SEARCH TEXT BASED ON LOCATIONS

A Thesis

by

WEIWEI ZHANG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Alexander Sprintson
Co-Chair of Committee,	Anxiao Jiang
Committee Member,	Srinivas Shakkottai
Head of Department,	Chanan Singh

December 2014

Major Subject: Computer Engineering

ABSTRACT

To satisfy the current need for finding queried information quickly, search engines, data mining systems, and many other applications have been in development in recent years. Some of those applications look for documents containing phrases of a particular topic, such as historical events from a certain time period. Among these applications, queries based on geographical data are receiving significant attention from the research community and industry. Therefore, this thesis studies text search based locations, which contributes to the Geographical Information Retrieval (GIR) systems.

In addition to the traditional applications of GIR systems, which are used for finding locations in documents, GIR can be applied to other fields as well. Firstly, it can retrieve location information in text and search for answers to questions of a spatial nature (such as “Where is College Station?”). Location information can improve presentation of the search results, for example, by presenting the search results on a map. GIR also adds to the field of spatial diversity search, which allows users to express preferences and constrain the search results to a particular geographical region. In addition, it finds related document based on location information from different sources of information and then represents the similarities graphically. In this way, the readers can visually see the data, helping them understand the document correlations in an intuitive way.

However, most of the previous research involves keyword searches in spatial databases instead of raw (unlabeled) text. Although there is some work on raw text processing, that work uses matching techniques, and limits the geographical range to small geographical regions such as a single country. Therefore, this thesis adopts a new clustering method, which utilizes a geographical dictionary to locate any place by its coordinates. This method reduces ambiguity and improves the accuracy over the previous research. This study also implements a new word-clustering method to detect a combination of topics in raw text. This method is more accurate than the latent Dirichlet allocation, a state of the art method based on a probabilistic model. In addition, a novel graphic illustration is utilized to visually represent the relevance ranking between documents.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my committee members. I would never have been able to finish my thesis without your guidance.

I would also like to express my deep thanks to Prof. Jiang, who helped guided my research in the past two years, and has been helpful in providing much assistance and support. I will never forget your patience, warm encouragement, valuable discussion, and comments on the entire research project.

In addition, I greatly appreciate Prof. Sprintson. You have been very helpful and provided much insightful advice.

I would also like to thank Prof. Shakkottai, for his thoughtful comments and suggestions during my defense.

Finally, I would like to thank my family, for all of the sacrifices that you have made while supporting me during graduate school. The past two years have not been an easy time, and it is your love and support that sustains and inspires me to always strive for my goals. I also appreciate all of my friends who have helped me in pursuing my master's degree by cheering me up, and sticking by my side during this challenging time.

NOMENCLATURE

LDA	Latent Dirichlet Allocation
HMM	Hidden Markov Model
GIR	Geographical Information Retrieval
GIS	Geographical Information System
NLP	Natural Language Processing
POS	Part-Of-Speech

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

FIGURE Page

LIST OF TABLES

1. INTRODUCTION

1.1 Motivation

Location-based information is useful for many applications such as search engines and data mining systems. Therefore, Geographical Information Retrieval (GIR) has received increasingly significant attention. Different from Geographic Information Systems (GIS), which extracts locations from a precise, map-based, structured representation, GIR gains geographical information from unstructured texts through natural language processing methods.

GIR is a particularly useful tool, as it provides different services such as spatial data management, analysis, retrieval, and visualization. GIR provides a bridge between the world of Geographic Information Systems (GIS) and Natural Language Processing (NLP) [1]. Furthermore, GIR allows users to express location preferences and to constrain the search results to a particular geographical area. It can be used to answer questions or provide services based on locations. In addition, different places in different sources of information can also be analyzed and compared to identify their relationships. Apart from traditional text data retrieval, GIR also provides an extra spatial clue to help understanding the documents. It introduces a novel method to directly visualize information, through maps and graphs. In this way, readers are able to interact with the data, and intuitively grasp the document relationships.

At the same time, it is helpful to find the most prominent topics in a collection of documents for modern search. The topic modeling provides a brief summary of documents and enables efficient processing of a large collection of documents, while preserving the essential statistical relationships for tasks, such as novelty detection, query relevance judgment, summarization, and text classification [2].

However, current research does not explore the location-based search methods in raw text and relevance ranking in depth, nor does it provide any studies that focus on the document similarities in locations and topics, especially from the graphic aspect.

Therefore, this research seeks to summarize and search text from the geographical view. This study adopts a new clustering method to retrieve locations and to rank document relevance. In addition, this thesis uses topic modeling as a supplement for the text classification and relevance ranking. Finally, a novel graphic illustration is implemented to indicate the relevance between different documents.

1.2 Background Review

Current geographical information retrieval consists of two main parts: geo-coding and geoparsing. After enriching location description from a postal address or a place name (geo-coding), or extracting and resolving the exact meaning of locations in the unstructured text (geoparsing), the location references are indexed for retrieval and search [3].

A large body of work has been done to solve keyword searching problems in GIR systems. The existing work can be divided into two areas. One is the detection of the geographical locations using keyword matching in a spatial database [4]. For instance, several works have investigated toponym disambiguation in geoparsing. For example, Buscaldi [5] compares different toponym disambiguation resources and implementes a conceptual density method. Other research has focused on the location query processing in web-based search engines, such as [6] and [7]. For instance, Zheng [8] evaluates a probabilistic system that uses the tweet content of a Twitter user to classify his/her city-level locations. The system relies on a classifier that identifies words in tweets within a local geographic scope. It also utilizes a lattice-based neighborhood smoothing model to refine the estimated results. Chen [9] discusses several efficient algorithms to integrate the query processing with textual criteria in geographical query searches. It uses the k-sweep algorithm, tile index algorithm, and space-filling inverted index to compute the exact score of the unions and intersections of query footprints. In [10], geographical indexing, which combines scope index and spatial index in labeled text, is described.

Another field in location searches involves searching for spatial-keyword queries based on a GIS-like database [11]. This kind of system uses a particular set of textual keywords to find the objects that are closest to a specific location. Various kinds of R-tree and R*-tree, which are balanced search trees and can organize any-dimensional data through a bounding box, have been extensively studied.

For instance, in [12], a hybrid indexing data structure called KR*-tree, which is the combination of R*-tree and inverted index, is used for processing spatial-keyword queries. It solves the performance bottlenecks and reduces the disk IOs by pruning text and space simultaneously in KR*-tree. Ian De Felipe [13] proposes the IR2-Tree structure, a combination of R-Trees and signature files techniques, to answer top-k spatial keyword queries, which specify both a location and a set of keywords. His work is further investigated in [14], which introduces the m-closest keywords (mCK) query. The mCK query specifies keywords without a particular location, and the keywords can be in multiple tuples instead of one result tuple.

Some research has combined the thematic and geographic queries as spatial-keyword queries [15]. For instance, a dynamic document ranking scheme is proposed in [16], which combines the two relevance scores to calculate a final weight. It also combines the method of dynamic weighted sum and the evidence using Dempster-Shafers theory.

For the topic modeling, the earliest technique was the manually built thesaurus. However, this technique is time consuming and can result in a substantial disagreement on the semantic classifications [17]. Therefore, topic extraction methods gradually have evolved and been divided into several different groups: manually built topic models, term association models, and latent mixture models [18].

A large collection of research has been conducted on each group. For example, to construct manually built topic models, predefined rules and common sense are used, and Xing Wei [19] manually constructs the topic models based on hand-crafted resources and smoothing of the queries with topic models.

Term similarity measures utilize the word similarity techniques, including linguistic-based analysis and vector-based similarity coefficient, to obtain the close terms, and then the words are grouped into clusters or topics. In a similar way, documents are classified by topics. In [20] and [21], documents are automatically categorized by meanings of words or concepts. These works use word hierarchy structures such as hypernyms and synonyms provided by WordNet.

Finally, the current latent mixture model is developed on the basis of Latent Semantic Analysis(LSA), Probabilistic LSA (PLSA) and Latent Dirichlet Allocation (LDA) [22]. The model combines word clustering and document clustering. In these methods, texts are reformulated (for example, expanded) to improve effectiveness in retrieval [18]. For instance, Wartena [23] detects topics without any prior knowledge of categories. Wartena employs the Jensen-Shannon divergence of probability distributions as the distance measurement and takes the term co-occurrence into account. Among the different methods [24], the most commonly explored method is LDA, which is a generative Bayesian model for text classification and collaborative filtering. For example, in [2], the topics in text are learned through the Bayes parameter estimation, which is based on variational methods and an expectationmaximization (EM) algorithm.

1.3 Major Contributions

To briefly summarize a large volume of information and retrieve useful information quickly, this thesis carried out the research on text searching based on locations. This work makes the following contributions.

1. A new clustering method is developed to extract the locations. This method analyzes the raw text through part-of-speech tagging, and uses WordNet (a lexical database for the English language) and a gazetteer (a geographical dictionary) to extract the worldwide locations. It overcomes the limitation of previous work in which spatial databases or labeled text are required for the keyword search. In addition, our method extends the geographical region to the global locations and lists longitude and latitude for each location through location clustering. In contrast, the previous research on raw text processing restricts the locations to small areas and lacks coordinate information.

2. Our clustering method solves the location ambiguity problem and achieves higher accuracy compared to the Stanford Named-Entity Recognizer. Ambiguity is a challenging precision problem in the field of natural language processing (NLP). Especially in geographical information retrieval, lack of precision in place names causes many issues. GIR ambiguity can be classified into two major types: geo/non-geo and geo/geo ambiguity [6]. Geo/non-geo ambiguity occurs when a place name also has a non-geographic meaning. For instance, Charlotte and Lafayette both refer to person's names and place names. Geo/geo ambiguity arises when distinct places have the same name. For instance, Paris can be the capital of France or a city in Texas, USA. Therefore, we apply different ambiguity elimination criteria to intelligently assign a unique meaning to each place name and resolve the location conflicts.

3. The common location search was extended to a more general document level. Most of the current research ranks documents by relevance to location queries. However, our method produces a ranked document list based on the information that user queries. In addition, because textual description is not as intuitive as a picture, locations are directly shown on a map.

4. This thesis introduces the concept of a knowledge graph, which helps to find useful information that supports the results but users never thought to ask, and explores collections and lists in a graphic way. This kind of graphic illustration provides a direct interaction between humans and data. In addition, this kind of graphic model enriches the way people learn new knowledge and review the old information.

5. Apart from the geographical information retrieval, this study also considers textual topics. We implemented a new method for topic modeling that uses synsets, a collection of words which share the same meanings, to cluster words into topics, and then gives documents a distribution of topics. This method proved to be more accurate than the commonly used topic modeling LDA method.

To summarize, this thesis consists of the following parts. It introduces a new algorithm that uses WordNet and gazetteer databases to extract worldwide locations from text, lists the longitude and latitude for each location and improves the accuracy of location extraction. The thesis also introduces a weight-based topic modeling method that can be compared with LDA. Finally, this research emphasizes the text relevance between different files and visualizes that relevance through a graph.

1.4 Overview of Work

In this study, we utilized information retrieval and graph technology to implement a new, effective system for textual searching and document relevance ranking. Different techniques were adopted to solve the problems such as location disambiguation, relevance ranking and document topic extraction.

In Chapter 2, the tasks and elementary functions of this work are briefly introduced. They are divided into three main parts: locations, topics, and relevance ranking.

Chapter 3 describes the implementation of the geographical information part of the system in detail. It explains how we solved data processing issues, such as term ambiguity that occurs when locations are retrieved from raw text and coordinates are selected. In addition, it discusses how centroid and K-means clustering methods can be used to reduce the computational complexity in the network and improve data processing performance. The accuracy of two methods is evaluated.

Similar to Chapter 3, Chapter 4 specifies the models for topic retrieval. It introduces the architecture and algorithms, and provides a detailed description of LDA and word-clustering models. A discussion of the analysis of the performance of topic modeling is presented.

Chapter 5 provides an overview of how the document relevance can be measured. Based on locations and topics, the relevant documents of a queried file can be found. The related concepts and algorithms are represented.

Chapter 6 focuses on the implementation of the location-based system. The resources in the system are introduced, and the summary of a file, including locations and topics, is shown. A novel graphic illustration with nodes and edges is also represented to present the document relevance ranking in a more intuitive way.

Finally, Chapter 7 provides an evaluation of the system model. It presents the study conclusions and possible directions for further development.

2. DESIGN OF LOCATION-BASED SEARCH SYSTEM

This chapter introduces the requirements and elementary functions of the location-based search system designed for this study, which primarily consists of location processing and topic modeling. Depending on the text base, the system outputs the information in the queried document. The document relevance is also measured and illustrated.

The text files are preprocessed through natural language processing, and then locations are extracted from the files. Ambiguity, which is the major problem in this part of the process, will be discussed in detail in Chapter 3. Each document involves more than one topic, and based on the probabilities, the system returns a list of topics for each document. All the information is indexed for higher efficiency.

After all the locations are extracted and indexed, they are also used for query searches. The user is able to search files by a list of locations. The documents that contain all the locations are ranked.

Finally, the system measures the document relevance in three terms: location relevance, topic relevance, and the linear combination of the location and topic relevance. Graphic illustration works as a visualization of document relevance. The system obtains the statistic data of the document relevance and returns a relationship graph. The graph has files as nodes and relevance as edges.

In summary, this thesis implements a novel location-based system. The architecture is shown in Figure 2.1. Using an input text database, the system returns the topic/information and gives access to the relevance ranking statistically or graphically.

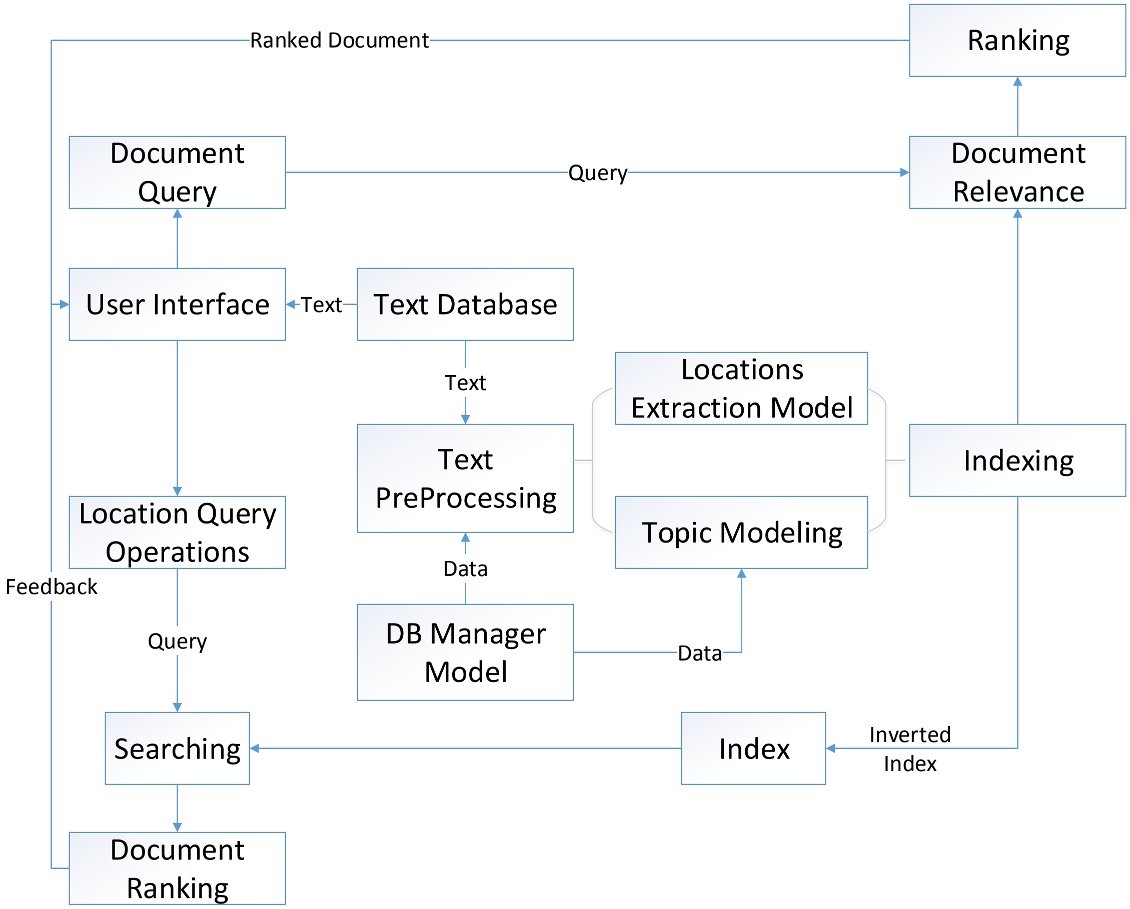


Figure 2.1: Design of Location Based System

3. ARCHITECTURE OF LOCATION BASED SEARCH SYSTEM

This chapter introduces the architecture and algorithms used in location extraction. After the unstructured text is processed and labeled, the user can extract geographical information from and eliminate the ambiguity for searches and rankings.To illustrate this process, Section 3.1 describes word processing, Section 3.2 presents a new location disambiguation method, and Section 3.3 explains the process for indexing all the locations in the files, and then discusses the performance of the architecture. Section 3.3 also examines the accuracy improvement compared with previous work. Figure 3.1 presents the entire process, including location extraction and disambiguation.

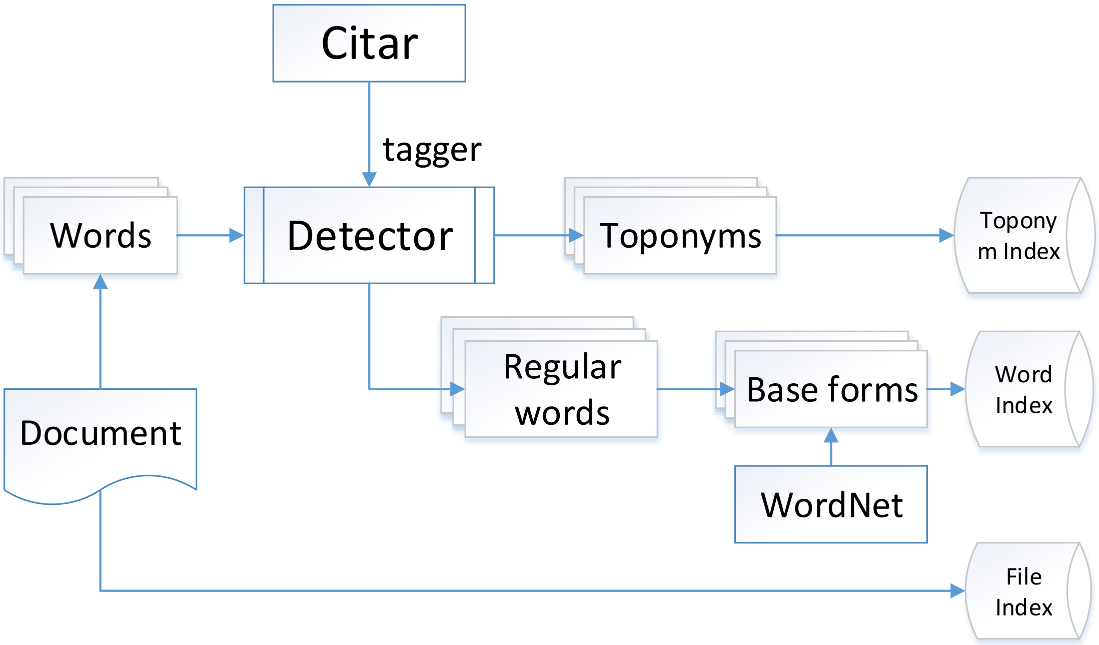


Figure 3.1: The Overall Structure of Location Indexing

3.1 Word Processing

In natural language, some words have different forms in different situations, such as “is” and “was”, but they carry the same meaning “is” in this case. The different forms increase the actual number of words in a file and thus need to be eliminated. However, many words are unable to settle their exact base form without additional information. For instance, “living” might refer to the noun form, which means the act or condition of a person or thing that lives, or to the present participle of the verb “live”. The meaning is decided by the sentence. Therefore, the files in the system must be preprocessed and the words are transformed to their basic forms according to the context of every word. After that, we can find out all the potential proper nouns that are most likely to be location names and disambiguate the location candidates.

3.1.1 Word Processing Resources

To transform every word to its base form, we need to assign part-of-speech taggers to the words and transform the words according to different word type requirements. We can use WordNet and Brown Corpus Taggers to implement the function.

3.1.1.1 Part-of-Speech Tagging

Part-of-speech tagging (POS tagging, or POST) [25] is also called word-category disambiguation or grammatical tagging. It marks up words through their definitions and the context in a text (corpus). POS tagging falls into two categories: rule-based and stochastic. This thesis uses the Brown Corpus, or Brown University Standard Corpus of Present-Day American English as a general standard for corpus linguistics tagging [26]. Words are assigned with parts of speech taggers, such as noun, verb, and adjective.

Many POS taggers using Brown Corpus are available now. Citar is a simple tagger used to mark English words with Brown Corpus standards. It is based on a trigram hidden Markov model (HMM). For instance, in the sentence “During 2007 the WDCS funded 32 conservation and research projects”, which was extracted from Wikipedia, Citar POS tagging tags each word in the sentence as “IN, CD, AT, NN, VBD, CD, NN, CC, NN, NNS”, which means noun, verb, preposition, cardinal numeral, and so on.

3.1.1.2 WordNet

WordNet is a large lexical database for English words [27]. The database is a free and greatly useful tool for natural language processing and computational linguistics.

The words in WordNet are connected with other words by forms and tenses. Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets), and each group expresses a distinct concept. The synsets in WordNet are interlinked through semantic relations. These kinds of semantic relations can be applied to all words in a synset, as they share the same meaning. For instance, the relations based on noun type are divided into hypernyms, synonyms, meronyms, hyponyms and holonyms. Furthermoew, WordNet provides a short, brief definition of words (gloss), as well as some general short sentences to illustrate the usage of the synset members. In this way, WordNet produces a meaningful word network with related words and concepts. It aims to produce an intuitively usable combination of dictionary, and supporting natural language processing and textual analysis applications, such as search engines. [28]

3.1.2 Finding Proper Nouns

WordNet and Citar work together to find out proper nouns. Since POS tagging tags all the contained words at the sentence level, files are divided into sentences. Then each word is tagged by POS taggers. Proper nouns are difficult to identify because they may be composed of several independent words. Making things even more complicated, there are no fixed rules regarding the types or forms of words for proper nouns, especially for location names. For instance, the two words “united” and “states” that compose “United States” are tagged as adjective and plural noun separately. How could the system understand that “United States” is a single word? Fortunately, the initials of most proper nouns are capitalized, thus, they could be a symbol of proper nouns and be further analyzed to settle the locations. Combining the above methods, the proper nouns are extracted and judged. If determined to be a proper noun, a word is kept in its original form.

In addition, words are also transformed into base forms by the morph function in WordNet for future use in topic modeling. Take the sentence “During 2007 the WDCS funded 32 conservation and research projects” as an example. After preprocessing, it is transformed into “during 2007 the WDCS fund 32 conservation and research project.”

Suppose Document O consists of n sentences: $S_1, S_2, ..., S_n$. The algorithm for word preprocessing is shown in Algorithm 1.

3.2 Disambiguation

During the process of retrieving proper nouns and finding the proper nouns location names, as described in the previous section, many garbage words (non-related to locations) and ambiguity such as geo/non-geo and geo/geo words appear, thus causing problems with ambiguity. Therefore, another step called disambiguation must occur. Disambiguation consists of two parts: deletion of all the non-location words and retrieval of the location coordinates. The architecture of this process is shown in Figure 3.2. WordNet and gazetteers can be used, and the chosen disambiguator retrieves the longitude/latitude of the location candidates.

Algorithm 1 Word Preprocessing

```
INITIALIZE set of proper nouns $PN = \emptyset$
for each sentence $S_i$ in document O do
    j=0
    while j < length of words in $S_i$ do
        get part-of-speech tag $t_{ij}$ with Citar
        if  $t_{ij}$ belongs to NOUN,ADV, ADJ, VERB then
            assign 1 - 4 to $t'_{ij}$ accordingly
        else
            $t'_{ij}$=0
        end if
        if j is end of possible proper noun $p_n$ with first capital letter then
            $PN = PN \cup \{p_n\}$
        end if
        j++
    end while
end for
```

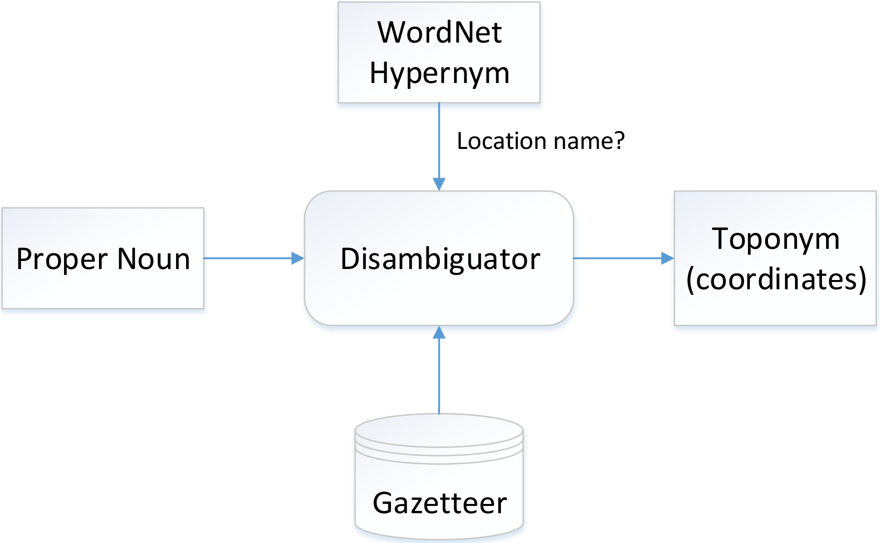
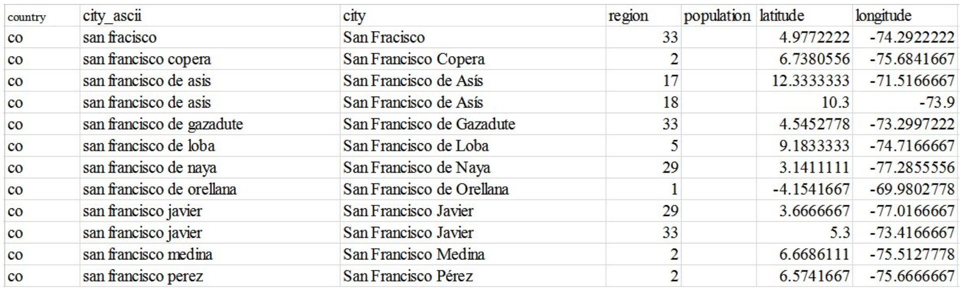


Figure 3.2: Process of Disambiguation

3.2.1 Gazetteers

A gazetteer is a geographical dictionary that includes an object's social statistics and physical features. Each entry in a gazetteer is associated with its location, coordinates, dimensions of peaks and waterways, and population, as well as some other information. Figure 3.3 shows an example of the gazetteer.



country	city_ascii	city	region	population	latitude	longitude
co	san fracisco	San Fracisco	33		4.9772222	-74.2922222
co	san francisco copera	San Francisco Copera	2		6.7380556	-75.6841667
co	san francisco de asis	San Francisco de Asís	17		12.3333333	-71.5166667
co	san francisco de asis	San Francisco de Asís	18		10.3	-73.9
co	san francisco de gazadute	San Francisco de Gazadute	33		4.5452778	-73.2997222
co	san francisco de loba	San Francisco de Loba	5		9.1833333	-74.7166667
co	san francisco de naya	San Francisco de Naya	29		3.1411111	-77.2855556
co	san francisco de orellana	San Francisco de Orellana	1		-4.1541667	-69.9802778
co	san francisco javier	San Francisco Javier	29		3.6666667	-77.0166667
co	san francisco javier	San Francisco Javier	33		5.3	-73.4166667
co	san francisco medina	San Francisco Medina	2		6.6686111	-75.5127778
co	san francisco perez	San Francisco Pérez	2		6.5741667	-75.6666667

Figure 3.3: Example of the Gazetteer

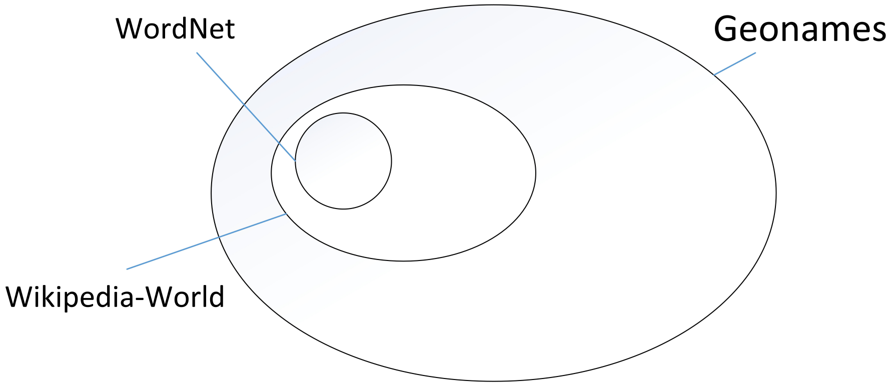


Figure 3.4: Gazetteers Comparison

Some of the available gazetteers are Geonames [29], World Gazetteer, and NGA

GEOnet Names Server. These resources contain different amounts of toponym coverage. Figure 3.4 compares the toponym coverage of different resources. Since WordNet, which has been discussed in previous section as a database for English words, involves some location names, we also include it here for comparison.

Geonames has the largest number of locations, while WordNet has the smallest. The coverages for WordNet and Geonames are listed in Table 3.1.

Table 3.1: Coverage of Different Gazetteers

Type	Name	Coverage
Gazetteer	Geonames	7,000,000
Ontologies	WordNet	2,188

Larger coverage means more ambiguity. For instance, "Beijing" has four references in Geonames, all of which are in different provinces in China, as Table 3.2 indicates.

Table 3.2: List of Beijings from Gazetteers

Code	Country	Name	Region code	Latitude	Longitude
427931	cn	Beijing	03	29.3464	116.199
427932	cn	Beijing	19	39.8825	123.912
427933	cn	Beijing	22	39.9289	116.388
427934	cn	Beijing	24	35.2092	110.733

The comparison of the toponym ambiguity shown in Table 3.3 illustrates that WordNet has the least ambiguity. Almost every location has a single meaning. However, WordNet has the disadvantage of fewer locations. Therefore, using data from

Table 3.3: Ambiguity of Gazetteers

	Unique names	References	Ambiguity ratio
Wikipedia (Geo)	180, 086	264, 288	1.47
Geonames	2, 954, 695	3, 988, 360	1.35
WordNet2.0	2, 069	2, 188	1.06

Geonames in combination with WordNet is the most effective method to eliminate ambiguity. A criteria is assigned to match the locations between the WordNet and Geonames databases, which allows the user to reduce the amount of possible ambiguity.

3.2.2 Removing Impossible Words

The first step to eliminate impossible words is to identify the meaning of all the proper nouns. The location names within WordNet are picked, and the chosen gazetteer is used to verify whether the proper nouns that are not contained in WordNet are locations. If a proper noun is not in the gazetteer, it is removed from the location candidate list. The next task involves finding the candidates' coordinates.

3.2.3 Coordinates Identification

The next step is to select the longitude/latitude for every location candidate. For each candidate, the chosen gazetteer returns a list of possible coordinates. If a location's meaning is confirmed, its longitude and latitude are settled. However, for the locations with ambiguity, further work must be done, using the two algorithms introduced below.

3.2.3.1 Centroid Algorithm

The locations in a document are always related to each other. They are likely to be in the same district. For instance, all the locations may be in the United States.

In this case, we would use this attribute to settle the coordinates. The geographical centroid of all the locations in a document must be calculated to identify the exact longitude/latitude of each location name. The procedure is as follows.

1. Get the coordinates of all the possible meanings of toponym t from the gazetteer , with geographical centroid set $C_t = \{c_{1t}, ..., c_{mt}\}$, and the meaning set $M_t, M = \{M_{t_1}, ..., M_{t_k}\}$.

2. Calculate the average centroid $\widehat{c}$.

3. Remove all the ambiguous points that are far from the centroid, for instance, $\sigma_{c_{it}} > \lambda\widehat{\sigma}$, then return to 2.

4. If no point is removed from 3, the distance from $\widehat{c}$ to any term in C_t is calculated.

5. Select c_{it} with minimum distance, and set it as the position of toponym t.

The centroid algorithm is shown in Algorithm 2.

For instance, consider the following text extracted from *National Geographic.*

“Charleston is brimming with art galleries, many of which are open to the public free of charge (for a complete list of galleries, click here). For some local heritage, stop by the Gallery Chuma located at 43 John Street. Chuma specializes in the art of the Gullah people. The Gullah are descendants of enslaved Africans who settled on the isolated barrier islands between Jacksonville, Florida, and Wilmington, North Carolina.”[30]

After preprocessing the paragraph, we get a simplified paragraph with less interference, along with some proper nouns, including Charleston, Gallery Chuma, John Street, Chuma, Gullah, The Gullah, Jacksonville, Florida, Wilmington, and North Carolina. WordNet and a gazetteer can be used to extract possible geographical names and their coordinates. There are five possible locations in the paragraph:

Algorithm 2 Centroid Disambiguation

```
INITIALIZE set of toponyms $T = \{t_1, ..., t_k\}$
for each toponym t in the set T do
    extract all the possible location coordinates from database, where the coordinate set is $C_t = \{c_{1t}, c_{2t}, ...c_{mt}\}$
end for
while true do
    Calculate the average centroid of all the coordinates of all the locations $\widehat{c}$
    if location name t contains ambiguity and one of its possible point $c_{it}$ is far from the centroid, $\sigma_k > \lambda\widehat{\sigma}$ then
        remove $c_{it}$
    end if
    if no point is removed then
        break
    end if
end while
for each toponym t in the set T do
    calculate the distances from $\widehat{c}$ to $c_{i_k t_k}$, Select $t_k$ with minimum distance $|\widehat{c} - c_{i_k t_k}|$
end for
return the toponym list
```

Charleston, Jacksonville, Florida, North Carolina, and Wilmington. Florida and North Carolina are settled first. Jacksonville and Wilmington are in these states, so they are also settled. However, there is ambiguity for other words, specifically Charleston in this case. According to WordNet, Charleston has two possible senses [31]:

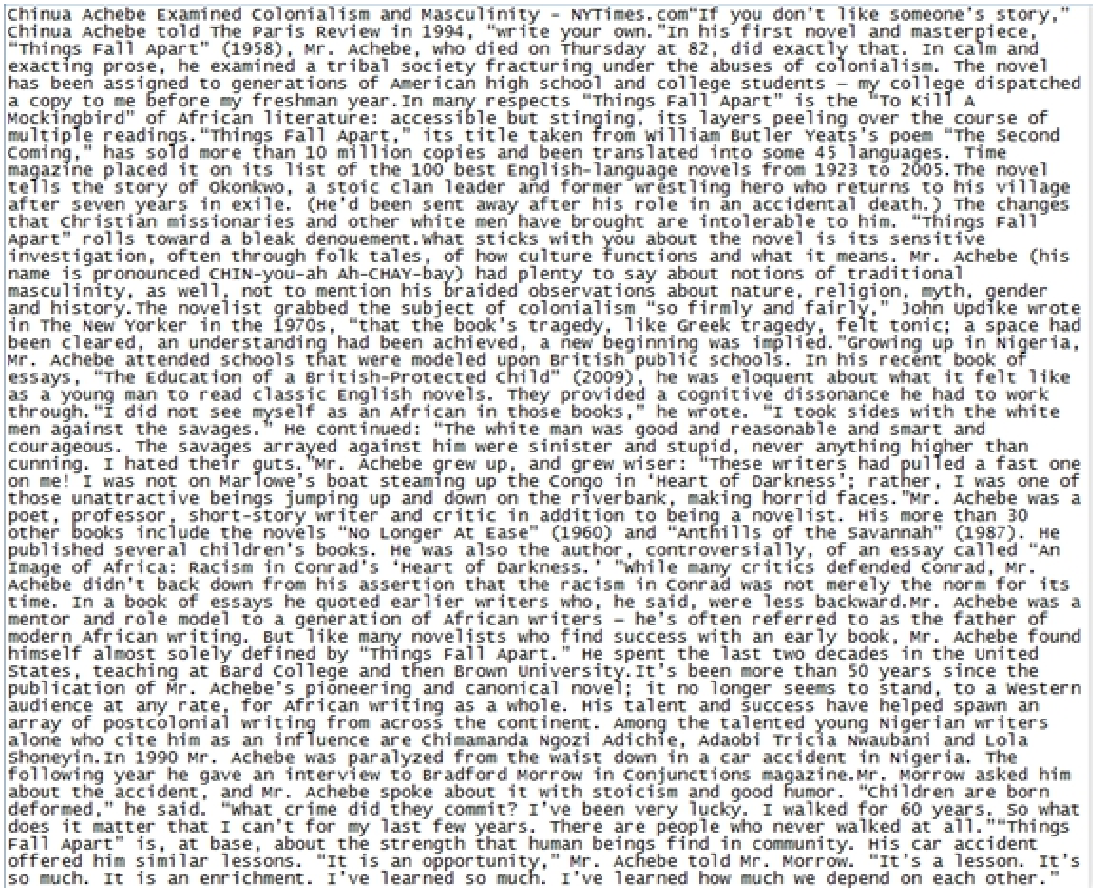
"1. Charleston, capital of West Virginia (state capital of West Virginia in the central part of the state on the Kanawha river)

2. Charleston (a port city in southeastern South Carolina). "

Next, the centroid algorithm should be applied to find the coordinates of all the location candidates and calculate the average centroid c. The distances of all the locations from the centroid and their standard deviations are computed. Since the possible meaning of Charleston, West Virginia, is farther from the centroid, it is eliminated. Therefore, Charleston, South Carolina, is selected and disambiguation is achieved.

3.2.3.2 K-Means Clustering Algorithm

Unfortunately, the centroid algorithm has some problems. Consider the file shown in Figure 3.5. It is a news article from the *New York Times* titled “Chinua Achebe Examined Colonialism and Masculinity”.



Chinua Achebe Examined Colonialism and Masculinity - NYTimes.com"If you don't like someone's story," Chinua Achebe told The Paris Review in 1994, “write your own."In his first novel and masterpiece, “Things Fall Apart" (1958), Mr. Achebe, who died on Thursday at 82, did exactly that. In calm and exacting prose, he examined a tribal society fracturing under the abuses of colonialism. The novel has been assigned to generations of American high school and college students – my college dispatched a copy to me before my freshman year.In many respects “Things Fall Apart" is the “To Kill A Mockingbird" of African literature: accessible but stinging, its layers peeling over the course of multiple readings.“Things Fall Apart," its title taken from William Butler Yeats's poem “The Second Coming," has sold more than 10 million copies and been translated into some 45 languages. Time magazine placed it on its list of the 100 best English-language novels from 1923 to 2005.The novel tells the story of Okonkwo, a stoic clan leader and former wrestling hero who returns to his village after seven years in exile. (He'd been sent away after his role in an accidental death.) The changes that Christian missionaries and other white men have brought are intolerable to him. “Things Fall Apart" rolls toward a bleak denouement.What sticks with you about the novel is its sensitive investigation, often through folk tales, of how culture functions and what it means. Mr. Achebe (his name is pronounced CHIN-you-ah Ah-CHAY-bay) had plenty to say about notions of traditional masculinity, as well, not to mention his braided observations about nature, religion, myth, gender and history.The novelist grabbed the subject of colonialism "so firmly and fairly," John Updike wrote in The New Yorker in the 1970s, “that the book's tragedy, like Greek tragedy, felt tonic; a space had been cleared, an understanding had been achieved, a new beginning was implied."Growing up in Nigeria, Mr. Achebe attended schools that were modeled upon British public schools. In his recent book of essays, “The Education of a British-Protected Child" (2009), he was eloquent about what it felt like as a young man to read classic English novels. They provided a cognitive dissonance he had to work through.“I did not see myself as an African in those books," he wrote. “I took sides with the white men against the savages." He continued: “The white man was good and reasonable and smart and courageous. The savages arrayed against him were sinister and stupid, never anything higher than cunning. I hated their guts."Mr. Achebe grew up, and grew wiser: “These writers had pulled a fast one on me! I was not on Marlowe's boat steaming up the Congo in 'Heart of Darkness'; rather, I was one of those unattractive beings jumping up and down on the riverbank, making horrid faces."Mr. Achebe was a poet, professor, short-story writer and critic in addition to being a novelist. His more than 30 other books include the novels “No Longer At Ease" (1960) and “Anthills of the Savannah" (1987). He published several children's books. He was also the author, controversially, of an essay called “An Image of Africa: Racism in Conrad's ‘Heart of Darkness.’ "While many critics defended Conrad, Mr. Achebe didn't back down from his assertion that the racism in Conrad was not merely the norm for its time. In a book of essays he quoted earlier writers who, he said, were less backward.Mr. Achebe was a mentor and role model to a generation of African writers – he's often referred to as the father of modern African writing. But like many novelists who find success with an early book, Mr. Achebe found himself almost solely defined by "Things Fall Apart." He spent the last two decades in the United States, teaching at Bard College and then Brown University.It's been more than 50 years since the publication of Mr. Achebe's pioneering and canonical novel; it no longer seems to stand, to a Western audience at any rate, for African writing as a whole. His talent and success have helped spawn an array of postcolonial writing from across the continent. Among the talented young Nigerian writers alone who cite him as an influence are Chimamanda Ngozi Adichie, Adaobi Tricia Nwaubani and Lola Shoneyin.In 1990 Mr. Achebe was paralyzed from the waist down in a car accident in Nigeria. The following year he gave an interview to Bradford Morrow in Conjunctions magazine.Mr. Morrow asked him about the accident, and Mr. Achebe spoke about it with stoicism and good humor. “Children are born deformed," he said. “What crime did they commit? I've been very lucky. I walked for 60 years. So what does it matter that I can't for my last few years. There are people who never walked at all.""Things Fall Apart" is, at base, about the strength that human beings find in community. His car accident offered him similar lessons. "It is an opportunity," Mr. Achebe told Mr. Morrow. "It's a lesson. It's so much. It is an enrichment. I've learned so much. I've learned how much we depend on each other."

Figure 3.5: A Piece of News

In the article, there are four toponym names in total, as Table 3.4 indicates.

Table 3.4: Location Attributes

Name	Congo	Savannah	United States	Nigeria
Longitude	15	-81.1	-97	8
Latitude	8	32.0833	38	10

Table 3.5: Location List from Gazetteer

Name	Longitude	Latitude
Nigeria	8	10
Congo	15	-1
United States	-97	38
Savannah	-81.3	19.2667
Savannah	-81.1	32.0833

Their possible coordinates extracted from a gazetteer are listed in Table 3.5.

Three of the locations have only one pair of longitude and latitude for each. However, there are two places named Savannah in the world. If calculated using the centroid algorithm, the one with the shortest distance to the average centroid is chosen, which is $81.3°W, 19.2667°N$, but this is the wrong answer.

Therefore, a new K-means algorithm that takes advantage of a K-means cluster to divide the location candidates into several clusters should be used.

In the new algorithm, all the locations are extracted from the document, as Algorithm 3 indicates. Each location candidate has a list of coordinates. If k locations are unique, they are settled first. All the locations are divided into k clusters, with one unambiguous location for each cluster. Next, the centroid of each cluster is computed, and all the ambiguous locations are put into the nearest cluster according to the distance between the location and the centroid. More specifically,the place population parameter r for coordinate t_k is considered, together with t_k's weight. Therefore, a new distance $d'_{c_{it}}$ is derived from the original distance $d_{c_{it}}$. $d'_{c_{it}} = \frac{w_{it}\theta}{\log r_{c_{it}}} d_{c_{it}}$, where

θ is a constant parameter. The centroids and deviations are updated until all the locations are settled.

Algorithm 3 K-means Disambiguation

```
INITIALIZE set of toponyms $\{t_1, ..., t_k\}$, cluster $T = \emptyset$
for each toponym t in the toponym set do
    extract all the possible location coordinates from database, where the coordinate set is $C_t = \{c_{1t}, c_{2t}, ...c_{mt}\}$
    if there is only one element in $C_t$ then
        put that element in T, increment cluster number by 1
    end if
end for
if cluster number = 0 then
    suppose K=1.
end if
Calculate the average centroid $\widehat{c}$ for each cluster k
while elements in clusters are different from previous do
    for each toponym t, put point $c_{it}$ into cluster k where $d(c_{it}, k)$ is minimum, where $d_{c_{it}} = w_{it}\theta|c_{it} - \widehat{c}|/\log(r_{c_{it}})$
    get average $\widehat{c}$ of cluster k, calculate the variance $V_k$ of cluster to $\widehat{c}$
    Remove all the points that are far from $\widehat{c}$, when $\sigma_k > \lambda\widehat{\sigma}$, update $\widehat{c}$.
end while
return the toponym list in the cluster set
```

Consider the example in Figure 3.5, which has been discussed previously. Using the K-means clustering algorithm, the toponyms are divided into three clusters: Nigeria, Congo and United States. Applying the K-means clustering algorithm, Savannah is found to have longitude $81.1°W$ and latitude $32.0833°N$, since it is nearer to the United States cluster. This outcome is the same as the actual result. Therefore, the K-means clustering algorithm is more credible in this case, as it classifies each toponym in a more accurate way.

3.3 Retrieval of Locations in Documents

The next step involves mapping the documents to a list of locations. The co-occurrence is also recorded in the inverted index, as Figure 3.6 shows.

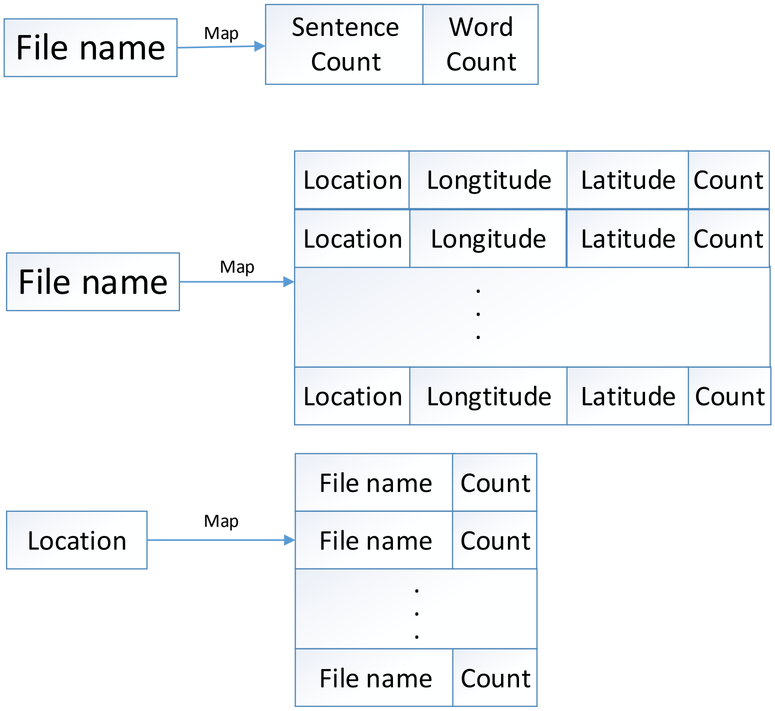


Figure 3.6: The Inverted Index

3.4 Location Search

Next, the location-based search engine is created. The user can search files based on the queried locations. The returning results are sorted by term frequency-inverse document frequency (tf-idf) using the following formula.

$$tfidf(d, D) = \prod_{t \in queries} tf(t, d) \times idf(t, D), \tag{3.1}$$

where

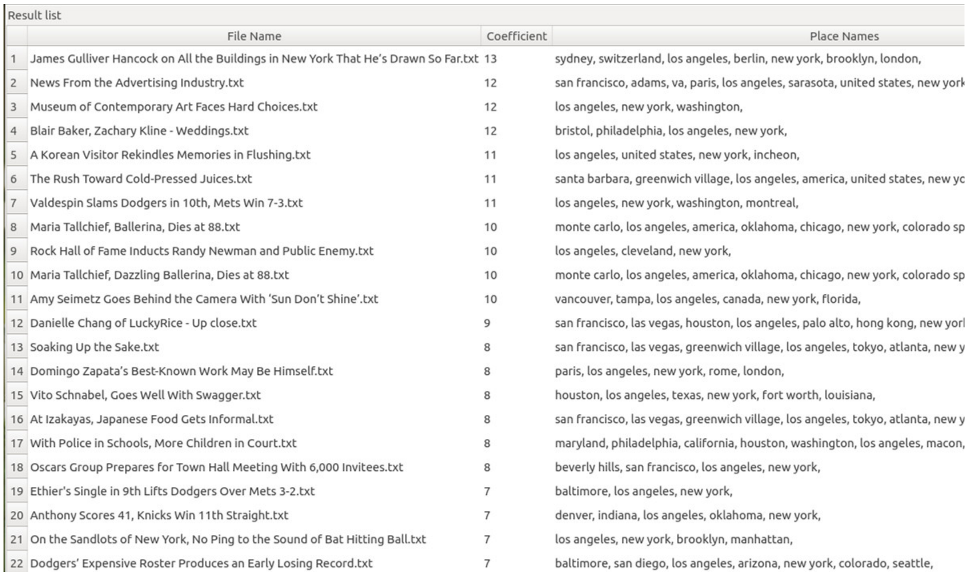
$$tf(t, d) = \log(f(t, d) + 1), \tag{3.2}$$

$$idf(t, D) = \log \frac{N}{1 + |\{d \in D : t \in d\}|}, \tag{3.3}$$

and t is a queried term, d is the document, and D is the corpus. While tf emphasizes the frequency a term occurs in a document, the inverse document frequency, idf, diminishes the weight of terms that appear frequently in a document set and increases the weight of terms that appear rarely.

The statistical data of tf-idf reflects how important each queried location is to a document in a collection of corpus, and also helps to control the issues created by some locations being more common than others.

Figure 3.7 shows an example of the location search result.



Result list

	File Name	Coefficient	Place Names
1	James Gulliver Hancock on All the Buildings in New York That He’s Drawn So Far.txt	13	sydney, switzerland, los angeles, berlin, new york, brooklyn, london,
2	News From the Advertising Industry.txt	12	san francisco, adams, va, paris, los angeles, sarasota, united states, new york
3	Museum of Contemporary Art Faces Hard Choices.txt	12	los angeles, new york, washington,
4	Blair Baker, Zachary Kline - Weddings.txt	12	bristol, philadelphia, los angeles, new york,
5	A Korean Visitor Rekindles Memories in Flushing.txt	11	los angeles, united states, new york, incheon,
6	The Rush Toward Cold-Pressed Juices.txt	11	santa barbara, greenwich village, los angeles, america, united states, new yo
7	Valdespin Slams Dodgers in 10th, Mets Win 7-3.txt	11	los angeles, new york, washington, montreal,
8	Maria Tallchief, Ballerina, Dies at 88.txt	10	monte carlo, los angeles, america, oklahoma, chicago, new york, colorado sp
9	Rock Hall of Fame Inducts Randy Newman and Public Enemy.txt	10	los angeles, cleveland, new york,
10	Maria Tallchief, Dazzling Ballerina, Dies at 88.txt	10	monte carlo, los angeles, america, oklahoma, chicago, new york, colorado sp
11	Amy Seimetz Goes Behind the Camera With ‘Sun Don’t Shine’.txt	10	vancouver, tampa, los angeles, canada, new york, florida,
12	Danielle Chang of LuckyRice - Up close.txt	9	san francisco, las vegas, houston, los angeles, palo alto, hong kong, new yorl
13	Soaking Up the Sake.txt	8	san francisco, las vegas, greenwich village, los angeles, tokyo, atlanta, new y
14	Domingo Zapata’s Best-Known Work May Be Himself.txt	8	paris, los angeles, new york, rome, london,
15	Vito Schnabel, Goes Well With Swagger.txt	8	houston, los angeles, texas, new york, fort worth, louisiana,
16	At Izakayas, Japanese Food Gets Informal.txt	8	san francisco, las vegas, greenwich village, los angeles, tokyo, atlanta, new y
17	With Police in Schools, More Children in Court.txt	8	maryland, philadelphia, california, houston, washington, los angeles, macon,
18	Oscars Group Prepares for Town Hall Meeting With 6,000 Invitees.txt	8	beverly hills, san francisco, los angeles, new york,
19	Ethier's Single in 9th Lifts Dodgers Over Mets 3-2.txt	7	baltimore, los angeles, new york,
20	Anthony Scores 41, Knicks Win 11th Straight.txt	7	denver, indiana, los angeles, oklahoma, new york,
21	On the Sandlots of New York, No Ping to the Sound of Bat Hitting Ball.txt	7	los angeles, new york, brooklyn, manhattan,
22	Dodgers’ Expensive Roster Produces an Early Losing Record.txt	7	baltimore, san diego, los angeles, arizona, new york, colorado, seattle,

Figure 3.7: Example of Location Search Result

3.5 Performance of Location Extraction

To describe the correctness of location extraction, F-measure in information retrieval should be analyzed [32]. The F-measure consists of precision and recall. Precision is the fraction of retrieved documents that are relevant to the findings:

$$precision = \frac{|\{relevantdocuments\} \cup \{retrieveddocuments\}|}{|\{retrieveddocuments\}|} \tag{3.4}$$

Recall in information retrieval is the fraction of the documents that are relevant to the query and are successfully retrieved:

$$recall = \frac{|\{relevantdocuments\} \cup \{retrieveddocuments\}|}{|\{relevantdocuments\}|} \tag{3.5}$$

The F-measure score combines precision and recall by their harmonic mean. It measures the accuracy of a test, as follows:

$$F = 2 \cdot \frac{precision \times recall}{precision + recall} \tag{3.6}$$

Table 3.6 illustrates the accuracy of the location coordinates for the centroid algorithm and K-means algorithm.

The accuracy of the K-means algorithm is about 93%, which is consistent with our expectations discussed in Section 3.2. Because the locations are scattered across the world, it does not make sense to simply assume that all the locations are near the centroid of the locations. Compared with the centroid method, K-means clustering is more common, as people often mention places by districts in human linguistic expression. Therefore, the K-means algorithm is used in the following chapters.

The accuracy of the names of locations can also be measured by comparing the data with the results from the Stanford Named Entity Recognizer (NER). NER [33]

Table 3.6: Coordinate Accuracy in Location Extraction

Method	Accuracy
1	68.89%
2	93.62%

Table 3.7: Accuracy Comparison of Location Extraction

Method	Toponyms	F-measure	Recall	Precision
NER	N/A	85.09%	76.97%	95.14%
Centroid Method	WordNet	86.27 %	99.10%	76.39%
Centroid Method	WordNet & Gazetteer	83.54%	76.74%	91.67%
K-means	WordNet	84.62 %	99.07%	73.61%
K-means	WordNet & Gazetteer	85.62%	82.05%	89.51%

is a text-processing task in information extraction. It classifies sequences of words in a text into predefined categories, such as person names, locations, and organizations, always by means of grammar-based techniques or statistical models. Stanford NER is an open-source JAVA implementation of NER produced by Stanford University to recognize names of things in text.

In the experiments conducted as part of this study, only the names of the locations were checked. The cases that used WordNet toponyms were compared with the cases that used both WordNet and a gazetteer.

The results in Table 3.7 indicates that the F-measure of the K-means algorithm is slightly higher than that of NER, while they have different recall and precision scores. Because the K-means algorithm processes the locations based on their coordinates, it is more effective and more accurate than NER. Results also shows that the location names extracted from WordNet contain less geo/non-geo information and the recall value is very high. However, restricted by the number of locations in the database, WordNet misses many toponyms. In contrast, when using the combined databases of WordNet and a gazetteer, the algorithm returns locations with higher precision at the cost of a lower recall score.

4. TOPIC INFORMATION RETRIEVAL

In natural language processing systems, it is useful to categorize the words. This kind of word clustering solves the problem of data scarcity and reduces the dimensions of words. However, there is no natural method to classify the similarity between words. For instance, it is hard to know the extent of similarity between the word "car" and "bus", or the word "car" and "fix". Thus, how should we define the similarity between different words? We would like to regard them as similar words if they are exchangeable to some extent or refer to the same thing.

In this study, words are clustered into the same topic if they have some degree of similarity. Then each document is given a distribution of topics [35]. The commonly used LDA method is implemented and compared with a new weight-based method using WordNet.

4.1 Term Group Association

LDA has been the prominent model for representing text corpora or other large collections of data as a mixture of various topics. As a generative probabilistic model, LDA iterates a set of data to explain the characteristics of untrained groups and reveal why some parts of the data are similar. These kinds of relationships are useful for basic tasks such as summarization, relevance, and similarity identification.

4.1.1 Process of LDA

In LDA, each word is modeled as a finite mixture over a set of topic probabilities so that the documents can be represented by a mixture of latent topics.[2]

LDA assumes a uniform Dirichlet prior distribution [22]. The generative process for each document O in a corpus D looks as follows:

1. Choose $N \sim Possion(\varepsilon)$.

2. Choose $\theta \sim Dir(\alpha)$.

3. For each of N words w_n in the document, choose a topic $z_k \sim Multinomial(\theta)$, and $p(w_n|z_k, \beta)$ is a multinomial probability conditioned on the topic z_k.

Suppose there is a set of K topics $\boldsymbol{z}$, and a set of N words $\boldsymbol{w}$. θ is a k-dimensional Dirichlet random variable based on the parameter α.

In such a model, the probability that the word w_n instantiates term t is:

$$p(w_n = t) = \sum_k p(w_n = t|z = k)p(z = k), \sum_k p(z = k) = 1 \tag{4.1}$$

Each mixture component $p(w_n = t|z = k)$ is a multinomial distribution over terms and is related to the latent topic $z = k$ in the text.

Given the parameters α and β, the joint distribution of a topic mixture $\boldsymbol{\theta}$ is given as:

$$p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta), \tag{4.2}$$

where $p(z_n|\theta)$ is simply θ_i for the unique i such that $z_n^i = 1$. The marginal distribution of a document, $p(\boldsymbol{w}|\alpha, \beta)$, is calculated by summing all the topics z and integrating over θ:

$$p(\boldsymbol{w}|\alpha, \beta) = \int p(\theta|\alpha) \prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta)p(w_n|z_n, \beta)d\theta \tag{4.3}$$

Finally, the probability of a corpus can be determined by taking the product of the marginal probabilities of documents:

$$p(\boldsymbol{D}|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \prod_{n=1}^{N} \sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_n|z_{dn}, \beta)d\theta_d \tag{4.4}$$

This generative process is shown in Figure 4.1. A document is separated into a stream of words, $\boldsymbol{w}$. The parameters α and β are the corpus-level variables. For each of the documents, a topic proportion, θ, is produced. Then the words in the document are set topics according to the topic specific mixture proportion, z, which is at the word level. Therefore, words are drawn with topic distribution and topics are sampled for the entire corpus. Finally, documents are assigned with multiple topics after iterations.

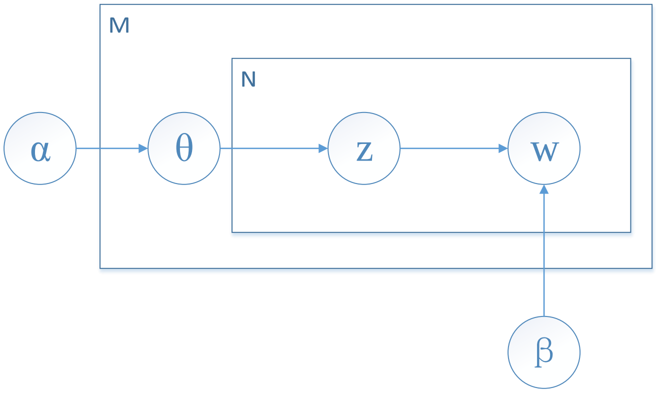


Figure 4.1: LDA Process

To solve the LDA problem, Gibbs sampling [34], which is based on the Monte Carlo simulation algorithm, is used [35].

The Gibbs sampling LDA algorithm is shown in Algorithm 4. After iterations, the topic distribution of each document and each word is settled.

To calculate the likelihood of word w_d in document m, define $\varphi_{k,t}$ as the probability distribution over a topic of each word, and $\theta_{m,k}$ as the topic mixture proportion for each document m. α and β are the model parameters as defined above.

Algorithm 4 Gibbs Sampling LDA in Topic Modeling

```
for each topic $k$ do
  sample mixture component $\varphi_k$
end for
for each document d do
  sample mixture proportion $\theta_d$ and document length $N_d$
  for each words $t_n$ do
    sample topic index $z_{d,k}$
  end for
end for
```

$$p(w_n = t|z = k) = \varphi_{k,t} = \frac{n_k^t + \beta_t}{\sum_t n_k^t + \beta_t} \tag{4.5}$$

$$p(z = k|d = m) = \theta_{m,k} = \frac{n_m^k + \alpha_k}{\sum_k n_m^k + \alpha_k} \tag{4.6}$$

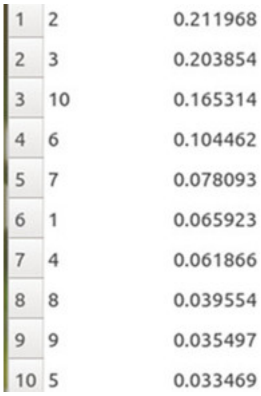
The likelihood of a word in a document of the test corpus $p(w_d)$ can be directly expressed as a function of the multinomial parameters:

$$p(w_d|d = m) = \prod_{n=1}^{N} \sum_{k=1}^{K} p(w_n = t|z = k)p(z = k|d = m) = \prod_{n=1}^{N} (\sum_{k=1}^{K} \varphi_{k,t}\theta_{m,k})^n \tag{4.7}$$

$$\log(p(w_d|d = m)) = \sum_{n=1}^{N} n_m^{(t)} \log(\sum_{k=1}^{K} \varphi_{k,t}\theta_{m,k}), \tag{4.8}$$

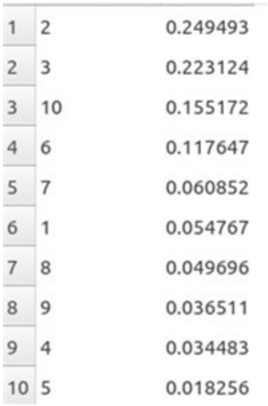
where $n_m^{(t)}$ is the occurrence of each term t in document m.

Figure 4.2 and Figure 4.3 show the topic distributions of two documents: “Yvonne Brill, a Pioneering Rocket Scientist, Dies at 88” and “Yvonne Brill, Rocket Scientist, Dies at 88” from the *New York Times*. They have similar topics.



1	2	0.211968
2	3	0.203854
3	10	0.165314
4	6	0.104462
5	7	0.078093
6	1	0.065923
7	4	0.061866
8	8	0.039554
9	9	0.035497
10	5	0.033469

Figure 4.2: Yvonne Brill, a Pioneering Rocket Scientist, Dies at 88



1	2	0.249493
2	3	0.223124
3	10	0.155172
4	6	0.117647
5	7	0.060852
6	1	0.054767
7	8	0.049696
8	9	0.036511
9	4	0.034483
10	5	0.018256

Figure 4.3: Yvonne Brill, Rocket Scientist, Dies at 88

4.1.2 Performance Measurement

To analyze the performance of the LDA modeling, the perplexity of the corpus is computed. This computation is a log likelihood value used to measure how a probability distribution predicts a sample. In natural language processing, the perplexity also evaluates the generalization performance of unseen data. Similar to entropy, a lower perplexity indicates better generalization performance.

$$perplexity(D) = \exp\,(-\frac{\sum_{m=1}^{M} \log(p(w_m))}{\sum_{m=1}^{M} N_m}), \tag{4.9}$$

where $\log(p(w_m))$ is retrieved by Equation (4.8).

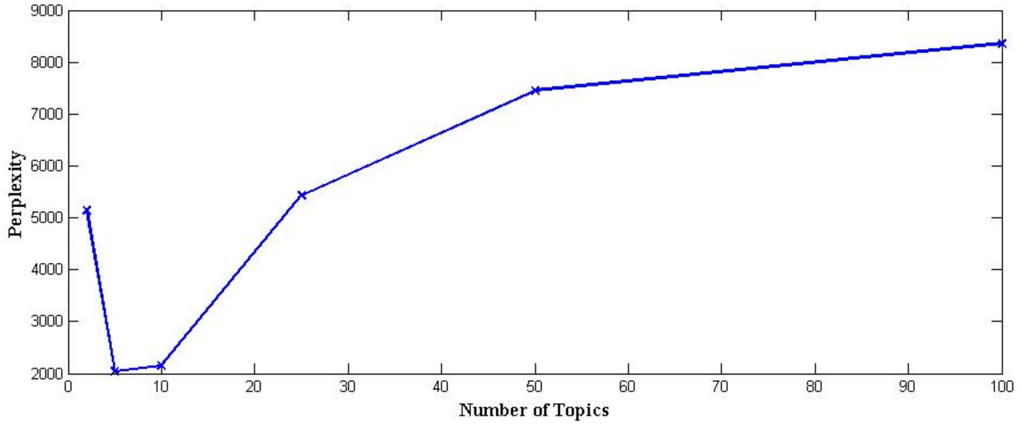


Figure 4.4: Perplexity of LDA on Different Numbers of Topics

In the experiments conducted as part of this study, we tested around 200 documents from the *New York Times*, and iterated 630 times. In the tests, α was 25.0 and β was 0.1. Figure 4.4 indicates the relationship between the topic number and the perplexity. When the topic number increases, the perplexity first decreases, and then increases. The performance of LDA is the best when the topic number is about 5 to 10. In this value range, the document topics achieves the best generalization.

4.2 Term Similarity Measure

Term similarity measurement, which groups words into clusters based on their similarity in meanings, is another popular technique in topic modeling.

This study takes advantage of synsets in WordNet to extract the topics of documents. Different super-subordinate relations are employed to group the terms in the document into several different bags of words. Each bag of words returns a topic, and the topic's probability is evaluated by its weight. The process is described in the following subsections.

4.2.1 Word Clustering

Specifically, synonym/antonym, hyponym, hypernym, and meronym sets in WordNet describe the lexical hierarchies among sequences of related words. This kind of lexical chain captures the cohesion of the text or corpus [36]. For instance, both "politics" and "government" lead to the same topic: government#2, governing#1, governance#2, government activity#1, and administration#5. These words are clustered together into the same category: "Politics".

In this study, different weights are assigned to different synsets according to the similarity levels between the words, such as meronyms, hypernyms, and synonyms. Examples are shown in Table 4.1. The co-occurrence of each term is also taken into consideration. Summing all the relations of the lexical chain produces the weight of every topic. For the word cluster shown in Figure 4.5, the score is equal to 11, since there is one hypernym relation and two synonym relations.

Table 4.1: Synset Weights

Type	Description	Weight
Meronyms	Member-of/ Has-a/ Part-of	2
Hypernyms/Hyponyms	Generalization/Specialization	3
synonyms	Same meaning	4

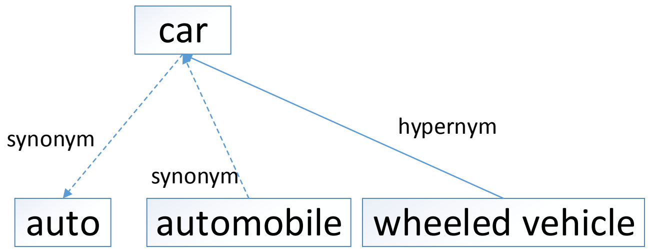


Figure 4.5: WordNet Lexical Chain

4.2.2 Topic Classification Based on WordNet

In the process of determining topic classification based on WordNet, we simply set 10 topics in advance and classify the documents using those 10 topics. The topic classification is extracted from the *New York Times* and included the following:

WEATHER, MILITARY, POLITICS, BUSINESS, EDUCATION, SCIENCE& TECHNOLOGY, HEALTH, SPORTS, ARTS & CULTURE, FASHION & SHOW

To get the topic classification, the terms in the document are categorized into one of the topics according to the synsets in WordNet. The weight of each topic is summed and the likelihood of each topic in a document is calculated by the ratio between the topic weight and the document length. The topic with the highest likelihood probability is considered the major topic of the document.

The algorithm we use is shown in Algorithm 5.

For example, take the document named "Art and Technology - A Clash of Cultures." from the *New York Times* as an example of an art and culture topic. The weight of each topic is listed in Table 4.2.

Algorithm 5 Topic Extraction Relevance

```
for each topic $t_n$ do
  use WordNet synset to extend to a topic set of words $T_{t_n}$
end for
for each sentence $S_i$ in document O do
  for each word $w'_{ij}$ in sentence $S_i$ do
    if $w'_{ij}$ belongs to NOUN,VERB then
      get hypernym/synonym/meronym list of $w'_{ij}$
      for each hypernym/synonym/meronym do
        add it in the topic list by its weight if it belongs to the defined topics
      end for
    end if
    for each topic $t_n$ do
      calculate the topic likelihood $p_n$
    end for
  end for
end for
```

Table 4.2: Likelihood of Topics

Topic	Weight
Weather	0
Military	0.00516129
Politics	0.0103226
Business	0.0490323
Education	0.00516129
Science & Technology	0.258065
Health	0
Sports	0.0129032
Arts	0.170323
Fashion	0.0154839

"Arts" has the highest weight, so the topic in this document is defined as arts, which is the same as the result that would be achieved through manual classification.

4.2.3 Performance

Table 4.3: Result Measurement

	In Topic	Not in Topic
In topic(system)	(1)	(2)
Not in topic(system)	(3)	(4)

To evaluate the performance of the WordNet-based topic modeling algorithm, this thesis specifies the F-measure. The results are divided into four parts according to the system topic and the measured topic list.

Using the data in Table 4.3, recall and precision were calculated as:

$$Recall = \frac{(1)}{[(1) + (3)]} \tag{4.10}$$

$$Precision = \frac{(1)}{[(1) + (2)]} \tag{4.11}$$

We categorized the documents extracted from the *New York Times* by the 10 predefined topics, and selected the topic with highest weight as the top one for each document. By computing the results calculated from the WordNet-based algorithm with the manually defined topics, we determined the recall and precision values, which are listed in Table 4.4.

The F-measure, which is the accuracy rate for the top topic, is about 87%.

Table 4.4: Accuracy of Major Topic Based on WordNet

F-measure	Recall	Precision
0.87	0.83	0.91

Table 4.5: Accuracy of Topic Modeling Method Based on WordNet

F-measure	Recall	Precision
0.89	0.83	0.96

Table 4.5 combines the accuracy of all 10 topics. All of the topic-related information was taken into account, as long as it was mentioned in the document. However, it was difficult to define the ratios of different topics in a file, so we simply calculated the recall and precision values based on whether or not a topic appears. The accuracy result is relatively high. For general topic modeling methods, the overall F-measure is around 0.8.

5. MEASUREMENT OF DOCUMENT RELEVANCE

In general, the relevance rankings of documents are computed using term frequency weighting and probability distribution. Each matching document is scored and ranked according to its similarity with the queries in a search engine.

In this study, the document-level relevance is considered and some smoothing methods are applied to measure the relevance ranking between documents.

5.1 Measurement in Relevance Ranking

There are many methods for similarity ranking, such as Jaccard indexing and dice coefficient. These methods use the minimum or maximum of the two elements to measure their similarity. Some measurements calculate the distance between the elements, such as Manhattan distance and Euclidean distance, as illustrated below.

Manhattan Distance:

$$distance(\vec{x}, \vec{y}) = \sum_{i=1}^{N} |x_i - y_i| \tag{5.1}$$

Euclidean Distance:

$$distance(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} \tag{5.2}$$

Vector space scoring is another widely used measurement in information retrieval. It compares the deviation of angles between vectors. The vectors are the representations of documents that capture the importance of terms in the documents. The similarity coefficient is determined by the magnitude between the normalized vectors with regard to the same information, and the overlapping area indicates similarity, as follows.

$$sim(x,y) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||} = \frac{\sum (x_i y_i)}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}, \tag{5.3}$$

Where $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors normalized by vector length. In document similarity, the vectors are word weights normalized by document length.

5.2 Document Relevance Based on Locations

In this thesis, after the locations are extracted in Chapter 3, the document relevance is measured by the location probability. To estimate the probability, smoothing is utilized [37]. Smoothing adjusts the maximum likelihood to compensate for data sparseness and makes the estimated language model more accurate.

5.2.1 Probability Smoothing

The likelihood of each location in a document can be retrieved using a smoothing method similar to Jelinek-Mercer. The Jelinek-Mercer method [38] is a popular smoothing method for large collections of documents in natural language models. The method adds some extra counts to each term in the collection. This kind of linear interpolation of the likelihood of terms helps to involve a non-zero probability to the words in a corpus and improves the accuracy of probability estimate.

The Jelinek-Mercer formula is as follows:

$$p_{ml}(w_i) = \frac{c(w_i)}{\sum_{w_i} c(w_i)}, \tag{5.4}$$

$$p_{interp}(w_i|w_{i-1}) = (1-\lambda)p_{ml}(w_i) + \lambda p_{ml}(w_i|w_{i-1}), \tag{5.5}$$

where λ is a smoothing parameter. Typically, $0 < \lambda < 1$. $p_{ml}(w_i)$ is the probability that i-th word appears, while $p(w_i|w_{i-1})$ is the probability that i-th word appears based on the (i-1)-th word.

In this approach, each document O in document set D is defined as a pair (loc, O), where loc is the list of locations in the document O. Then the smoothing method is applied, as follows:

$$\widehat{p}(t|doc) = (1-\lambda)p(t|doc) + \lambda\frac{tf(t, doc)}{tf|Coll|}, \tag{5.6}$$

where $\widehat{p}(t|doc)$ is the probability of location t in the doc, $p(t|doc)$ is the likelihood estimate of t in document doc, depending on the occurrences of all the locations in doc, $tf(t, doc)$ is the occurrence of location t in document O and $tf|doc|$ is the sum of co-occurrences of all the locations in the document, $tf|Coll|$ is the number of occurrences of location t in the collection Coll of D, $\frac{tf(t,doc)}{tf|Coll|}$ is the likelihood estimate of t in collection Coll, and λ is a smoothing coefficient to control the influence of the two parts in the Jelinek-Mercer method. The smoothing coefficient should be smaller than 0.5 so that the result depends more on the document itself. In this study, λ is 0.1.

Consider the same example presented in Section 3.2. Suppose Charleston appears seven times in document collection in total. In the document, it appears one time and there are five locations in all. Then {loc}={Charleston, Florida, North Carolina, Jacksonville and Wilmington}, tf(Charleston,O)=1, tf$|doc| = 5$, $|Coll| = 7$. Therefore, the probability of Charleston is calculated by $0.343 - 0.057\lambda$.

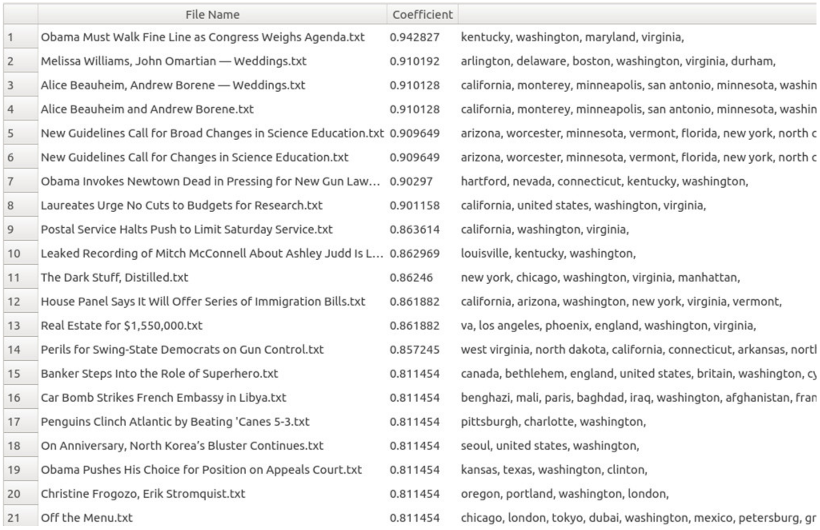
5.2.2 Document Ranking Based on Locations

The likelihood probabilities of the locations in a queried document is compared with other documents to determine the documents' similarity rankings [39].

Given a location list U_{loc} where $\{t_n \in loc\}$ and a queried document O, the location-based similarity between two documents is calculated by the vector space method, as shown below:

$$P(O'|O) = \frac{\sum\left(\widehat{p_{t_n}}\widehat{p'_{t_n}}\right)}{\sqrt{\sum \widehat{p_{t_n}}^2}\sqrt{\sum \widehat{p'_{t_n}}^2}}, \tag{5.7}$$

where O' is a targeting document to compare with O. $\widehat{p_{t_n}}$ is the likelihood of term t_n in O, and $\widehat{p'_{t_n}}$ is the likelihood of term t_n in O'.



	File Name	Coefficient	
1	Obama Must Walk Fine Line as Congress Weighs Agenda.txt	0.942827	kentucky, washington, maryland, virginia,
2	Melissa Williams, John Omartian — Weddings.txt	0.910192	arlington, delaware, boston, washington, virginia, durham,
3	Alice Beauheim, Andrew Borene — Weddings.txt	0.910128	california, monterey, minneapolis, san antonio, minnesota, washin
4	Alice Beauheim and Andrew Borene.txt	0.910128	california, monterey, minneapolis, san antonio, minnesota, washin
5	New Guidelines Call for Broad Changes in Science Education.txt	0.909649	arizona, worcester, minnesota, vermont, florida, new york, north c
6	New Guidelines Call for Changes in Science Education.txt	0.909649	arizona, worcester, minnesota, vermont, florida, new york, north c
7	Obama Invokes Newtown Dead in Pressing for New Gun Law…	0.90297	hartford, nevada, connecticut, kentucky, washington,
8	Laureates Urge No Cuts to Budgets for Research.txt	0.901158	california, united states, washington, virginia,
9	Postal Service Halts Push to Limit Saturday Service.txt	0.863614	california, washington, virginia,
10	Leaked Recording of Mitch McConnell About Ashley Judd Is L…	0.862969	louisville, kentucky, washington,
11	The Dark Stuff, Distilled.txt	0.86246	new york, chicago, washington, virginia, manhattan,
12	House Panel Says It Will Offer Series of Immigration Bills.txt	0.861882	california, arizona, washington, new york, virginia, vermont,
13	Real Estate for $1,550,000.txt	0.861882	va, los angeles, phoenix, england, washington, virginia,
14	Perils for Swing-State Democrats on Gun Control.txt	0.857245	west virginia, north dakota, california, connecticut, arkansas, nortl
15	Banker Steps Into the Role of Superhero.txt	0.811454	canada, bethlehem, england, united states, britain, washington, cy
16	Car Bomb Strikes French Embassy in Libya.txt	0.811454	benghazi, mali, paris, baghdad, iraq, washington, afghanistan, fran
17	Penguins Clinch Atlantic by Beating 'Canes 5-3.txt	0.811454	pittsburgh, charlotte, washington,
18	On Anniversary, North Korea's Bluster Continues.txt	0.811454	seoul, united states, washington,
19	Obama Pushes His Choice for Position on Appeals Court.txt	0.811454	kansas, texas, washington, clinton,
20	Christine Frogozo, Erik Stromquist.txt	0.811454	oregon, portland, washington, london,
21	Off the Menu.txt	0.811454	chicago, london, tokyo, dubai, washington, mexico, petersburg, gr

Figure 5.1: Rankings of Similar Dcuments Using Locations

Figure 5.1 shows an example of the similarity ranking between documents. The queried document contains Washington, Kentucky, and Virginia. The results are sorted.

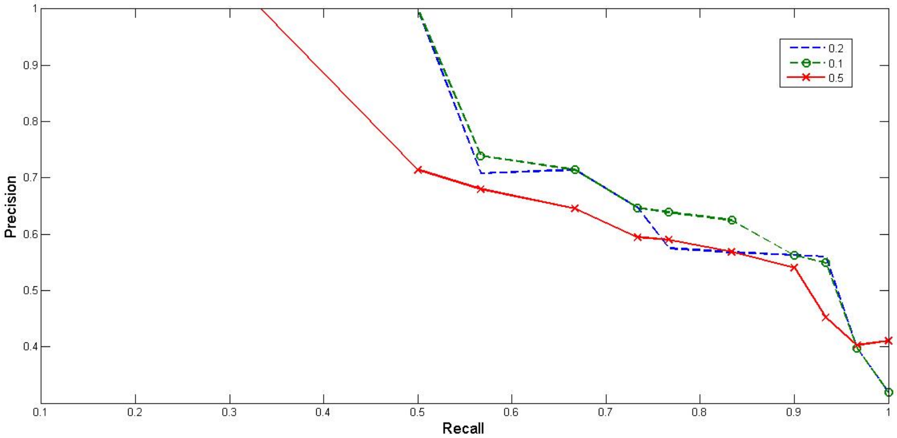


Figure 5.2: Recall/Precision Curve of Different Jelinek-Mercer Smoothing Parameters

Figure 5.2 illustrates the influence of different Jelinek-Mercer smoothing parameters on the recall/precision curve. When λ is smaller, the probabilities of the locations in the queried document have greater impact on the results than those in the text corpus. In other words, the relevance ranking depends on the document itself instead of the document collection. As a result, the recall/precision value rises when λ decreases.

5.3 Document Relevance Based on Topic Modeling

Next, this thesis analyzed the relevance ranking from the topic modeling aspect. The similarities based on the LDA method and term similarity measurements were studied and compared. The processes are described below.

5.3.1 Jensen-Shannon in LDA Topic Modeling

To estimate the likelihood probability of the LDA model, the Kullback Leibler (KL) divergence is used [40]. It is a standard function for divergence between normalized probability distributions. The KL method depends on the distribution divergence of topics. In the following the KL divergence function, $P = \{p_j | 0 < j \leq T\}$ and $Q = \{q_j | 0 < j \leq T\}$. There are two distribution lists.

$$D(p,q) = \sum_{j=1}^{T} p_j \log_2 \frac{p_j}{q_j}, \tag{5.8}$$

where $D(p,q)$ is the cross entropy or difference between the expected values of the probabilities P and Q, and is bounded by 0 and 1.

Unfortunately, the KL divergence has the disadvantage of asymmetry. It is undefined when $p_j \neq 0$ and $q_j = 0$. Therefore, another advanced measure must be introduced, the Jensen-Shannon distribution [41]. Jensen-Shannon takes the average of two symmetric KL divergences and is more smoothing and convincing:

$$JS(p,q) = \frac{1}{2}[D(p,(p+q)/2) + D(q,(p+q)/2)]. \tag{5.9}$$

To define document relevance based on LDA topic modeling, Jensen-Shannon divergence is used in this thesis, as follows:

$$P(O'|O) = \frac{1}{2}[\sum_{j=1}^{T} p_j \log_2 \frac{2p_j}{q_j + p_j} + \sum_{j=1}^{T} q_j \log_2 \frac{2q_j}{q_j + p_j}], \tag{5.10}$$

where p_i and q_i are the corresponding probabilities of each topic in document O and document O' separately. The relevance result is sorted in descending order. Because Jensen-Shannon divergence compares the mutual information between two mixture distributions, less entropy indicates more information; in other words, the two documents are more similar.

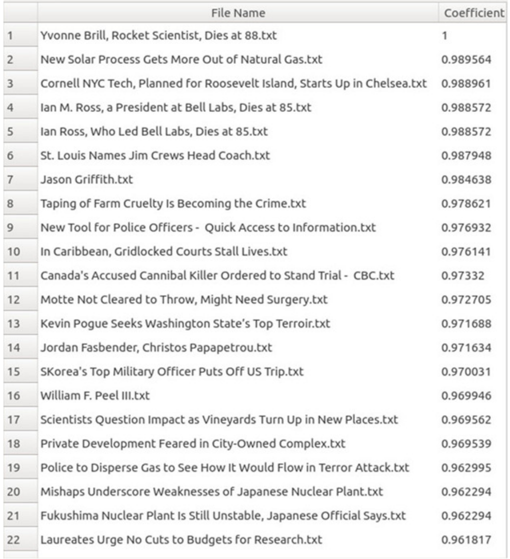
5.3.2 Vector Space in Topic Modeling

For WordNet-based term similarity measurements, the vector space model is applied to determine the likelihood for relevance ranking. Given a topic list Q =

$\{t_l | l \in 1, 2, ..n\}$ and a document O, the probability of document O' is

$$P(O'|O) = \frac{\sum(\widehat{t_l}\widehat{t'_l})}{\sqrt{\sum \widehat{t_l}^2}\sqrt{\sum \widehat{t'_l}^2}} \tag{5.11}$$

Figure 5.3 shows an example of relevance ranking based on WordNet evaluation. The relevance similarity is sorted in ascending order. Because the vector space calculates the cosine value of two vectors, a larger value means a closer relationship exists between the vectors, and the two documents are more similar.

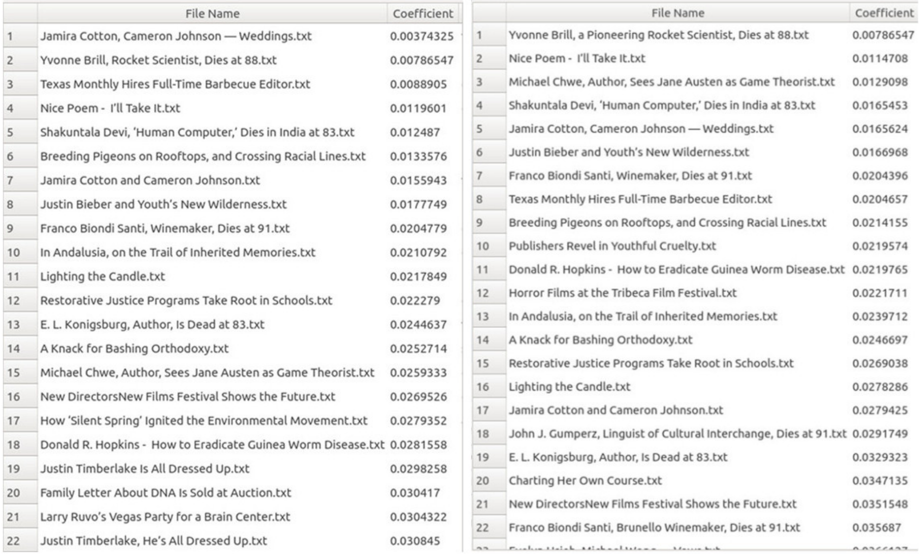


	File Name	Coefficient
1	Yvonne Brill, Rocket Scientist, Dies at 88.txt	1
2	New Solar Process Gets More Out of Natural Gas.txt	0.989564
3	Cornell NYC Tech, Planned for Roosevelt Island, Starts Up in Chelsea.txt	0.988961
4	Ian M. Ross, a President at Bell Labs, Dies at 85.txt	0.988572
5	Ian Ross, Who Led Bell Labs, Dies at 85.txt	0.988572
6	St. Louis Names Jim Crews Head Coach.txt	0.987948
7	Jason Griffith.txt	0.984638
8	Taping of Farm Cruelty Is Becoming the Crime.txt	0.978621
9	New Tool for Police Officers - Quick Access to Information.txt	0.976932
10	In Caribbean, Gridlocked Courts Stall Lives.txt	0.976141
11	Canada's Accused Cannibal Killer Ordered to Stand Trial - CBC.txt	0.97332
12	Motte Not Cleared to Throw, Might Need Surgery.txt	0.972705
13	Kevin Pogue Seeks Washington State's Top Terroir.txt	0.971688
14	Jordan Fasbender, Christos Papapetrou.txt	0.971634
15	SKorea's Top Military Officer Puts Off US Trip.txt	0.970031
16	William F. Peel III.txt	0.969946
17	Scientists Question Impact as Vineyards Turn Up in New Places.txt	0.969562
18	Private Development Feared in City-Owned Complex.txt	0.969539
19	Police to Disperse Gas to See How It Would Flow in Terror Attack.txt	0.962995
20	Mishaps Underscore Weaknesses of Japanese Nuclear Plant.txt	0.962294
21	Fukushima Nuclear Plant Is Still Unstable, Japanese Official Says.txt	0.962294
22	Laureates Urge No Cuts to Budgets for Research.txt	0.961817

Figure 5.3: Rankings of Similar Documents Using Topics

5.3.3 Comparison

To compare the similarity ranking performance of the two methods above, this study performed the experiments on two similar documents extracted from the *New York Times*. They are "Yvonne Brill, a Pioneering Rocket Scientist, Dies at 88." and "Yvonne Brill, Rocket Scientist, Dies at 88." The similarities for "Yvonne Brill, a Pioneering Rocket scientist, dies at 88." are shown in Figure 5.4. Figure 5.5 describes the results for "Yvonne Brill, Rocket Scientist, Dies at 88." The left slide shows the similarity ranking using LDA, while the right slide shows the WordNet method.

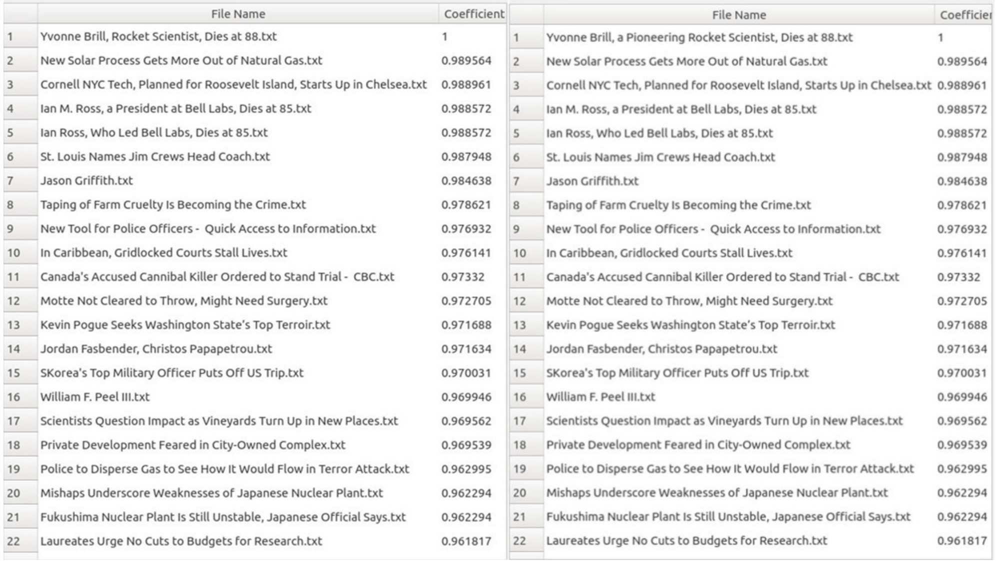


	File Name	Coefficient
1	Jamira Cotton, Cameron Johnson — Weddings.txt	0.00374325
2	Yvonne Brill, Rocket Scientist, Dies at 88.txt	0.00786547
3	Texas Monthly Hires Full-Time Barbecue Editor.txt	0.0088905
4	Nice Poem - I'll Take It.txt	0.0119601
5	Shakuntala Devi, 'Human Computer,' Dies in India at 83.txt	0.012487
6	Breeding Pigeons on Rooftops, and Crossing Racial Lines.txt	0.0133576
7	Jamira Cotton and Cameron Johnson.txt	0.0155943
8	Justin Bieber and Youth's New Wilderness.txt	0.0177749
9	Franco Biondi Santi, Winemaker, Dies at 91.txt	0.0204779
10	In Andalusia, on the Trail of Inherited Memories.txt	0.0210792
11	Lighting the Candle.txt	0.0217849
12	Restorative Justice Programs Take Root in Schools.txt	0.022279
13	E. L. Konigsburg, Author, Is Dead at 83.txt	0.0244637
14	A Knack for Bashing Orthodoxy.txt	0.0252714
15	Michael Chwe, Author, Sees Jane Austen as Game Theorist.txt	0.0259333
16	New DirectorsNew Films Festival Shows the Future.txt	0.0269526
17	How 'Silent Spring' Ignited the Environmental Movement.txt	0.0279352
18	Donald R. Hopkins - How to Eradicate Guinea Worm Disease.txt	0.0281558
19	Justin Timberlake Is All Dressed Up.txt	0.0298258
20	Family Letter About DNA Is Sold at Auction.txt	0.030417
21	Larry Ruvo's Vegas Party for a Brain Center.txt	0.0304322
22	Justin Timberlake, He's All Dressed Up.txt	0.030845

	File Name	Coefficient
1	Yvonne Brill, a Pioneering Rocket Scientist, Dies at 88.txt	0.00786547
2	Nice Poem - I'll Take It.txt	0.0114708
3	Michael Chwe, Author, Sees Jane Austen as Game Theorist.txt	0.0129098
4	Shakuntala Devi, 'Human Computer,' Dies in India at 83.txt	0.0165453
5	Jamira Cotton, Cameron Johnson — Weddings.txt	0.0165624
6	Justin Bieber and Youth's New Wilderness.txt	0.0166968
7	Franco Biondi Santi, Winemaker, Dies at 91.txt	0.0204396
8	Texas Monthly Hires Full-Time Barbecue Editor.txt	0.0204657
9	Breeding Pigeons on Rooftops, and Crossing Racial Lines.txt	0.0214155
10	Publishers Revel in Youthful Cruelty.txt	0.0219574
11	Donald R. Hopkins - How to Eradicate Guinea Worm Disease.txt	0.0219765
12	Horror Films at the Tribeca Film Festival.txt	0.0221711
13	In Andalusia, on the Trail of Inherited Memories.txt	0.0239712
14	A Knack for Bashing Orthodoxy.txt	0.0246697
15	Restorative Justice Programs Take Root in Schools.txt	0.0269038
16	Lighting the Candle.txt	0.0278286
17	Jamira Cotton and Cameron Johnson.txt	0.0279425
18	John J. Gumperz, Linguist of Cultural Interchange, Dies at 91.txt	0.0291749
19	E. L. Konigsburg, Author, Is Dead at 83.txt	0.0329323
20	Charting Her Own Course.txt	0.0347135
21	New DirectorsNew Films Festival Shows the Future.txt	0.0351548
22	Franco Biondi Santi, Brunello Winemaker, Dies at 91.txt	0.035687

Figure 5.4: Rankings of Similar Files for Document 1

In the experiments, "Yvonnes" ranked at the top of the lists. For the LDA method, the similarity is measured by the exact words. In contrast, the WordNet-based modeling compares the words by their meanings. Therefore, in the similarity list returned by WordNet, the files are more likely in the same field (science in this case) than those using LDA.



	File Name	Coefficient
1	Yvonne Brill, Rocket Scientist, Dies at 88.txt	1
2	New Solar Process Gets More Out of Natural Gas.txt	0.989564
3	Cornell NYC Tech, Planned for Roosevelt Island, Starts Up in Chelsea.txt	0.988961
4	Ian M. Ross, a President at Bell Labs, Dies at 85.txt	0.988572
5	Ian Ross, Who Led Bell Labs, Dies at 85.txt	0.988572
6	St. Louis Names Jim Crews Head Coach.txt	0.987948
7	Jason Griffith.txt	0.984638
8	Taping of Farm Cruelty Is Becoming the Crime.txt	0.978621
9	New Tool for Police Officers - Quick Access to Information.txt	0.976932
10	In Caribbean, Gridlocked Courts Stall Lives.txt	0.976141
11	Canada's Accused Cannibal Killer Ordered to Stand Trial - CBC.txt	0.97332
12	Motte Not Cleared to Throw, Might Need Surgery.txt	0.972705
13	Kevin Pogue Seeks Washington State's Top Terroir.txt	0.971688
14	Jordan Fasbender, Christos Papapetrou.txt	0.971634
15	SKorea's Top Military Officer Puts Off US Trip.txt	0.970031
16	William F. Peel III.txt	0.969946
17	Scientists Question Impact as Vineyards Turn Up in New Places.txt	0.969562
18	Private Development Feared in City-Owned Complex.txt	0.969539
19	Police to Disperse Gas to See How It Would Flow in Terror Attack.txt	0.962995
20	Mishaps Underscore Weaknesses of Japanese Nuclear Plant.txt	0.962294
21	Fukushima Nuclear Plant Is Still Unstable, Japanese Official Says.txt	0.962294
22	Laureates Urge No Cuts to Budgets for Research.txt	0.961817

	File Name	Coefficier
1	Yvonne Brill, a Pioneering Rocket Scientist, Dies at 88.txt	1
2	New Solar Process Gets More Out of Natural Gas.txt	0.989564
3	Cornell NYC Tech, Planned for Roosevelt Island, Starts Up in Chelsea.txt	0.988961
4	Ian M. Ross, a President at Bell Labs, Dies at 85.txt	0.988572
5	Ian Ross, Who Led Bell Labs, Dies at 85.txt	0.988572
6	St. Louis Names Jim Crews Head Coach.txt	0.987948
7	Jason Griffith.txt	0.984638
8	Taping of Farm Cruelty Is Becoming the Crime.txt	0.978621
9	New Tool for Police Officers - Quick Access to Information.txt	0.976932
10	In Caribbean, Gridlocked Courts Stall Lives.txt	0.976141
11	Canada's Accused Cannibal Killer Ordered to Stand Trial - CBC.txt	0.97332
12	Motte Not Cleared to Throw, Might Need Surgery.txt	0.972705
13	Kevin Pogue Seeks Washington State's Top Terroir.txt	0.971688
14	Jordan Fasbender, Christos Papapetrou.txt	0.971634
15	SKorea's Top Military Officer Puts Off US Trip.txt	0.970031
16	William F. Peel III.txt	0.969946
17	Scientists Question Impact as Vineyards Turn Up in New Places.txt	0.969562
18	Private Development Feared in City-Owned Complex.txt	0.969539
19	Police to Disperse Gas to See How It Would Flow in Terror Attack.txt	0.962995
20	Mishaps Underscore Weaknesses of Japanese Nuclear Plant.txt	0.962294
21	Fukushima Nuclear Plant Is Still Unstable, Japanese Official Says.txt	0.962294
22	Laureates Urge No Cuts to Budgets for Research.txt	0.961817

Figure 5.5: Rankings of Similar Files for Document 2

Figure 5.6 represents the recall/precision curve using the LDA and WordNet-based term similarity method. As shown, the WordNet method has higher precision than LDA when the recall values are the same. Therefore, it performs better in relevance ranking.

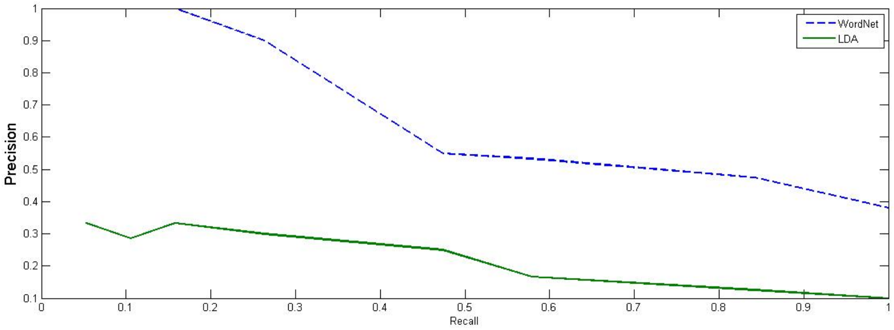


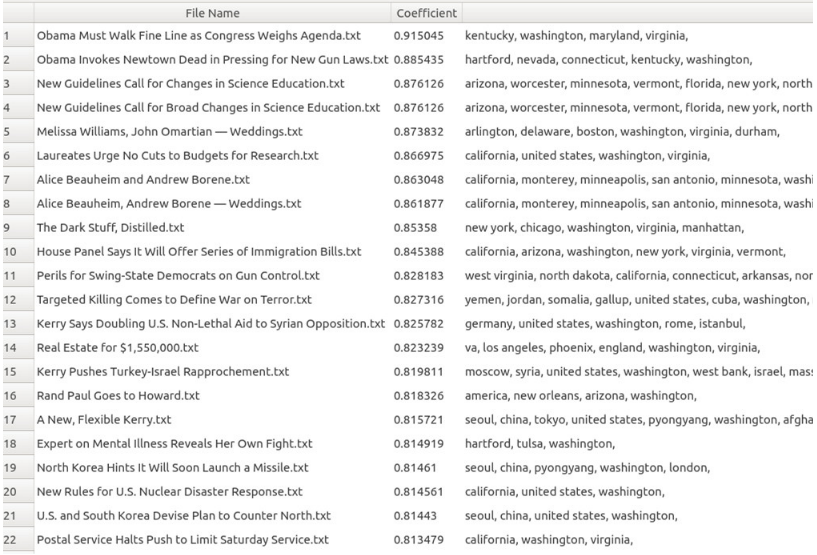
Figure 5.6: Recall/Precision Curve of Topic Modeling

5.4 Document Relevance

To calculate the documents' relevance, both geographical relevance P_g and topic relevance P_t are summed by linear combination. The topic relevance is based on WordNet modeling. λ is the coefficient between P_g and P_t.

$$P = \lambda(P_g) + (1 - \lambda)(P_t) \tag{5.12}$$

Since this study is evaluating the document relevance mainly by locations, λ is expected to be larger.



	File Name	Coefficient	
1	Obama Must Walk Fine Line as Congress Weighs Agenda.txt	0.915045	kentucky, washington, maryland, virginia,
2	Obama Invokes Newtown Dead in Pressing for New Gun Laws.txt	0.885435	hartford, nevada, connecticut, kentucky, washington,
3	New Guidelines Call for Changes in Science Education.txt	0.876126	arizona, worcester, minnesota, vermont, florida, new york, north
4	New Guidelines Call for Broad Changes in Science Education.txt	0.876126	arizona, worcester, minnesota, vermont, florida, new york, north
5	Melissa Williams, John Omartian — Weddings.txt	0.873832	arlington, delaware, boston, washington, virginia, durham,
6	Laureates Urge No Cuts to Budgets for Research.txt	0.866975	california, united states, washington, virginia,
7	Alice Beauheim and Andrew Borene.txt	0.863048	california, monterey, minneapolis, san antonio, minnesota, washi
8	Alice Beauheim, Andrew Borene — Weddings.txt	0.861877	california, monterey, minneapolis, san antonio, minnesota, washi
9	The Dark Stuff, Distilled.txt	0.85358	new york, chicago, washington, virginia, manhattan,
10	House Panel Says It Will Offer Series of Immigration Bills.txt	0.845388	california, arizona, washington, new york, virginia, vermont,
11	Perils for Swing-State Democrats on Gun Control.txt	0.828183	west virginia, north dakota, california, connecticut, arkansas, nor
12	Targeted Killing Comes to Define War on Terror.txt	0.827316	yemen, jordan, somalia, gallup, united states, cuba, washington,
13	Kerry Says Doubling U.S. Non-Lethal Aid to Syrian Opposition.txt	0.825782	germany, united states, washington, rome, istanbul,
14	Real Estate for $1,550,000.txt	0.823239	va, los angeles, phoenix, england, washington, virginia,
15	Kerry Pushes Turkey-Israel Rapprochement.txt	0.819811	moscow, syria, united states, washington, west bank, israel, mas
16	Rand Paul Goes to Howard.txt	0.818326	america, new orleans, arizona, washington,
17	A New, Flexible Kerry.txt	0.815721	seoul, china, tokyo, united states, pyongyang, washington, afgha
18	Expert on Mental Illness Reveals Her Own Fight.txt	0.814919	hartford, tulsa, washington,
19	North Korea Hints It Will Soon Launch a Missile.txt	0.81461	seoul, china, pyongyang, washington, london,
20	New Rules for U.S. Nuclear Disaster Response.txt	0.814561	california, united states, washington,
21	U.S. and South Korea Devise Plan to Counter North.txt	0.81443	seoul, china, united states, washington,
22	Postal Service Halts Push to Limit Saturday Service.txt	0.813479	california, washington, virginia,

Figure 5.7: Rankings of Similar Documents

Figure 5.7 represents an example of the similarity results when combing the topic similarity and location similarity. This study supposes $\lambda = 0.7$.

6. IMPLEMENTATION OF TEXT SEARCHING BASED ON LOCATIONS

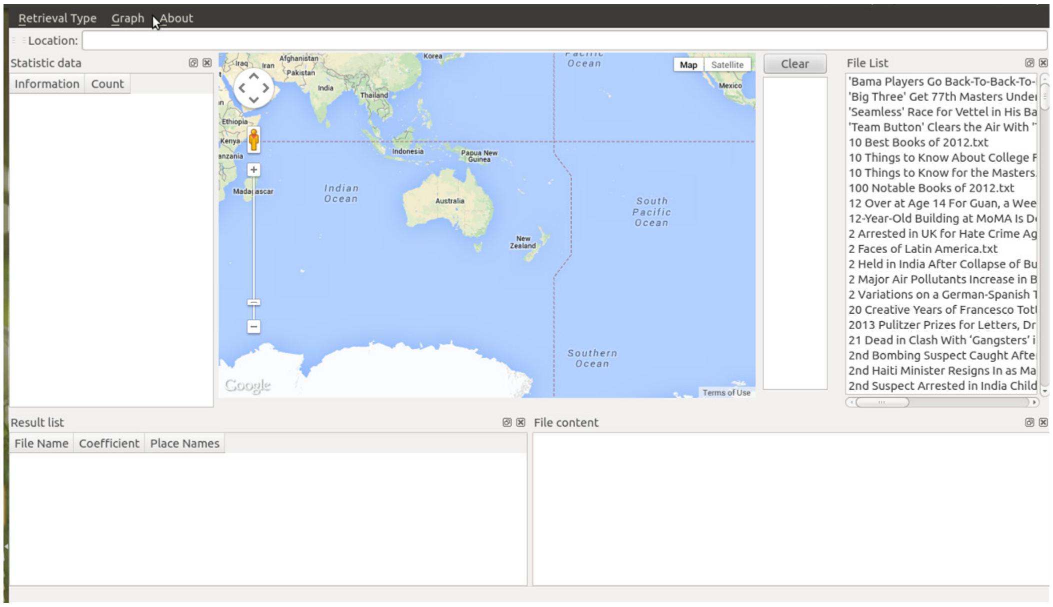
This chapter provides an outline of the location-based search system used in this study and describes the major parts of system implementation. The implementation involves more than 3,000 news articles from the *New York Times.* The system is implemented using QT, which is a cross-platform application and the UI framework of C++ or QML. In addition, WordNet, Citar and Vis Javascript Library are supplemented within the system. The topics and locations are indexed first.

The following sections describe the overall process for system implementation, which consists of several main functions:

1. Summary of file content, containing topic and location information.
2. File search based on locations.
3. Document relevance.
4. Graphic illustration of related files.

6.1 General View

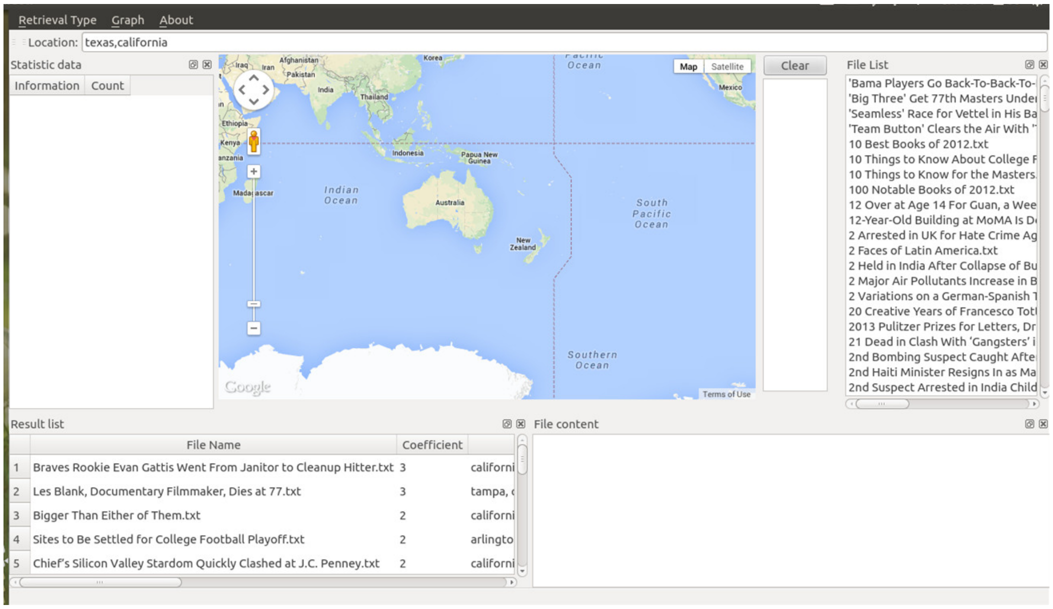
The interface of the software is shown in Figure 6.1.



'Bama Players Go Back-To-Back-To-
'Big Three' Get 77th Masters Under
'Seamless' Race for Vettel in His Ba
'Team Button' Clears the Air With "
10 Best Books of 2012.txt
10 Things to Know About College F
10 Things to Know for the Masters.
100 Notable Books of 2012.txt
12 Over at Age 14 For Guan, a Wee
12-Year-Old Building at MoMA Is D
2 Arrested in UK for Hate Crime Ag
2 Faces of Latin America.txt
2 Held in India After Collapse of Bu
2 Major Air Pollutants Increase in B
2 Variations on a German-Spanish T
20 Creative Years of Francesco Tott
2013 Pulitzer Prizes for Letters, Dr
21 Dead in Clash With 'Gangsters' i
2nd Bombing Suspect Caught Afte
2nd Haiti Minister Resigns In as Ma
2nd Suspect Arrested in India Child

Figure 6.1: Implementation of Location Based Engine

6.1.1 Searching Based on Locations



Location: texas,california

'Bama Players Go Back-To-Back-To-
'Big Three' Get 77th Masters Under
'Seamless' Race for Vettel in His Ba
'Team Button' Clears the Air With "
10 Best Books of 2012.txt
10 Things to Know About College F
10 Things to Know for the Masters.
100 Notable Books of 2012.txt
12 Over at Age 14 For Guan, a Wee
12-Year-Old Building at MoMA Is D
2 Arrested in UK for Hate Crime Ag
2 Faces of Latin America.txt
2 Held in India After Collapse of Bu
2 Major Air Pollutants Increase in B
2 Variations on a German-Spanish T
20 Creative Years of Francesco Tott
2013 Pulitzer Prizes for Letters, Dr
21 Dead in Clash With 'Gangsters' i
2nd Bombing Suspect Caught Afte
2nd Haiti Minister Resigns In as Ma
2nd Suspect Arrested in India Child

	File Name	Coefficient	
1	Braves Rookie Evan Gattis Went From Janitor to Cleanup Hitter.txt	3	californi
2	Les Blank, Documentary Filmmaker, Dies at 77.txt	3	tampa, c
3	Bigger Than Either of Them.txt	2	californi
4	Sites to Be Settled for College Football Playoff.txt	2	arlingto
5	Chief's Silicon Valley Stardom Quickly Clashed at J.C. Penney.txt	2	californi

Figure 6.2: Query Search by Locations

First, the user is able to search documents by a list of locations, as Figure 6.2 shows. The system returns a list of documents that contain all the locations that are being queried. The documents are sorted by tf-idf values.

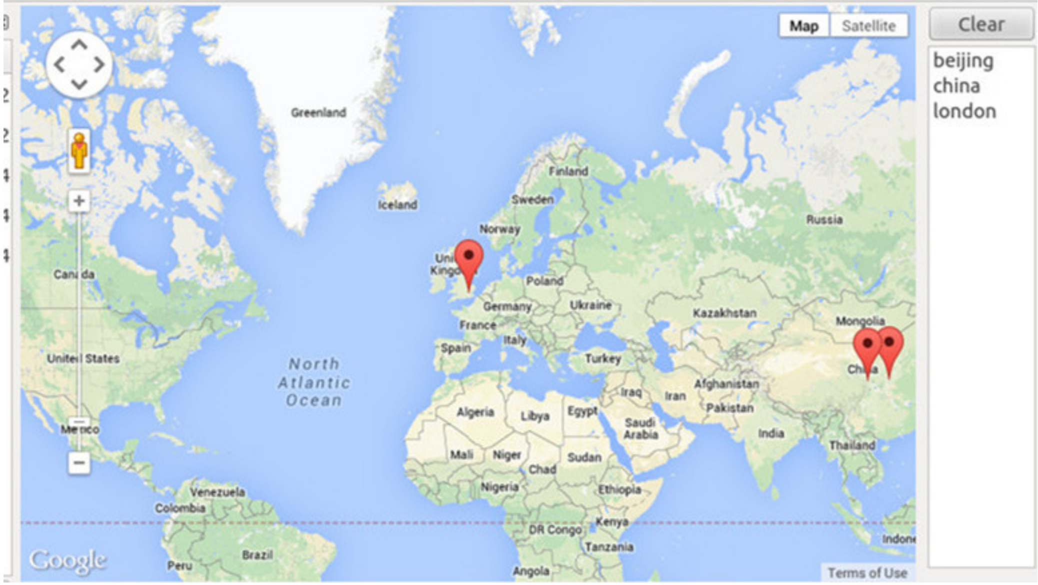
Apart from presenting the document content when clicking on a document, the implementation has a map plugin to describe the coordinates of locations.

For the map implementation, there are several major software applications available, such as QGIS [42] and ArcGIS [43]. There are also some application programming interfaces (APIs) such as Google Maps API and Qt Mobility API.

Esri's ArcGIS is a commercially available software suite that includes three desktop versions with varying levels of complexity. It also has mobile and web components. However, it is restricted by a license, and each installation version requires a licensing key. QGIS is a GIS suite of software that has a desktop option, along with mobile and web components. It is open source and freely downloadable, so there is no license concern. QGIS has a faster startup time than ArcGIS, but the API is quite unstable and old compared to others. Concerning web APIs, Qt Mobility API is produced and developed by NOKIA, and the service is not guaranteed. In contrast, Google Maps API [44] is complete and stable, as long as there is a web connection.

In light of the above, this study uses Google Maps API for the map implementation. The longitude/latitude of every found location is passed to the Google Maps API and marked on the map. As Figure 6.3 illustrates, the locations in a file are all marked on the map and can be cleared using the "Clear" button.

With the map marker available, the user can better understand which locations are mentioned in the documents. This feature also enhances document summarization and searches.



beijing
china
london

Figure 6.3: Map of Locations

6.1.2 Text Classification Types

In the location-based search system software, different types of classification can be chosen using the "Retrieval Type" button in the menu. There are three choices, as Figure 6.4 illustrates.

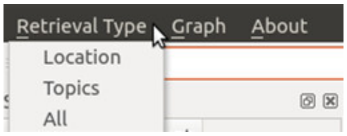
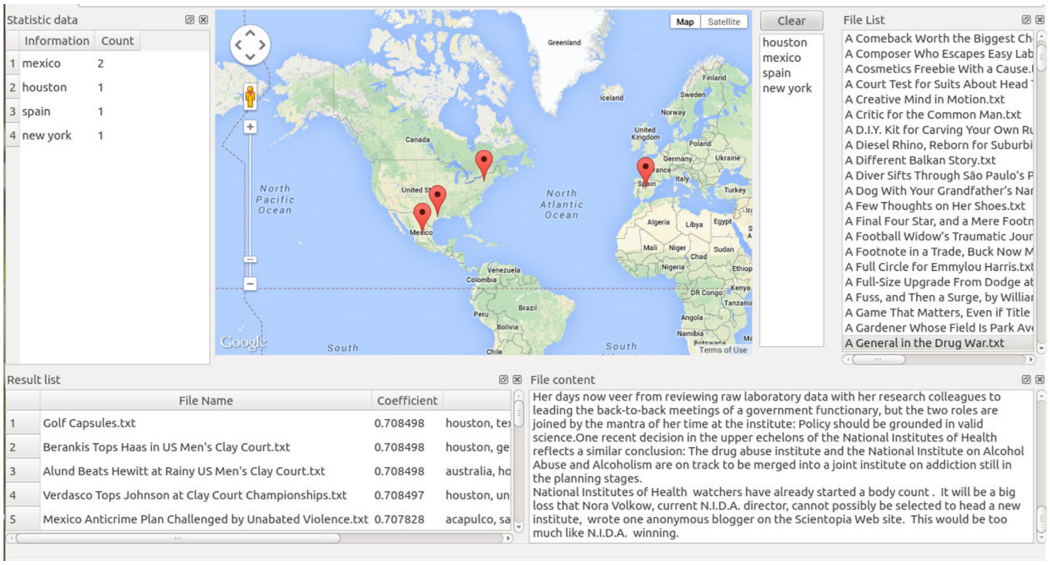


Figure 6.4: Menu Selection

If, for example, the user picks "Location" from the menu, the chosen file's locations are summarized and also marked on the map, which is presented in Figure 6.5. The statistical data list the locations in the document with their co-appearances. In addition, similar documents are sorted according to locations and displayed in the result widget, as discussed in Chapter 5.



	Information	Count
1	mexico	2
2	houston	1
3	spain	1
4	new york	1

houston
mexico
spain
new york

A Comeback Worth the Biggest Ch
A Composer Who Escapes Easy Lab
A Cosmetics Freebie With a Cause.
A Court Test for Suits About Head
A Creative Mind in Motion.txt
A Critic for the Common Man.txt
A D.I.Y. Kit for Carving Your Own Ru
A Diesel Rhino, Reborn for Suburbi
A Different Balkan Story.txt
A Diver Sifts Through São Paulo's P
A Dog With Your Grandfather's Nar
A Few Thoughts on Her Shoes.txt
A Final Four Star, and a Mere Footn
A Football Widow's Traumatic Jour
A Footnote in a Trade, Buck Now M
A Full Circle for Emmylou Harris.txt
A Full-Size Upgrade From Dodge at
A Fuss, and Then a Surge, by Williar
A Game That Matters, Even if Title
A Gardener Whose Field Is Park Ave
A General in the Drug War.txt

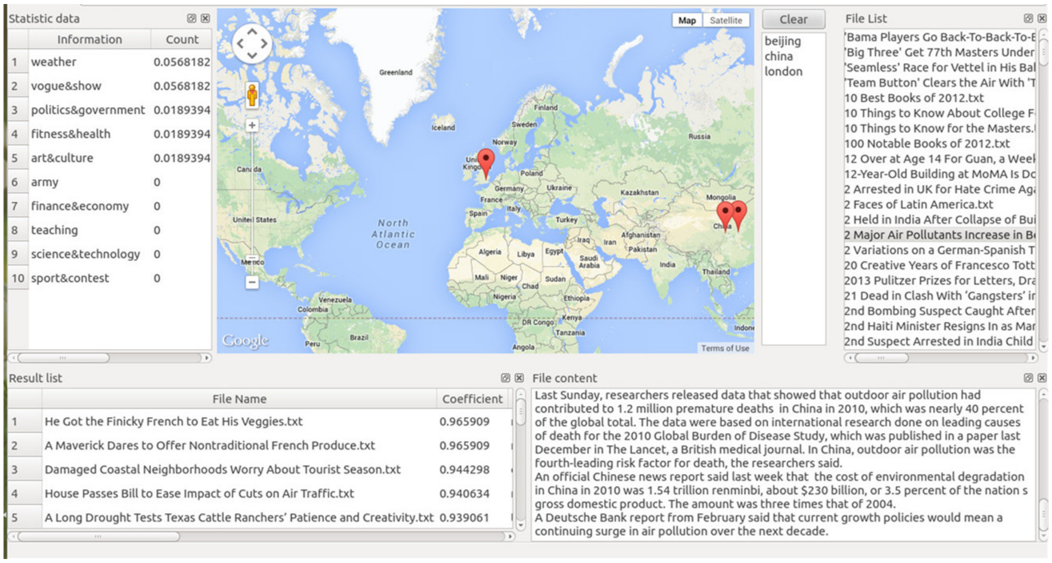
	File Name	Coefficient	
1	Golf Capsules.txt	0.708498	houston, tex
2	Berankis Tops Haas in US Men's Clay Court.txt	0.708498	houston, ge
3	Alund Beats Hewitt at Rainy US Men's Clay Court.txt	0.708498	australia, ho
4	Verdasco Tops Johnson at Clay Court Championships.txt	0.708497	houston, un
5	Mexico Anticrime Plan Challenged by Unabated Violence.txt	0.707828	acapulco, sa

Her days now veer from reviewing raw laboratory data with her research colleagues to leading the back-to-back meetings of a government functionary, but the two roles are joined by the mantra of her time at the institute: Policy should be grounded in valid science.One recent decision in the upper echelons of the National Institutes of Health reflects a similar conclusion: The drug abuse institute and the National Institute on Alcohol Abuse and Alcoholism are on track to be merged into a joint institute on addiction still in the planning stages.
National Institutes of Health watchers have already started a body count . It will be a big loss that Nora Volkow, current N.I.D.A. director, cannot possibly be selected to head a new institute, wrote one anonymous blogger on the Scientopia Web site. This would be too much like N.I.D.A. winning.

Figure 6.5: Location

Similarly, if the "Topics" type is selected, the probability of each topic within the topic list in a document is returned. As Figure 6.6 shows, the statistical data represent the selected document's topics. Documents are sorted and listed according to their topic similarity to the selected file.

In addition, Figure 6.7 displays the results when the user chooses the "All" type. The documents are classified and sorted by the combination of topic model and toponym similarity. The topic and location summary is shown in the statistical data.



Statistic data

	Information	Count
1	weather	0.0568182
2	vogue&show	0.0568182
3	politics&government	0.0189394
4	fitness&health	0.0189394
5	art&culture	0.0189394
6	army	0
7	finance&economy	0
8	teaching	0
9	science&technology	0
10	sport&contest	0

beijing
china
london

File List

'Bama Players Go Back-To-Back-To-E
'Big Three' Get 77th Masters Under
'Seamless' Race for Vettel in His Bal
'Team Button' Clears the Air With 'T
10 Best Books of 2012.txt
10 Things to Know About College F
10 Things to Know for the Masters.
100 Notable Books of 2012.txt
12 Over at Age 14 For Guan, a Weel
12-Year-Old Building at MoMA Is Dc
2 Arrested in UK for Hate Crime Ag
2 Faces of Latin America.txt
2 Held in India After Collapse of Bui
2 Major Air Pollutants Increase in B
2 Variations on a German-Spanish T
20 Creative Years of Francesco Tott
2013 Pulitzer Prizes for Letters, Dra
21 Dead in Clash With 'Gangsters' ir
2nd Bombing Suspect Caught After
2nd Haiti Minister Resigns In as Mar
2nd Suspect Arrested in India Child

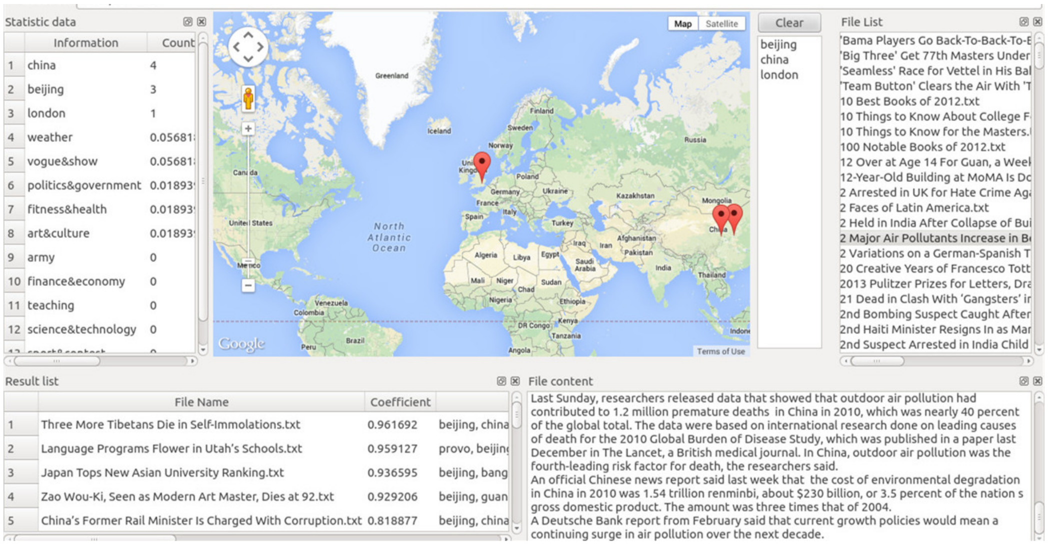
Result list

	File Name	Coefficient
1	He Got the Finicky French to Eat His Veggies.txt	0.965909
2	A Maverick Dares to Offer Nontraditional French Produce.txt	0.965909
3	Damaged Coastal Neighborhoods Worry About Tourist Season.txt	0.944298
4	House Passes Bill to Ease Impact of Cuts on Air Traffic.txt	0.940634
5	A Long Drought Tests Texas Cattle Ranchers' Patience and Creativity.txt	0.939061

File content

Last Sunday, researchers released data that showed that outdoor air pollution had contributed to 1.2 million premature deaths in China in 2010, which was nearly 40 percent of the global total. The data were based on international research done on leading causes of death for the 2010 Global Burden of Disease Study, which was published in a paper last December in The Lancet, a British medical journal. In China, outdoor air pollution was the fourth-leading risk factor for death, the researchers said.
An official Chinese news report said last week that the cost of environmental degradation in China in 2010 was 1.54 trillion renminbi, about $230 billion, or 3.5 percent of the nation s gross domestic product. The amount was three times that of 2004.
A Deutsche Bank report from February said that current growth policies would mean a continuing surge in air pollution over the next decade.

Figure 6.6: Topic



Statistic data

	Information	Count
1	china	4
2	beijing	3
3	london	1
4	weather	0.05681
5	vogue&show	0.05681
6	politics&government	0.01893
7	fitness&health	0.01893
8	art&culture	0.01893
9	army	0
10	finance&economy	0
11	teaching	0
12	science&technology	0
13	sport&contest	0

beijing
china
london

File List

'Bama Players Go Back-To-Back-To-E
'Big Three' Get 77th Masters Under
'Seamless' Race for Vettel in His Bal
'Team Button' Clears the Air With 'T
10 Best Books of 2012.txt
10 Things to Know About College F
10 Things to Know for the Masters.
100 Notable Books of 2012.txt
12 Over at Age 14 For Guan, a Weel
12-Year-Old Building at MoMA Is Dc
2 Arrested in UK for Hate Crime Ag
2 Faces of Latin America.txt
2 Held in India After Collapse of Bui
2 Major Air Pollutants Increase in B
2 Variations on a German-Spanish T
20 Creative Years of Francesco Tott
2013 Pulitzer Prizes for Letters, Dra
21 Dead in Clash With 'Gangsters' ir
2nd Bombing Suspect Caught After
2nd Haiti Minister Resigns In as Mar
2nd Suspect Arrested in India Child

Result list

	File Name	Coefficient	
1	Three More Tibetans Die in Self-Immolations.txt	0.961692	beijing, china
2	Language Programs Flower in Utah's Schools.txt	0.959127	provo, beijin
3	Japan Tops New Asian University Ranking.txt	0.936595	beijing, bang
4	Zao Wou-Ki, Seen as Modern Art Master, Dies at 92.txt	0.929206	beijing, guan
5	China's Former Rail Minister Is Charged With Corruption.txt	0.818877	beijing, china

File content

Last Sunday, researchers released data that showed that outdoor air pollution had contributed to 1.2 million premature deaths in China in 2010, which was nearly 40 percent of the global total. The data were based on international research done on leading causes of death for the 2010 Global Burden of Disease Study, which was published in a paper last December in The Lancet, a British medical journal. In China, outdoor air pollution was the fourth-leading risk factor for death, the researchers said.
An official Chinese news report said last week that the cost of environmental degradation in China in 2010 was 1.54 trillion renminbi, about $230 billion, or 3.5 percent of the nation s gross domestic product. The amount was three times that of 2004.
A Deutsche Bank report from February said that current growth policies would mean a continuing surge in air pollution over the next decade.

Figure 6.7: Combination of Locations and Topics

The system implementation clearly presents the major ideas of a particular document. Users are able to quickly get information without reading all the content in the document.

6.2 Graphic Illustration of Document Relevance

To visualize the document relationships, this thesis is inspired by the idea of the Google knowledge graph [45] and uses the graphic illustration to show the document relevance. The Google knowledge graph [46] is a knowledge base produced by Google. It gathers information from a variety of sources on the Internet, provides a list of similarities to other related sites, and offers some structured and detailed information about the topic. All the information is shown at the top of the search page. The users can resolve queries without navigating to other sites. Therefore, it enhances the performance of the search engine greatly. In this thesis, the idea of connection exploration in the Google knowledge graph is adopted. Furthermore, a graph is used to assemble the data and visualized dynamically.

The relevance graph is drawn by JavaScript [47] to show nodes and edges in a graph dynamically with the help of Vis Library. Clicking on the "knowledge graph" button in the document menu opens another window for graphic illustration of document relevance. For a particular document that is put in the search box, the visualized graph tree pops up to display its relevant files.

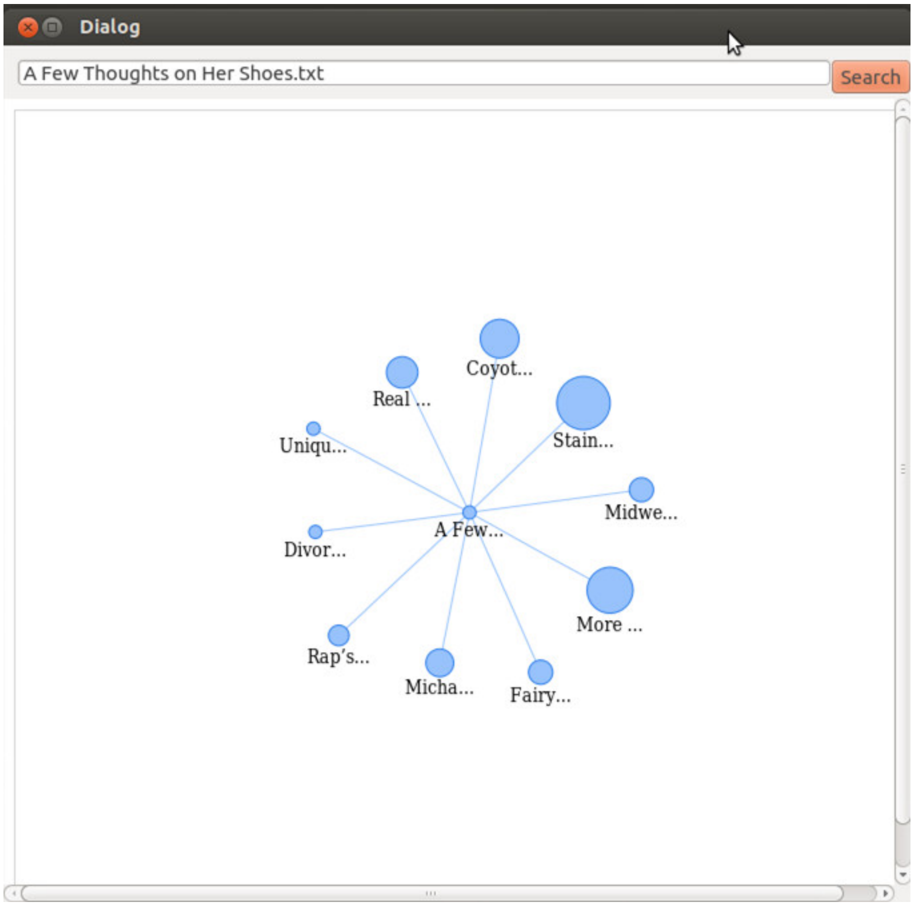


Figure 6.8: Graphic Illustration of Document Relevance

An example of the relevance graph is shown in Figure 6.8. In the relevance graph, nodes represent documents. A larger node indicates more toponyms in the document. An edge represents the similarity between two documents. The similarity degree is inversely proportional to the distance between the two nodes.

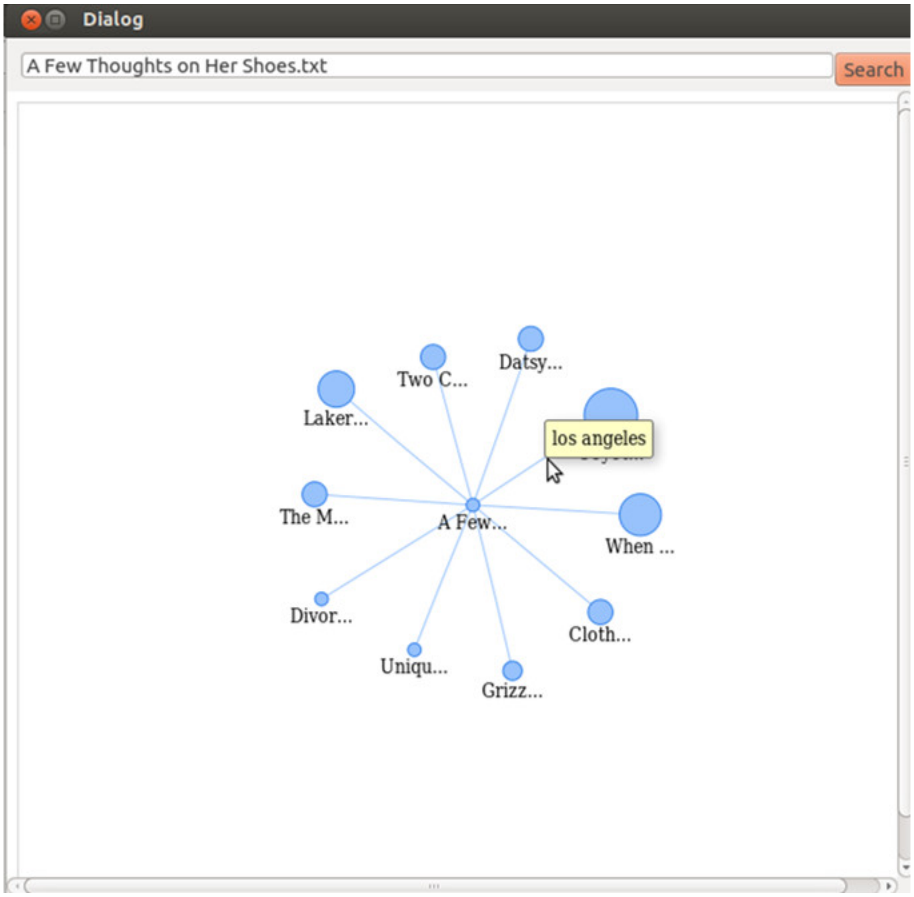


Figure 6.9: Same Locations Shown in Edge Values

As Figure 6.9 indicates, The same locations of two files are shown in the edge values. Double-clicking on the node causes the node to expand, and related documents are shown in the graph. Figure 6.10 shows an example.

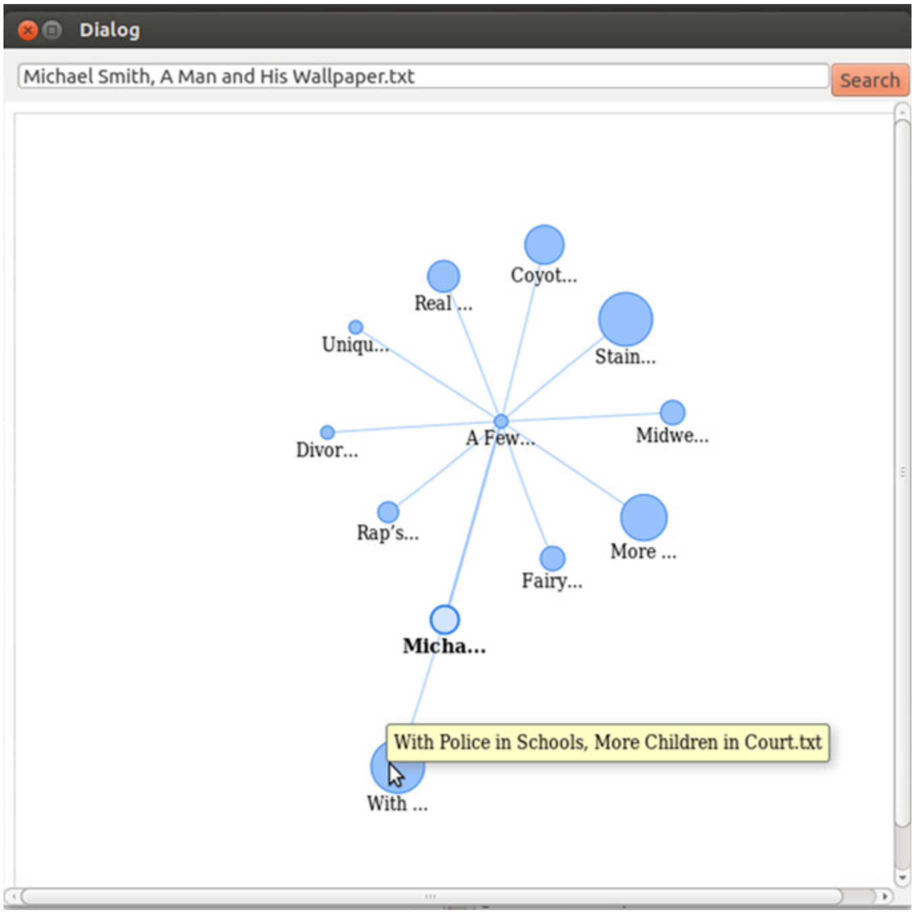


Figure 6.10: Node Expansion

Using this method, the document network is visualized interactively and vividly. It introduces another creative angle from which to present the information searching and classification in document relevance ranking, especially from the geographical aspect.

7. CONCLUSION

This study focused on the location indexing, topic summarizing, and relevance ranking, all of which can be used for searching queries and finding document simialrities.

7.1 Summary of Work

Motivated by the need for a more effective method of location retrieval and document relevance ranking in large data sources, this thesis developed a new method for text searching and ranking of document similarities. This study also took advantage of map view and graphic illustration to represent text-based geographical information and document relevance more intuitively.

In this thesis, the exploration of the text search based on locations is presented from the following aspects:

1. Location extraction and searching.
2. Topic modeling.
3. Document relevance based on locations and topics.
4. Graphic illustration of document relevance.

This study used POS tagger, WordNet, and a gazetteer to address the problems in location extraction and indexing. The POS tagger and WordNet were used to identify proper nouns, and a gazetteer provided all the necessary information about locations. Two algorithms were used for location disambiguation. The results of experiments showed that K-means clustering method, which was derived from the centroid algorithm, outperformed the centroid method. One of the reasons for this finding is that the centroid method gathers all the location candidates into a single cluster and thus leads to the incorrect classification of the locations, while the K-means clustering algorithm avoids this phenomenon through separating the clusters. The K-clustering method is more accurate in finding location names and settling their coordinates. Our approach performed even better, with higher accuracy and more detailed information about locations compared with Stanford Named-Entity Recognition, which uses the machine learning method.

We evaluated two algorithms for topic modeling: term group association (Latent Dirichlet Allocation) and term similarity measure. LDA is a commonly used method for topic modeling. The perplexity experiment that we conducted showed that the results with around 5 to 10 topics performed best. For term similarity measurement, we developed a new word-clustering method that uses WordNet to define topics at a limit of 10 and then classifies the words by these topics. We conducted experiments on these two methods to evaluate the F-measure. The method based on WordNet proved to be more accurate than LDA.

Finally, the document relevance was also explored by means of vector space and graphs. The linear combination method was first used to combine the location relevance and topic relevance together, and proved to be a simple but efficient way to rank document similarities. The method also involves a coefficient to add the global factor of each term and improves the accuracy of estimation. Then the text relevance is represented through a graphic illustration, which provides an interactive and novel way for users to view the document relationships.

7.2 Further Study

This thesis introduces the idea of text summarizing and ranking according to the relationships between different documents, which can be further studied in the future.

Location-based search engines can reflect their worth in question answering. Users can find some specific features and even the similarities of the given locations, for instance, the features of Lyon and Paris that are in common. Therefore, some research should be carried out to settle the location features through textual and spatial searches.

Greater improvement in location-based searches can be achieved by recognizing features such as organizations and person's names. These features can provide more detailed information for a text corpus.

Finally, our research involves a manually-input text corpus, which can be extended to a web-based engine in the future and do not require the manual input. Compared to the current research we have conducted, the web-based engines search for any text a user requires through the Internet.

REFERENCES

[1] H. Schutze, C. D. Manning. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, USA, 1999.

[2] D. M. Blei, A. Y. Ng, M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[3] Wikipedia, http://en.wikipedia.org/wiki/Geographic_information_retrieval, *Geographical Information Retrieval*, Jul. 2012.

[4] H. Li, R. K. Srihari, C. Niu, W. Li. InfoXtract location normalization: a hybrid approach to geographic references in information extraction. In *Proc. HLT-NAACL 2003 Workshop on Analysis of Geographic References*, vol. 1, pp. 39-44, Alberta, Canada, 2003.

[5] D. Buscaldi. *Toponym Disambiguation in Information Retrieval.* PhD thesis, Polytechnic University of Valencia, Valencia, Spain, Oct. 2010.

[6] E. Amitay, N. harEi, R. Sivan, A. Soffer. Web-a-where: geotagging web content. In *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 273-280, UK, Jul. 2004.

[7] L. Backstrom, E. Sun, C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proc. International Conference on World Wide Web*, pp. 61-70, Raleigh, North Carolina, USA, Apr. 2010.

[8] Z. Cheng, J. Caverlee, K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proc. ACM International Conference on Information and Knowledge Management*, pp. 759-768, Toronto, Canada, Oct. 2010.

[9] C. Chen, T. Suel, A. Markowetz. Efficient query processing in geographic web search engines. In *Proc. ACM SIGMOD International Conference on Management of Data*, pp. 277-288, Chicago, Illinois, USA, Jun. 2006.

[10] B. Martins, M. J. Silva, L. Andrade. Indexing and ranking in Geo-IR systems. In *Proc. ACM Workshop on Geographic Information Retrieval*, pp. 31-34, Bremen, Germany, Nov. 2005.

[11] H. Li, R. S. Srihari, C. Niu, W. Li. Location normalization for information extraction. In *Proc. International Conference on Computational linguistics*, vol. 1, pp. 1-7, Taipei, Taiwan, Aug. 2002.

[12] R. Hariharan, B. Hore, L. Chen, S. Mehrotra. Processing Spatial-Keyword (SK) queries in geographic information retrieval (GIR) systems. In *19th International Conference on Scientific and Statistical Database Management*, Banff, Canada, Jul. 2007.

[13] I. D. Felipe, V. Hristidis, N. Rishe. Keyword search on spatial databases. In *Proc. IEEE International Conference on Data Engineering*, pp. 656-665, Cancun, Mexico, Apr. 2008.

[14] D. Zhang, Y. Chee, A. Mondal,A. Tung, M. Kitsuregawa. Keyword search in spatial databases: towards searching by document. In *Proc. IEEE International Conference on Data Engineering*, pp. 688-699, Shanghai, China, Mar. 2009.

[15] G. Cong, C. S. Jensen, D. Wu. Efficient retrieval of the top-k most relevant spatial web objects. In *Proc. of the VLDB Endowment*, vol. 2, no. 1, pp. 337-348, Lyon, France, Aug. 2009.

[16] B. Yu, G. Cai. A query-aware document ranking method for geographic information retrieval. In *Proc. ACM Workshop on Geographical Information Retrieval*,

pp. 49-54, Lisbon, Portugal, Nov. 2007.

[17] K. S. Jones. *Automatic Keyword Classification for Information Retrieval.* Butterworths, London, UK, 1971.

[18] X. Wei. *Topic Models in Information Retrieval.* ProQuest, Aug. 2007.

[19] X.Wei, W. B. Croft. Investigating retrieval performance with manually-built topic models. In *Proc. Large-Scale Semantic Access to Content*, pp. 333-349, Pittsburgh, USA, May, 2007.

[20] X. Peng, B. Choi. Document classifications based on word semantic hierarchies. In *Proc. International Conference on Artificial Intelligence and Applications*, pp. 362-367, Innsbruck, Austria, Feb. 2005.

[21] Z. Elberrichi, A. Rahmoun, M. A. Bentaalah. Using WordNet for text categorization. *The International Arab Journal of Information Technology*, Vol. 5, No. 1, pp. 16-24, Jan. 2008.

[22] M. Girlomi, A. Kaban. On an equivalence between PLSI and LDA. In *Proc. ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pp. 433-434, Toronto, Canada, Jul. 2003.

[23] C. Wartena, R. Brussee. Topic detection by clustering keywords. In *Database and Expert Systems Application*, pp. 54-58, Turin, Italy, Sep. 2008.

[24] D. Newman, A. Asuncion, P. Smyth, M. Welling. Distributed algorithms for topic models. *The Journal of Machine Learning Research*, vol. 10, pp. 1801-1828, 2009.

[25] T. Brants. TnT – a statistical Part-of-Speech tagger. In *Proc. Conference on Applied Natural Language Processing*, pp. 224-231, Seattle, Washington, USA, Apr. 2000.

[26] G. D. Kennedy. *An Introduction to Corpus Linguistics*, Longman, UK, 1998.

[27] G. A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, vol. 38, no. 11, pp. 39-41, 1995.

[28] C. Fellbaum. *WordNet: an Electronic Lexical Database.* MIT Press, Cambridge, USA, 1998.

[29] GeoNames, http://www.geonames.org/about.html, Jul. 2014.

[30] National Geographic, http://travel.nationalgeographic.com/travel/city-guides/free-charleston-traveler/, *Free Things to Do in Charleston*, Jul. 2014.

[31] Charleston in WordNet, http://wordnetweb.princeton.edu/perl/webwn?s=Charleston, Jul. 2014.

[32] Van Rijsbergen. *Information Retrieval (2nd edition).* Butterworths, London, UK, 1979.

[33] J. D. Burger, J. C. Henderson, W. T. Morgan. Statistical named entity recognizer adaptation. In *Proc. Conference on Natural Language Learning*, vol. 20, pp. 1-4, Taipei, 2002.

[34] J. R. Finkel, T. Grenager, C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling, In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 363-370, Ann Arbor, Michigan, USA, Jun. 2005.

[35] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, vol. 55, no. 4, pp. 77-84, Apr. 2012.

[36] G. Ercan, I. Cicekli. Using lexical chains for keyword extraction. *Information Processing and Management: an International Journal*, vol. 43, no. 6, pp. 1705-1714, Nov. 2007.

[37] K. Janowicz, M. Raubal, W. Kuhn. The semantics of similarity in geographic information retrieval. *Journal of Spatial Information Science*, vol. 2, pp. 29-57, 2011.

[38] C. Zhai, J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, vol. 22, no. 2, pp. 179-214, Apr. 2004.

[39] F. Mata. iRank: ranking geographical information by conceptual, geographic and topologic similarity. In *Third International Conference GeoS 2009*, pp. 159-174, Mexico City, Mexico, Dec. 2009.

[40] M. Steyvers, T. Griffiths. Probabilistic topic models. In *Handbook of Latent Semantic Analysis*, Lawrence Erlbaum Associates, Mahwah, USA, 2007.

[41] J. Jin. Divergence measures based on the shannon entropy. *IEEE transactions on information theory*, vol. 37, no. 1, pp. 145-151, Jan. 1991.

[42] QGIS, http://www.qgis.org, Sep. 2013.

[43] ArcGIS, http://www.esri.com/software/arcgis, Feb. 2014.

[44] G. Svennerberg, *Beginning Google Maps API 3*, Apress Berkely, USA, 2010.

[45] Google knowledge graph, http://www.google.com/insidesearch/features/search/knowledge.html, Aug. 2012.

[46] S. Amit. *Introducing the Knowledge Graph: Things, Not Strings*. Official Blog (of Google), http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html, May 2012.

[47] S. Vercruysse, M. Kuiper. WordVis: JavaScript and animation to visualize the WordNet relational dictionary. *Advances in Intelligent Systems and Computing*, vol. 179, pp. 137-145, 2013.