Exploring Trends of Cancer Research Based on Topic Model

Mingliang Cui, Yanchun Liang*, Yuping Li, Renchu Guan*
College of Computer Science and Technology, Jilin University,
Changchun, 130012, China
{Yanchun Liang ycliang@jlu.edu.cn; Renchu Guan guanrenchu@jlu.edu.cn}

Abstract. Cancer research is of great importance in life science and medicine and attracts research funds of thousands of millions dollars each year. With the explosion of biomedical research papers, it becomes more and more necessary to show the research trend in this spotlight area. In this paper, to provide a straightforward research atlas for the top killer cancers, Latent Dirichlet Allocation (LDA) is performed on the massive quantities of biomedical literatures. Moreover, Gibbs Sampling is used to make assessment on the parameters of the LDA model. The proposed evaluation carried out under multiple conditions with different Ks (the number of topics) for the top five cancers in recent five years. Additionally, a biomedical topic model was generated with the LDA model and delicate analysis was performed on the basis of that in order to explore the trending topic in cancer research. It can help the biology and medicine doctors quickly catch the frontiers of the cancer study, improve and expand their research programme, especially in today's era of "big data".

Key words: Topic Model, LDA, Gibbs Sampling, Topic Analysis, Cancer Research Trend

1 Introduction

1.1 Significance

Cancer is of great threat to human health, scientists and Medicians have been continuously looking for effective treatment to conquer it, however, cancer research is a very challenging field in today's life science. Over the past 5 years cancer research has diverged enormously, partly based on the quickly development of biotechnology and bioinformatics. It is not easy to summarize recent trends for different cancers' study, and identify where the new findings are and therapies might come from [1]. Under this circumstance, we choose an appropriate machine learning method—topic model—to explore the trend in cancer research, specifically, the topics in cancer research papers.

The topic model is a probabilistic model of text mining appeared in recent years [2]. It is an algorithm that can discover the topic structure hidden in large-scale data. In topic model, the vocabulary items is visible while the topic structure hidden. In order

to reduce the dimensions of the feature vector space, texts are usually mapped into the topic space via the topic model. It is different from the traditional vector space model, which just simply considers each document as a sample and each word as a feature. Instead, it maps a high dimensional frequency space to a lower dimensional topic space. Moreover, the topic model can capture the semantic information, which can reveal that the latent relations among documents. It also can effectively solve the polysemy, synonym and other problems, which has the vital significance in document feature extraction and content analysis. However, using a topic model (i.e. Latent Dirichlet Allocation) to analyze the trends of cancers has not been reported.

The rest of the article is organized as follows: we start in section 2 with a brief review of a topic model named as Latent Dirichlet Allocation and its related work. In section 3, the algorithm of Gibbs Sampling is introduced, and the framework of our exploration is described in detail. It presents the experimental methodology and results on Medline dataset in Section 4. At last, conclusions and future work are depicted in Section 5.

2 Background

In 1998, Professor Papadimitriou proposes LSI (Latent Semantic Indexing) model [3], which can be seen as the origin of the Topic model. In 1999, Professor Hofmann put it forward to pLSI (Probabilistic Latent Semantic Indexing) model [4] and LDA (Latent Dirichlet Allocation) is a generalization of pLSI, which adds Dirichlet Bias generating prior distribution model, proposed by Blei in 2003[5].

LDA has been widely applied in information retrieval, text mining, Natural Language Processing fields and has become a research hotspots recently [6-8]. In this paper, LDA model is introduced to analyze the Medline data especially on cancer research. In the LDA model, a document is generated as follows [9]:

1. First, generate the topic distribution of the document *i* by sampling from the dirichlet distribution α.

2. Second, generate the topic of the word *j* in the document *i* by sampling from the topic distribution.

3. Then generate the word distribution of the topic by sampling from the Dirichlet distribution β.

4. Finally generate the word *j* in the document *i* by sampling from the word distribution.

Which is similar to the binomial distribution Beta distribution [10] is the conjugate prior probability distribution [11], while the Dirichlet distribution is a polynomial distributed conjugate prior probability distribution.

The structure diagram of LDA model is shown in the Figure.1 (similar to the Bayesian network structure [12]):

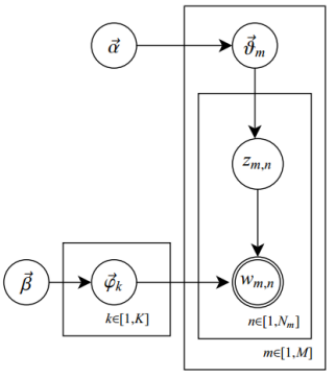

Fig. 1. LDA model structure diagram which visualize the generation process using the LDA

Hyper parameters are subject to Dirichlet distribution. We explain the symbol used to describe the model in Table.1. Now we have a text representation of the probability:

$$p(\vec{\mathrm{w}}_m, \vec{\mathrm{z}}_m, \vec{\theta}_m, \Phi \mid \vec{\alpha}, \vec{\beta}) = \prod_{n=1}^{N_m} p(\mathrm{w}_{m,n} \mid \vec{\varphi}_{z_{m,n}}) p(\mathrm{z}_{m,n} \mid \vec{\theta}_m) \cdot p(\vec{\theta}_m \mid \vec{\alpha}) \cdot p(\Phi \mid \vec{\beta}) \quad \textbf{(1)}$$

Table 1. A table of symbols used to describe the LDA model.

Symbol	Meaning
M	document collection
K	topic collection
V	term collection
N_m	length of the document *m*
α	dirichlet prior of the topic distribution, hyper parameter
β	dirichlet prior of the term distribution, hyper parameter
α_i	*i* component of α
β_i	*i* component of β
θ	topic distribution generated by $Dir(\alpha)$
θ_m	document m distribution $p(z\|d=m)$
φ_m	term distribution generated by $Dir(\beta)$
φ_k	term distribution of topic *k* $p(t\|z=k)$
z	topic number mapping to φ_z in topic collection K
$w_{m,n}$	word *n* of the document *m*
$z_{w_{m,n}}$	topic of word *n* of the document *m*
Φ	generated topic collection $\Phi = \{\varphi_k\}_{k=1}^{K}$
n_m	topic vector of document *m*
n_k	term vector of topic *k*
n_k^t	*t* component of n_k
n_m^k	*k* component of n_m
$n_{m,\neg i}^k$	ditto
$n_{k,\neg i}^t$	ditto
$\hat{L}$	the evaluation parameter of the model

The convergence of LDA is a more critical issue. We use the convergence function to solve it. Under the given model conditions, we choose the appearance probability of samples as the evaluation criterion of the model. The performance of the LDA model:

$$p(\vec{\mathrm{w}} \mid \mathrm{M}) = \prod_{m=1}^{M} p(\vec{\mathrm{w}}_m \mid \vec{\theta}_m, \Phi) \tag{2}$$

$$= \prod_{m=1}^{M} \prod_{n=1}^{N_m} \sum_{k=1}^{K} \left(\frac{n_k^t + \beta_t}{\sum_{t=1}^{V} n_k^t + \beta_t} \cdot \frac{n_m^k + \alpha_k}{\sum_{k=1}^{K} n_m^k + \alpha_k}\right) \cdot \tag{3}$$

For the convenience of the calculation, we use log transform to the equation (1), denoted as $\hat{L}$. Along with the iterative constantly, $\hat{L}$ is used to determine the model convergence [13].

$$\hat{L} = -\sum_{m=1}^{M} \sum_{n=1}^{N_m} \log_2 \left(\sum_{k=1}^{K} \left(\frac{n_k^t + \beta_t}{\sum_{t=1}^{V} n_k^t + \beta_t} \cdot \frac{n_m^k + \alpha_k}{\sum_{k=1}^{K} n_m^k + \alpha_k}\right)\right) \tag{4}$$

3 Main Work

3.1 Core Method

In fact, LDA is one of the probabilistic. Data are generally divided into two parts, visible variables and latent variables. It is believed in the topic model that the data is produced by the generation process, which defines the joint probability distribution of visible random variables and latent random variables. For many modern probabilistic models, including Bayesian statistics, the priori probability calculation is extremely difficult. So the core research objective of modern probability modelling is to do everything possible to obtain an approximate solution. Random sampling is a kind of methods for solving the approximate solution with good performance. This article describes a method commonly used sampling MCMC (Markov Chain Monte Carlo) [14] and Gibbs Sampling algorithm [15], Gibbs Sampling algorithm has been widely used in modern Bayesian analysis.

MCMC methods have Gibbs Sampling algorithm and Metropolis-Hastings (MH) [16] algorithm commonly used sampling methods. Gibbs Sampling algorithm is a special case of MH algorithm, Gibbs Sampling from a high-dimensional space are sampled separately for each dimension, and gradually get higher dimensional sampling points, making sampling difficult to reduce.

N-dimensional Gibbs Sampling:

1. Random initialization $\{ x_i : i = 1, 2, \cdots, n \}$

2. For $t = 0, 1, 2, \cdots$ loop in sampling

$X_1^{(t+1)} \sim p(\mathrm{x}_1 \mid \mathrm{x}_2^t, \mathrm{x}_3^t, \cdots, \mathrm{x}_n^t)$

$$\cdots$$
$$X_j^{(t+1)} \sim p(\mathrm{x}_j \mid \mathrm{x}_1^{(t+1)}, \cdots, \mathrm{x}_{j-1}^{(t+1)}, \mathrm{x}_{j+1}^{t}, \cdots, \mathrm{x}_n^{t})$$
$$\cdots$$
$$X_n^{(t+1)} \sim p(x_n \mid x_1^{t+1}, x_2^{t+1}, \cdots, x_{n-1}^{t+1})$$

3.2 Complete Process

We choose the top 5 cancer from the top 10 deadliest cancer published by the LiveScience [17], which are Breast Cancer, Lung Cancer, Pancreatic Cancer, Prostate Cancer and Colon Cancer. We search and download the research paper related to these cancers from NCBI PubMed from 2010 to 2014 separately [18].

Then we make a pre-process to extract the title and abstract for each paper and make some hyphens for the entity name for a better segmentation. After that we use WordNet [19] to stem the text in order to get an exact description for the topic. Before the document input into model, we will do the pre-processing for the document, thereby obtaining the document term matrix. It can be seen by term matrix that how many documents in corpus, how many word terms and how frequently each term appears in a document. The inputs of the model are the document collection M, the number of topics K, and the hyper parameters α, β.The topic of a number K we need to specify its value according to the experience, we want to take advantages of K solution, the need for repeated experiments, and then to carry on the value according to the different K value under the situation of convergence. After repeated experiments of LDA convergence on the data, we get a reasonable value of K=100. The hyper parameter α, β is the Dirichlet of the prior distribution. In fact they have smooth effect on data. Because there is no supervision information too much, we assign the hyper parameters an empirical value $\alpha_k = 0.5, \beta_t = 0.1$, tending to take symmetry value [20].

Take the dataset of Pancreatic Cancer in 2010 as an example, the number of the abstract is 2088, includes 54004 words. The number of topic K is set to 100, 200, 300, 400, 500, considered for model convergence condition.

According to the evaluation function of the convergence $\hat{L}$ mentioned in Section 2, which shows the cost of compression of the text using the model, the smaller the better. Figure.2 below respectively under different K values for the iterative convergence condition of iteration times of 400 and 1000.

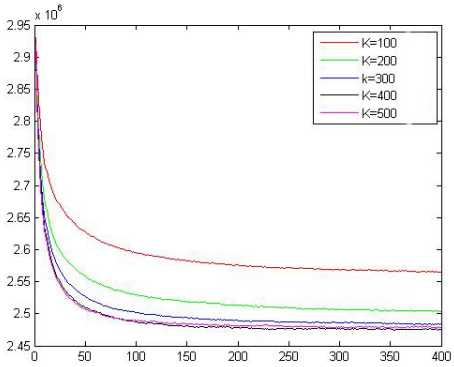

Fig. 2. The 400 iteration convergence condition of the LDA Gibbs Sampling model. X-axis is iterations, Y-axis is results of convergence function.

We further analysis the output of the LDA Gibbs Sampling, particularly in the topic words file (this file contains topic words most likely words of each topic). We first transform the topic-word matrix to a topic word vector representing the selection and the frequency of topic words to describe one cancer research. Then we use feature scaling normalization [21] (formula 5) to deal with the word vector, in order to compare them between different years and different cancers.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{5}$$

We measure the similarity and differences between the 5 cancer research by their word vector cosine coefficients similarity [22] (formula 6).

$$\text{simlarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} Ai \times Bi}{\sqrt{\sum_{i=1}^{n}(Ai)^2} \times \sqrt{\sum_{i=1}^{n}(Bi)^2}} \tag{6}$$

Unlike the cosine coefficients which give a numerical description on the trend and topic of the 5 cancers, the common words are easier for readers to have an intuition on the 5 cancers topics. We do the both analysis of the results so that we may get a comprehensive answer.

4 Experiments and Discussion

To examine the behavior and the performance of LDA, the experiments are illustrated on the widely used Medline dataset. In order to straightforwardly compare the trend and topic resulted from the LDA model, the visual results to get the entire recognition of the topic words of different cancers are shown, which uses the Word Cloud tool supported by Tagul.com [24]. And we also draw the trend of cancer research in different years, supported by Plot.ly [25].

4.1 Experimental Setup

The publicly available Medline dataset is provided by NCBI PubMed, which contains a title and an abstract for each paper on the five fatal cancers. In order to get a better visual understanding of the data, the 5 cancers with different colors (in details, breast cancer blue, colon cancer orange, lung cancer green, pancreatic cancer red, prostate cancer purple) are listed in Table 2. An outline of the selected dataset is shown below (actually we collect the 2014 data before October 2014, that is why it seems all the 2014 paper decrease from the previous years):

Table 2. A table reports the statistics of papers on the 5 cancer published from 2010 to 2014

Cancer	2010	2011	2012	2013	2014
Breast	11558	12810	13908	**14619**	11421
Colon	4484	4852	5386	**5498**	4305
Lung	7609	8530	9615	**10572**	8896
Pancreas	2088	2367	2667	**2860**	2468
Prostate	6365	7165	7759	**8137**	6787

4.2 Experiment Results

After introducing the Gibb Sampling and transforming the topic-word matrix to the vector, we draw all the five cancer topic words in the recent five years. A snapshot of the collection of the topic words clouds are shown in Figure.3.

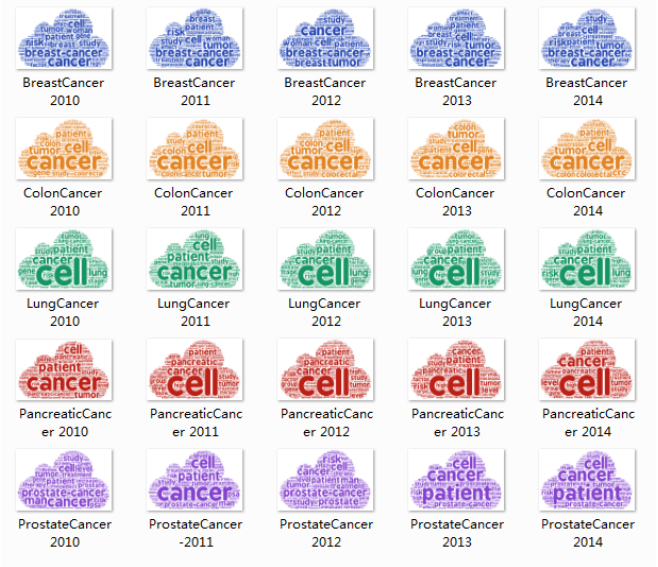

Fig. 3. A snapshot about the topic words for each cancer research for the 5 years. Each row is a type of cancer, and each column is a certain year. The 5 cancers are drawn with different colors. Also, the more frequently the word appears, the more striking the word displays.

Filtering the topic words for each cancer into the common words collection, we can easily discover the main topic for each cancer and new findings and focus for each year. Take breast cancer as an example shown below:

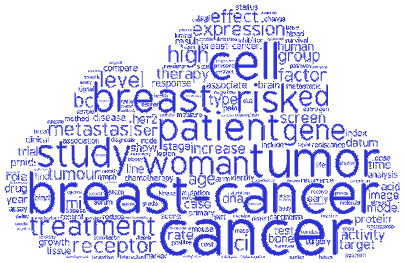

Fig. 4. The word cloud of the common topic words for breast cancer research during 2010-2014.

In the Figure.4 above, we can find out the big words in it like cell, breast cancer, tumor, patient, woman and so on, these words are simply the most common used to elaborate breast cancer, which means if you talk about breast cancer and you just can't avoid mentioning them. They are obviously from the different topics, therapy, factor, risk and other aspects of the breast cancer. The common words are defined as the words which appear every year for certain cancer research during 2010-2014. They are shown in Figure.5 below, and the shape of the words represent their frequency.

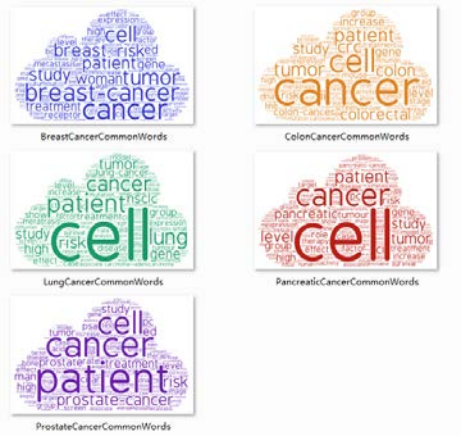

Fig. 5 The word cloud of the common topic words for the top 5 deadliest cancer research during 2010-2014.

4.3 General Comparison and Discussions

The Figure.6 above shows the contrast result of the vector cosine coefficient between two cancers in the same year. The higher coefficient, the more similar. We can easily find out that all the similarities contrast with colon cancer (see colon&lung, colon & pancreatic, breast&colon, colon&prostate) are higher than the other cancer pairs. It indicates colon cancer research is more related with the others. It is because that: lung cancer may easily metastasize to colon; colon cancer and pancreatic cancer are both belong to lower digestive cancers; lack of exercise and sitting for long time are the common causes of breast and colon cancer; long-term androgen deprivation therapy for prostate cancer may increase the risk of colon cancer.

And we find it interesting that almost all the five cancer research diverse from each other in 2012, because they have a concave at the time in the figure. Fig.7 shows the topic words changes for each cancer research during the recent five years. We can learn from the figure that breast cancer, lung cancer, prostate cancer, these 3 cancer research only change a little, and pancreatic cancer research changes a lot. It is because that the pancreatic incidence rate increase fast recently and it is named as “the king of cancer” because the lowest survival rate. Many scientists engage in the research of pancreatic cancer.

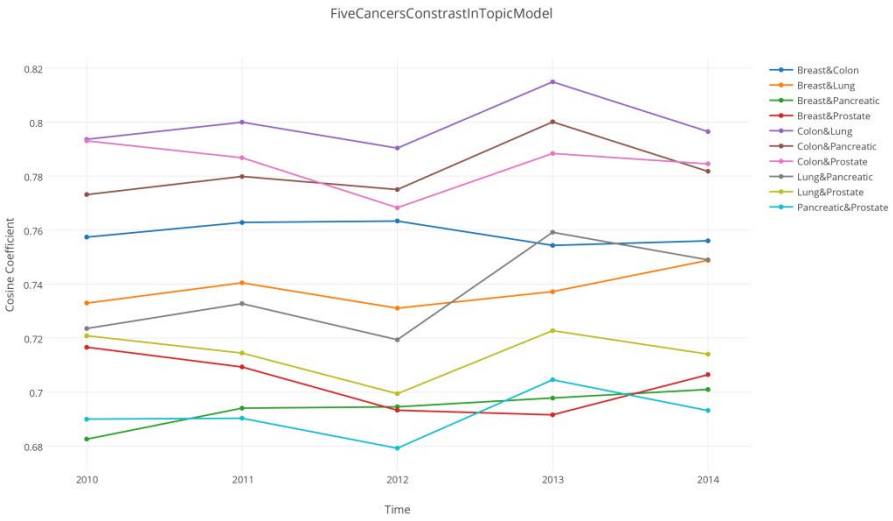

Fig. 6. All five cancers intercomparsion every year during 2010-2014. X-axis is year time, Y-axis is cosine coefficient.

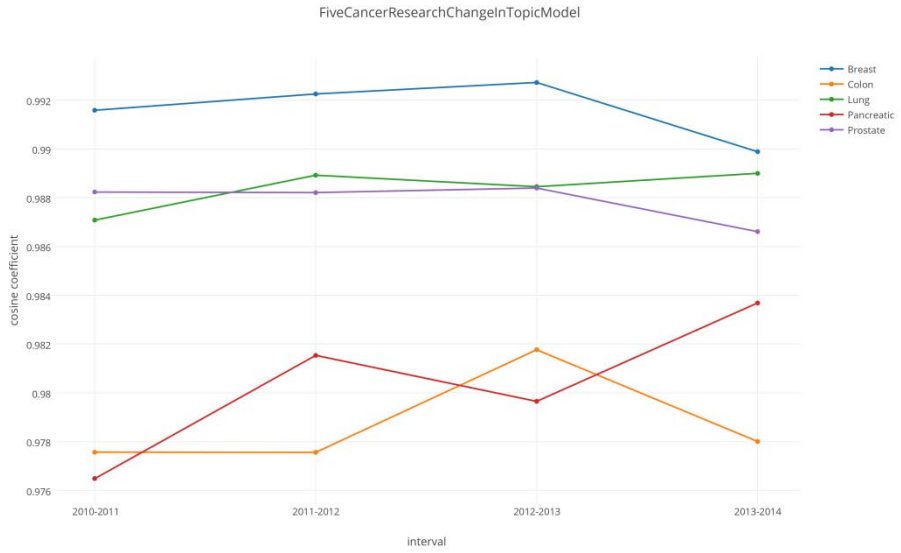

Fig.7. All five cancers change by calculating the topic vector similarity with their former year. X-axis is interval, Y-axis is cosine coefficient.

5 Conclusions

In this paper, we first apply LDA Gibbs Sampling model on the analysis of the top 5 deadliest cancer research trends, which is extended from cosine coefficient using the vector space model after transforming topic-word matrix into topic word vector. Then we generate the common topic word collection for each cancer research, in order to get the trending topic words. We further explore the trend by comparing and contrasting topic words for each cancer and their cosine coefficients. It is found that the trending topic words for the 5 cancers research from 2010 to 2014, which are depicted in the words clouds. Moreover, it is found that numerical trends of the four cancers research are as follows: breast cancer, lung cancer, and prostate cancer are of little change, but pancreatic cancer is changing a lot in the recent five years. But what cause all five cancer research diverse in 2012, how to visualize the allocation of the topics for each cancer research, and how to make a computational evaluation for the trend results, are all what need to be explored in our future work.

Acknowledgments.

The authors are grateful to the support of the NSFC (61272207, 61472158, 61103092) and the Science Technology Development Project from Jilin Province (20130522106JH, 20140520070JH）

References

1. Cao, Y., DePinho, R., Ernst, M., Vousden, K.: Cancer research: past, present and future. Nature Reviews Cancer. 11, 749–754 (2011).
2. Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P.: The author-topic model for authors and documents. Proceedings of the 20th conference on Uncertainty in artificial intelligence. 487-494 (2004).
3. Papadimitriou, C. H., Tamaki, H., Raghavan, P., & Vempala, S.: Latent semantic indexing: A probabilistic analysis. Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. 159-168 (1998).
4. Hofmann, T.: Probabilistic latent semantic indexing. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. 50-57 (1999).
5. Blei, D. M., Ng, A. Y., & Jordan, M. I.: Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022 (2003).
6. Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M.: Hierarchical dirichlet processes. Journal of the american statistical association 101.476 (2006).
7. Mcauliffe, Jon D., David M. Blei.: Supervised topic models. Advances in neural information processing systems. 121-128 (2008).
8. Petinot, Yves, Kathleen McKeown, and Kapil Thadani.: A hierarchical model of web summaries. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. 670-675. Association for Computational Linguistics (2011).
9. Diebolt, Jean, Christian P. Robert.: Estimation of finite mixture distributions through Bayesian sampling. Journal of the Royal Statistical Society. Series B (Methodological). 363-375 (1994).
10. S. Kotz, N. Balakrishnan, N. L. Johnson: Dirichlet and Inverted Dirichlet Distributions. Continuous Multivariate Distributions, Models and Applications. New York: Wiley, 2000., United States (2004).
11. Lindley, Dennis V.: The use of prior probability distributions in statistical inference and decision. Proc. 4th Berkeley Symp. on Math. Stat. and Prob. 453-468 (1961).
12. Beal, Matthew James.: Variational algorithms for approximate Bayesian inference. PhD diss. University of London (2003).
13. Heinrich, Gregor.: Parameter estimation for text analysis. Technical report. (2005).
14. Gilks, Walter R.: Markov chain monte carlo. John Wiley & Sons, Ltd. (2005).
15. Porteous, Ian, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, Max Welling.: Fast collapsed gibbs sampling for latent dirichlet allocation. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. 569-577 (2008).
16. Chib S., Greenberg E.: Understanding the metropolis-hastings algorithm. The american statistician. 49(4): 327-335 (1995).

17. Chan, A.: The 10 Deadliest Cancers and Why There's No Cure, http://www.livescience.com/11041-10-deadliest-cancers-cure.html.
18. Ncbi.nlm.nih.gov,: Home - PubMed - NCBI, http://www.ncbi.nlm.nih.gov/pubmed.
19. University, P.: About WordNet - WordNet - About WordNet, http://wordnet.princeton.edu/.
20. Griffiths, T., Steyvers: Finding scientific topics. Proceedings of the National Academy of Sciences. 101, 5228–5235 (2004).
21. Wikipedia,: Normalization, http://en.wikipedia.org/wiki/Normalization.
22. Wikipedia,: Cosine similarity, http://en.wikipedia.org/wiki/Cosine_similarity.
23. O'Rourke, Norm, R. Psych, Larry Hatcher.: A step-by-step approach to using SAS for factor analysis and structural equation modeling. Sas Institute. (2013).
24. Tagul.com,: Tagul - Gorgeous word clouds, http://tagul.com.
25. Plot.ly,: Plotly, https://plot.ly/.