DEVELOPING TOPICAL CLUSTERING OF LIBYAN'S TWEETS USING CARROT^2 FRAMEWORK

Moammar Mohamed Abdalmjied, Selo Sulistyo, Adhistya Erna Permanasari
Department of Electrical Engineering and Information Technology
Faculty of Engineering of Gadjah Mada University
Email: 1985mejo11@gmail.com, selo@ugm.acid, adhistya@ugm.ac.id

Abstract

Today, Twitter in addition has become a way of life also has become one of the sources of data in a very large number that can be accessed by anyone and anytime as long as it is connected to the internet. Data twitter is now widely used for various purposes, e.g. for the purposes of the survey. As we know that today many organizations, both government and non-government have been utilizing Twitter as a data source. In this study, developed an application to analyze data or tweets twitter committed by the Libyan people, where most of the tweets are written in letters and Arabic. The results of the analysis are expected to be topics of conversation among the Libyan people on Twitter.

Development is done by utilizing Carrot^2 Framework, where the stages of analysis and design are done by using the object-oriented approach utilizing UML. In addition, the application also utilizes several open libraries, such as Apache Lucene is useful for the process of indexing and retrieval and also Twitter4J tweets, a useful library API to access tweets from Twitter.

The resulting application is then tested with multiple queries, and the results of these tests can then be calculated degree of accuracy in the system generate topics of conversation. With 10 queries, system generates 197 topics, which 185 topics are relevant while the 12 topics considered irrelevant. So the accuracy of the system can be calculated and the results obtained 93.91. It can be concluded that the accuracy of the system in generating relevant topics is 93.91%.

Introduction

Today, twitter has become life style. It is through twitter that People all over the world can communicate and share their opinions and thoughts. The twitter's user can post million of posts in short messages every day. I would argue that tracking all messages and conversations that posted by friends as a whole is almost tedious and impossible.

During Libya's revolution, social media such as facebook and twitter have become the Libyans means of communications to express their protest against their government. As tension in the region increased and social media continued to play a significant role in the escalation of the Libyan leader, Muammar Gaddafi, he cautioned his citizens to hinder their Facebook usage. Activist groups used the social network as well as Twitter to call for reform and support the efforts of Egypt's own digital revolutionaries [1]. After the fall of the regime of Muammar Gaddafi, the Libyan people continue to use twitter to make voice for their aspirations, criticizing the government and calling for a better life expectancy based on peace and justice.

With the increasing number of Twitter's user in Libya, then the study to find out what Libyan talk about in Twitter have become important factor to be. By knowing what topics that Libyan talk on Twitter, it can be seen on the current issues that occur in the middle of the Libyan people. These issues can be very useful for Government or private institutions in Libya for a variety of interests. So far, studies have not been conducted to determine issues that arise in Libya by investigate the Libyan tweets on Twitter.

Carrot2 Framework is an Open Source framework that provides several integrated component that allows us to develop clustering rapidly. These components are open libraries,

such as apache lucene that very useful and powerful for indexing text and text retrieval that can process text in a large amount very quickly. Carrot2 also supported some clustering algorithms that have been integrated in it, such as Suffix Tree Algorithm (STC), K-Means, and LINGO.

There are some research that compared thee algorithm. S. Osinski & D. Weiss (2003) on a research titled "Conceptual Clustering using LINGO Algorithm: Evaluation on Open Directory Project Data" concluded that that Lingo does a decent job at separating topics present in search results and labels them informatively compared with STC and K-Means [2]. Other research in 2013 by R. Mahalakshmi & V. L. Praba titled "A Relative Study on Search Results Clustering Algorithms - K-means, Suffix Tree and LINGO" show that Lingo produces high cluster diversity [3].

According to the facts presented in the research, in this paper we will analyze the conversation of Libyan's tweets on Twitter. By utilizing the Carrot2 Framework and Apache Lucene; an application will be built to cluster topics of the conversation. The main reason underlying the use of the Carrot2 Framework and Apache Lucene is due to both of them support the analysis of text written in Arabic and most of Libyan send tweets on Arabic.

Literature Review

There are several studies that have been done before associated with this study. On 2011 K.D. Rosa et al [4] presented a study on automatically clustering and classifying Twitter messages, also known as "tweets", into different categories, inspired by the approaches taken by news aggregating services like Google News. The experiment with ways of clustering tweets into six predefined topics: News, Sports, Entertainment, Science, Technology, Money, and "Just for Fun". To define the clusters, they leverage popular hash-tags that appear in the tweets for a given topic as a sort of gold standard label used by the different clustering and categorization algorithms.

The study's results suggest that the clusters produced by traditional unsupervised methods can often be incoherent from a topical perspective, but utilizing a supervised methodology that utilize the hash-tags as indicators of topics produce surprisingly good results. This study also offer a discussion on temporal effects of the methodology and training set size considerations. Lastly, the study also describe a simple method of finding the most representative tweet in a cluster, and provide an analysis of the results.

A study titled "Clustering Users in Twitter Based on Interests" in 2012 also discuss about clustering on twitter's data ot tweets [5]. This study investigate the problem of clustering users in Twitter based on their interests. Solving this problem is very important in many different fields, such as user recommendation, personalized services, viral marketing, etc.

To address this problem, the author first calculate user similarity leveraging both textual contents and social structure, according with Twitter's role, not only a news media but also a social network. These features include tweet text, URLs, hashtags, following relationship and retweeting relationship, all of them are closely correlated with user's interests. Then by using user similarity as a measure to cluster users. To assess effectiveness of our method, we propose clustering metrics "average number of following links per user in per cluster".

Experimental results show that our method can successfully cluster users in Twitter, and get a much better performance than random selection.

Experimental results show that the method can successfully cluster users in Twitter, and performs much better than random selection. From a side view, our experiment also shows that users in the dataset of Twitter can be approximately categorized into 400 clusters.

Other study that also associated with the study that will be done is a study titled "Delineating Real-Time Events by Identifying Relevant Tweets with Popular Discussion Points" by M.A. H. Khan et al (2013) [6]. This study proposed a n unsupervised method for recommending search users a set of tweets that best delineate an ongoing public event. The proposed graph-based retrieval algorithm is based on a hypothesis that the discussion points that are common among majority of event-relevant tweets are motivated by the important occurrences of the ongoing event. Hence, by identifying the popular discussion points in a collection of event-relevant tweets, and retrieving tweets comprising those discussion points, it is possible to outline real-time events. Further perform topical clustering on the relevant tweets before applying the retrieval algorithm on each topic cluster, so that, users interested in a particular aspect of the event can dig deeper into the search results returned for that particular cluster. Evaluation performed on about 270,000 relevant tweets generated during a real-world event reveals that, the tweets recommended by the proposed model could delineate the proceeding of the event with high precision and recall and could also outperform two intuitive and competitive baseline models.

The difference between the study to be conducted with the studies described above are on the methodologies used to solve the problem and the data that will be analyzed. The study to be conducted will be focused on data of Libyan's tweets. By utilizing Carrot2 Framework an application will be developed.

Research Method

1. Analysis

System Architecture

Figure 1 show the system architecture that will build in this study. From Figure 1 can be seen that Twitter's data or tweets accessed from Twitter based on the query from administrator utilizing Twitter4J library. The Twitter data then will be processing utilizing Apache Lucene Library to produce index that will be store into the system's storage. In the other side, when user submit a query in order to get the topical cluster of tweets, the query will processed on feature extraction process. This process will access data from storage. Feature resulted from this process will be processed on the next step in order to produce cluster and cluster's content. At last, the result will be presented to user as topical cluster of tweets.

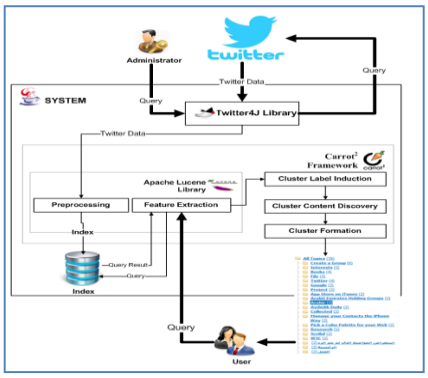

Figure 1 System Architecture

System Requirement Analysis

Table 1 presents a list of system requirement for each user that involved in the system. There

Table 1 System Requirement List

Actor	Requirement
Administrator	1) Administrator can log into the system. 2) Administrator can log out from the system. 3) Administrator can manage tweets or twitter's data. In process of manage tweets also include several use cases, namely: a) Get tweets from location b) Load tweets from index c) View tweet list
user	User can process an analysis The main purpose of this system development is to produce an application that can be used to analyze and generate topics of tweets in the form of clusters. To process an analysis, there are several use cases that will be included, namely: a) Submit query b) Load tweets from index based on query c) Clustering query result d) Display clusters e) Display statistic chart of tweet's topic

are two actors in the system, namely administrator and user. System's requirement for each actor have been presented on Table 1.

1) Design

Figure 2 shows the use case diagram of the system. From Figure 2 can be seen that there are two actors in the system, administrator and user.

Administrator's role is to manage tweets. While the user is using the system for the purpose of obtaining a particular analysis where the analysis is based on their needs using tweets stored in the index.

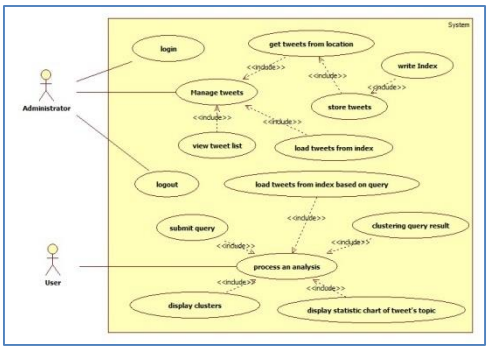

Figure 2 Use case diagram

Refer Figure 2 it also can be seen that the administrator access of twitter tweets from a particular location. In this case, a particular method of accessing is used, so the tweets coming from specific locations can be accessed in real time and stored in the index. Therefore, the use case get tweets from location also will include a use case store index that also will ultimately involve the use case write index into the index store and write tweets into index.

In his role managing tweets, then the administrator has the authority to enter into the system, in this case must first pass through the login process. If the administrator can log into the system, the administrator can also log out of the system.

The role of the actor in the specific user is interested parties to use the system to perform a specific analysis based on their needs, which in this case is done by inserting a specific query. As already detailed in Table 1, that the analysis results, there are stages that will be done of the system. These stages are the steps to do in doing clustering using Lingo algorithm. Own analysis results will be displayed in the form of a specific visualization, for example: charts, tree, or any other form. Visualization is intended to allow a

user to read and understand the topics generated from this analysis easily.

The clustering result then will be send to displayCluster class to create cluster display in tree view form.

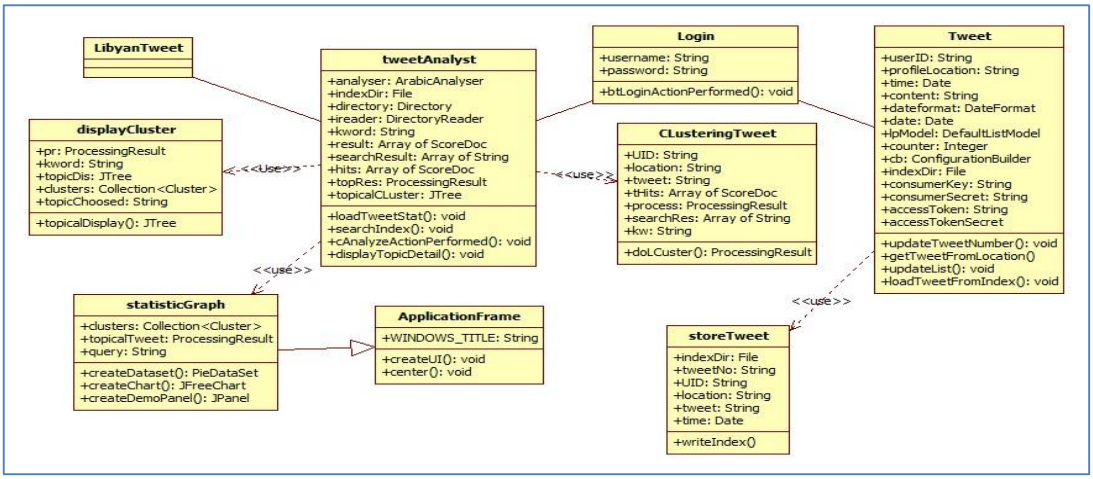

Figure 3 Class diagram

Class diagram of the system to be developed is shown by Figure 3. From Figure 3 can be seen the relation between classes of the system. As described on table 4.3 that LibyanTweet is the main class that will executed first when the system run. This main class will call tweetAnalyst class, and the relation between these classes is association. To perform tweets analysis and display the result in tree view and chart, tweetAnalyst uses these classes:

1) CLusteringTweet

   tweetAnalyst use this class to perform clustering of tweets. Based on the query entered by users, system will load query result or search result that stored in array of string which will be forwarded into CLusteringTweet as a parameter. And then CLusteringTweet will cluster using doCluster() method that will send processingResult as a return value to tweetAnalyst.

2) displayCluster

3) statisticGraph

   The analysis result also will send to this class in order to create cluster's display in a chart. This class inherit the properties and behavior of ApplicationFrame. The relation type between statisticGraph and ApplicationFrame is generalization.

Furthermore, tweetAnalyst also has an association relationship with login in which Tweet class has a dependency on it. In order to manage tweets, Tweet class will use storeTweet to write tweets into index.

2. Result and Discussion

Result

The purpose of this system development is to analyze the tweets sent by Twitter users who are in the areas of Libya, which of course most of the tweets were written in Arabic language and writing. Therefore, in the analysis, the authors need library that support Arabic language text analysis and in Arabic writing. The analysis is meant here is how to determine the topics in tweets using the clustering method. Carrot2 Framework is a solution that can address this

problem, in which in addition to providing a variety of useful features to make text clustering, also support the analysis of text written in Arabic language and writing.

To perform tweets analysis, there are classes that will be used which have specific function. These classes are as follows:

1) tweetAnalyst.java

This class is a class that inherits properties and behaviors of JFrame class in Java. The JFrame class is used as a top-level container for the graphical components of a graphical user interface (GUI). Figure 4 shows the GUI's display of tweetAnalyst.java.

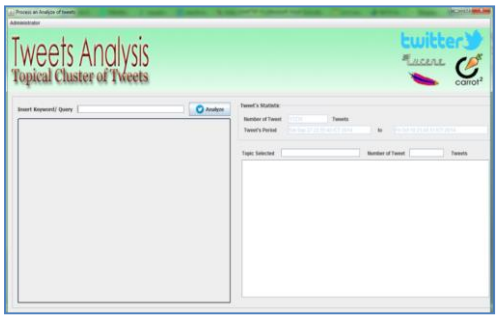

Figure 4 GUI's display of tweetAnalyst.java

2) CLusteringTweet.java

CLusteringTweet.java is the most important class in the clustering process to generate topics of conversation in the tweets from the search results. Search results transmitted by tweetAnalyst.java then will be stored into variable called searchRes. It is would then be converted into a List of the Document called documents. Document is a class that included in Carrot2 Framework which is also a type of data, which serves to store the data sources to analyzed.

3) clusterDisplay.java

This class used to produce TreeView of clusters. To produce TreeView, this class has a method called topicalDisplay() that will return value with type JTree. JTree is a GUI component of java to display list in a TreeView. The ProcessingResult that sent via parameter then stored in a variable called pr. First on the top of this class, a Collection of Cluster called clusters has defined. Statement "clusters = pr.getClusters();" used to access and store all clusters in variable pr into variable clusters. After all cluster stored, then there will a looping to determine every cluster as node of tree. So that the TreeView will display the label formed cluster, and the cluster labels are generated topics.

4) statisticGraphic.java

In these developed applications, in addition to the analysis results are displayed in the TreeView, is also presented in chart form. In the implementation, statisticGraphic was extending ApplicationFrame class. This mean that statisticGraphic will inherits properties and behavior of ApplicationFrame class. ProcessingResult parameter than will be processed to produce dataset. A dataset is a data source to create a chart. Figure 5 shows a pie chart of analysis result based on the specific query entered by user.

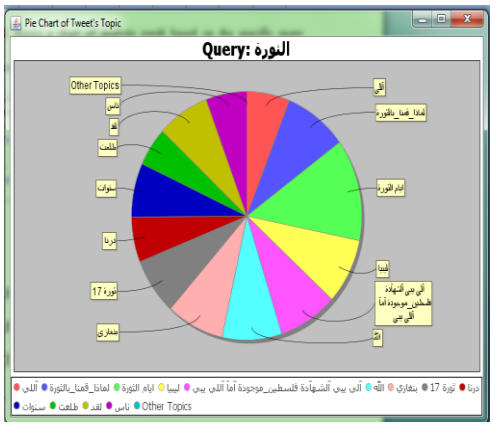

Figure 5 Pie chart

5) ApplicationFrame.java

ApplicationFrame.java is a class to create frame to display the chart generated. This class extends JFrame class of java, a GUI component to create frame form.

Discussion

a) Analysis of the Result

Analysis of result will be conducted by analyze several topics resulted by system based on specific query and compare the result by consider the real condition during the period of accessing tweets from twitter. Table 4.1 presents tweets access information about tweets that has been stored in index.

Table 2 Tweets access information

Number of tweets	17.289 tweets
Start date	27th September 2014
End Date	14th October 2014

Here is an analysis of the topics generated based on specific queries:

1) Query : "قبلية اشتباكات" or "Tribal Clashes"

This query resulted 17 clusters or topics. To view cluster/ detail, just click on the node of TreeView, and the detail will be displayed on right side of the form as shown in Figure 6.

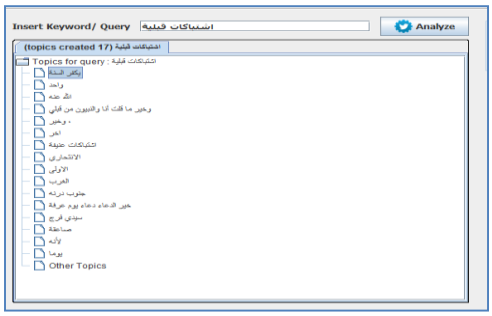

Figure 6 TreeView of Topical Cluster with query "قبلية اشتباكات"

2) Query : "الأضحى عيد" or "Eid Al-Adha"

There are a lot of tweets that talked about this. Using this query, system resulted 53 topics as shown in Figure 7.

Based on the discussion of some topics generated from the queries above, it can be concluded that most of the topics that are produced can be said to be relevant to the query is entered. However, there are some topics that felt irrelevant and has no meaning. See Figure 4.18 for example. Topic "RayanBanoun" when we click on the node of TreeView, can see the detail of this topic as shown in Figure 4.19. This topic is

Figure 7 TreeView of Topical Cluster with query " **الأضحى عيد** "

generated because the tweets were found contained the word "RayanBanoun" or "RayanBanoun". Character "@" will be eliminated by Carrot2 Framework when passing through the preprocessing stage. Actually RayanBanoun is the profile name of the twitter user. And if we look carefully, we can see that twitter user with the profile name "RayanBanoun" is mentioned several times by other twitter users.

Other examples can also be seen in Figure 4.18. In the TreeView there is a topic "هههههههه" which if translated into English as "Hahahaha". It is an expression of laughter. And of course not relevant when used as a topic. Just like the topic "RayanHanoun", this topic generated because the

Table 3 Summary of testing result

No	Query (Arabic)	English	Number of topic	Number of relevant topic	Number of irrelevant topic
1	قبلية اشتباكات	Tribal Clashes	17	17	0
2	الأضحى عيد	Eid Al-Adha	53	50	3
3	الشباب حركة	Youth Movement	18	16	2
4	المعلومات تقنية	Information Technology	10	9	1
5	المتحدة الامم	United Nation	7	7	0
6	الابن القذافي	Gaddafi's Son	22	20	2
7	الممتاز الدوري	Premier League	20	19	1
8	طرابلس	Tripoli	29	27	2
9	والثقافة التعليم	Education and Culture	8	8	0
10	الاجتماعية الاعلام وسائل	Social Media	13	12	1
TOTAL			197	185	12

tweets are grouped in this topic contains the word "هههههههه".

Framework Carrot2 actually define the words that are considered as the stopword to be removed when passing through the preprocessing stage. But words like "هههههههه" of course not included as a stopword. For those reasons, the resulting system still needs to be improved so that the preprocessing stage is necessary to add another stage to filter out words that are not included in the stopword defined by Carrot2 Framework. For example, when a user in a tweet mentioning the profile name of the other user, it should be called the profile name be removed when passing through the stages of preprocessing. So that is stored in the index should just tweet the contents of the user.

b) Accuracy Testing

Taking into account the discussion conducted in the segment analysis of results above, this segment will be measured on how the accuracy of the system in generating relevant topics. Measurement of accuracy is done by assessing each topic generated by several queries and then determines whether the topic is relevant or not.

$$Accuracy = \frac{number\ of\ relevant\ topics}{number\ of\ overall\ topics} x\ 100$$

$$= \frac{185}{197} x100 = 93.91\%$$

Accuracy percentage will be calculated based on the ratio between the numbers of topics relevant to the numbers of topics that are not relevant.

Table 3 presents list of queries that will be used to measure the accuracy of the system also the summary of testing result by each query.

From Table 4.3 showed that the overall number of topics is 197 topics, where 185 topics are considered relevant and 12 irrelevant topics. From this data, the level of accuracy can be calculated as follows:

So it can be concluded that the accuracy of the system in analyzing tweets and generating relevant topics is 93.91%.

3. Conclusion

Social media no doubt has brought enormous influence in many aspects of life.

At the time of the revolution in Libya, social media such as twitter and other has become effective media for interaction and communication in performing the movement. Even at this point too, passed a few years after the revolution, social media such as twitter has become the Libyan's lifestyles. Libyan people use twitter for various reasons, such as lifestyle, as a medium to communicate and interact, as a medium to

express ideas and concepts, as the media to criticize government policies, etc.

Based on the above facts, in this study developed a system or application that can be used to analyze tweets or twitter the data sent by the Libyan people. The result of this analysis is in the form of topics being discussed by the Libyan community on twitter by a certain issue. By knowing the various topics of conversation on a particular issue that is going on, then this will be a boon for those who need this kind of analysis, such as government, non-governmental organizations and others.

In the stages of analysis and system design, object-oriented approach used utilizes the Unified Modeling Language (UML). At the analysis stage, is the determination of the architecture of the system to be built and analyzed what the functional requirements that must be met by the system. Furthermore, based on the results of the analysis, the system design was developed. The system design is done by modeling the system into UML diagrams, such as use case diagrams, activity diagrams, sequence diagrams, and class diagrams.

The process of developing this system utilizes libraries provided by the open source libraries such as Carrot2 Framework, Apache Lucene, and Twitter4J. Twitter4J useful in accessing tweets from twitter. While Apache Lucene is useful in the process of tweets indexing and tweets retrieval based on user's query. And Carrot2 Framework tweets used to process search results clustering, where clustering is done using the Lingo algorithm provided by this framework. Both Apache Lucene and Carrot2 Framework already support the processing and analysis of text in Arabic language and writing. This is the reason of the use of Apache Lucene and Carrot2 Framework. As we know that most of the Libyan use Arabic language and writing.

The resulting system was then tested by entering certain queries and then performed an analysis of the topics generated. It can be concluded that the topics generated still found irrelevant topics. This is because there are certain words contained in the tweets that are not included in the Arabic stopword list, even though the words are in fact can be considered as a stopword. So the system needs to be developed further so that the preprocessing stage, these words can be eliminated.

Considering the above, then in the discussion phase is necessary to test the accuracy of the system by measuring the level of relevance of the topics generated. Tests conducted on topics generated by using 10 queries. With 10 queries, system generates 197 topics, which 185 topics are relevant while the 12 topics considered irrelevant. So the accuracy of the system can be calculated and the results obtained 93.91. It can be concluded that the accuracy of the system in generating relevant topics is 93.91%.

References

[1] M. McHugh. (2011). "Libya Inspired By Egyptian Revolution, Uses Social Media in Midst of Protests". http://www.digitaltrends.com/social-media/libya-inspired-by-egyptian-revolution-uses-social-media-in-midst-of-protests/#!4SWD7.

[2] S. Osinski & D. Weiss. (2004). "Conceptual Clustering using LINGO Algorithm: Evaluation on Open Directory Project Data".

http://www.cs.put.poznan.pl/dweiss/site/publications/download/iipwm-osinski-weiss-2004-lingoeval.pdf

[3] R. Mahalakshmi & V. L. Praba. (2013). "A Relative Study on Search Results Clustering Algorithms - K-means, Suffix Tree and LINGO". International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-6, August 2013.

[4] K.D. Rosa, R. Shah,B. Lin, A. Gershman & R. Frederking. (2011). "Topical Clustering of Tweets". http://www.cs.cmu.edu/~encore/sigir_swsm2011.pdf

[5] Anonymous. (2011). "Clustering Users in Twitter Based on Interests". http://www.nlpr.ia.ac.cn/2011papers/gnhy/nh4.pdf.

[6] M.A. H. Khan, G. Liu, D. Bollegala and K. Sezaki. (2013). "Delineating Real-Time Events by Identifying Relevant Tweets with Popular Discussion Points". http://arnetminer.org/doc/paper%206.pdf.