Hot news from twitter feeds

Nasrin Baratalipour
nasrin.baratali@gmail.com

Abstract

In this document, I analyzed the tweets of some important news agencies. To do so, I group tweets about the same event into one cluster and select the best tweets from every such cluster. I used the MALLET topic modeling package to identify and label the related clusters of the tweets. I also recognize tweets that are hot and urgent, independent of their clusters, and ranked them based on their degree of importance. To find interesting tweets, I considered the number of times a tweet has been retweeted and how many times this tweet has been "favorited" by by Twitter users.

1 Dataset

I used Twitter4j library to download the tweets of a specific twitter feed in a specific range of time. So, I created a twitter dataset downloaded from the provided list of twitter feeds from Dec. 12th to Dec. 14th. The twitter dataset contains 9081 records. Each record stores a tweet with its corresponding attributes (e.g. retweeted count, favourite count, etc.).
You can find the twitter dataset in the address: `data/tweets`. The codes that I used to download and store the tweets are in following address:
`https://github.com/Nasriin/HotNewsClustering/blob/master/News360/src/main/java/downloading/TweetsDownloader.java`

The Twitter4j library gives the possibility to stream the tweets as well. I wrote a code for streaming tweets but because lack of time I could not complete it. You can find the code in the following address: `https://github.com/Nasriin/HotNewsClustering/blob/master/News360/src/main/java/downloading/TwitterStreamer.java`

2 Clusters of Tweets

To detect cluster of a tweet, I first used hashtags of the tweets. However, I realized 8306 from 9081 tweets does not have any hashtages, so I used another approach for the clustering task. You can find a list of hashtages and the number of tweets used those hashtages, extracted from our twitter dataset in a CSV file in address: `data/hashtags.csv`.
I applied the topic modeling concept in my second attempt for the clustering task. For our current setting, we can assume that each cluster is equivalent to the topics and each tweet represents a document in the topic modeling terminology. For this purpose, I used MALLET topic model package. The MALLET topic model package assumes that the documents are a mixture of topics and each topic can be represented by a distribution over words. More specifically, given the distribution $P(z)$ over the cluster z in a particular document, and the distribution $P(w|z)$ for the probability distribution over the words given the topic z:

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j) \times P(z_i = j) \tag{1}$$

where $P(w_i)$ is the probability of the i-th word of the document, and T is the number of topics. MALLET also imposes a Dirichlet prior on the parameters of the multinomial distributions $P(w|z)$

and $P(z)$. Here, the parameters of Dirichlet prior are assumed to be equal to α and β for $P(w|z)$ and $P(z)$ respectively.

The value of T (i.e. the number of topics) can affect the interpretability of the results. if T is too small, the output topics will typically be a very broad topic and will not be useful; whereas if T is too large, then many topics will be uninterpretable and may contain idiosyncratic word combinations. There are different approaches to choose the number of topics such as [1] and [3]. Considering the time limit of the project, I did not use such approaches and instead I tried with different numbers between 5 to 100 and finally used 20 for the number of topics. Based on my analysis of topics, I believe that the final 20 topics are not broad and can be easily interpreted.

For the hyperparameters α and β, I used the rule of thumb $\alpha = 50/T = 1$ and $\beta = 0.01$ proposed by [2]. Finally, I set the number of iteration of Gibbs sampling to 2000. Note that all stop words were removed before identifying documents' topics. Figure 1 shows the distribution of the 20 extracted topics over the dataset.

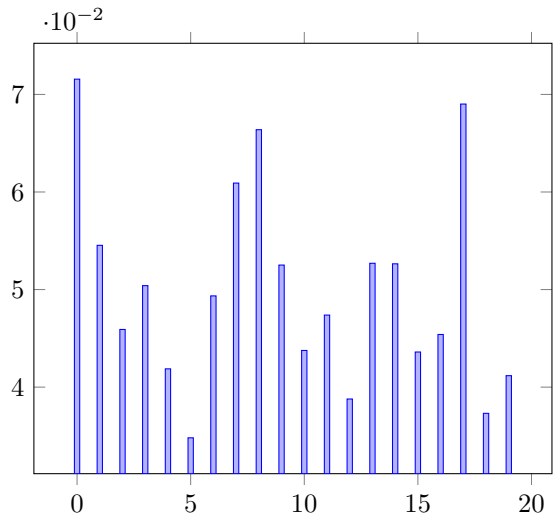

Figure 1: The distribution of 20 extracted topics.

The extracted topics can be useful for different applications. For example, since the meta-data of tweets contains the date that tweets were tweeted, the frequency of each topic at the specific time can be monitored to find the period of time a topic became hot. In this report, the main focus is to recognize the tweets when something extraordinary happens. Therefore, I cluster tweets about the same topic using topic modeling.

2.1 Hot clusters

In this project, a cluster is determined as a hot cluster if it contains tweets with high number of retweeted and favorite count. In this regard, I defined a hotness degree that is calculated by summing up the retweeted count and favourite count of each tweet in a cluster. The clusters and their calculated hotness degrees are presented in the Table 2. The clusters are sorted by their hotness degrees in descending order. In Addition to clusters and the hotness degree, the words representing each clusters have been shown in the Table 2. The representative words have been obtained by topic modelling approach that is used for clustering. As it is shown the table, the highest hot/urgent/important cluster belongs to the cluster of tweets talking about Syria.

2.2 Hot Tweets in each Cluster

To recognize the most urgent/hot tweets in each cluster I used retweeted count and favorite count of each tweet. In this regard, I calculated the hotness degree of each tweet in a cluster by summing up the retweeted count and the favorite count. I reported two most urgent/hot tweets of five clusters in the Table 3. You can find the list of all clusters with their five most important tweets in the address:

Cluster	HD	Representative words
(0)	*258406*	*aleppo (375) syrian (97) civilians (73) rebels (68) syria (60) deal (50) people (50)*
(1)	*124155*	*trump (370) donald (220) secretary (151) rex (145) tillerson (140) state (134) trump's (110)*
(2)	*56255*	*man (91) police (71) year-old (42) christmas (33) holiday (32) million (29) gift (26)*
(3)	*44662*	*billion (38) people (36) killed (29) yahoo (26) group (26) bombing (24) accounts (23)*
(4)	*40020*	*golden (87) nominations (65) globe (53) usatodaylife (49) awards (45) land (43) list (39)*
(5)	*50706*	*years (40) christmas (38) world (25) ago (19) life (19) truth (18) sex (16)*
(6)	*42892*	*trailer (27) fast (26) scientists (26) women (23) secret (22) eating (19) watch (19)*
(7)	*41319*	*patriots (93) star (75) indyfootball (67) wars (55) rogue (52) monday (39) night (37)*
(8)	*36611*	*social (55) care (41) council (38) tax (37) david (28) planet (26) money (25)*
(9)	*42726*	*car (37) women (34) uber (32) self-driving (30) president (27) cars (23) police (21)*
(10)	*62168*	*strike (84) rail (80) southern (76) christmas (65) santa (48) boy (41) strikes (39)*
(11)	*36058*	*thicke (63) alan (59) growing (37) pains (33) dies (32) legal (29) newscomauhq (24)*
(12)	*54944*	*win (30) health (28) donald (24) bill (23) trump's (22) white (21) real (20)*
(13)	*46079*	*brexit (100) trump (45) theresa (43) tech (41) minister (39) labour (28) pmqs (25)*
(14)	*51313*	*woman (51) people (42) fire (37) christmas (34) oakland (31) teen (27) star (27)*
(15)	*40795*	*front (53) page (46) tomorrowspaperstoday (35) skypapers (32) daily (28) bbcpapers (27)*
(16)	*256091*	*trump (404) donald (210) election (122) russian (92) kanye (71) west (67) hacking (59)*
(17)	*44363*	*finale (33) guardian sport (29) dead (29) woman (26) walking (24) baby (23) england (21)*
(18)	*52182*	*study (30) finds (30) city (29) water (24) emoji (19) top (18) tax (16)*
(19)	*52092*	*fed (61) time (46) year (43) rates (42) interest (38) day (36) rate (30)*

Figure 2: 20 clusters along with their hotness degree and the clusters representative words. The word HD in the second column label stands for hotness degree.

`data/clusterHotTweets.txt`. It is worth mentioning that the function that I have defined to degree the hotness and popularity of a tweet is based on two parameters: favorite count and retweeted count. However, time is also an important factor that can affect the hotness degree of a tweet. Unfortunately, because the lack of time I did not include it in the defined popularity function.

Cluster	Hotness Degree	Tweet
0	*7855.39*	*This little boy is the newest face of OshKosh B'gosh's holiday ads after initially being turned down by a talent ag*
0	*2988.12*	*"Santa, can you help me?" This Santa fulfilled a dying boy's wish to meet him, and the child passed away in his arm*
1	*1975.94*	*Reversing course, the EPA says fracking can contaminate drinking water*
1	*1384.30*	*The EPA has concluded that fracking has contaminated drinking water in some circumstances*
2	*3821.19*	*Clinton jokes at Reid portrait unveiling: "After a few weeks of taking selfies in the woods, I thought it would be*
2	*1336.34*	*Japanese aviary park has a unique way of celebrating Christmaswith Santa-costumed penguins.*
3	*3305.27*	*Energy Dept. rejects Trumps request to name climate change workers, who remain worried*
3	*1608.45*	*Trumps team asked the Department of Energy for names of all employees who have attended climate change conferences*
4	*7776.26*	*BREAKING: Donald Trump remains winner of Wisconsin following statewide recount showing few changes in vote totals.*
4	*5449.08*	*Pres. Obama: "I'd like everybody to just please join me in thanking what I consider to be the finest vice president*

Figure 3: The most urgent tweets of five clusters.

3 Hot Tweets

Independent of clusters, I used retweeted count and favorite count to rank the tweets in the downloaded twitter dataset. So, The tweets are ranked based on how they are hot, urgent and worth to show them to users. As I mentioned in the previous section, time is an important element that can help to determine the hotness of a tweet. I have shown the highest ten important tweets in the Table 4. You can find the list of ranked tweets along with their corresponding hotness degree in the following address:
`data/rankedTweets.txt`

Hotness Degree	Tweet
46097	*Save Aleppo. Save humanity. Residents of East Aleppo are giving their final messages to the world. https://t.co/Hzd4VWp0wC*
33818	*We've been friends for a long time: Kanye West and President-elect Trump appear together at Trump Tower https://t.co/KIdFwkVTGr*
11496	*This little boy is the newest face of OshKosh B'gosh's holiday ads after initially being turned down by a talent ag https://t.co/LG0rdFFVkc*
10680	*"The battle for Aleppo has reached its final stage, and residents are posting desperate goodbye messages https://t.co/xxGnVL7Ghg"*
10281	*RT @BBCEarth: Every turtle that was seen or filmed by the PlanetEarth2 crew was collected and put back into the sea.*
9561	*"Dear world, why are you silent?': Desperate pleas from inside Aleppo https://t.co/1RvPlSDB7a https://t.co/XYwbJibFTD"*
7948	*BREAKING: Donald Trump remains winner of Wisconsin following statewide recount showing few changes in vote totals.*
7593	*RT @BBCWorld: Battle for Aleppo: Final goodbyes from a city under siege in Syria https://t.co/EuG4vHsQzl https://t.co/FL3nCCreL5*
7496	*Aleppo is being destroyed by the silence of the world' https://t.co/O8vRi4auQi https://t.co/CGxzzDZWLC*
7311	*"JUST IN: Alan Thicke has died at age 69, the actors publicist confirms to @ABC News. https://t.co/xCjfckaezm"*

Figure 4: The highest 10 important tweets in the downloded Twitter dataset.

You can also find the codes for topic modeling, selecting hot clusters/tweets in the following address:
`https://github.com/Nasriin/HotNewsClustering/tree/master/News360/src/main/java/clustering`

References

[1] Michal Rosen-Zvi et al. "The author-topic model for authors and documents". In: *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. AUAI Press. 2004, pp. 487–494.

[2] Mark Steyvers and Tom Griffiths. "Probabilistic topic models". In: *Handbook of latent semantic analysis* 427.7 (2007), pp. 424–440.

[3] Yee Whye Teh et al. "Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes." In: *NIPS*. 2004, pp. 1385–1392.