Open Access Library Journal
2019, Volume 6, e5893
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Research on Political Trend of North Korea Based on Big Data Text Mining Method

Hongyi Li[1], Zhezhi Jin[2*]

[1]Department of Mathematics, Yanbian University, Yanji, China
[2]Department of Economics and Management, Yanbian University, Yanji, China
Email: *780831601@qq.com

How to cite this paper: Li, H.Y. and Jin, Z.Z. (2019) Research on Political Trend of North Korea Based on Big Data Text Mining Method. *Open Access Library Journal*, **6**: e5893.
https://doi.org/10.4236/oalib.1105893

Received: October 31, 2019
Accepted: November 24, 2019
Published: November 27, 2019

Abstract

With the method of text mining, this paper takes the related data of Rodong Sinmun in the recent ten years as the research object, extracts the hot topics and carries on the trend analysis. With the special attribute of his speech media, this paper analyzes the political issue. By extracting nearly one million news text data, their topic content is analyzed, combining with LDA topic model, and using K-means clustering algorithm. Aiming at the limitations of the traditional K-means algorithm, it is solved on the pre-built big data analysis platform, and the structure and content of the theme are analyzed in detail. In the end, the political theme and the trend of public opinion in recent years are derived. In terms of application, it is of great significance to analyze and study Korean big data text.

Subject Areas

Statistics

Keywords

Text Mining, K-Means Algorithm, LDA Topic Model, Public Opinion Analysis

1. Introduction

With the development of network technology and the advancement of various technical devices, as well as the increase in online media content, we have created more opportunities to collect public and media data. Therefore, the next data that travel through media services, news, and other media have become a new source of deep public heart. And with the increase of data and the development of text mining technology, the importance [1] of public opinion analysis technology to extract the main contents and core themes contained in the text also begins to emerge.

Public opinion analysis is mainly used by enterprises or various departments in various fields for the purpose of collecting public opinions on social events or political focus, corporate strategy or marketing, and product preference, and making decisions. Therefore, the existing research on public opinion analysis is mainly based on commodity reviews or film reviews, which take opinions as emotional language. It focuses on the method of judging the extremes of emotions and the method of constructing the emotional language dictionary [2] [3].

Different from the above, the analysis is not based on non-words or articles, but the themes extracted from the text. In a large amount of datum that can be used for timing analysis, it's considered that news materials composed of a certain amount of text in each time quantum are suitable, so it is selected as research data. Then, the sentiment of media revealed in the news report is regarded as an opinion and analyzed as one of the viewpoints.

To this end, reports related to politics were collected from nearly a decade of North Korean Rodong news data. In terms of text clustering, Dhillon [4] used the K-means algorithm and used cosine similarity to calculate the distance between texts. Text has unstructured and high-dimensional features, and in order to extract topics from each text, the topic modeling method is used. The topic modeling algorithm is a statistical method that analyzes a large number of vocabularies used in text data, discovers their themes, and analyzes how these topics relate to each other and how they change over time.

In order to analyze all of these extracted topics, it is divided into three steps. Firstly, the wordcloud visualization method is used to sort the core topics extracted from the word frequency, and the frequency of occurrence of each topic in the text and the importance in each text can be visualized and compared. Secondly, using the hierarchical clustering method to analyze the correlation of each core topic in the text data, even if the content of the core topic in some texts is similar, the correlation differences among the various topics can also be seen intuitively according to the hierarchical clustering method. Finally, in view of the limitations [5] [6] [7] of traditional K-means algorithm, the literature [8] [9] [10] combines K-means algorithm with LDA topic model, and based on the built large data analysis platform, makes more accurate and detailed analysis of the theme in text data. There will be some different phenomena appearing for different time distributions of common themes. It is proved by the method of time series analysis whether the subject-based analysis system is feasible.

2. Related Theory

2.1. K-Means Algorithm

The core idea of K-means algorithm is to divide the data objects into different clusters through iteration to minimize the objective function, so that the generated clusters are as compact and independent as possible. It is very important to measure the distance between data objects in the process of clustering. The euclidian distance is generally used, and other can also be used for measurement.

The method is as follows:

1) Determine the value of K. (K is the number of clusters.)

2) Randomly select the centroid $C=\left\{c^{(1)},c^{(2)},\cdots,c^{(K)}\right\}$.

3) For each object, calculate the distance from each centroid, and assign each individual to the cluster centroid closest to them, so as to determine the class membership of all objects.

4) Calculate the average of all individuals assigned to the same centroid $\mu=\left\{m^{(1)},m^{(2)},\cdots,m^{(K)}\right\}$.

5) Set the average value as the new centroid
$C=\left\{c^{(1)},c^{(2)},\cdots,c^{(K)}\right\}\leftarrow\mu=\left\{m^{(1)},m^{(2)},\cdots,m^{(K)}\right\}$.

Repeat the above three stages until the individual no longer changes the cluster to which he belongs, that is, the relationship between the clusters does not change.

If expressed mathematically, the cluster based on the centroid can be expressed as $C_j=\{i|i\in\{1,\cdots,n\}$ *s.t. the closest centroid from* $x^{(i)}$ *is* $c^{(j)}\}$, where C_j is a set of all indicators, and the cost function based on the centroid is expressed as follows:

$$cost\left(C_1,C_2,\cdots,C_K,c^{(1)},c^{(2)},\cdots,c^{(K)}\right)=\sum_{j=1,\cdots,K}\sum_{i\in C_j}\left\|x^{(i)}-c^{(j)}\right\| \tag{1}$$

Faced with a large amount of data, the advantages [5] of K-means algorithm are still obvious. As one of the most classical algorithms. Firstly, the algorithm is simple and easy to implement and debug. Secondly, the functions of the algorithm are very intuitive and can optimize intra-cluster similarity. But the biggest disadvantage of K-means algorithm is that we need to specify the number of clusters K that we expect to generate. It has limitations in some applications, and it is also true in text mining. Because it is impossible to know the value of the optimization K, the optimal clustering effect cannot be obtained in the process of text data analysis.

2.2. LDA Topic Model

LDA (Latent Dirichlet Allocation) is a generative probabilistic model for text data. As a kind of theme model, LDA model not only has the simple features of the model, but also can reduce the level of data, and has the advantage of being able to produce a consistent theme in the sense of inclusion. And it is one of the most useful models for text analysis.

In the model, to generate a document, firstly, select $N\sim Poisson\left(\xi\right)$ and $\theta\sim Dir\left(\alpha\right)$, then select a topic $z_n\sim Multinomial\left(\theta\right)$ for each w_n of N word, and then select a word w_n *from* $p(w_n|z_n,\beta)$*, a multinomial probability conditioned on the topic* z_n*.*

Among them, α and β are values established in units of corpus, and N and θ

are values determined in units of text. θ represents the probability of generating a specific word in each theme (two-dimensional matrix), N is the length of the text, θ is the weight value of each topic on the related question, and z_i is the topic vector for the i-th word in the document. In this model, the number of topics is fixed at k, and θ and z_i are vectors of length k. Simply put, it is the parameter of a certain text. Whenever a word is filled from the front, a topic will be selected from θ, and then the generated document will be modeled by selecting the word from the theme.

For the speculation of the parameters, the distribution of Dirichlet should be explained. If the content of the above description is expressed in the form of a mathematical formula, it will be as follows:

$$p\left(z_1,\cdots,z_N\right)=\int p\left(\theta\right)\left(\prod_{n=1}^{N}p\left(z_n\middle|\theta\right)\right)\mathrm{d}\theta \tag{2}$$

$$p\left(w,z\right)=\int p\left(\theta\right)\left(\prod_{n=1}^{N}p\left(z_n\middle|\theta\right)p\left(w_n\middle|z_n\right)\right)\mathrm{d}\theta \tag{3}$$

The former is the subject of generating text, the latter is the subject of the text and generation of words, and the z representing the subject or content of the text is the conditional probability of θ.

$$\underbrace{p\left(\theta\middle|data\right)}_{posterior}\propto\underbrace{\ell\left(data\middle|\theta\right)}_{likelihood}\underbrace{\tilde{p}\left(\theta\right)}_{prior} \tag{4}$$

The goal is to infer θ from the text content. If you know how the *posterior* (Posterior probability) occurred in the above equations are distributed, this work is easier. And the necessity of conjugate prior arises. In Equation (4), if *posterior* and *prior* (Prior probability) follow the same distribution, *prior* will become the conjugate prior of *likelihood*, and *likelihood* is the multiple distribution of θ. Dirichlet distribution is the conjugate prior of the multi-distribution.

We can think that the LDA topic model is to some extent simulate the subject classification of the human brain for the corpus. The category construction process carried out by the researcher in the content analysis process. The model assumes that the words selected by the author in each article are through the following process: a topic is selected with a certain probability and a certain word is selected from the topic with a certain probability. So we do LDA modeling for corpus, which is to dig out different themes from the corpus and analyze them, In other words, LDA provides a convenient way to quantify the research topic. In view of the limitations of the above K-means algorithm, the combination [7] [8] [9] of LDA topic model and K-means algorithm is used to optimize the clustering effect in the processing process of large text data.

3. Specific Application of Text Mining

The next step is to analyze the political data of north Koran Rodong news on Transwarp big data platform, the data are based on political news data from 2009 to 2019.

However, in the process of public opinion analyses, if the news data of nearly ten years are combined and analyzed, it is impossible to analyze its theme and trend changes, and it has no practical significance. Therefore, in order to find out the distribution of its theme and trend, the method of the year-by-year analysis and integration of the discussion can more accurately determine the change of its trend. Due to space limitations, it is impossible to display the results of each year. However, in order to show its accuracy and representativeness, the data of the years before and after it were selected and analyzed by using 2012 as the dividing line. As for why the news data of 2012 was chosen, because it is a political analysis of the public opinion, 2012 is the upper year of Kim Jong-un. Compared with other years, after analysis, the trend can be seen more clearly and accurately.

3.1. Wordcloud Visualization

As one of the more common expressions in visualization technology, wordcloud can display the hot topics and keywords in data very intuitively. Based on the principle of frequency statistics, wordcloud can quickly and easily understand the central ideas and keywords that text data wants to express, and does not require a lot of time and effort to read all the data. The R language used in this analysis, and the Wordcloud package is used.

Analyze the text data of each year from 2009 to 2019, and sort out the results based on the frequency of words, and arrange them in descending order. The words obtained are "People", "the United States", "Revolution", "Construction", "Unity", "Kim Jong il", "Peace", "Development", "Kim Jong-un", "Park Geun-hye". These words and the most representative and appear more frequently. And select the data of three representative years, using the wordcloud to visualize. You can visually see the trend of each core word, the results are shown in Figure 1.

Figure 1 is the result of the data from 2008, 2012, and 2017. It can be intuitively seen that the core words with the highest frequency of occurrence in each year are different. It can be analyzed that the news of each year is very obvious. Among the 2008 news data, the words such as "Leader", "Great", "People", "Revolution", and "Kim Jong il". The words with higher frequency in 2012 are "People", "Revolution", "the United States", "Kim Jung-un", "Construction". In 2017, the core words with high frequency are "Park Guen-hye", "the United

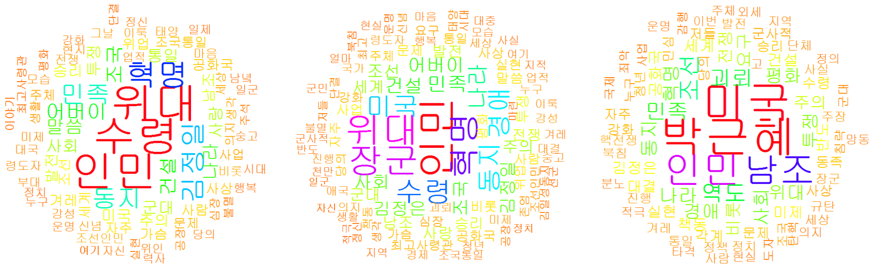

Figure 1. Wordcloud map of North Korea political data.

States", "South Korea", and "Puppet".

Judging from the core words that have emerged in the past three years, we can also intuitively understand the changes in the focus and trends of the news In 2008, when Kim-Jong il was in power. According to the core words appearing in the worldcloud map, it can be seen that in the news content of the year, the praises of Kim Jong il and the related content of the beautiful words are the vast majority. In 2012, with the rise of Kim Jong-en, the new leader of North Korea, the frequency of his name has grown tremendously, and he has become more and more important about the revolution and construction. It can also be seen between North Korea and the United States. The intersection is also getting more and more frequent. In 2017, according to the wordcloud map, it can be clearly seen that Park Geun-hye is the most frequently seen one, so that it can be known that since the South Korean President Park Guen-hye wa impeached in 2017. Most of the content of the North Korean news reports were that Park Guen-hye was impeached. From other core words, the North Korean news has a critical attitude toward South Korea and Park Guen-hye's party. Therefore, according to the method of wordcloud visualization. It is also possible to intuitively understand the hotspots of the text data and its changing trends.

3.2. Hierarchical Clustering

However, it is not enough to only understand the hotspots and core words of the data. In order to conduct more in-depth analysis, it is necessary to analyze the correlation between the core words in the news data and explore the relationship between them. It is not enough to rely on wordcloud, but to use the method of text clustering. Among them, the first step is to use the hierarchical clustering method.

Hierarchical clustering is same as the previously described K-means clustering. And it is much simpler. When only simple text clustering is performed, the effect of hierarchical clustering is very obvious. Because clustering is based on the principle of distance metrics, the relationship between words in text data can be seen very intuitively. The maximum number of words displayed in each hierarchical clustering map can be set by itself. This time, using the news data of 2012 and 2017 for hierarchical clustering, and set the maximum number of words that can be displayed to 30. And analyze the relationship changes between the core words that appear in each year.

As shown in Figure 2, the news data of two representative years are selected for clustering. The left figure shows the hierarchical clustering chart of 2012 data, and the right figure shows the clustering chart of 2017. From the text clustering map of 2012, the most important core words are "Kim Jong-un", "the United States", "People", "Construction", and "Revolution". Moreover, it can be seen from the cluster diagram that the emerging words such as "Youth", "Society", "System", "Army", "Victory" are words with increasing frequency.

Although the wordcloud map is more intuitive, it is impossible to see all the

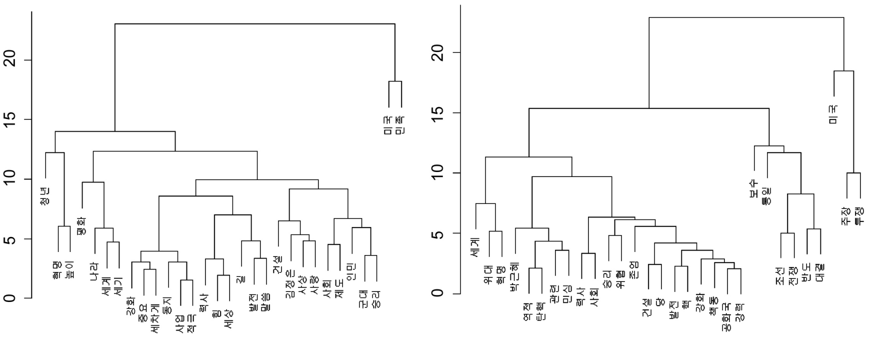

Figure 2. Hierarchical cluster map of North Korea political data.

core words. This is also the limitation of subjection. However, in the hierarchical clustering diagram, its effect is more obvious. You can see that the closer they are in each branch, the closer the distance is, the stronger the correlation between them is. For example, in the left figure, “Military” and “Victory”, “Peace” and “Country”, “Youth” and “Revolution”, “Society” and “Institution” are all strongly related words. In the cluster map of 2017, the words with the highest frequency are “Park Geun-hye”, “Conservative”, “Unified”, “impeachment” and other words. It can be seen that the main content of the news report in 2017 is the impeachment of Park Geun-hye. From the perspective of clustering, the correlation of clustering, the correlation between “Park Geun-hye” and “Impeachment”, “History” and “Society”, “the United States” and “Struggle” are relatively strong. It is also possible to analyze the correlation between hotspots and core words in 2017 news reports.

3.3. K-Means Clustering Based on LDA Model

In the end, in order to make a more accurate and detailed analysis of the news text data, it is necessary to use K-means clustering algorithm. In the above content, K-means algorithm is described and introduced in detail, but as mentioned above, K-means algorithm has limitations. So it is necessary to combine the LDA topic model to solve the problem that the optimal cluster number cannot be formulated. Compared with the traditional K-means clustering algorithm, the subject content of news data is more accurately judged and analyzed, thus obtaining better clustering effect.

In order to verify the effect, the 2017 news data was processed and analyzed, and it was displayed through LDAvis visualization. The results are shown in **Figure 3**.

LDAvis as an interactive analysis method may not be able to show its advantages better through the above diagram. But it can also be seen very intuitively that the analysis results are very outstanding. Through the LDA modeling method, the optimal cluster number is 10. And combined with the K-means clustering algorithm, the clustering results are shown above.

The detailed analysis results are sorted out and displayed in the form of a text table. The results are shown in **Table 1**.

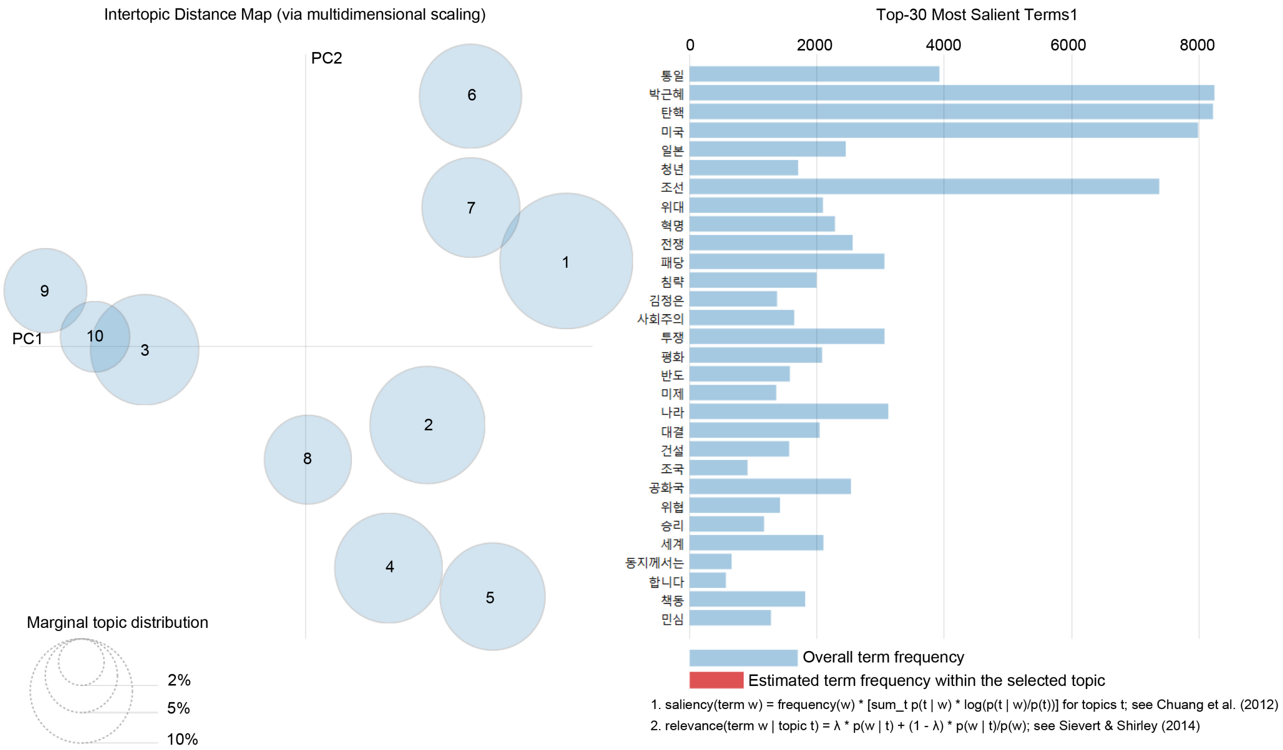

Figure 3. The LDAvis of North Korea political data (2017).

Table 1. LDA topic analysis results.

Core Theme	Related Words
Park Geun-hye	Impeachment, Party, Confrontation, Conservative, Republic
US	Japan, Country, World, Instigation, War, Aggression
Great	Revolution, Socialism, Victory, Kim Jong-un, Construction
Republic	Peace, Threat, Army, War, Peninsula, UN, Autonomy
North Korea	War, US, Emperor, Nuclear War, THAAD, Aggression
Impeachment	Park Geun-hye, Struggle, North Korea, Advocate, Society
Struggle	Impeachment, Park Geun-hye, Society, History, Conservative
Unity	Motherland, Peace, North-South Relationship, History, Republic
Youth	Revolution, Construction, Socialism, Kim Il Sung, Intensive
Kim Jong-en	North Korea, Leader, Construction, Korean Workers Party, Chairman

As seen from the above table, the number of optimized clusters set before is 10. Among the results, there are 10 core topics. Because the *K* value in K-means algorithm, that is, the optimal number of clusters, in the process of text analysis, represents the number of core topics in the text data. By combining the LDA topic model with K-means algorithm and using the LDAvis method, the text data can be more accurately screened and analyzed. So that the clustering effect can be further improved.

4. Conclusions

This paper makes use of the text mining method, based on the political related data of North Korean Rodong news on the big data analysis platform, screens the core themes and carries out detailed public opinion analysis. Firstly, wordcloud visualization is carried out on text data. As one of the most intuitive visualization methods, the core words in the text data are screened by wordcloud visualization method, and the core content and ideas of the text are analyzed. Then, hierarchical clustering method is used to analyze the correlation between the core words in the text data and the changes of the relationship between the core words appearing each year. Finally, combining the LDA topic model with K-mean algorithm not only solves the problems of the traditional K-means algorithm, but also exploits the advantages of the K-means algorithm. Through LDAvis method, more detailed and accurate text mining analysis is carried out for text data. Through these three methods, this paper obtained the nature and distribution change of the theme in the text, and analyzed the position and viewpoint of the text producer on the issue, as well as the position of the media and the public on the topic of the hot content of public opinion.

Through a study combining exploration with application, this paper wants to find out whether text mining based on topic modeling technology has reasonable public opinion analysis function. From this point of view, it is different from traditional text mining research. And from the perspective of demonstrating whether the subject-based public opinion analysis program is feasible, the answer is yes. In addition, it is considered that the topic model algorithm is simply used to extract topics from a large amount of text data. It is important to refine the aspects of individual subject analysis functions by analyzing the structure and content of individual topics. By analyzing the structure and content of individual topics, it is important to refine the aspects of individual subject analysis functions.

Funding

This work is supported by Jilin province education department "13th Five-Year" science and technology research project (Project Number: JJKH20191119KJ).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Likas, A., Vlassis, N.J. and Verbeek, J. (2003) The Global K-Means Clustering Algorithm. *Pattern Recognition*, **36**, 451-461. https://doi.org/10.1016/S0031-3203(02)00060-2

[2] Noh, Y., Kim, T., Jeong, D.-K. and Lee, K. (2019) Trend Analysis of Convergence Research Based on Social Big Data. *Journal of The Korea Contents Association*, **19**,

135-146.

[3] Kim, M., Koo, C. and Sohn, B. (2019) A Study on the Effectiveness of Education Welfare Priority Support Program through Text Mining. *Korean Journal of Youth Studies*, **26**, 313-332. https://doi.org/10.21509/KJYS.2019.02.26.2.313

[4] Dhillon, I.S. and Modha, D.S. (2001) Concept Decompositions for Large Sparse Text Data Using Clustering. *Machine Learning*, **42**, 143-175. https://doi.org/10.1023/A:1007612920971

[5] Juan, Z., Xiong, Z.Y. and Zhang, Y.F. (2006) Multi-Center Clustering Algorithm Based on Max-Min Distance Method. *Journal of Computer Applications*, **26**, 1425-1427.

[6] Wang, Y. and Tang, J. (2014) High-Efficiency K-Means Optimal Clustering Number Determination Algorithm. *Journal of Computer Applications*, **34**, 1331-1335.

[7] Sun, J.G., Liu, J. and Zhao, L.Y. (2008) Clustering Algorithms Research. *Journal of Software*, **19**, 48-61. https://doi.org/10.3724/SP.J.1001.2008.00048

[8] Wei, J. (2018) Research on Improved Algorithm Based on K-Means Clustering Algorithm. *Information and Communications*, **2018**, 14-15.

[9] An, J.Y., An, G.G. and Shi, Z.Q. (2015) An Improved K-Means Text Clustering Algorithm. *Transducer and Microsystem Technologies*, **34**, 130-133.

[10] Wang, C.L. and Zhang, J.X. (2014) Application of Improved K-Means Algorithm Based on LDA in Text Clustering. *Journal of Computer Applications*, **34**, 249-254.