Eluvio DS Challenge

Junliang Yu
(858)250-9868
yujl@ucsd.edu

May 2019

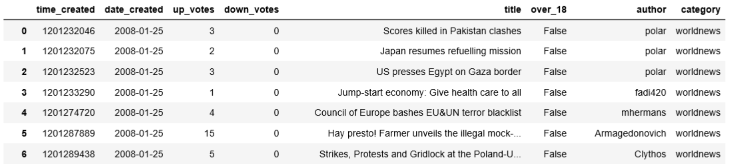
1 Data preprocessing

The original data is of size (509236, 8). As shown in Figure 1, time_created is Unix Time Stamp format. Up_votes and down_votes give the number of likes and dislikes from readers. Title column is the title of the article. Over_18 represents if it is inappropriate for under 18. Besides there are the author and the category of the article.

Time_created contains information of date and time, so it is a relatively standard reference of time. Date_created is also date information, but it contradicts with the Unix Time Stamp, maybe because the date here is recorded in different time zone. Thus Date_created column is dropped. Meanwhile, down_votes, category, and over_18 hold to little information, thus they are put aside. Time_created column can be convert to GMT time stamp every 2 hours, weekday, season, and year information. Add them to the data frame. Weekday ranges from 0 to 6 corresponding to Monday to Sunday. Season ranges from 0 to 3 for Spring to Winter.

2 trends of news and upvotes

First look at some overall characteristics and get a sense of how the data look like. Plot the number of news and totle upvotes by season, in Figure 2.



	time_created	date_created	up_votes	down_votes	title	over_18	author	category
0	1201232046	2008-01-25	3	0	Scores killed in Pakistan clashes	False	polar	worldnews
1	1201232075	2008-01-25	2	0	Japan resumes refuelling mission	False	polar	worldnews
2	1201232523	2008-01-25	3	0	US presses Egypt on Gaza border	False	polar	worldnews
3	1201233290	2008-01-25	1	0	Jump-start economy: Give health care to all	False	fadi420	worldnews
4	1201274720	2008-01-25	4	0	Council of Europe bashes EU&UN terror blacklist	False	mhermans	worldnews
5	1201287889	2008-01-25	15	0	Hay presto! Farmer unveils the illegal mock-...	False	Armagedonovich	worldnews
6	1201289438	2008-01-25	5	0	Strikes, Protests and Gridlock at the Poland-U...	False	Clythos	worldnews

Figure 1: Original data

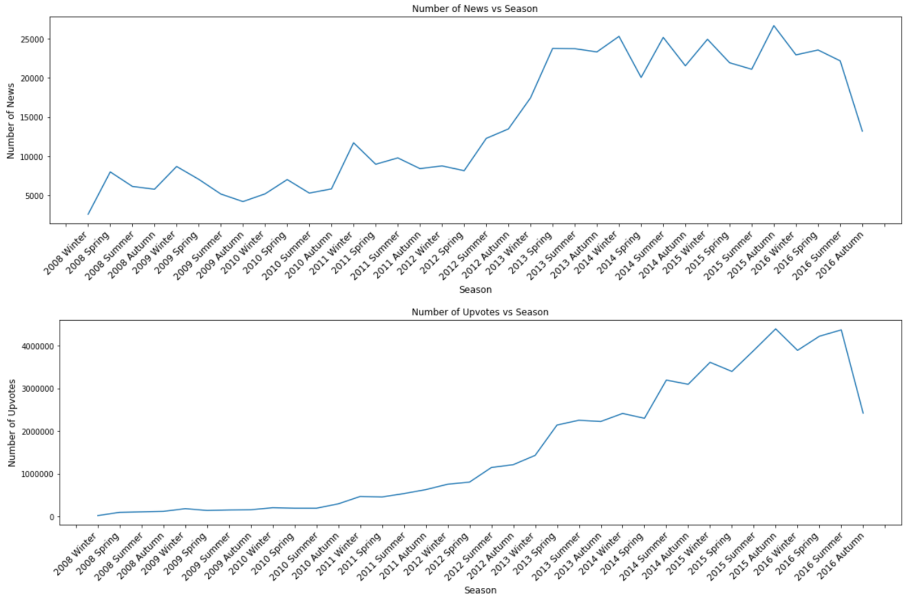


Figure 2: Trend of news amount and upvotes.

Both news amount and upvotes have been climbing up since 2008. The news amount reached its steady state in 2013 Spring. This is because before 2013 the world were getting closer and closer due to the development of Internet. Increasing number of people jumped out of the traditional media into new media, making people faster and easier to produce news articles. The steady point shows that the transmission capacity of the Internet has surpassed the speed news are created.

Upvotes keeps going up regardless of the incomplete data of the last time stamp. This is because the sources of new media, like news apps, are getting mature and delicate. People can check the news whenever and wherever, meanwhile be kept by the fineness of these apps. Additionally, upvoting is a way of expressing oneself and building self-identification for social attributes of the Internet. It is well possible that the number of upvotes will keep growing in the future.

3 Optimal publication time

This part will give suggestions to those who want to maximize the upvotes of their articles. If you are going to publish your article, what time would be the best?

From Figure 3, most people publish their articles at 6 to 8 on workdays. However, after calculating the heatmap of average upvotes, it is a little counterintuitive that the best time to publish your article is at 4 on weekends and 16 on Saturday. 6am on workday is probably not a good choice.

4 Text based clustering analysis

Now let's dive into the content, or the title of the article.

4.1 Corpus construction

First build the corpus, which is the basis of text analysis. Convert all text to lowercase. Tokenize and stem all strings of titles. Make a dictionary mapping from stems to tokens as the vocabulary frame.

4.2 Vectorization with Tf-idf

Combine 2 stopwords set from nltk.corpus.stopwords.words('english') and sklearn.feature_extraction.text.ENGLISH_STOP_WORDS. Set the minimum ratio to 10^{-3}. Use stem for token. Utilize 3-gram to generate features. The result of vectorization is a matrix of size (509236, 1814).

4.3 K-means clustering on titles

After testing a bunch of number, it is found out that set number of clusters to 20 would bring a good performance. Add the cluster number to the data frame as a new column. Now the data frame looks like Figure 4. Extract the most frequent 3 keywords of each cluster, and the results are shown in Figure 5

Traverse every title in a cluster, to make sure that there exists stems in the title matching the keywords of the cluster. The matching ratio of every cluster is in Table 1

Table 1: Matching ratio of every cluster

cluster number	0	1	2	3	4	5	6	7	8	9
matching ratio	1	1	1	1	0.085	1	1	1	1	1
cluster number	10	11	12	13	14	15	16	17	18	19
matching ratio	1	1	1	1	1	1	1	1	1	1

The result of clustering is pretty good except Cluster 4. The size of Cluster 4 is pretty large, which indicates that it is not completely clustered.

4.4 PCA on points for visualization

Reduce the dimension of the matrix to 2. Now we can plot the points in 2D figure and see their clusters in Figure 6.

5 Prediction for clusters

Say we want to know the trend of 'attacks'. It belongs to Cluster 17. Group all the news in Cluster 17 by season and plot the line plot as in Figure 7.

The frequency of attacks and kill reached its peak in 2015 Autumn, but started to decline afterwards. This demonstrates the world overall is living more peaceful than before. This is also an honor of the dedicated intelligence people and police to maintain the world in order and keep us safe.

6 Important and hot news

Another interesting topic is to find the important topics and hot topics. The difference between this two is that important topics is what the media report most frequently, and hot topics are what people are most engaged in, or rather, got most upvotes. The word cloud analysis is made to see how topics differ between the two, as shown in Figure 8.

We can see that media focus on keywords like China, Russia, and Israel that is very related to international situation. Whereas for people, what they are interested in are marijuana, Snowden and Trump. People pay more attention to their own life and politics that are close to them.

7 Most popular author

This analysis is significant if you are looking for writers to serve your purpose. Sort all author by average upvotes can tell us something (see Figure 9)

The most popular author is named Neosporin. Dig out what he/she is good at. Check out his/her most frequent keywords to get to know its specialty in Figure 10. He/she is a professional in topics related to Canada, treaty, and democracy.

8 Recent keywords

This can help us know what's going on right now, and give us insights on latent possibility. Assume that today is the last date in the form, then in the past 7 days, the most important topics and hot topics are shown in Figure 11.

Trump has frequently been in news. Still, the media and the people concern about different things.

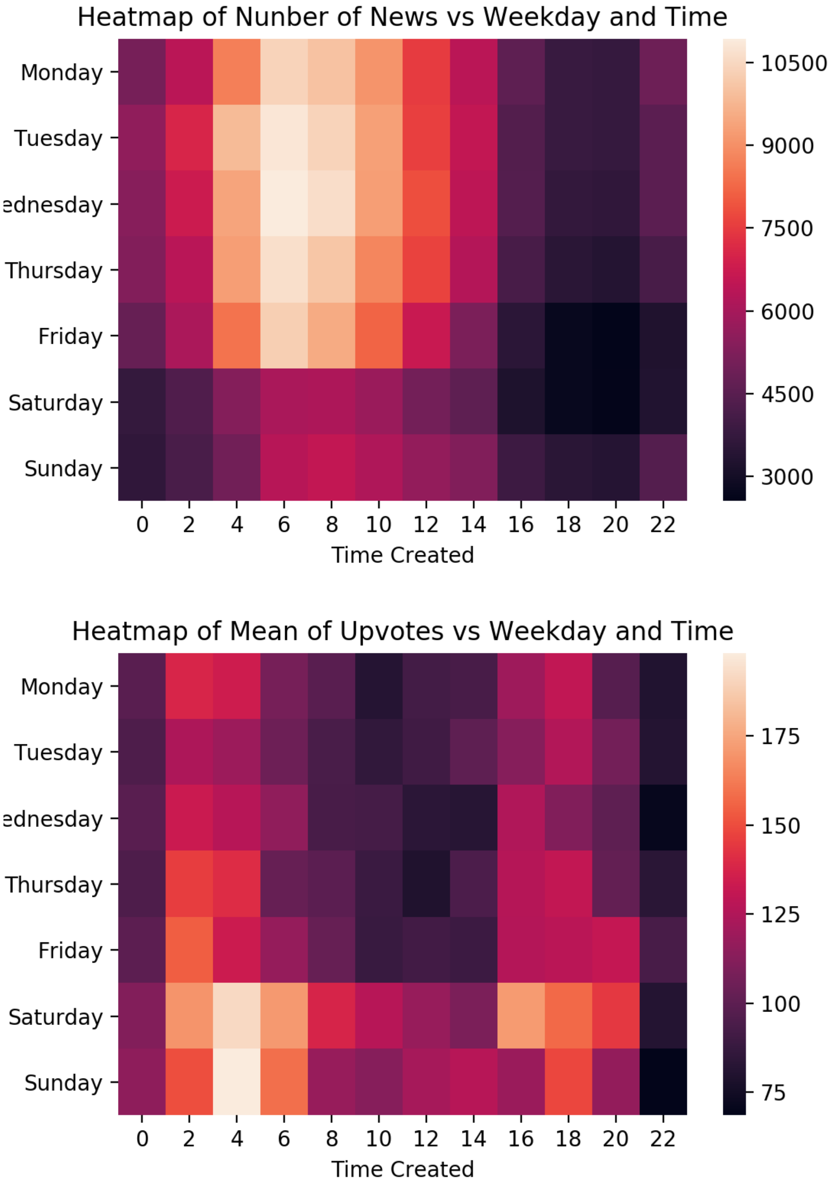
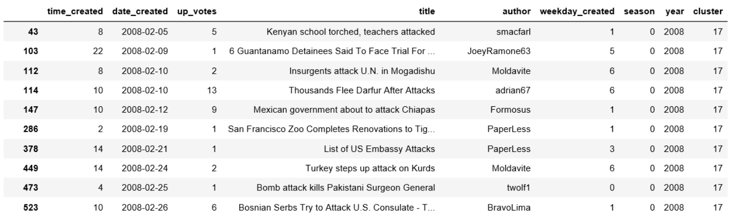
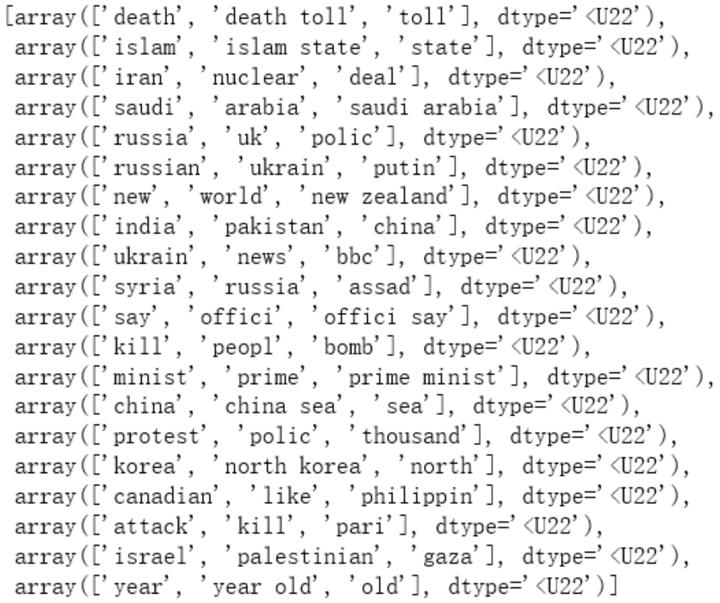


Figure 3: Heatmap of news amount and average upvotes.



	time_created	date_created	up_votes	title	author	weekday_created	season	year	cluster
43	8	2008-02-05	5	Kenyan school torched, teachers attacked	smacfarl	1	0	2008	17
103	22	2008-02-09	1	6 Guantanamo Detainees Said To Face Trial For ...	JoeyRamone63	5	0	2008	17
112	8	2008-02-10	2	Insurgents attack U.N. in Mogadishu	Moldavite	6	0	2008	17
114	10	2008-02-10	13	Thousands Flee Darfur After Attacks	adrian67	6	0	2008	17
147	10	2008-02-12	9	Mexican government about to attack Chiapas	Formosus	1	0	2008	17
286	2	2008-02-19	1	San Francisco Zoo Completes Renovations to Tig...	PaperLess	1	0	2008	17
378	14	2008-02-21	1	List of US Embassy Attacks	PaperLess	3	0	2008	17
449	14	2008-02-24	2	Turkey steps up attack on Kurds	Moldavite	6	0	2008	17
473	4	2008-02-25	1	Bomb attack kills Pakistani Surgeon General	twolf1	0	0	2008	17
523	10	2008-02-26	6	Bosnian Serbs Try to Attack U.S. Consulate - T...	BravoLima	1	0	2008	17

Figure 4: New data frame



```
[array(['death’, 'death toll’, 'toll’], dtype=’<U22’),
 array(['islam’, 'islam state’, 'state’], dtype=’<U22’),
 array(['iran’, 'nuclear’, 'deal’], dtype=’<U22’),
 array(['saudi’, 'arabia’, 'saudi arabia’], dtype=’<U22’),
 array(['russia’, 'uk’, 'polic’], dtype=’<U22’),
 array(['russian’, 'ukrain’, 'putin’], dtype=’<U22’),
 array(['new’, 'world’, 'new zealand’], dtype=’<U22’),
 array(['india’, 'pakistan’, 'china’], dtype=’<U22’),
 array(['ukrain’, 'news’, 'bbc’], dtype=’<U22’),
 array(['syria’, 'russia’, 'assad’], dtype=’<U22’),
 array(['say’, 'offici’, 'offici say’], dtype=’<U22’),
 array(['kill’, 'peopl’, 'bomb’], dtype=’<U22’),
 array(['minist’, 'prime’, 'prime minist’], dtype=’<U22’),
 array(['china’, 'china sea’, 'sea’], dtype=’<U22’),
 array(['protest’, 'polic’, 'thousand’], dtype=’<U22’),
 array(['korea’, 'north korea’, 'north’], dtype=’<U22’),
 array(['canadian’, 'like’, 'philippin’], dtype=’<U22’),
 array(['attack’, 'kill’, 'pari’], dtype=’<U22’),
 array(['israel’, 'palestinian’, 'gaza’], dtype=’<U22’),
 array(['year’, 'year old’, 'old’], dtype=’<U22’)]
```

Figure 5: Keywords for every cluster

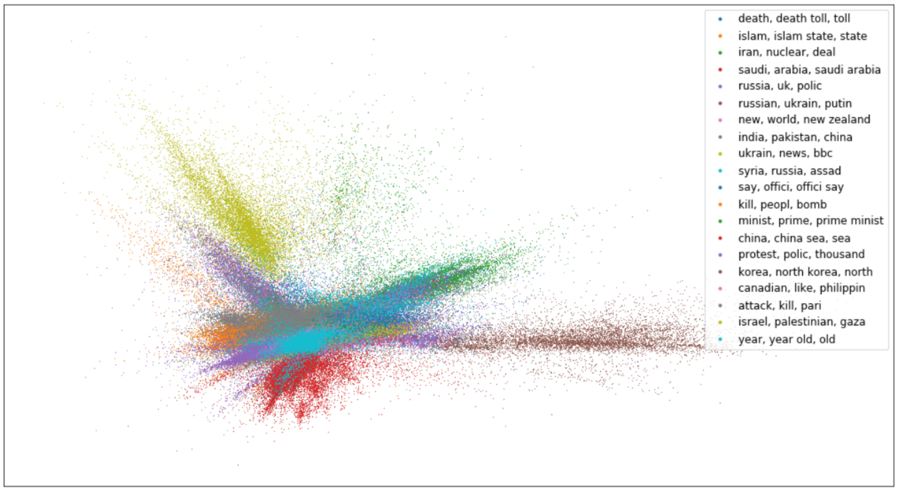


Figure 6: Clusters of all data points in 2D.

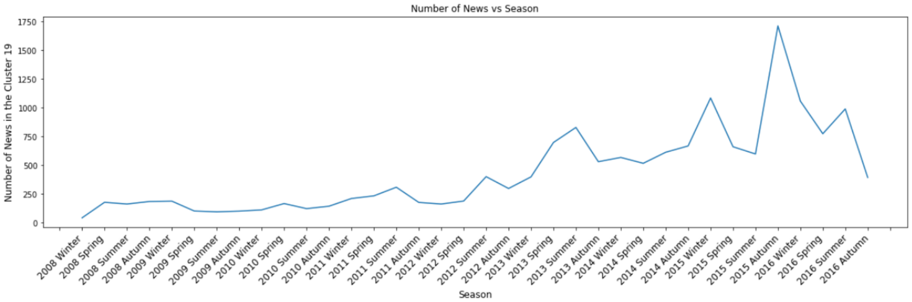


Figure 7: Trend of news amount of Cluster 17.

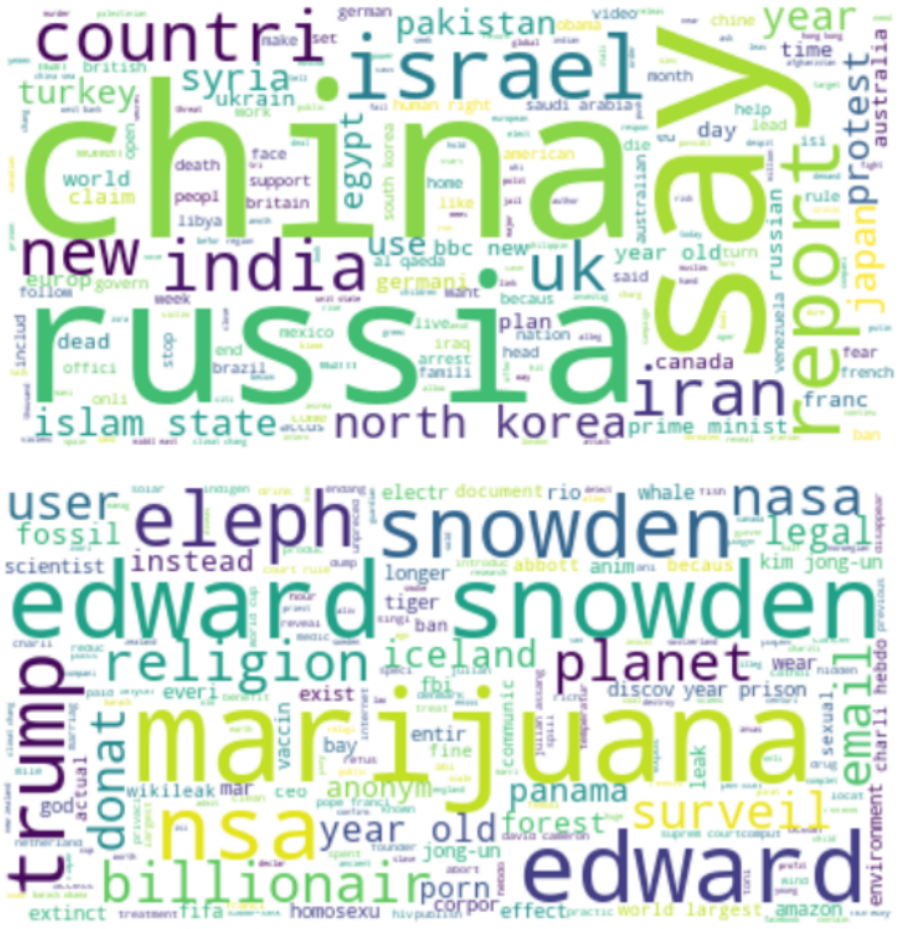
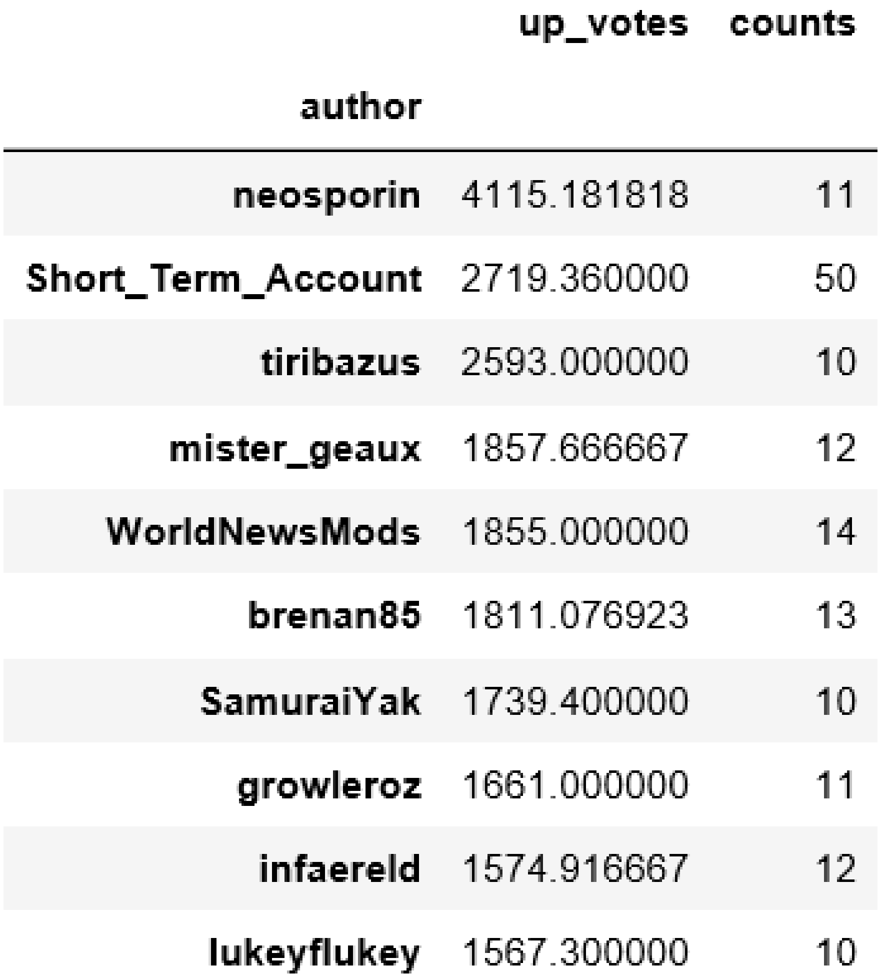


Figure 8: Important news and hot news.



	up_votes	counts
author		
neosporin	4115.181818	11
Short_Term_Account	2719.360000	50
tiribazus	2593.000000	10
mister_geaux	1857.666667	12
WorldNewsMods	1855.000000	14
brenan85	1811.076923	13
SamuraiYak	1739.400000	10
growleroz	1661.000000	11
infaereld	1574.916667	12
lukeyflukey	1567.300000	10

Figure 9: Most popular authors (at least 10 publilcations).

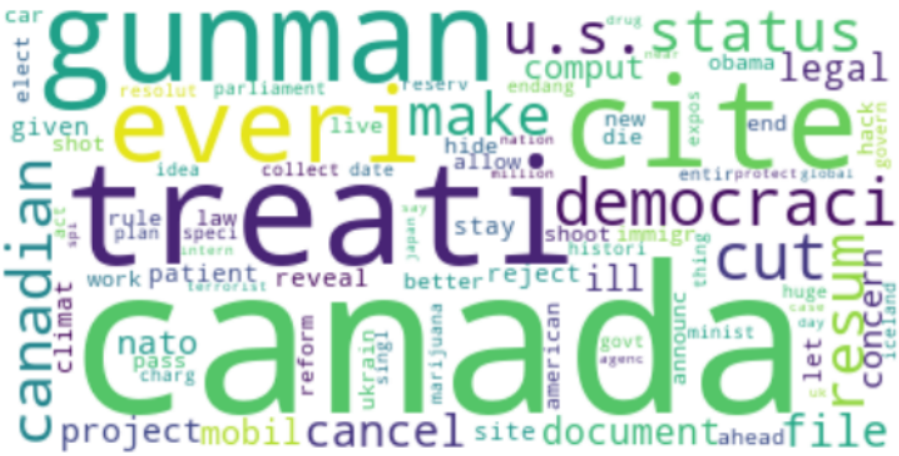


Figure 10: The most frequent keywords of the author.

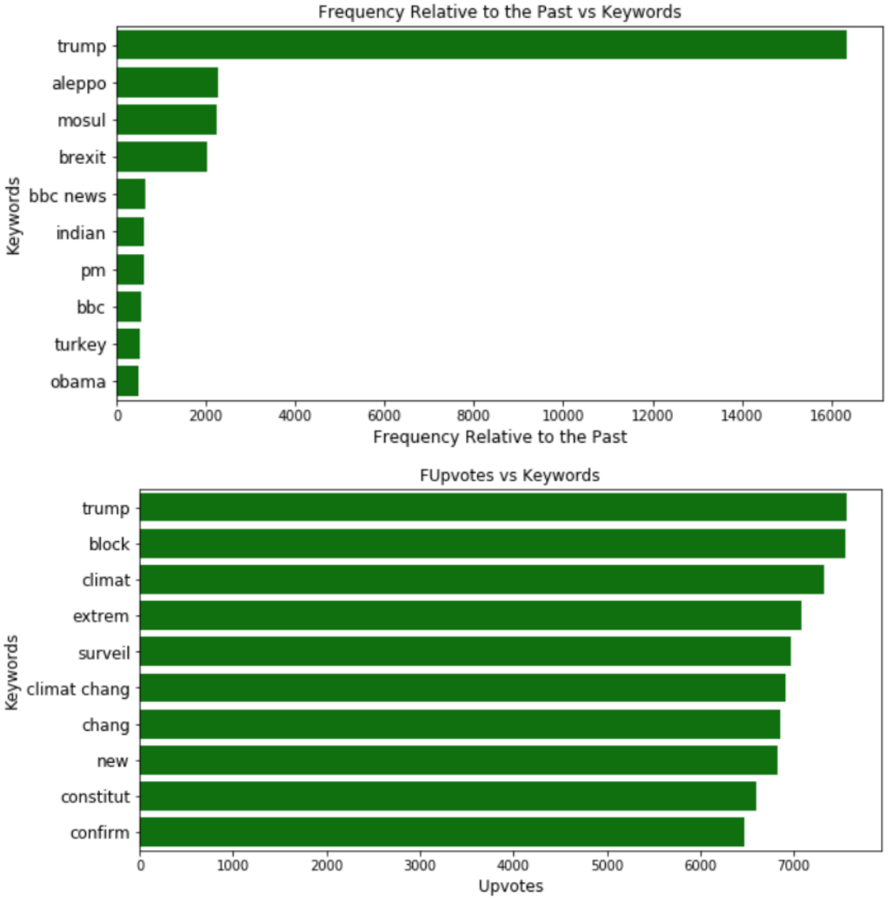


Figure 11: Important news and hot news in the past 7 days.