Transactions on Machine Learning
and Data Mining
Vol. 8, No. 2 (2015) 101-109
ISSN 1864-9734
ISBN: 879-3-942952-39-2

ibai Publishing
www.ibai-publishing.org

Design and Considerations of a Searching Software

Haiyi Zhang

Jodrey School of Computer Science, Acadia University, Wolfville, Nova Scotia, Canada
Haiyi.zhang@acadiau.ca

Abstract. In order to help Financial analysts and investors to make sound decisions for their businesses and investments, large archives of news articles could be analyzed to develop a more comprehensive view of the economic consequences from information about what events happen and the economic effects from these events. We try to develop a systematic solution for finding the economic effects of temporal events.

Keywords: Big Data Analystics, Searching, Text Mining, Machine Learning.

1 Introduction

Numerous news articles are generated daily on the web. If we could adopt a systematic method to analyze these articles, it could help us better understand the events that happened and the relationship between these events. This research intends to develop a systematic approach, in order to find the relationship between economics and the event of user interest.

In this paper, topic model and statistical language model are applied to retrieve the related documents of an event. Subsequently how to find the economic articles related to the user specified event remains a difficult problem.

Previous methods have been developed to reference between events using features of a document. Traditional features extraction methods include Document Frequency, Mutual Information, Information gain etc [6] [7] [8]. These methods are used to extract features from a single document. This paper proposes a new method combining TDIDF and Information Gain Ratio to extract features from a collection of documents.

In order to do event reference from user specified event to economic event, Normalized Feature Referencing is proposed and analyzed. Instead of manually choosing the number of features used for event reference, this method will automatically choose the number of features depending on the probability distribution of the features. In this way, not only the limitation of manual work is eliminated, it achieves a better reference performance for various feature probability distributions.

2 Background

Topic models are a suite of algorithms that uncover the hidden thematic structure in document collections. These algorithms help us develop new ways to search, browse and summarize large archives of texts [1]. In order to analyze large amount of data collected, topic model [3] is applied in this paper. More precisely, Latent Dirichlet Allocation (LDA)[2], a form of topic model, is adopted. A graphical model describing LDA is shown as below:

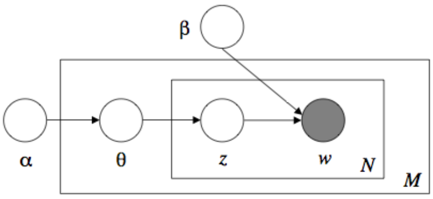


Fig. 1: Latent Dirichlet Allocation Model [2]

where α is the Dirichlet prior over per-document topic proportion; θ is the per-document topic distribution; z is the word to topic assignment in a document; β is the topic distribution over the text corpus and w is the observed word. The generative mathematic process of LDA is as below:

$$P(B,\Theta,Z,W) = \prod_{i=1}^{k} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) (\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|B, z_{d,n})) \quad (1)$$

By applying Gibbs sampling or variational inference, the posterior variables can be approximated.

TFIDF is short for term frequency inverse document frequency [4]. It is a statistical indicator to measure the importance of a term to a document in a corpus. The word importance increases as the frequency of the word in the document increases but is offset by the frequency of the word in the corpus. The TF (Term frequency) part calculates the frequency of a term in a single document. The IDF (Inverse Document Frequency) part diminishes the weight of terms that generally occurs in the collection of documents and increases the weight of terms that occurs rarely. For example, assume without removing stop words, the term “the” occurs frequently in almost all documents. If only term frequency is used, “the” will be important to every document. Therefore, by incorporating inverse document frequency, the importance of the term “the” will be decreased. In this paper, information gain ratio is used to calculate the IDF part.

In classification problems, we want to know which attribute is most useful in discriminating the classes to be classified. Information gain tells us how important of a given attribute to classification. The higher the information gain, the better the attribute. Information gain is calculated as below:

$$IG(S,x) = H(S) - H(S|x) \quad (2)$$

where $H(S)$ is the parent entropy of the system S and $H(S|x)$ is the entropy given variable x. $IG(S)$ is the information gain by using the variable x.
Furthermore, Information gain ratio is calculated as below:

$$IGR = \frac{IG}{Entropy(Parent)} \tag{3}$$

3 System Design

A system architecture needs to be derived to describe the structure, behavior and relationship among the components of the system. In addition, a detail arrangement of the elements needs to be provided to satisfy the functional requirements.

The system architecture is shown in Figure 2. It contains four primary components: preprocessing, training, query and event reference. As we can see, each component is dependent on another component. The input dataset for training requires preprocessing in advance; the query module is based on the LDA model and language model trained in the training module; and event reference module uses the related documents retrieved by the query.Preprocessing is one of the key component in this framework. The effectiveness of the topic model and subsequent query performance are greatly influenced by different combinations of preprocessing methods. Main content extraction from HTML document is an essential function, which removes irrelevant boilerplate and leave only the text of main content; tokenization breaks a stream of text into tokens and the training for our models is based on these tokens; Stop words removal and stemming are optional functions that can bring improvement of performance of the models to fulfill business objectives. There are other application dependent preprocessing methods. For example, nnecessary documents will be removed if a document contains less than a certain amount of tokens, such as 25 tokens; The document level of occurrence of a word is counted, if the word occurs in most of the documents like more than 95 percent of the document, or it occurs very rarely in the document collection, in either case, it will be removed from the vocabulary.

In the training module, LDA model and statistical language model will be trained by using pre-processed dataset as input. In LDA, posterior hidden variables θ, β and z will be computed and stored. The language model will produce all relevant statistics related to the corpus such as term frequency in corpus and term frequency in a document.

The query module accomplishes retrieving documents of an event based on the combination of trained LDA model and language model.

Eventually, the relationship between event and economics will be set up by feature extraction from the related documents of an event, then applying normalized feature referencing using the extracted features.

4 System Manual

4.1 User Guide

Following the design of the system and system architecture, the user interface consists of three main parts: preprocessing, topic modelling and event reference.

In the preprocessing function group as shown in figure 3, user can preprocess the document collection using the interface in the following steps:

- First, the source directory can be selected by clicking on the input field under the "Source Directory" label. The directory selected contains all the html documents that previous crawled by a web crawler.
- Second, four preprocessing techniques can be checked. These techniques include extracting main content from HTML files, snowball stemming, removing stop words and removing documents that have less than a certain number of words. The technique of extracting content from HTML is disabled because it is mandatory to adopt this technique in order to make the system work.
- Third, user can click the "START" button to kick out the preprocessing process. The process of preprocessing is displayed on the right side panel under the "Output" label.

Once the preprocessing is finished, the user can switch to the second function group "Topic Modelling", which contains topic model training, topic model training results and query (event search). The user is able to finish the event search function by the following sequence of steps:

- Under the train label, user can click "START TRAINING" to start training the topic model based on the preprocessed text corpus.
- After the training is complete, the "SHOW TOPICS" button can be clicked to show the topics and the corresponding word distribution in topics. The "SHOW DOCUMENTS" button is used to show the topic distribution in each document.
- Under the "Query" label, a query string can be specified by the user in the input field under the "Query String" label. An average probability threshold for selecting the related documents is displayed in the input field under "Avg. Probability Threshold" label. This figure can be adjusted by the user to fit different text corpus.
- Finally, the user can click the "START QUERY" button to start searching for related documents to the query. Eventually, the document names and probabilities will be displayed on the output panel to the right.

Since the related documents of an event (query) are retrieved, we can now use the third function group "Reference" to find the economic articles related to the event.

- First of all, the economic directory under the "Economic Directory" label can be clicked to select the directory that contains all the economic HTML documents.
- Under the label of "Feature Extraction", two parameters can be input by the user. The default value for "Maximum Features Allowed" is 15 and is recommended to remain unchanged. This figure set the maximum number of extracted features from the related documents of an event. The "Feature Probability Threshold" is the threshold value to pick the features, which can be adjusted by the user.
- After the parameters are set, the user can click "EXTRACT FEATURES" button to extract the features from the related documents of an event.
- Under the "Reference" label, the confidence value can be set to find the related economic articles to an event. The default value is 0.75.
- Eventually, the "START REFERENCE" button can be clicked to start the referencing process. The progress of processing is shown in the output panel on the right.

5 Conclusion and Future Work

We have developed a software system that is able to retrieve documents of an event based on user interest as well as finding the economic effects of the event. The system demonstrates its ability to reasonably accurately retrieve documents using topic modelling and statistical language model. A novel method regarded as Normalized Feature Referencing is proposed and implemented in order to find the economic effect of user specified event. The experiment results proves the effectiveness of the proposed method.

Many future works can be done to further improve the system. An online method could be applied to feed the news articles continuously to train the topic model and language model; Natural language processing could also be researched and applied in order to achieve a better performance of event reference.

References

1. David M. Blei, *Probabilistic topic models*. Magazine Communications of the ACM, Volume 55 Issue 4, April 2012, Pp 77-84 (2012)
2. David M. Blei, Andrew Ng, *Latent Dirichlet Allocation*. The Journal of Machine Learning Research, Volume 3, 3/1/2003, Pp 993-1022 (2003)
3. Ian H. Witten, Eibe Frank, Mark A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition.*, Morgan Kaufmann Publishers is an imprint of Elsevier, 30 Corporate Drive, Suite 400, Burlington, MA 01803, USA (2011)
4. Jos M. Bernardo, *Bayesian Statistics*. International Encyclopedia of Statistical Science, pp 107-133 (2011)
5. Jie Tang, Jing Zhang, *ArnetMiner: extraction and mining of academic social networks*. Proceeding KDD '08 Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, Pp 990-998 (2008)
6. Yi Cai, *Event Relationship Analysis for Temporal Event Search*. Database Systems for Advanced Applications, Lecture Notes in Computer Science Volume 7826, pp 179-193 (2013)
7. Yiming Yang, Jan O. Pedersen *A Comparative Study on Feature Selection in Text Categorization*. ICML '97 Proceedings of the Fourteenth International Conference on Machine Learning Pp 412-420 (1997)
8. David M. Blei, *Topic models.*, https://www.cs.princeton.edu/ blei/topicmodeling.html, September 1 (2009)

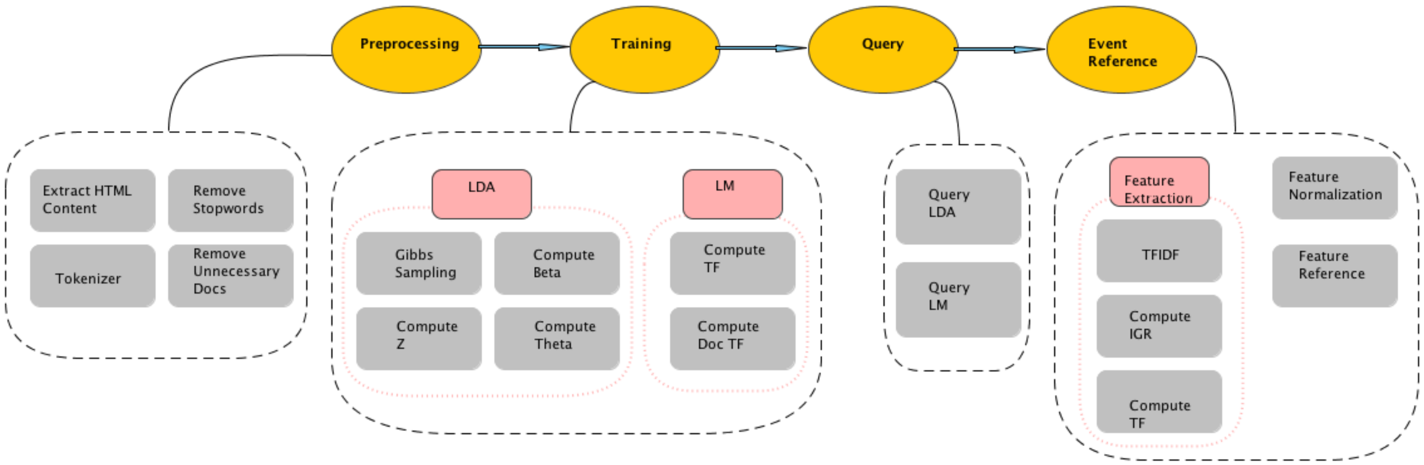
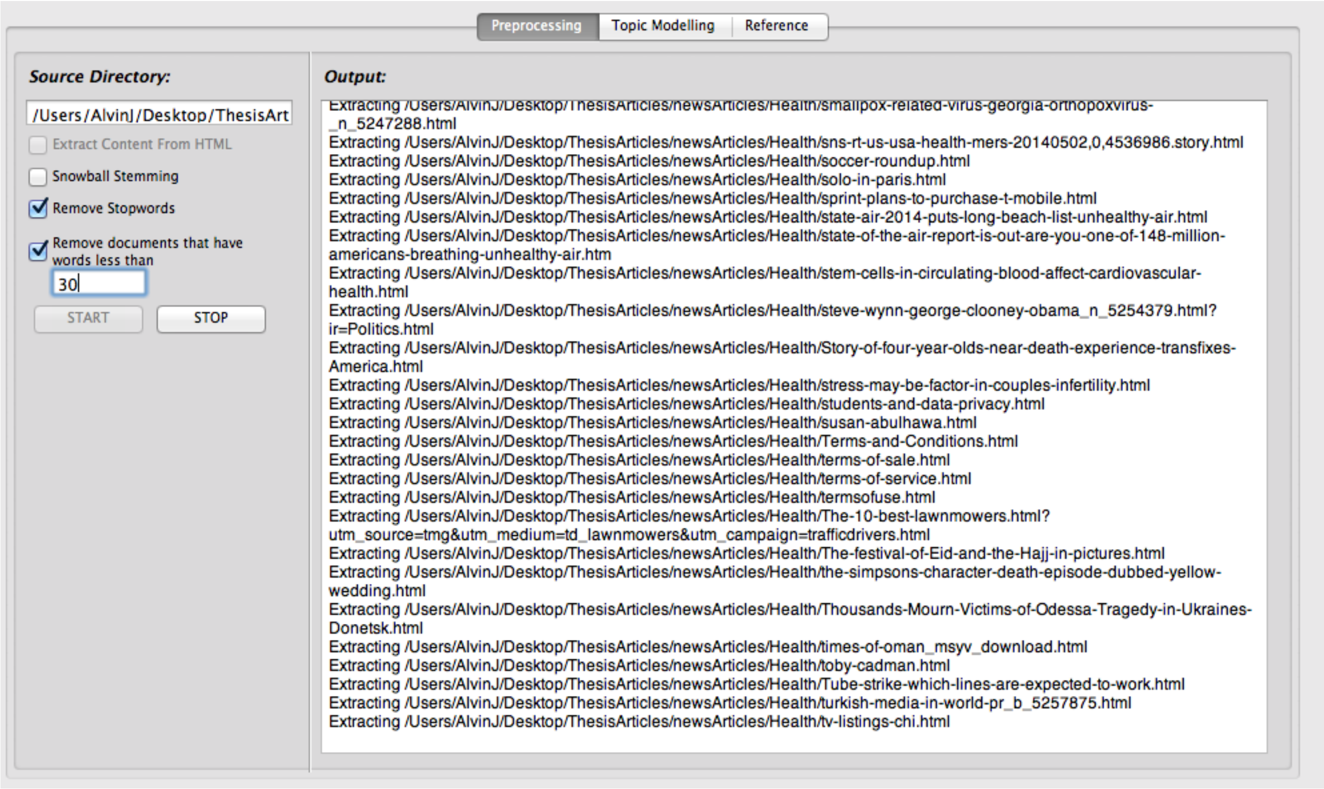


Fig. 2: System Architecture



Source Directory:

/Users/AlvinJ/Desktop/ThesisArt

□ Extract Content From HTML

□ Snowball Stemming

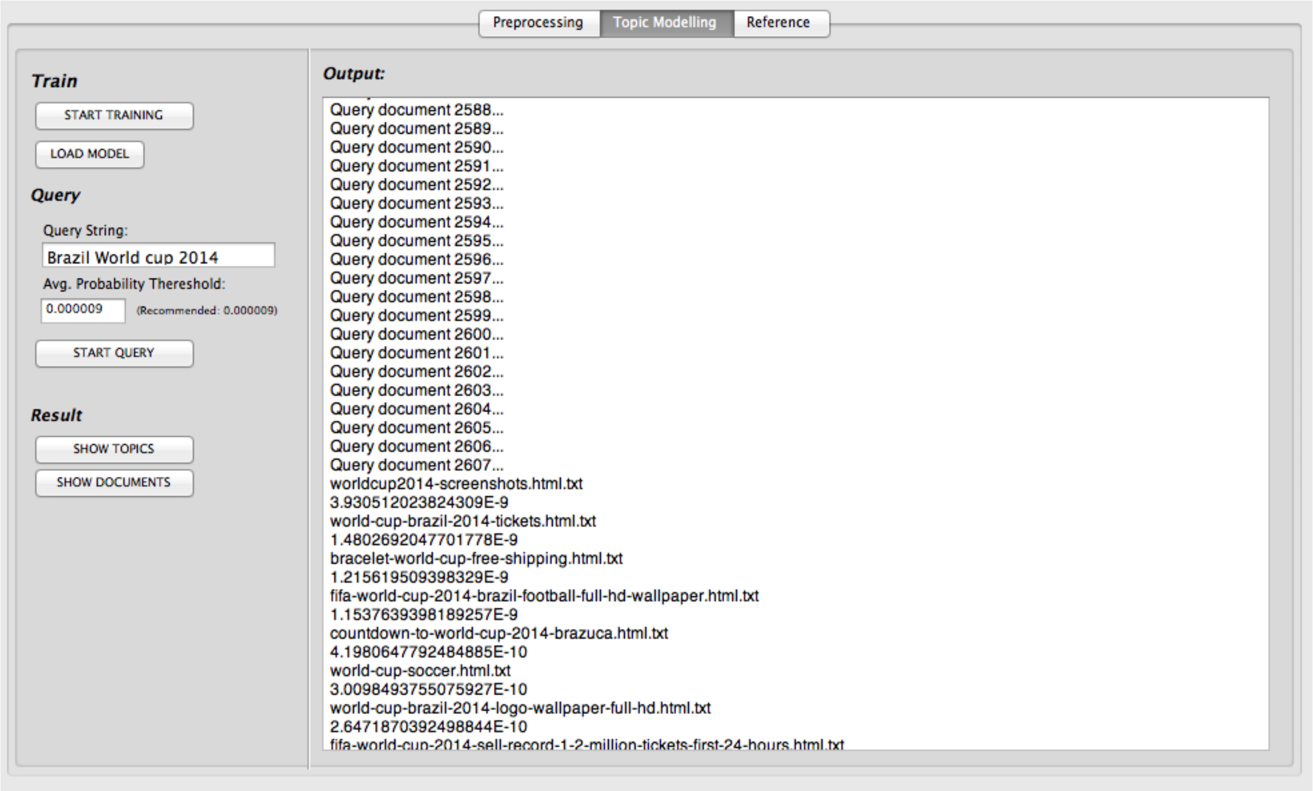
☑ Remove Stopwords

☑ Remove documents that have words less than 30

Output:

```
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/smallpox-related-virus-georgia-orthopoxvirus-
_n_5247288.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/sns-rt-us-usa-health-mers-20140502,0,4536986.story.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/soccer-roundup.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/solo-in-paris.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/sprint-plans-to-purchase-t-mobile.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/state-air-2014-puts-long-beach-list-unhealthy-air.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/state-of-the-air-report-is-out-are-you-one-of-148-million-
americans-breathing-unhealthy-air.htm
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/stem-cells-in-circulating-blood-affect-cardiovascular-
health.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/steve-wynn-george-clooney-obama_n_5254379.html?
ir=Politics.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/Story-of-four-year-olds-near-death-experience-transfixes-
America.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/stress-may-be-factor-in-couples-infertility.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/students-and-data-privacy.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/susan-abulhawa.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/Terms-and-Conditions.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/terms-of-sale.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/terms-of-service.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/termsofuse.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/The-10-best-lawnmowers.html?
utm_source=tmg&utm_medium=td_lawnmowers&utm_campaign=trafficdrivers.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/The-festival-of-Eid-and-the-Hajj-in-pictures.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/the-simpsons-character-death-episode-dubbed-yellow-
wedding.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/Thousands-Mourn-Victims-of-Odessa-Tragedy-in-Ukraines-
Donetsk.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/times-of-oman_msyv_download.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/toby-cadman.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/Tube-strike-which-lines-are-expected-to-work.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/turkish-media-in-world-pr_b_5257875.html
Extracting /Users/AlvinJ/Desktop/ThesisArticles/newsArticles/Health/tv-listings-chi.html
```

Fig. 3: Preprocessing Example of Text Corpus

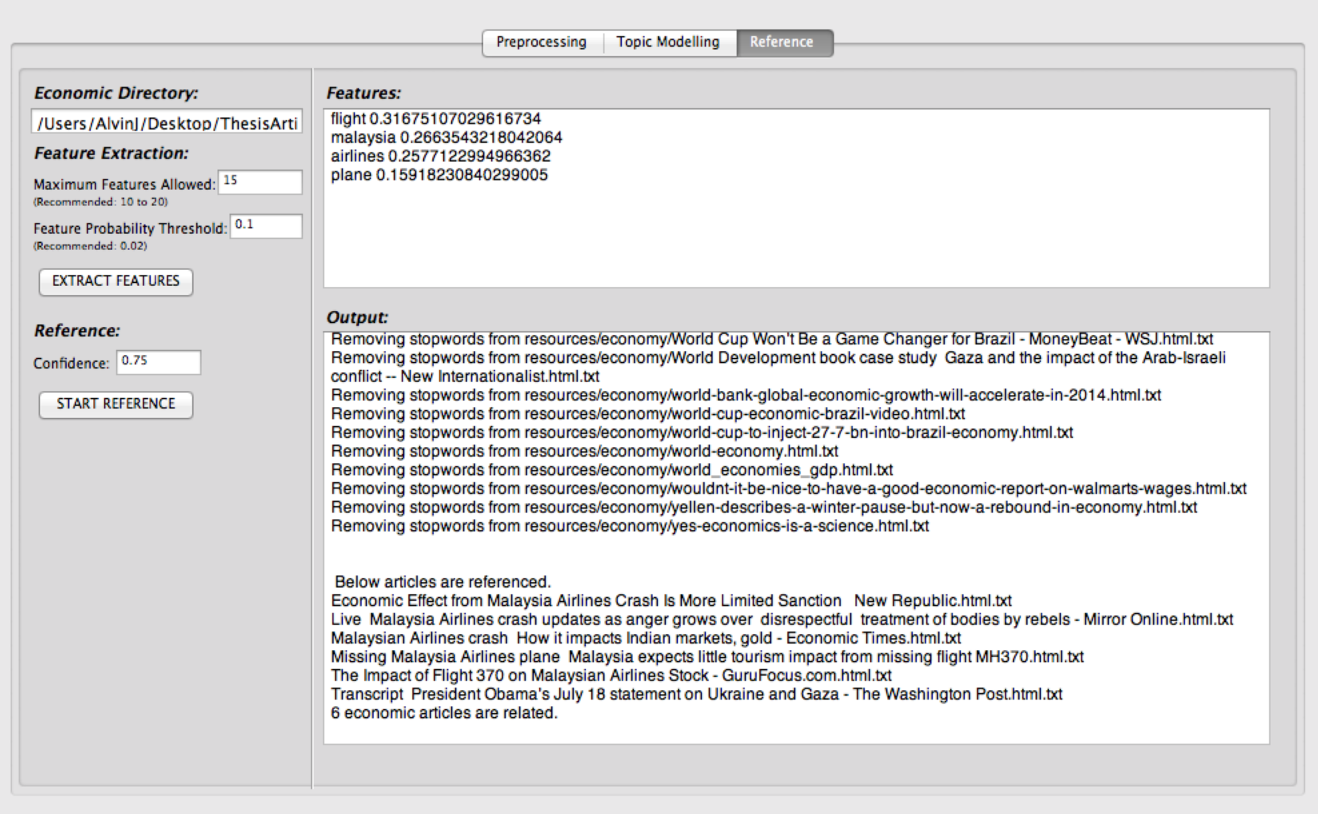


Query String: Brazil World cup 2014

Avg. Probability Thereshold: 0.000009 (Recommended: 0.000009)

```
Query document 2588...
Query document 2589...
Query document 2590...
Query document 2591...
Query document 2592...
Query document 2593...
Query document 2594...
Query document 2595...
Query document 2596...
Query document 2597...
Query document 2598...
Query document 2599...
Query document 2600...
Query document 2601...
Query document 2602...
Query document 2603...
Query document 2604...
Query document 2605...
Query document 2606...
Query document 2607...
worldcup2014-screenshots.html.txt
3.930512023824309E-9
world-cup-brazil-2014-tickets.html.txt
1.4802692047701778E-9
bracelet-world-cup-free-shipping.html.txt
1.215619509398329E-9
fifa-world-cup-2014-brazil-football-full-hd-wallpaper.html.txt
1.1537639398189257E-9
countdown-to-world-cup-2014-brazuca.html.txt
4.1980647792484885E-10
world-cup-soccer.html.txt
3.0098493755075927E-10
world-cup-brazil-2014-logo-wallpaper-full-hd.html.txt
2.6471870392498844E-10
fifa-world-cup-2014-sell-record-1-2-million-tickets-first-24-hours.html.txt
```

Fig. 4: Example of Query in Topic Modelling



Economic Directory:

/Users/AlvinJ/Desktop/ThesisArti

Feature Extraction:

Maximum Features Allowed: 15
(Recommended: 10 to 20)

Feature Probability Threshold: 0.1
(Recommended: 0.02)

Reference:

Confidence: 0.75

Features:

```
flight 0.31675107029616734
malaysia 0.2663543218042064
airlines 0.2577122994966362
plane 0.15918230840299005
```

Output:

```
Removing stopwords from resources/economy/World Cup Won't Be a Game Changer for Brazil - MoneyBeat - WSJ.html.txt
Removing stopwords from resources/economy/World Development book case study  Gaza and the impact of the Arab-Israeli
conflict -- New Internationalist.html.txt
Removing stopwords from resources/economy/world-bank-global-economic-growth-will-accelerate-in-2014.html.txt
Removing stopwords from resources/economy/world-cup-economic-brazil-video.html.txt
Removing stopwords from resources/economy/world-cup-to-inject-27-7-bn-into-brazil-economy.html.txt
Removing stopwords from resources/economy/world-economy.html.txt
Removing stopwords from resources/economy/world_economies_gdp.html.txt
Removing stopwords from resources/economy/wouldnt-it-be-nice-to-have-a-good-economic-report-on-walmarts-wages.html.txt
Removing stopwords from resources/economy/yellen-describes-a-winter-pause-but-now-a-rebound-in-economy.html.txt
Removing stopwords from resources/economy/yes-economics-is-a-science.html.txt


 Below articles are referenced.
Economic Effect from Malaysia Airlines Crash Is More Limited Sanction   New Republic.html.txt
Live  Malaysia Airlines crash updates as anger grows over  disrespectful  treatment of bodies by rebels - Mirror Online.html.txt
Malaysian Airlines crash  How it impacts Indian markets, gold - Economic Times.html.txt
Missing Malaysia Airlines plane  Malaysia expects little tourism impact from missing flight MH370.html.txt
The Impact of Flight 370 on Malaysian Airlines Stock - GuruFocus.com.html.txt
Transcript  President Obama's July 18 statement on Ukraine and Gaza - The Washington Post.html.txt
6 economic articles are related.
```

Fig. 5: Example of Event Reference