A DOCUMENT EXPLORING SYSTEM ON LDA TOPIC MODEL FOR WIKIPEDIA ARTICLES

Zhou Tong[1] and Haiyi Zhang[2]

Jodrey School of Computer Science, Acadia University, Wolfville, NS, Canada

ABSTRACT

A Large number of digital text information is generated every day. Effectively searching, managing and exploring the text data has become a main task. In this paper, we first present an introduction to text mining and LDA topic model. Then we deeply explained how to apply LDA topic model to text corpus by doing experiments on Simple Wikipedia documents. The experiments include all necessary steps of data retrieving, pre-processing, fitting the model and an application of document exploring system. The result of the experiments shows LDA topic model working effectively on documents clustering and finding the similar documents. Furthermore, the document exploring system could be a useful research tool for students and researchers.

KEYWORDS

topic model, LDA, Wikipedia, exploring system

1. INTRODUCTION

As computers and Internet are widely used in almost every area, more and more information is digitized and stored online in the form of news, blogs, and social networks. Since the amount of the information is exploded to astronomic figures, searching and exploring the data has become the main problem. Our research is intended to design a new computational tool based on topic models using text mining techniques to organize, search and analyse the vast amounts of data, providing a better way understanding and finding the information.

2. BACKGROUND

2.1. Text Mining

Text mining is the process of deriving high-quality information from text [1]. Text mining usually involves the process of structuring the input text, finding patterns within the structured data, and finally evaluation and interpretation of the output. Typical text mining tasks include text categorization, text clustering, document summarization, keyword extraction and etc. In this research, statistical and machine learning techniques will be used to mine meaningful information and explore data analysis.

DOI : 10.5121/ijma.2016.8401

2.2. Topic Modelling

In machine learning and natural language processing, topic models are generative models, which provide a probabilistic framework [2]. Topic modelling methods are generally used for automatically organizing, understanding, searching, and summarizing large electronic archives.

The “topics” signifies the hidden, to be estimated, variable relations that link words in a vocabulary and their occurrence in documents. A document is seen as a mixture of topics. Topic models discover the hidden themes through out the collection and annotate the documents according to those themes. Each word is seen as drawn from one of those topics. Finally, A document coverage distribution of topics is generated and it provides a new way to explore the data on the perspective of topics.

2.3. Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar [3]. LDA has made a big impact in the fields of natural language processing and statistical machine learning and has quickly become one of the most popular probabilistic text modelling techniques in machine learning.

Intuitively in LDA, documents exhibit multiple topics [4]. In text pre-processing, we exclude punctuation and stop words (such as, "if", "the", or "on", which contain little topical content). Therefore, each document is regarded as a mixture of corpus-wide topics. A topic is a distribution over a fixed vocabulary. These topics are generated from the collection of documents [5]. For example, the sports topic has word "football", "hockey" with high probability and the computer topic has word "data", "network" with high probability. Then, a collection of documents has probability distribution over topics, where each word is regarded as drawn from one of those topics. With this document probability distribution over each topic, we will know how much each topic is involved in a document, meaning which topics a document is mainly talking about.

A graphical model for LDA is shown in Figure 1:

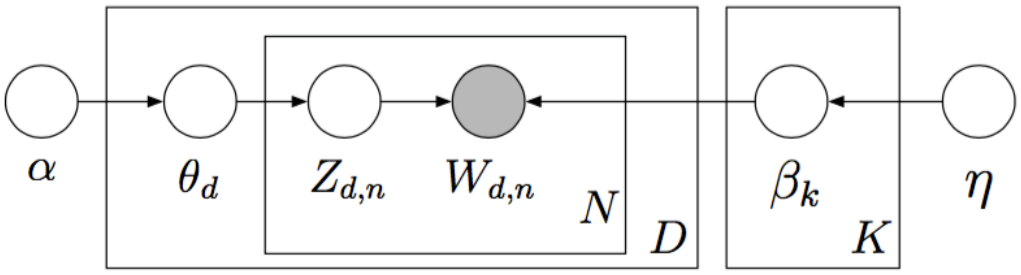

Figure 1. Graphic model for Latent Dirichlet allocation

As the figure illustrated, we can describe LDA more formally with the following notation. First, α and η are proportion parameter and topic parameter, respectively. The topics are $\beta_{1:K}$, where each β_k is a distribution over the vocabulary. The topic proportion for the d th document are θ_d, where $\theta_{d,k}$ is the topic proportion for topic k in document d. The topic assignments for the d th

document are Z_d, where $Z_{d,n}$ is the topic assignment for the nth word in document d. Finally, the observed words for document d are w_d, where $w_{d,n}$ is the nth word in document d, which is an element from the fixed vocabulary.

With this notation, the generative process for LDA corresponds to the following joint distribution of the hidden and observed variables [6]:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) (\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n})).$$

Notice that this distribution specifies a number of dependencies. The topic assignment $Z_{d,n}$ depends on the per-document topic distribution θ_d; and the word $w_{d,n}$ depends on all of the topics $\beta_{1:K}$ and the topic assignment $Z_{d,n}$.

2.4. Jensen-Shannon Divergence

In probability theory and statistics, the Jensen-Shannon divergence is a popular method of measuring the similarity between two probability distributions. It is also known as information radius or total divergence to the average. It is based on the Kullback-Leibler divergence. The square root of the Jensen-Shannon divergence is a metric often referred to as Jensen-Shannon distance [7].

For discrete probability distributions P and Q, Kullback-Leibler divergence of Q from P is defined to be:

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}.$$

So, the Jensen-Shannon divergence of Q from P is defined by:

$$JSD(P \parallel Q) = \frac{1}{2} D_{KL}(P \parallel M) + \frac{1}{2} D_{KL}(Q \parallel M),$$

where $M = \frac{1}{2}(P+Q)$.

Jensen-Shannon divergence measures the similarity between two distributions. By applying Jensen-Shannon divergence to the topic assignment for the dth document θ_d, it will allow us to measure the distance and similarity between each document.

3. MODEL TRAINING

In order to implement LDA topic model and design a document exploring system, we will train the LDA topic model, which consists of data retrieving, pre-processing and fitting the model. The experiment data is simple Wikipedia document collection. By applying these steps of experiment,

we can get a clear look at how LDA topic model works under practical circumstance, and the fitted model will be applied to implement a document exploring system.

3.1. Training Overview

In order to get better performance and reduce the processing time, the running environment of this research is the Elastic Compute Cloud (EC2) from Amazon Web Service (AWS). It allows users to rent virtual computers and provides many options to customize [8]. The coding environment is R programming language, which is widely used for statistical computing. R is an open source tool and is provided for various operating systems. One of the good features in R is the capability to extend by installing user-created packages. These packages largely enhance the functionality of R language [9]. In this experiment, I will use many R packages to process the data, such as **tm** and **topicmodels**.

The experiment basically consists of three parts - data retrieving, pre-processing, and model training. The first step is retrieving the data from Wikipedia, but the format can not be processed directly by the algorithm. We need a series of pre-processing steps to make the raw text data clean and usable. It includes parsing the text data, formatting to the right style, tokenization, stopwords removing, and stemming. The pre-processing can transfer the raw text data into a document-term matrix which is required by the bag of words assumption. Then we will apply the text data to fit the LDA topic model and the fitted model will be used for clustering, calculating similarity and assembled for a document exploring system.

3.2. Data Retrieving

Wikipedia is a free-access, free-content Internet encyclopaedia, supported by non-profit Wikimedia Foundation. It has millions of articles for people to search, explore or even edit. In this experiment, the text data is from simplified Wikipedia (English version) with over 200,000 articles. We downloaded the free backup XML file and parsed the document content via R package XML. Then a sample of 2000 simple Wikipedia articles is randomly selected as the experiment data.

3.3. Data Pre-processing

Data pre-processing, also called data cleaning, is one of the most important steps of this experiment. The purpose of pre-processing is to simplify the data, eliminating as much as possible language dependent factors. Articles are written in natural language for human to understand. But in text mining, those data are not always easy for computers to process. In this experiment, there are three steps in text cleaning [10]:

- Tokenization: a document is treated as a string, removing all the punctuations and then partitioned into a list of tokens.
- Removing stop words: stop words such as “the”, “if”, “and” ... are frequently occurring but no significant meanings which need to be removed.
- Stemming word: stemming word that converts different word form into similar canonical form. For example, computing to compute, happiness to happy. This process reduces the data redundancy and simplifies the later computation.

Besides those steps, for Wikipedia articles only, we also applied several steps to better clean the text data, including removal of URLs, attached files, XML labels and some special words like " " (as a space in mark-up language).

3.4. Fitting the Model

The training process requires R package **topicmodels** with its package dependencies (**tm** and others) to be loaded. An LDA model of simplified English Wikipedia on a sample of 2000 articles each with more than 100 characters, returned after 2000 iterations of Gibbs sampling, with $K = 50$ topics, and Dirichlet hyper-parameters $\beta = 0.1$ and $\alpha = 50 / K$.

The output of the fitted model are 50 topics and the per-document topic proportion θ_d. Here is a part of the results from the fitted model. Table 1 shows 5 of 50 topics after the model is trained, where top ten terms are listed for each topic. With LDA training, the terms in the same topic tend to be similar. Formally speaking, they are highly associated. For example, topic 2 is about music, topic 17 is about computer science and topic 50 is about language. These topics provide a way to search documents and explore between topics which is a new way finding the document especially for researching.

Table 1. A few selected topics generated from Wikipedia topic distribution

Topic 2	Topic 13	Topic 17	Topic 33	Topic 50
music	day	computability	planet	language
dance	january	internet	earth	english
compose	april	system	moon	word
instrument	march	window	star	use
vienna	december	program	sun	mean
classic	may	opera	system	alphabet
austria	july	web	galaxy	letter
beethoven	september	software	solar	verb
play	october	explore	ring	write
sound	novemeber	use	saturn	greek

Besides 50 topics, LDA topic model also processes documents into a set of 50 topic proportions, also called per-document topic distribution. Simply speaking, each document can be represented as a combination of topics with different proportions. Here is an example of Article *Biology* in Figure 2.

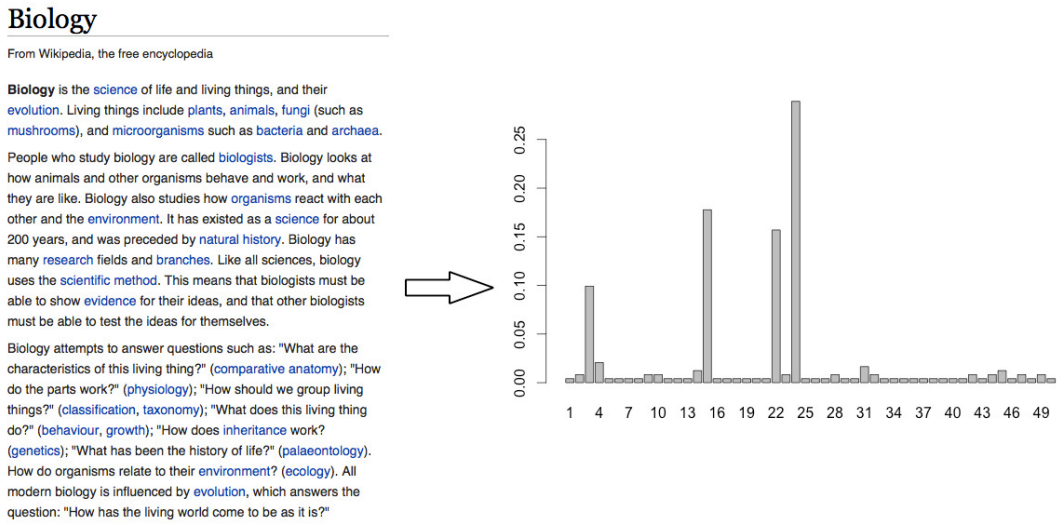

Figure 2. Topics Distribution of Article *Biology*

The original article is shown in the left part of Figure 2, which can be also accessed on simple Wikipedia website online. After the model is trained, we can get a series of article topics, which is bar plotted in the right part of Figure 2. In total of 50 topics, we can easily find there are several topics with obviously high probabilities, where top three topics are topic 24, topic 15 and topic 22. The three high proportion topics can represent the main topics of this article. Table 2 shows the terms in these 5 topics with 10 terms each. If we take a look at Table 2, the topic terms, such as "cell", "body", "science", or "bird", are highly associated with the Article Biology, which makes sense because these topics are what the article is mainly talking about.

Table 2. Top 5 Topic Terms of Article *Biology*

Topic 24	**Topic 15**	**Topic 22**
cell	universe	bird
body	science	animal
disease	study	live
organ	book	special
human	philosophiae	fish
cancer	philosopher	dog
system	human	insect
virus	question	cat
blood	categorical	eat
cause	theory	hair

So just like this example, every article of the whole collection is represented as a set of probabilities over 50 topics. This is the core data of our model, where we can do all sorts of analysis and implementations.

4. DOCUMENT EXPLORING SYSTEM

The goal of this research is to design a document exploring system for better researching through documents according to a certain topic. Figure 3 shows the system workflow of the design. After fitting the model, we will do the following three steps: document clustering, calculating the similarity and assembling HTML files.

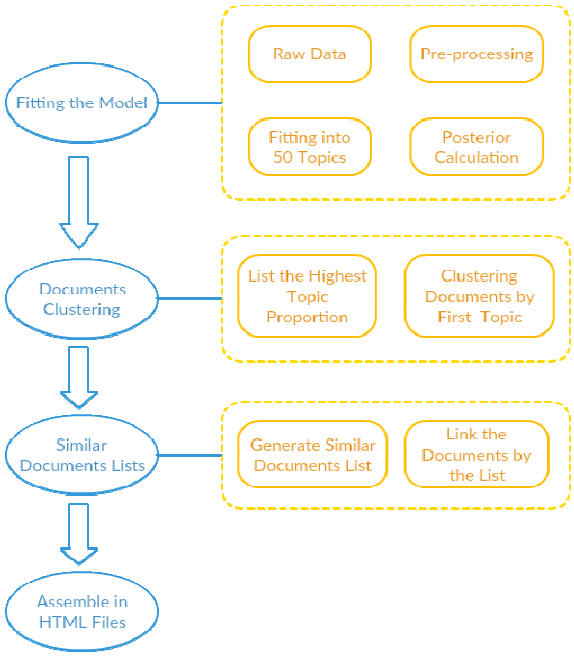

Figure 3. Workflow of Document Exploring System

4.1. Document Clustering

The method of document clustering with LDA topic model is to find the highest topic proportion and assign the document into that topic. As a previous example of Article Biology, it will assign Biology to the highest proportion topic, which is topic 24. If we apply this method for the whole collection, documents will be clustered into 50 groups.

Table 4. Top 10 Shortest Distance of Article *Light*

Titles	Topic (cluster)	Titles	Topic (cluster)
April	Topic 13	Computer Science	Topic 17
August	Topic 13	Google	Topic 17
December	Topic 13	Apple Macintosh	Topic 17
Brazil	Topic 47	Galaxy	Topic 33
China	Topic 47	Mars	Topic 33
Breakfast Sausage	Topic 10	Earth	Topic 33
Berry	Topic 10	Chinese	Topic 50
Fruit	Topic 10	Grammar	Topic 50

If we take a look at Table 4, compared with the titles and the topics, we can find the clustering algorithm works quite well. Figure 4 shows the histogram of the clustering result in the whole collection, which is the number of documents in each topic. We can find an extremely high cluster, topic 3, which contains more than 250 documents. Compared with other 49 clusters, topic 3 is not a well clustered topic. If we have a look at the topic terms in topic 3, which are "people", "use", "can", "also" and so on. These terms are content neutral terms, which are hard to decide to a certain topic. To fix this problem, we have two solutions. First one is to remove these terms like stopwords, then re-train the model and re-cluster documents. The other one, which is what we chose to do, is using two option clustering.

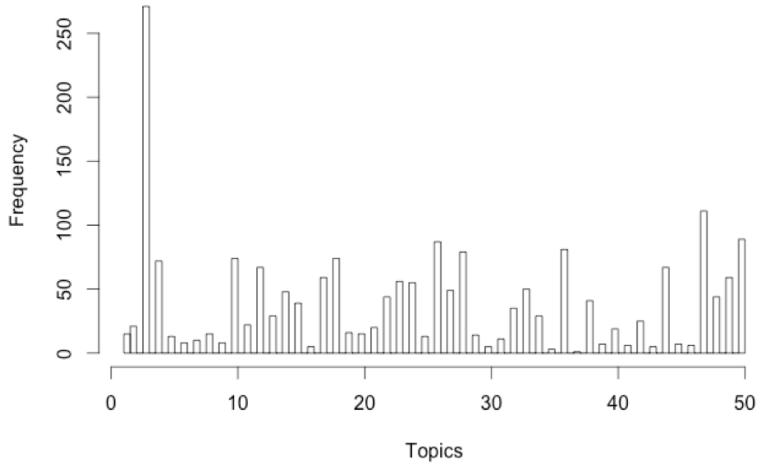

Figure 4. Histogram of Document Topic Clustering

Since LDA topic model provides a set of 50 topics proportion, we can also use the second highest proportion to decide the cluster, given that the highest proportion does not perform well. In Table 5, we list the results of two option clustering for some documents belonging to topic 3. The results make more sense than the original one. The reason behind this method is that for most documents, the topic proportions are not equally distributed. The highest three are pretty dominated by about half of the percentage and the rest 47 topics share the other half. So it makes sense that we can use the highest two or highest three topics to represent a document.

Table 5. Clustering with Two Options

Title	First Option	Second Option	Top Terms in Second Option
Abbreviation	Topic 3	Topic 50	language, english, word
Application	Topic 3	Topic 17	computability, internet, system
Coin	Topic 3	Topic 34	money, econometrics, companies
Human Body	Topic 3	Topic 24	cell, body, disease, organ
Harbour	Topic 3	Topic 27	engine, ship, boot, make
Nature	Topic 3	Topic 15	universe, science, study, book
Statistics	Topic 3	Topic 14	number, mathematic, line

4.2. Document Similarity

The method of calculating document similarity with LDA topic model is to compare the topic proportions. The measurement is JS distance, which we have talked about in Section 2.4. JS distance is a popular way to measure the similarity between two distributions. The topic proportions of a document can be treated as a distribution over 50 topics. We can calculate the JS distance between every two documents.

4.2.1. Document Similarity of Article Constitution

For each document, we have calculated the JS distance with every other document in the collection. We use the Article Constitution as an example. Table 6 shows the top 5 closest JS distance of Article Constitution. Article 99 has a distance of 0, because it is Constitution itself. The following ones are Article 983, Article 1473, Article 897, and Article 1016, whose titles are Freedom of Speech, Head of State, Citizenship and Election.

Table 6. Top 5 Closest JS Distance of Article *Constitution*

Articles	Article 99	Article 983	Article 1473	Article 897	Article 1016
JS Distance	0.0000000	0.2271741	0.3404084	0.3881467	0.4583745

4.2.2. Human Judgment Comparisons

How is the performance? The performance of the document similarity is hard to decide. Everyone may hold a different opinion on which document is the most similar to another. It depends on many factors, time, location, background, and personal educations. Comparing to finding a standards of measuring document similarity, we would rather get a general opinion on which document is more similar. What we actually want is a comparison between the human judgment and the LDA topic model.

So we made a questionnaire to ask a small group of people to rate the most similar documents for a given one. The questionnaire structure are five groups of documents. Each group has a given document and four choice documents. People will be asked to read the given document first and then rate the four choice documents by similarity with their own standards. With the results of the human rating, we also applied a voting method, which is widely used to select the MVP award in sports. For four documents, the first option gets four points, the second option gets three points, the third option gets four points and the fourth option gets one point. Then we will calculate the points for every document and order them. Table 7 shows the comparisons between LDA topic model and human judgment.

Given Document	**Constitution**
Readers	Freedom of Speech, Head of State
LDA Topic Model	Freedom of Speech, Head of State
Given Document	**Health**
Readers	Physical Exercise, Medicine
LDA Topic Model	Physical Exercise, Medicine
Given Document	**Molecule**
Readers	Polymer, Organic Compound
LDA Topic Model	Organic Compound, Polymer

Given Document	**Roman**
Readers	9 (year of 9 A.D.), Denarius
LDA Topic Model	9 (year of 9 A.D.), Naples
Given Document	**Euro**
Readers	Dollar, Continent
LDA Topic Model	Dollar, GDP

The results show that LDA topic model has a great match with human judgment. The first two groups, LDA topic model has a perfect match with two most similar documents. The following group, LDA topic model has a reverse of the answer. The last two groups, LDA topic model gets the first one correct. To conclude this part, human judgment is a way to evaluate the performance of topic models. However, this method is limited with the size of the experiment groups and may have different results in different people. Even with the same people, we can have a different outcome in a different time frame. But the idea of the method and the way to evaluate may be useful or inspiring to other researchers.

4.3. System Design

We will combine the clustering results and document similarity to assemble the document exploring system for Wikipedia articles. The system is based on HTML webpages and divided into three layers: index page, topic page and document page, which are shown in Figure 5.

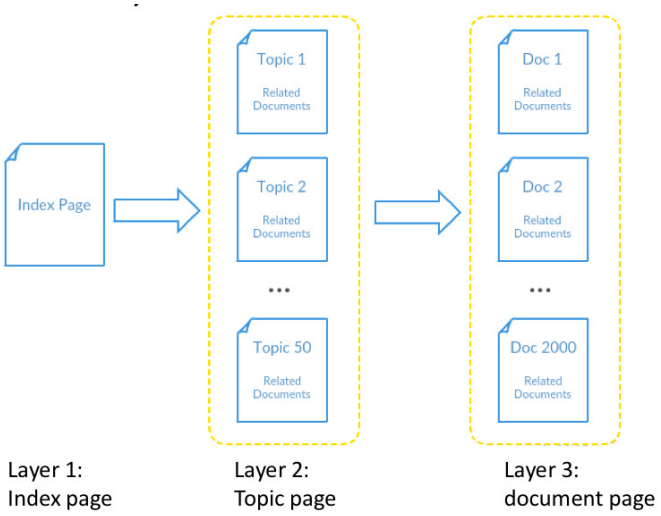

Figure 5. System Layers

The first layer is the index page including 50 topics and its topic terms. This is the entry of the system, and users will choose one of the topics to start exploring. After clicking a certain topic, users will go to the second layer: topic page. Users can have a closer look at the topic terms, along with the related documents of this topic. Figure 6 shows the screenshots of index page and the topic page.

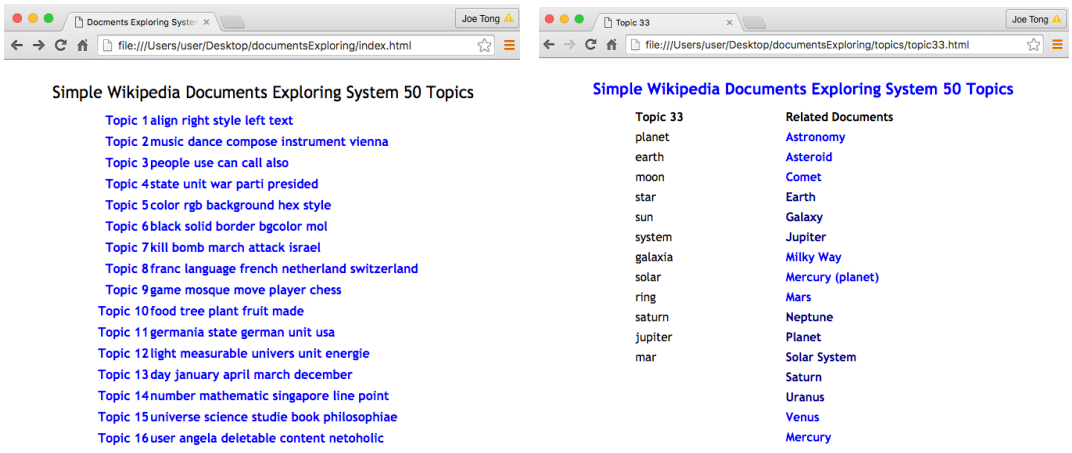

Figure 6. System Index Page and Topic Page

If users click a certain document for example Mars, which is shown in the left of Figure 7. It is the third layer, a document page of Mars. This page shows related topics of Mars and also related documents of Mars. Users are able to exploring through topics and documents according to a topic. If the title of Mars on the top is clicked, the system will open a new page linked to the article in Simple Wikipedia. That is shown on the right of Figure 7.

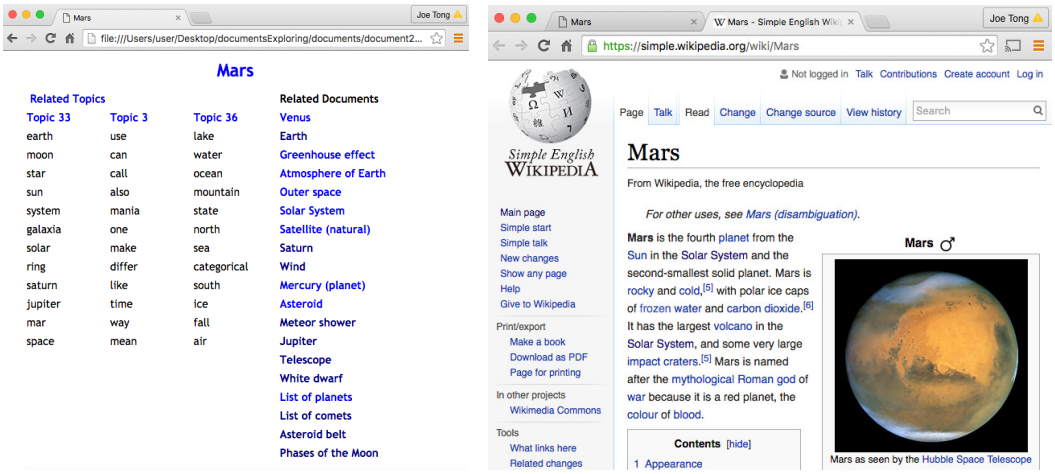

Figure 7. System Document Page and Simple Wikipedia Page

5. CONCLUSIONS AND FUTURE WORK

In this paper, we briefly talked about the background knowledge of LDA topic model and its working principles. Then we explained in depth how to apply LDA topic model to a document collection by doing experiments on Simple Wikipedia articles. The experiments include all necessary steps of data retrieving, pre-processing, fitting the model and evaluations. The result of the experiments shows LDA topic model working effectively on document clustering and finding similar documents. Based on LDA topic model, we designed a document exploring system which

allows users to organize and explore the documents by topics where related documents are easier to find and access.

With the experiment data and the conclusions, a number of future works can be done for further research and experiment:

- From the perspective of this research, upgrading the computing power and enlarging the sample database on Wikipedia or even a collection of scientific papers could be a very good future project. Wikipedia or papers are an actual knowledge base. They are well worth mining and the results could be the key elements to build a super intelligent robot. This is one of the reasons why we chose Simple Wikipedia documents to be the sample in this research.
- Applying data mining techniques to social network is a hot new direction. At the beginning of the research, we have done some experiments and work on applying topic models to Twitter. Considering the length of the paper and the limitation of the time, we decided to exclude the part of in this paper. Social network data mining is not only important for computer science, but also for social science and business. It is a great future work to follow and actually an interesting topic that we are doing next.
- In terms of topic models, a well designed and widely approved evaluation method is one of the most significant directions of the research.
- Some data mining scientists have already made progress on topic model for image processing. If further development is achieved, the topic model could be a complete package of document processing

REFERENCES

[1] Martin Ponweiser (2012) *Latent Dirichlet Allocation in R*, Vienna University of Business and Economics.
[2] Bettina Grun, Kurt Hornik (2011) "topicmodels: An R Package for Fitting Topic Model", Journal of Statistical Software Vol. 40, No. 13.
[3] Qi Jing (2015) *Searching for Economic Effects of User Specified Event Based on Topic Modelling and Event Reference*, Jordery School of Computer Science, Acadia University.
[4] David M.Blei (2012) "Probabilistic Topic Models", Communications of the ACM Vol. 55, No. 4, pp77-84.
[5] David M.Blei, John D. Lafferty (2006) "A Correlated Topic Model of Science", Annals of Applied Statistics Vol. 1, No. 1, pp17-35.
[6] David M. Blei and John D. Lafferty. Topic models. In Text Mining: Classification, Clustering, and Applications, pages 71–94, 2009.
[7] Jianhua Lin (1991) "Divergence Measures Based on the Shannon Entropy", IEEE Transactions on Information Theory Vol. 37, No. 1, pp145-151.
[8] Amazon Web Services. "Amazon ec2 - virtual server hosting". URL https://aws. amazon.com/ec2/.
[9] The R Foundation. "What is R?". URL https://www.r-project.org/about. html.
[10] Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, Khairullah Khan (2010) "A Review of Machine Learning Algorithms for Text-Document Classification", Journal of Advances in Information Technology Vol. 1, No. 1, pp4-20.

AUTHORS

Zhou Tong is currently a master student of computer science at Acadia University, Canada. His research is focusing on text mining.

Haiyi Zhang received his MS degree in 1990 from the Computer Science department of New Jersey Institute of Technology of USA, and his Ph.D in 1996 from Harbin Institute of Technology in China. He was a post-doctor in information department of ABO, Finland in 2000. His research interests are machine learning, data mining. He has more than 50 academic papers published. Currently he is an associate professor at Acadia University, Canada.