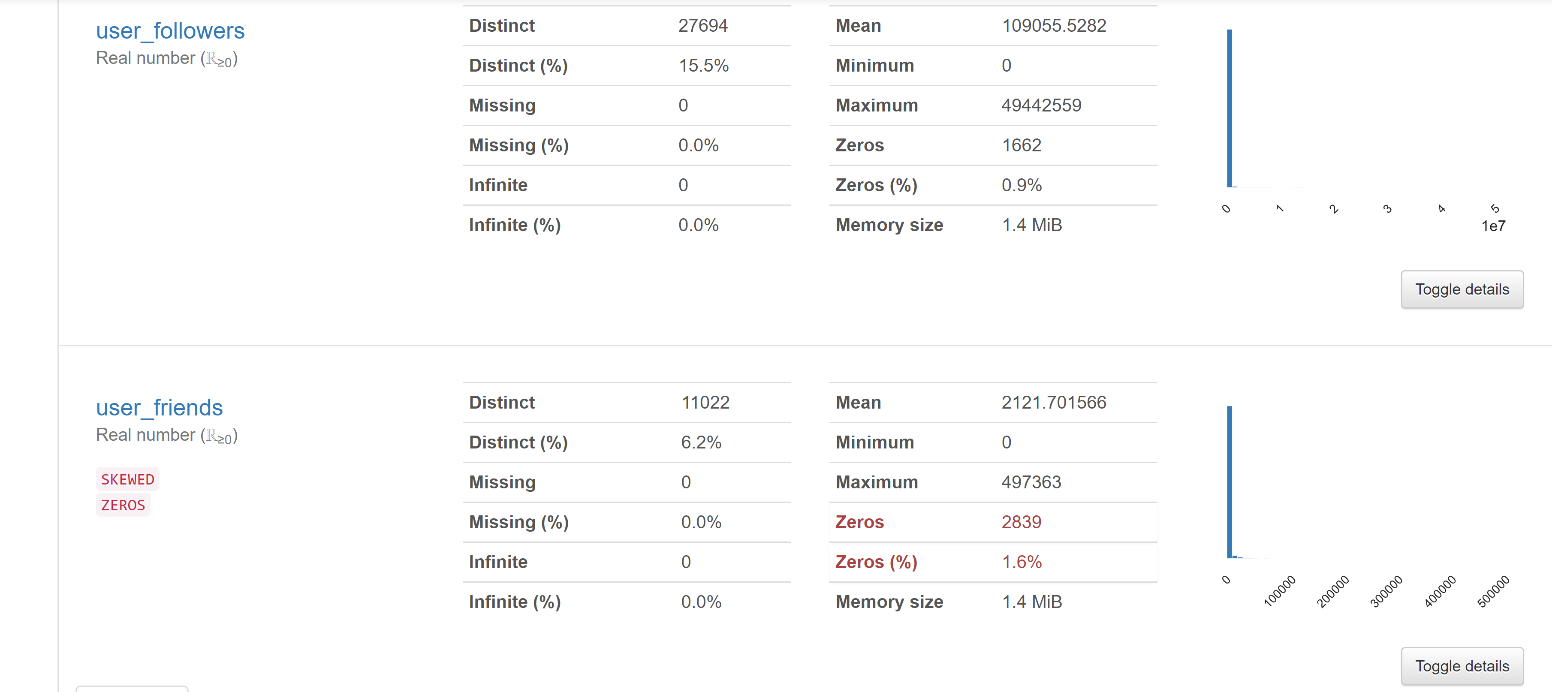
### Missing Values

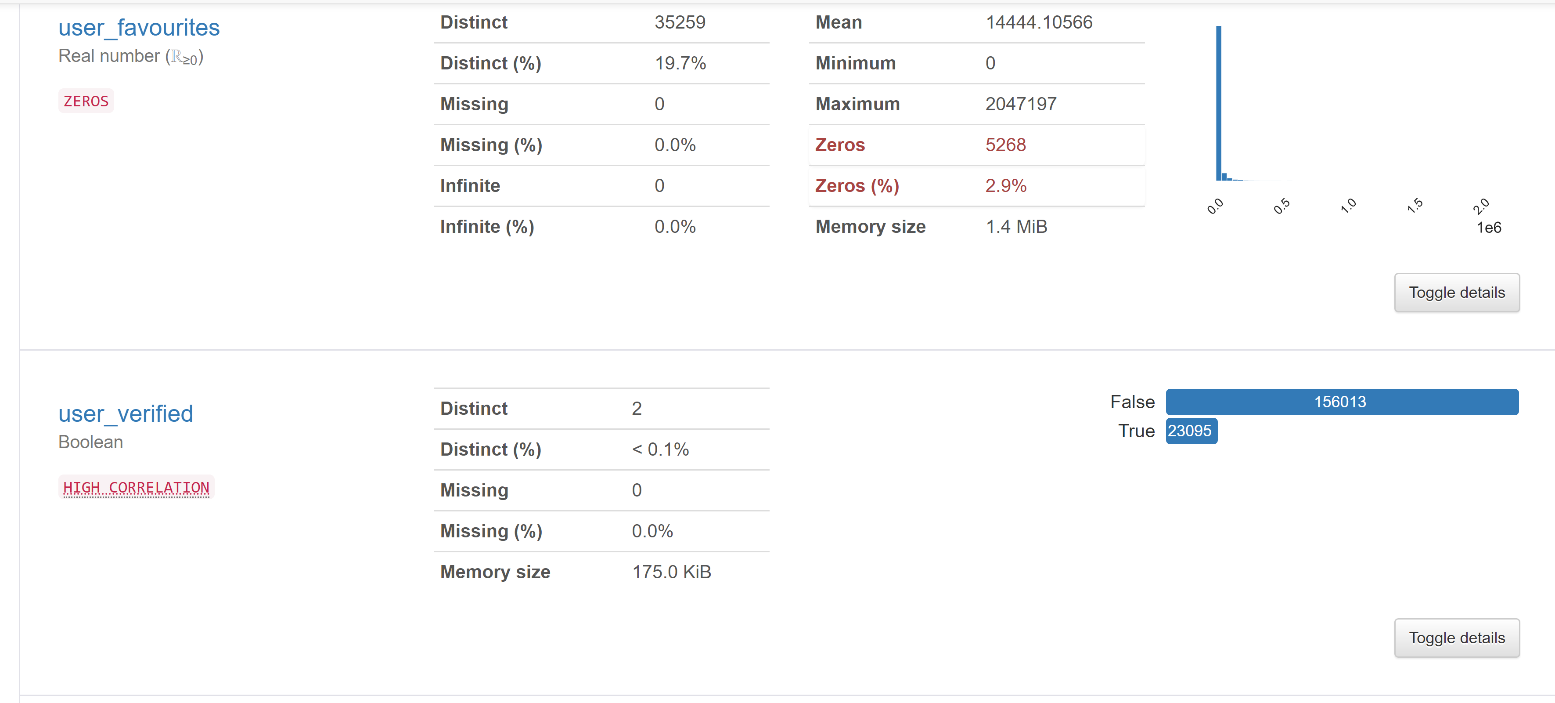


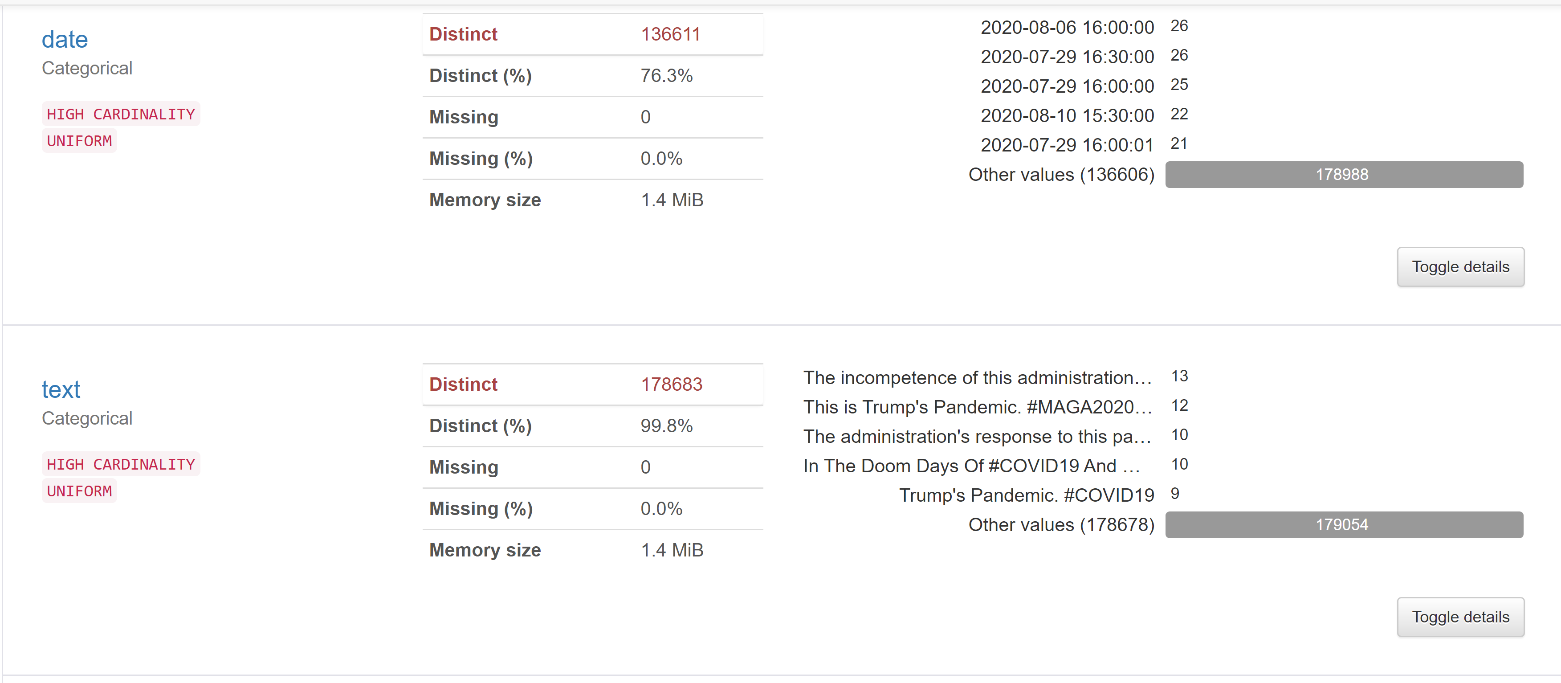
### Variables

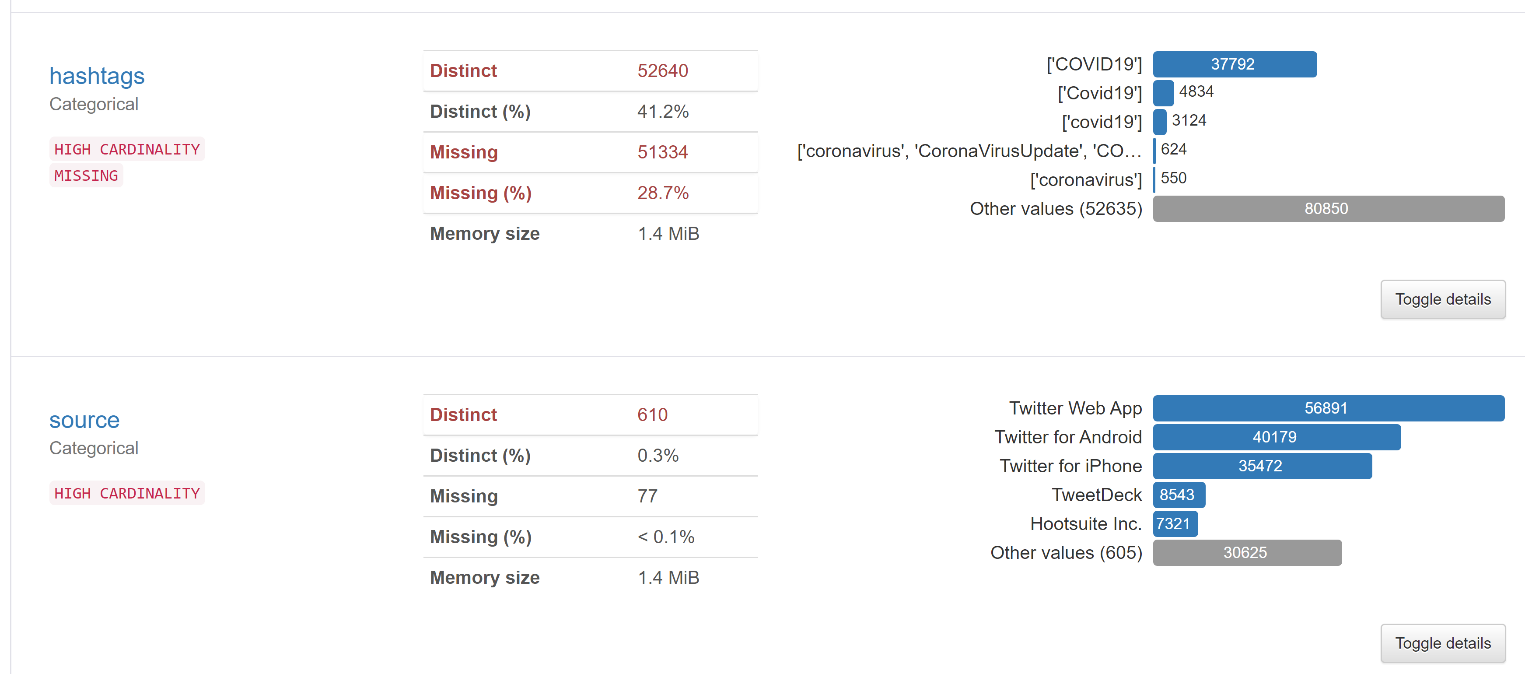




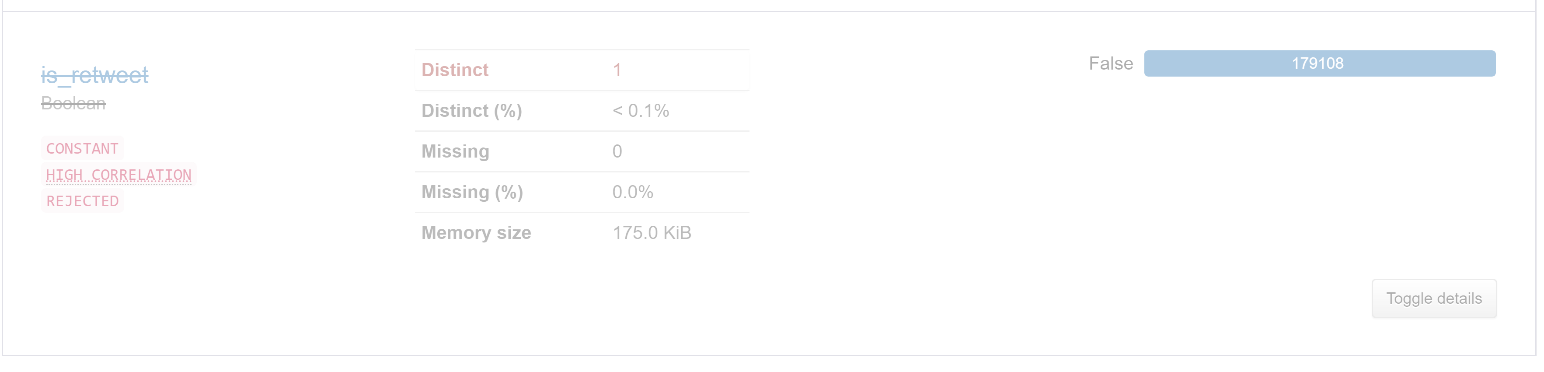






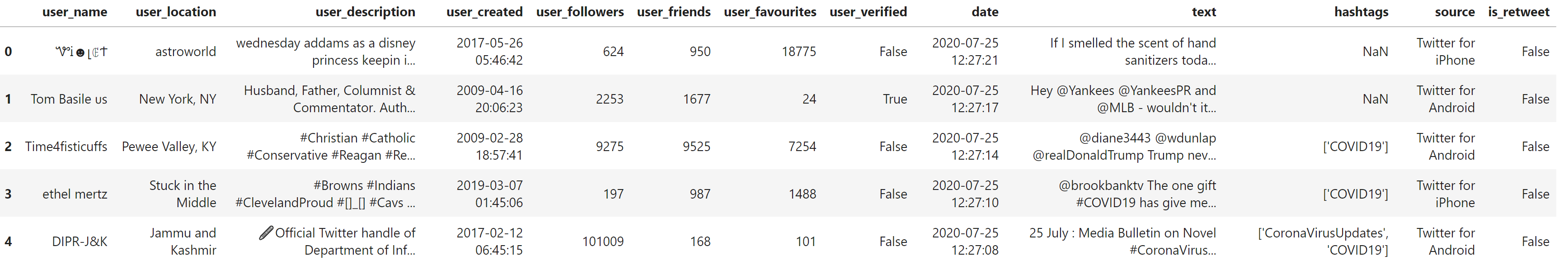


Retweets are removed from the dataset, so there is only one distinct value in this column. This column can be deleted.

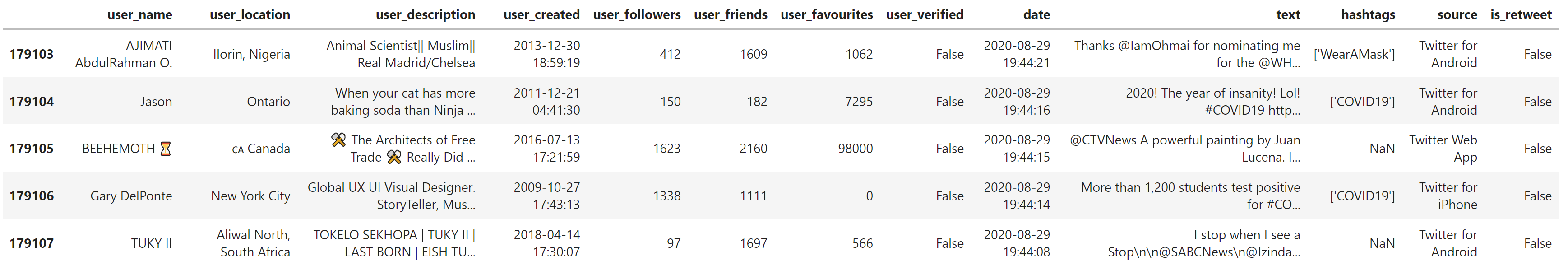


## Data Visualization and Summarization

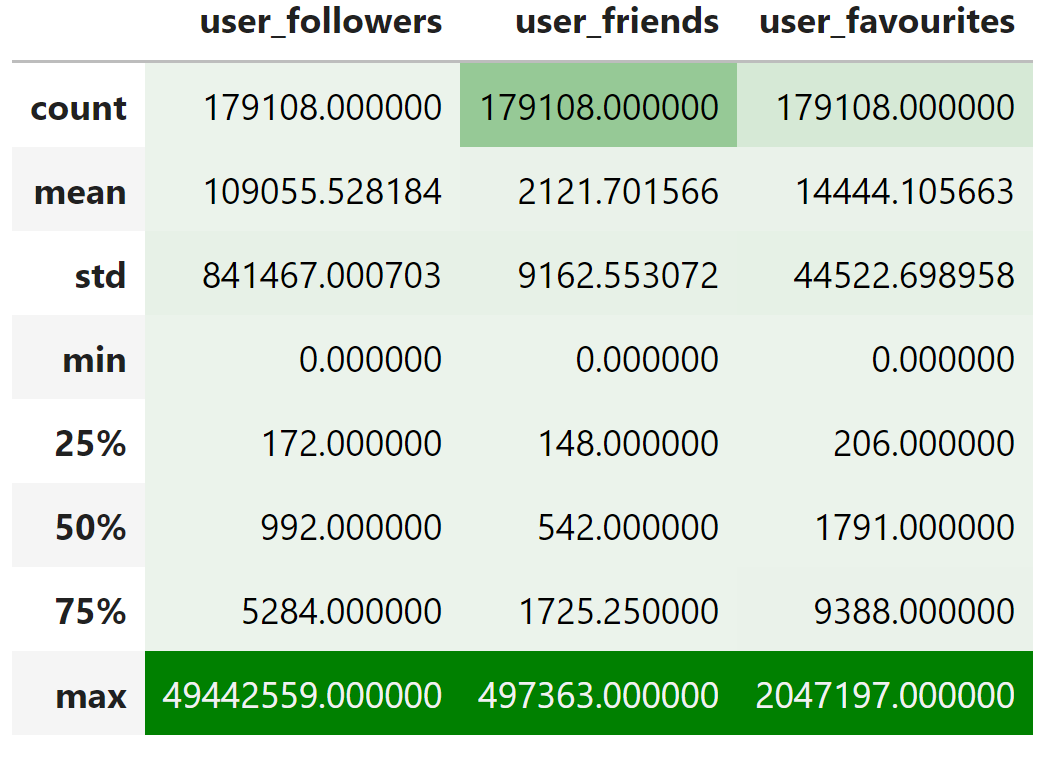
The first few records in the dataset:



The last few records in the dataset:



Summary of the numerical values in the dataset:

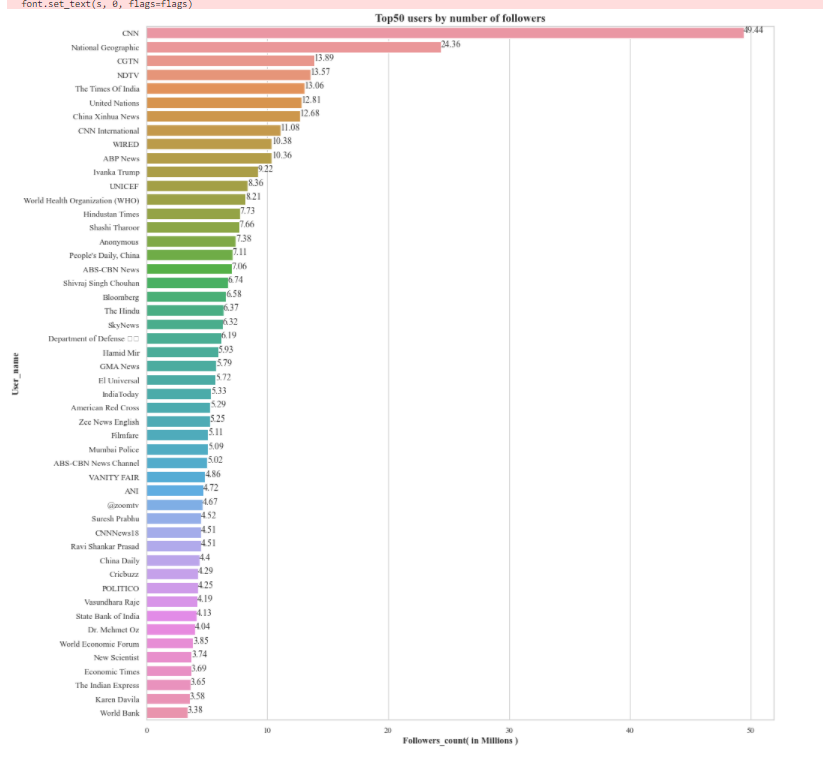


Percentage of missing values by columns:

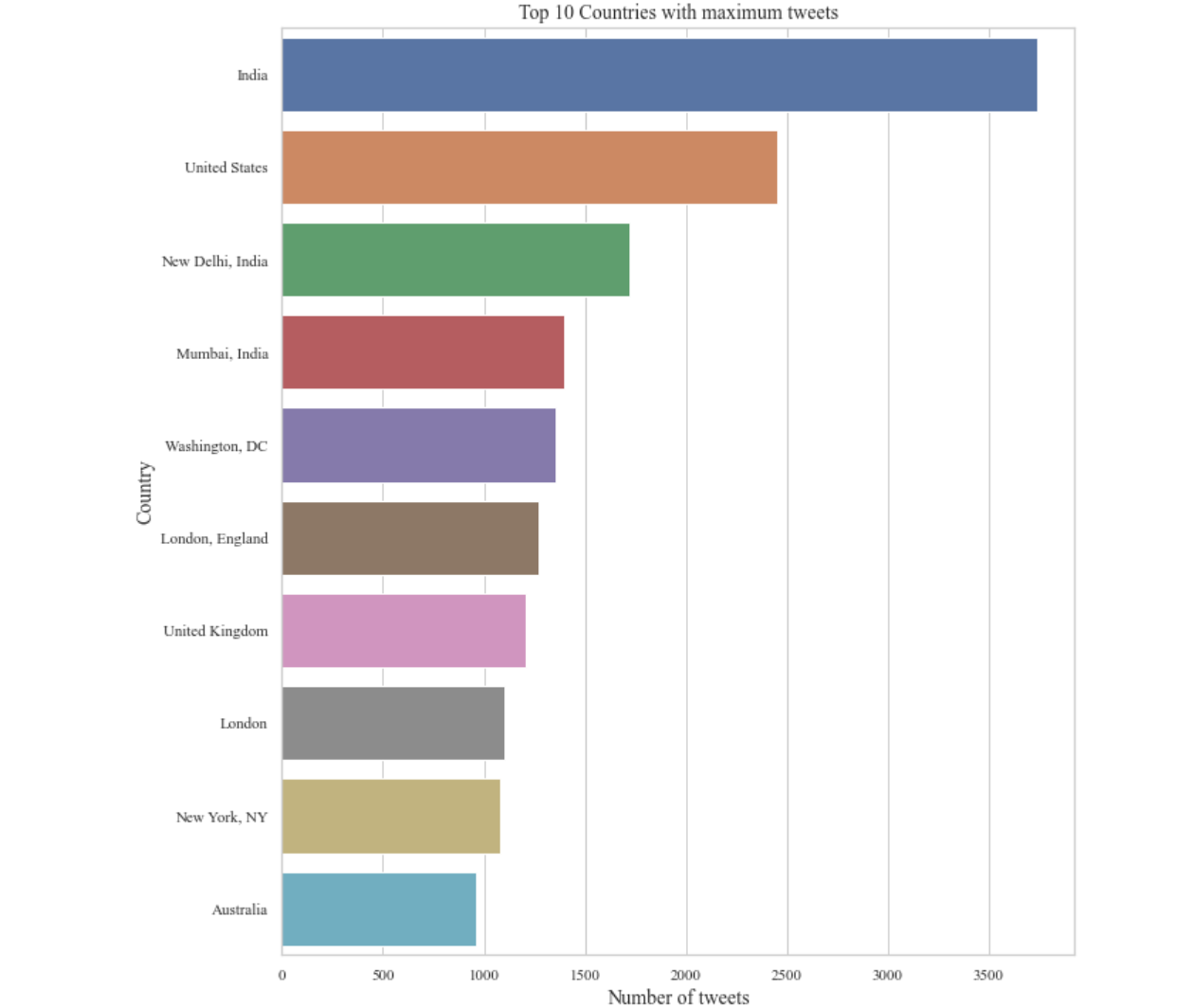
The hashtags have the maximum number of missing values.



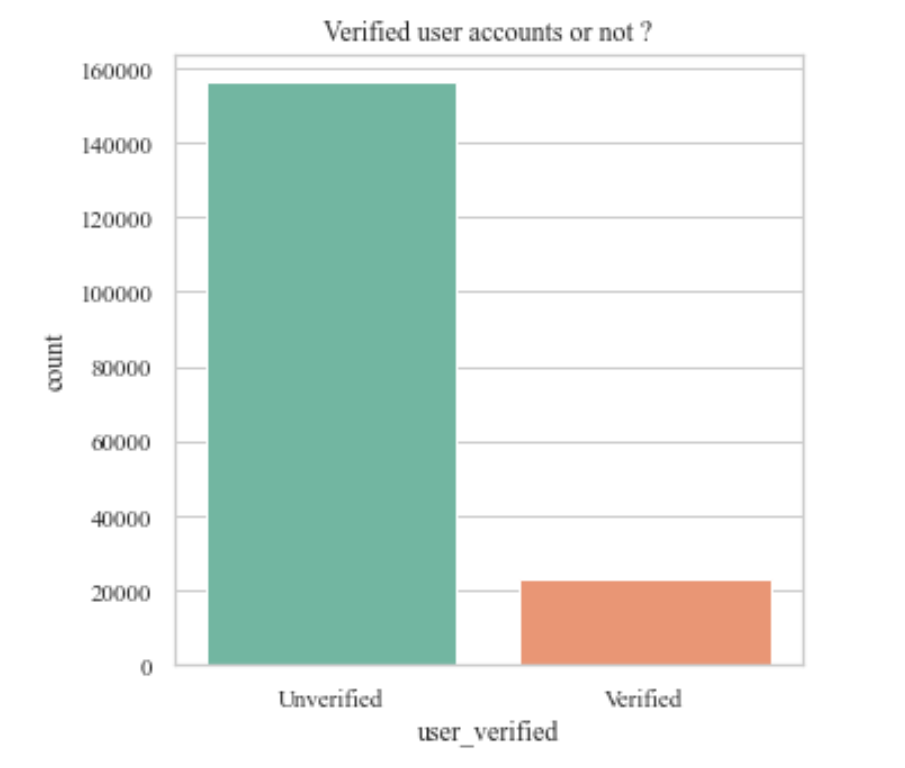
The top 50 users by number of followers:



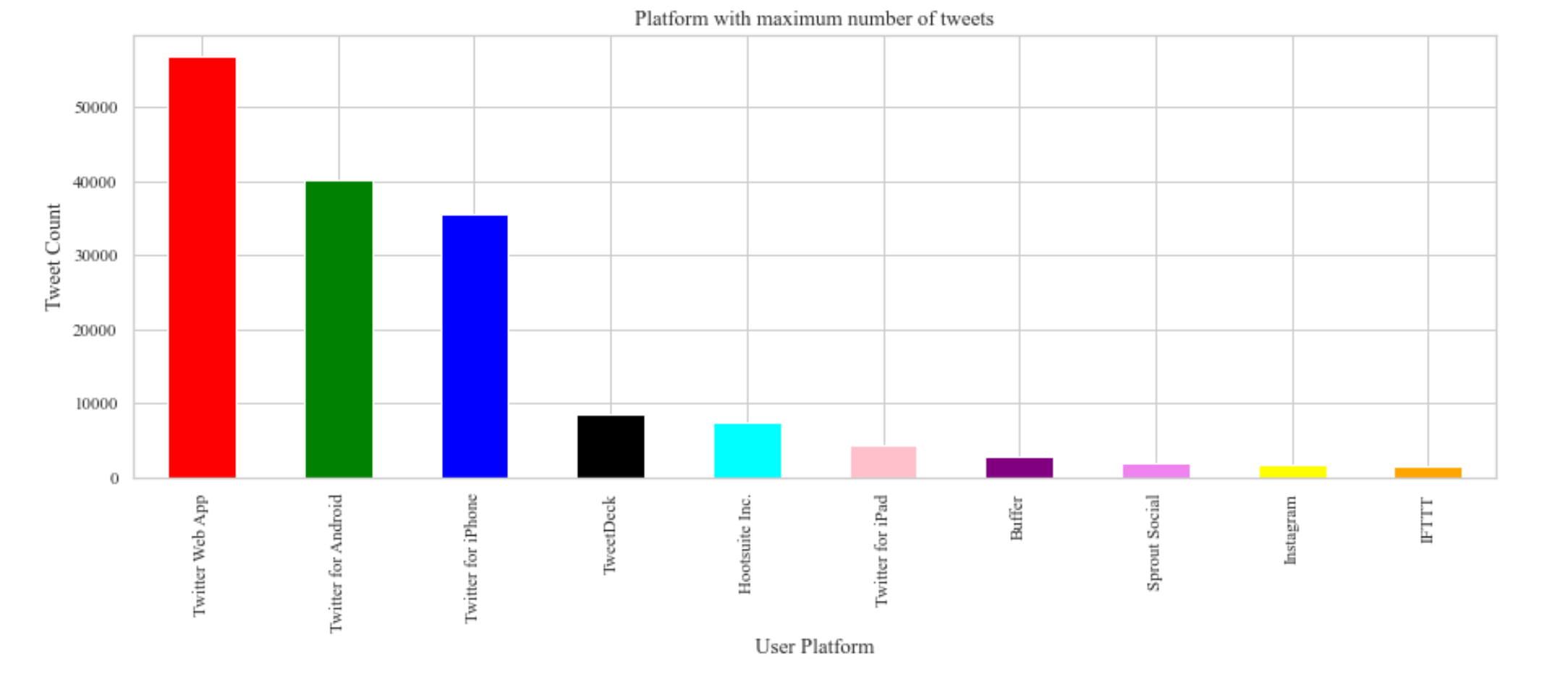
Top 10 countries by number of tweets:



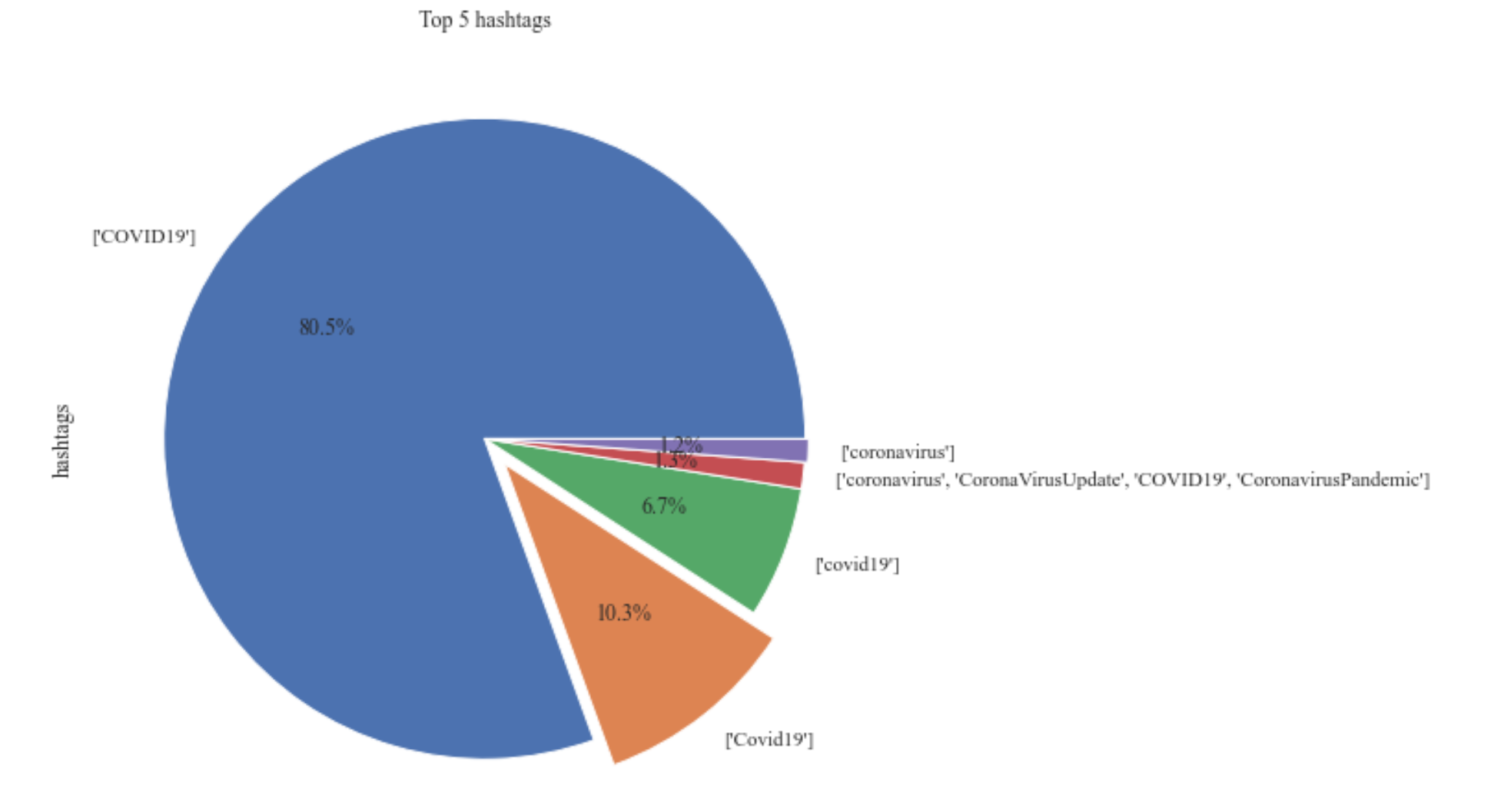
The number verified vs non verified users:



Platform with the maximum number of tweets:



Top 5 hashtags:

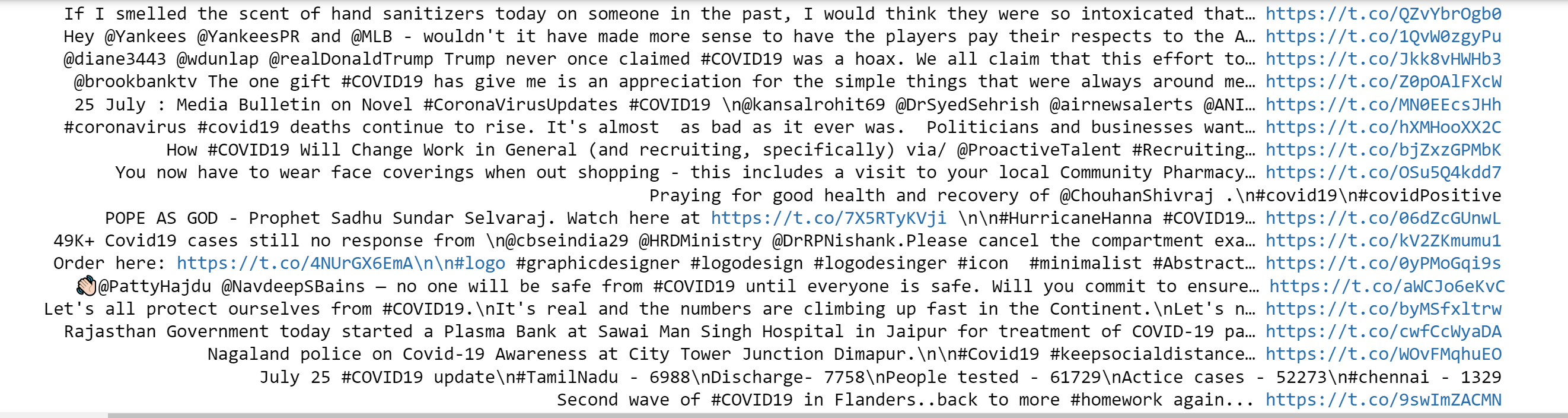


## Data Cleaning and Processing

The raw tweets consist of hashtags, @ mentions, special characters, emojis and URLs. The following steps have been done to clean the text column of the dataset. The cleaned text is added to the dataset as the clean\_text column.

1. Each tweet is split into individual words
2. The words are converted to lower case
3. Contractions are replaced with their full form. For example, didn’t is replaced with did not.
4. URLs are removed
5. @ Mentions are removed
6. Text in square brackets is removed
7. Special characters are removed
8. The word tokens are created using a twitter specific tokenizer
9. The lemmatization of the tokenized words is done

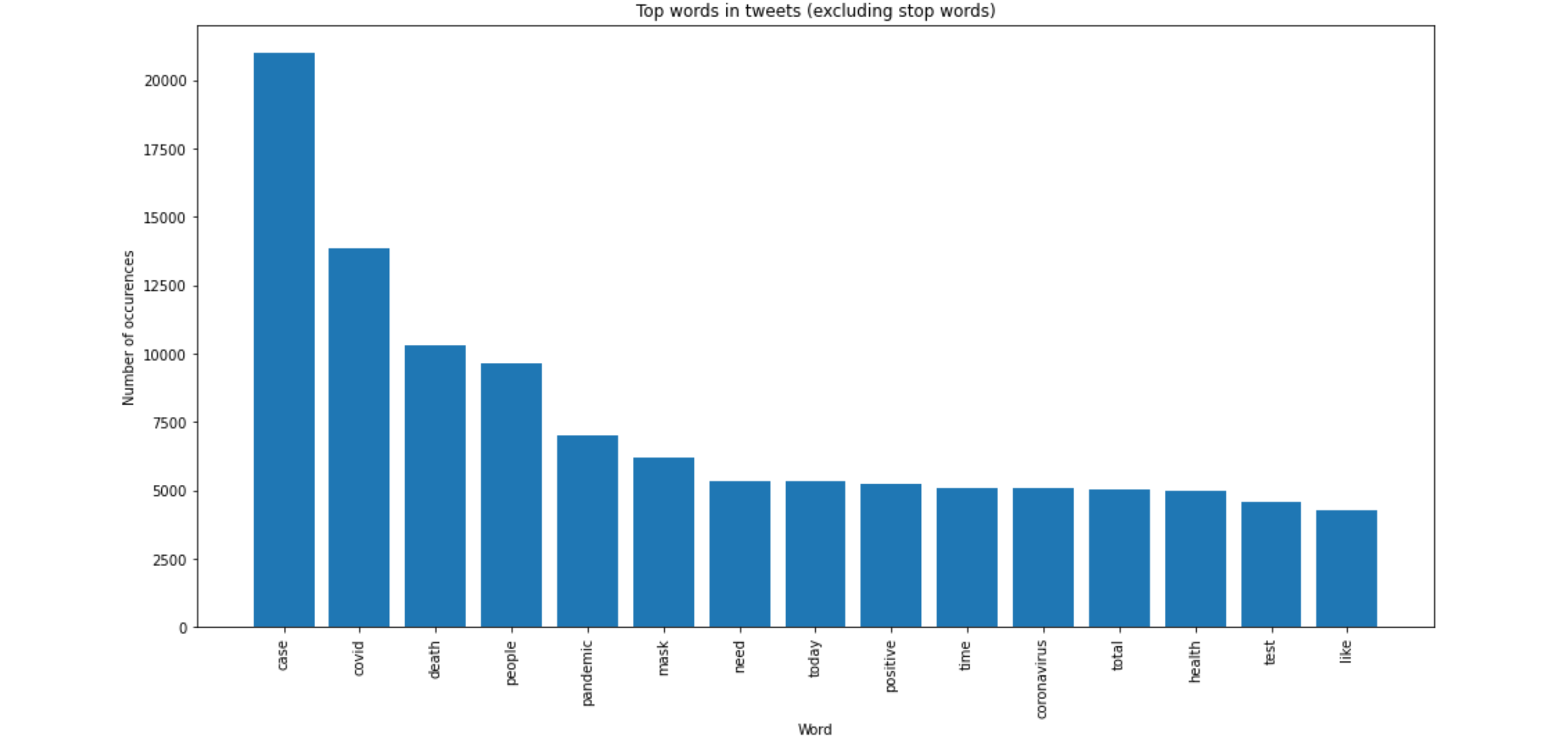
The original tweets text:



The cleaned tweets text:



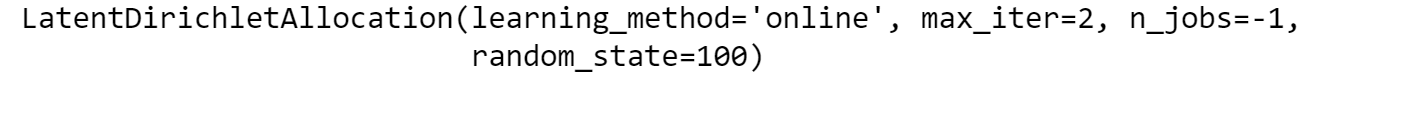
Top words in the tweets after the text has been cleaned:



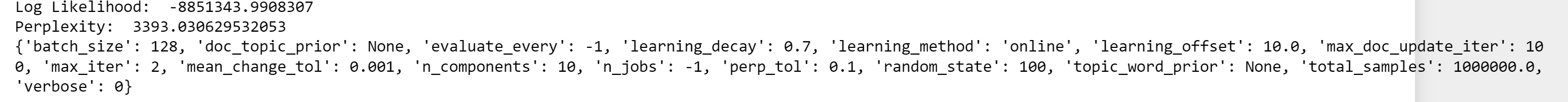
## Model and Testing

The cleaned tweets text is vectorized using a count vectorizer. The Latent Dirichlet Allocation method is used to build the topic model. The number of topics was set to ten. The maximum number of iterations was set to 2. The log likelihood and perplexity of the model are obtained from the parameters of the model. The higher the value of log likelihood better is the model. Lower the value of perplexity better is the model.

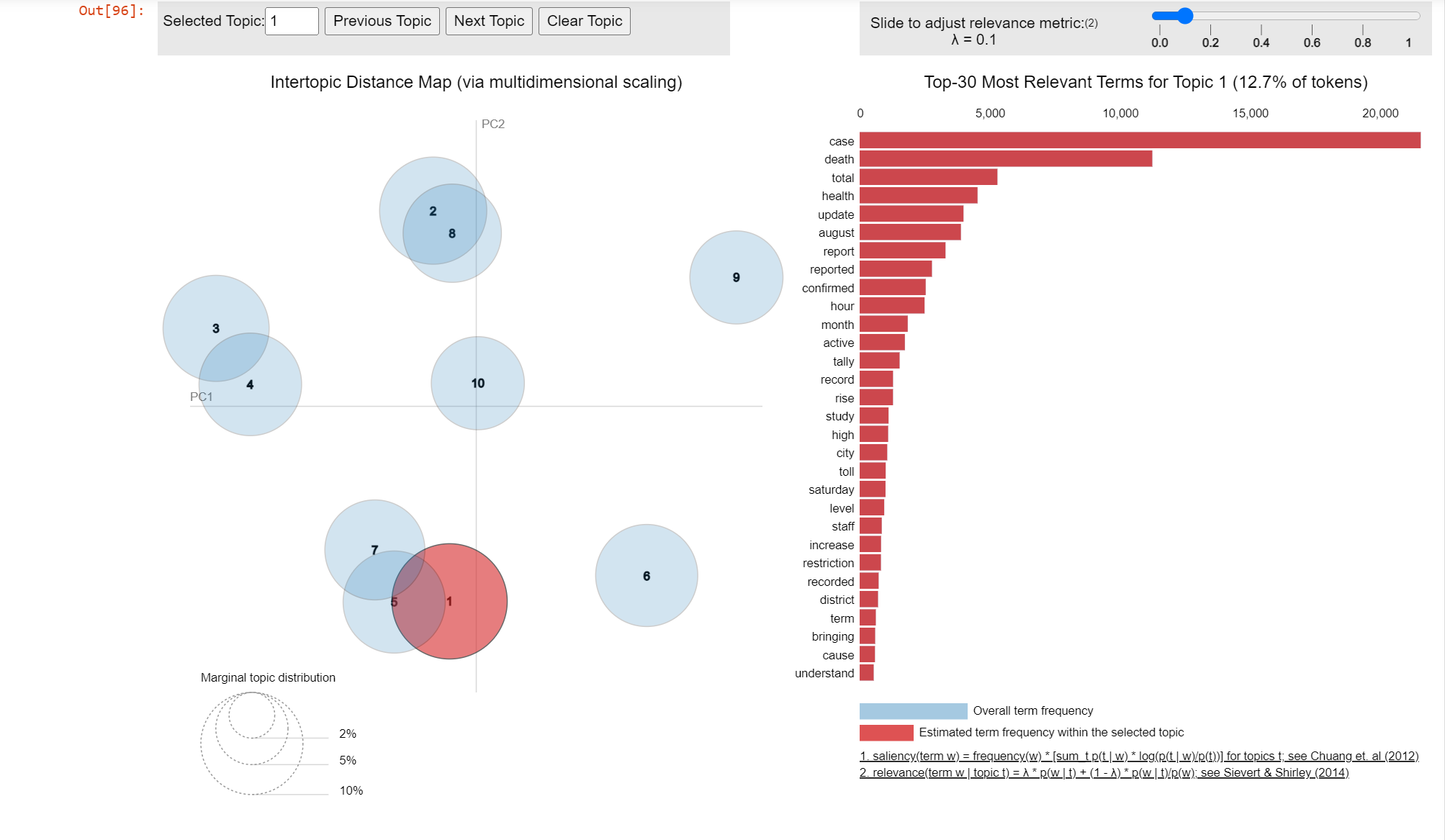
Model:



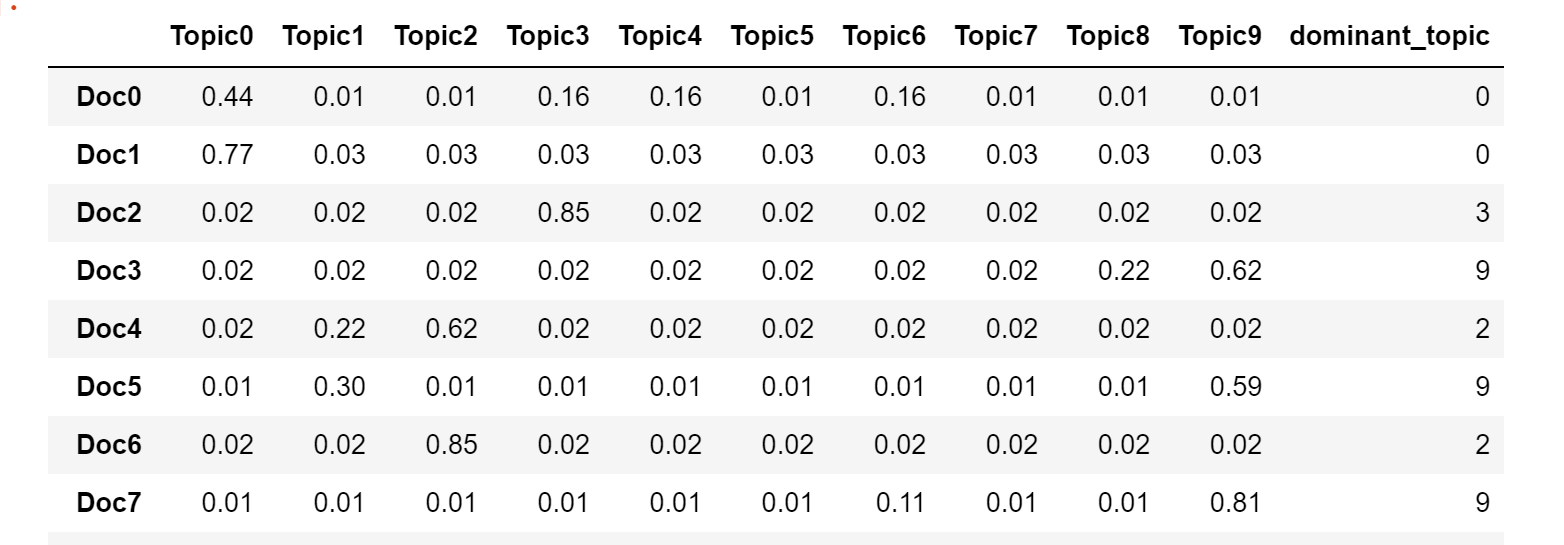
Parameters of the model:



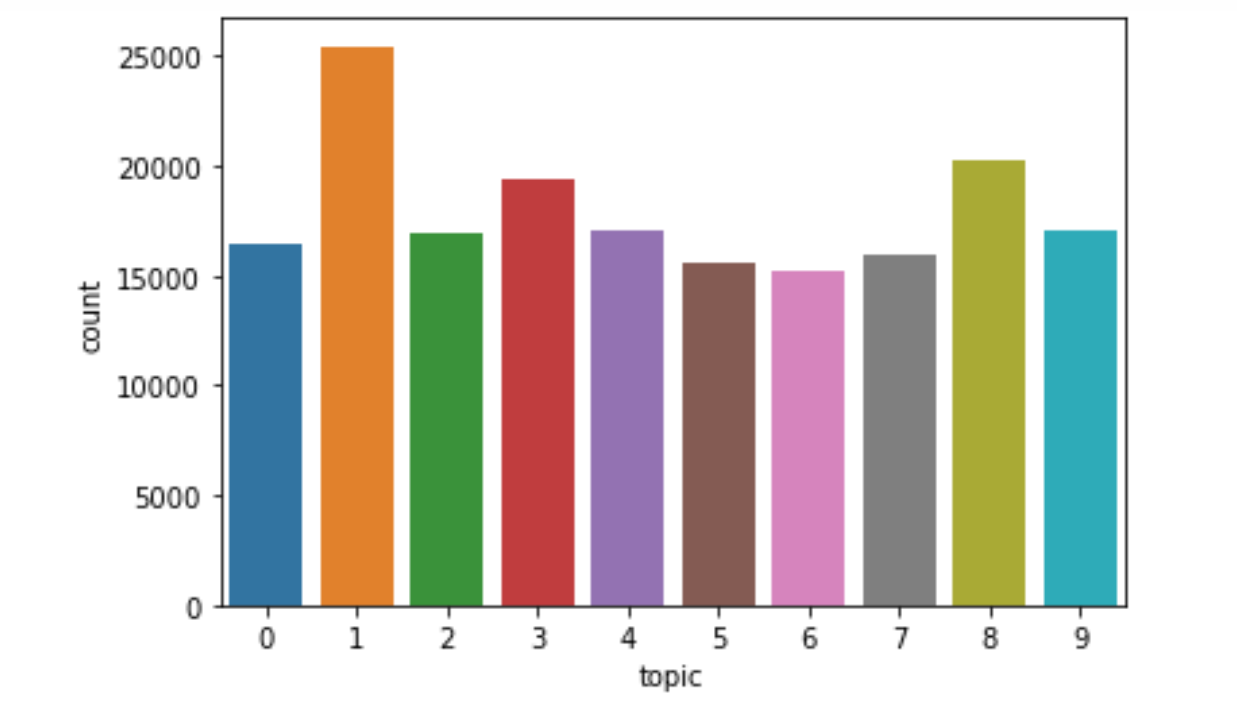
The topics and the keywords for each topic are visualized:



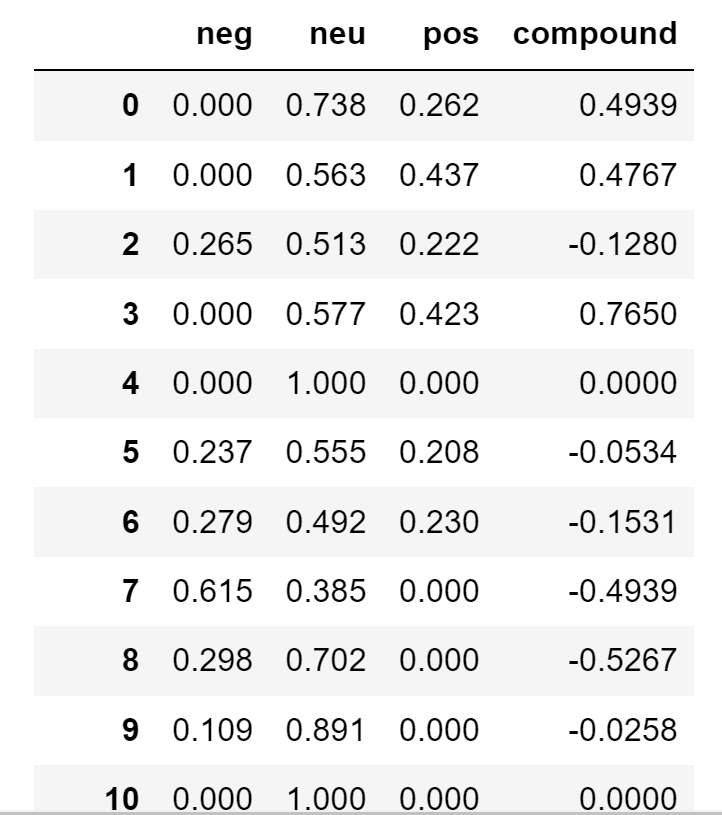
Using the model, each clean text is assigned a dominant topic.



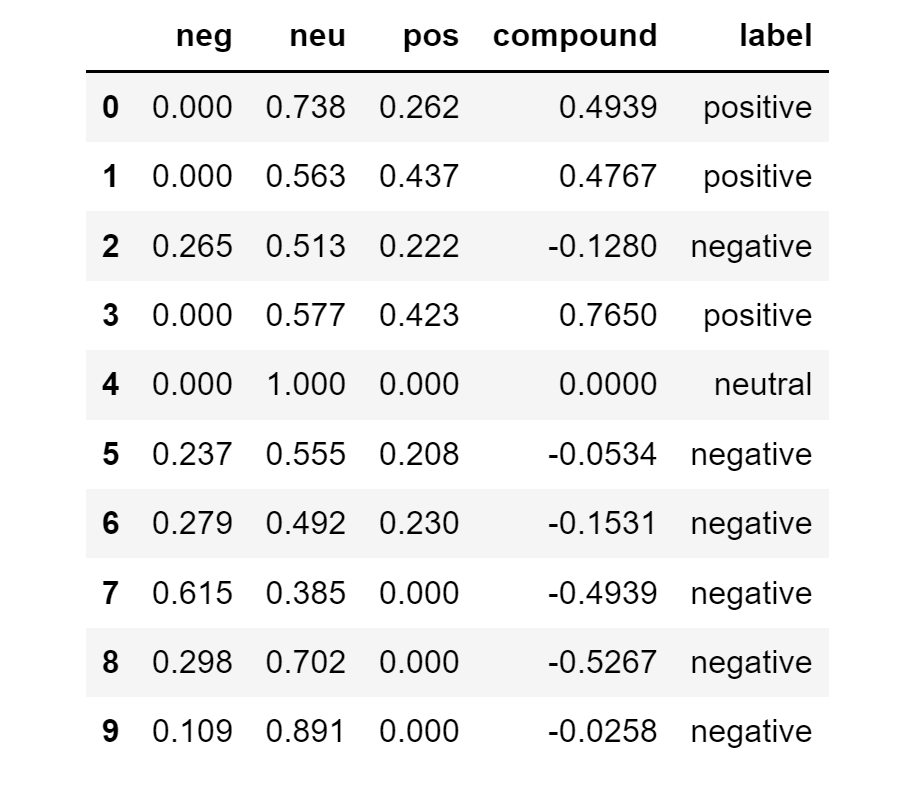
The number of documents in each topic:



Sentiment Analysis is done within each dominant topic. The following is the sentiment classification done using the Sentiment Analyser for topic 0.



A sentiment label is assigned to each tweet. If the compound value of the sentiment is 0, we assign neutral to the tweet as the sentiment. Greater than zero is assigned a sentiment of positive and less than 0 is assigned a sentiment of negative.



The total number of neutral, negative and positive tweets is calculated for the topic.

