Recommendation
Algorithm on Academic Research
Final Report
Chan Erik
Supervisor: Dr. Yiu Siu Ming
# Table of Content

TABLE OF CONTENT ..........................................................................................2
ABSTRACT ...........................................................................................................4
ACKNOWLEDGEMENT .......................................................................................4
ABBREVIATIONS ...............................................................................................4
BACKGROUND .....................................................................................................5
OBJECTIVES .........................................................................................................7
DIVISION OF LABOUR ........................................................................................8
INTRODUCTION OF THE STATE OF THE ART ..............................................8
  NLP ....................................................................................................................8
  TEXT RANK AND PAGE RANK .........................................................................8
  TF-IDF ..............................................................................................................10
  LSA AND LSI ....................................................................................................10
  LDA ..................................................................................................................10
IMPLEMENTATION ...............................................................................................11
  SEARCHING AND RANKING .........................................................................11
  RECOMMENDATION SYSTEM .........................................................................11
EXPERIMENTS .....................................................................................................13
RESULTS ...............................................................................................................13
  ACCURACY ......................................................................................................13
    1st Approach ................................................................................................15
    2nd Approach ................................................................................................17
    3rd Approach ................................................................................................19
  SPEED AND SCALABILITY .............................................................................21
    1st Approach ................................................................................................21
    2nd Approach ................................................................................................21
    3rd Approach ................................................................................................21
DIFFICULTIES AND LIMITATIONS .................................................................22
CONCERNS .........................................................................................................23
  SECURITY .........................................................................................................23
  LEGAL ...............................................................................................................23
CONCLUSION AND FUTURE WORKS ...............................................................23
Abstract

In this project, an algorithm that recommends academic research resources is to be implemented. A server-side database structure for citation storage is designed and citations are generated automatically for target resources. We also discuss the idea different recommendation systems and analysis of their applicability to recommending research resources.

Acknowledgement

We would like to express our appreciation to Dr. Yiu Siu Ming, our supervisor in the project, for his time and effort in guiding us in overcoming many difficulties and discovering new areas that we had not explored. Furthermore, we would also like to acknowledge with much appreciation the role of Mr. Patrick Desloge, who gave guidance to our project presentation and the writing of this report that has improved our presentation and writing skills thanks to his comments and advices.

Abbreviations

NLP - Natural Language Processing
TF-IDF - Term Frequency – Inverse Document Frequency
LSA – Latent Semantic Analysis
LSI – Latent Semantic Indexing
LDA – Latent Dirichlet Allocation
Background

Since the invention of the World Wide Web, doing good research on the Internet is an essential part in writing any essays and articles as information on the Internet has a wide coverage on different topics. Several difficulties arise when writing essays. Writers often need to spend a lot of time on searching and filtering for relevant resources on their research topic. Google search and Google Scholar are very popular tools for search for academic research resources. They can significantly reduce searching time once the user had found the right keyword to search for. Despite that, time is still spent on scanning the articles and filtering useful texts. Second, while technological advancements had allowed us to write and do research with the same device, the process of researching and writing had been largely separated. Integrating the research process into the writing had become a time-consuming task. Tools like Zotero and Endnote exist to help manage the texts which users find and store from websites. They also provide auto-generated citations. However, these tools do not integrate the process of searching and writing as searching is still largely separate from the writing workflow. Developing software with algorithms that can automatically understands your current writing and provide adequate resources on the Internet can greatly reduce the time required on researching and integrating the research with the writing.

There are several popular tools for academic research, including Google Scholar, Zotero and EndNote. Table 1 shows a summary of advantages and disadvantages of each of the tools. Google scholar is one of the most popular searching tools. It can rank research papers so that users’ searching time is greatly reduced. However, users still need to scan the whole article to determine if it is useful and locate the useful texts. Moreover, the writing and research process is separated. The general approach by most people is to use Google Scholar to search for relevant resources, open new tab for each web page, and then write their essays. This approach has a big disadvantage: writers must spend time on revising their materials again during writing. Even if Google Scholar is combined with Zotero and EndNote, the above two problems still cannot be solved.
<table>
<thead>
<tr>
<th>Tools</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
</table>
| Google Scholar | ● Search for academic research resources quickly  
● Search results are ranked by number of citations, helping users get the most relevant resources | ● Lack a tool to highlight and store the relevant research resources found  
● Lack a suggestion of text. Users need to scan the whole article to get useful text and see whether the result is appropriate |
| Zotero      | ● Automatically senses the content in a web browser and allows users to store the content | ● Lack a search function to search for relevant resources |
| EndNote     | ● Manage PDF academic research resources  
● Search research resources in PDF form in online database  
● Users can look at comment made on PDF online | ● Lack a method to rank resources  
● Lack searching of internet materials |

Table 1: Summary of current academic research tools

The deliverable of this project will be an integration of functions provided by the above-mentioned tools, together with functions that recommend relevant texts and integrate writing and searching process. Table 2 shows a comparison of the deliverable and other research tools. The core advantage of the deliverable is that it reduces essay writing time significantly by recommending most relevant resources and texts and this function is not in any of the other research tools.
This project will be built on top of a developing web application named GistNote which provides support of highlighting and storing websites’ texts. At this stage, GistNote can allow users to highlight any texts on a website. The highlighted texts are stored and can later be referred when users write essays. With the highlight data from users, it allows the searching of database to find relevant resources.

<table>
<thead>
<tr>
<th></th>
<th>Google/Google Scholar</th>
<th>Zotero</th>
<th>EndNote</th>
<th>Our deliverable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search for relevant website</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Search for research paper</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Search results are ranked</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Auto-generated citations</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Allow users to store useful results</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Recommend relevant text</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Recommend resources when writing</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 2: Comparison of project deliverable and other research tools

**Objectives**

This project is going to achieve five things.
1. Structuring a database to store users’ highlights for searching.
2. Developing an algorithm or integrating existing algorithms to search for relevant recommendations on research resources from keywords.
3. Ranking existing cited research resources by an algorithm.
4. Developing an algorithm to understand users’ essay and recommend resources accordingly automatically.
5. Generating citations in different styles.

These algorithms and functions are mainly developed for use in web application GistNote. The integration of the algorithms and the user-interface of the extension depend on the development progress of GistNote. (not a part of this project)
Division of Labour

This project is a joint effort between me, Erik Chan, and my fellow classmate Lui Hoi Ching (Vincent). In order to achieve the objectives mentioned above, our effort is split between the 5 objectives. Objectives 1 and 5 are mainly achieved by the work of Vincent, while objective 4 is mainly achieved by me. Objectives 2 and 3 are joint efforts. Therefore, this individual report will focus on objective 2, 3 and 4.

Introduction of the State of the Art

NLP

Recommendation generation requires the program to understand the context of the current writing. Algorithms that tries to achieve this falls into the field of NLP. More specifically, automatic summarization and keyword extraction or generation given a certain length of text is the tasks most relevant to the project. The current technology in NLP is still in development, and most solutions involve the use of machine learning.

TextRank and PageRank

TextRank is an unsupervised machine learning algorithm derived from Google search ranking algorithm PageRank. (figure 1) TextRank utilize the co-occurrence of words as edges between nodes of words (figure 2). [1]

Since TextRank is based on word co-occurrence. Linguistic properties can be ignored, which will be both advantageous and disadvantageous. One of the main advantage is that it is highly portable to other languages, which will be tremendously helpful when GistNote is expanding to other markets around the World.
Figure 1: Representation of the Google PageRank algorithm. Site rankings are represented by the size of the nodes and links are represented by the edges.

Figure 2: Example of a co-occurrence graph of a body of text
TF-IDF

TF-IDF is the widely popular statistic that somewhat accurately reflect the weight, calculated as show in figure 3, of a word or term in a document. It is based on simple frequency analysis of words and terms with a corpus. It has been in use for 30 years and allows many search engines to accurately score and rank the relevance of documents to the queries. It also allows accurate detection of stopwords based on the inverse document frequency part.

\[ w_{i,j} = tf_{i,j} \times \log \left( \frac{N}{df_i} \right) \]

*tf*$_{i,j}$ = number of occurrences of *i* in *j*

*df*$_i$ = number of documents containing *i*

*N* = total number of documents

Figure 3: Calculation of term weight using TF-IDF

LSA and LSI

Using a matrix of word count or weight (usually by TF-IDF), LSA can groups documents using similar words and words that appear in similar documents. [2] Given a word, the trained model can provide similar words that appear together with some other words. Implementing LSA into document retrieval, i.e. LSI, creates semantic search from blurry keywords.

However, different from TextRank, LSA disregard the distance of appearances of words and utilize a simple bag of words model. Not only are associations between words lost, multiple meanings of a single word may be erroneously grouped together generating unwanted results.

LDA

LDA was developed from the probabilistic version of LSA(LSI). LDA models documents as a mix of topics, and can infer topic probability distribution of previously unseen documents. [3]

LDA is very powerful in modelling topics however, the topics created from the trained model include probability distribution of words that are not easily interpretable by humans, as we will show in our tests.
Implementation

Searching and Ranking
For objective 2 and 3, we had decided to use Apache’s Solr engine. The reasons are:
1. As our project continues, the user base of GistNote also grew tremendously. The scalability of our search engine had become increasingly important to us. Solr’s performance had been spectacular and would definitely help the future expansion.
2. Solr supports indexing and searching sites and attachments, especially PDFs, which is a frequent research resource.
3. Solr is a popular engine, and is well supported by the community.

Recommendation System
There are two main approaches in developing a user friendly and accurate recommendation system to recommend relevant research resources, namely collaborative filtering and content-based filter. In our scenario of recommending users extra research resources from the resources that the user is currently using and the topic they are writing about, both approaches require a deep understanding of the context of the current writing and resource. According to Beel, J., Gipp, B., Langer, S. et al., in other research recommender systems from research articles from 2000-2016, 55% approaches applied content-based filtering while 18% applied collaborative filtering. [4]

- Content-based filtering
Based on the content that the user is or was accessing, recommending similar resources by keywords was our first simple idea and implementation. More specifically, automatic summarization and keywords extraction or generation given a certain length of text, and inspecting the frequency of the keywords in the candidates to retrieve and rank the information. Utilizing the existing unsupervised machine learning algorithm TextRank, derived from Google search ranking algorithm PageRank, we were able to obtain acceptable results of important keywords within a document by the co-occurrence of tokens. In depth evaluation of the performance is in the following chapter of this report.
Collaborative filtering

Feeding the keyword extraction result to the information retrieval system we developed and described above is a simple system that worked well. The next step was to include semantic search using LSA in the similarities score between documents in our search algorithm to allow retrieval of information with blurry keywords. However, we felt the possibility and were encouraged by our supervisor Dr. Yiu to go deeper in developing the recommendation algorithm. Using machine learning to model either user behaviour in choosing research resource or the resource chosen for certain topics are the two main possibilities. Considering that a single user can be researching multiple completely unrelated topics, and previously researched topics of a user does not necessarily affect future research topics, modelling user behaviour simply does not provide us with the accuracy required to make the experience user friendly. Topic modelling, on the other hand, is useful in allowing accurate retrieval and access to resources with similar topics that other users have researched.

A logic step to start topic modelling is to use the popular Latent Dirichlet Allocation. We considered three options in the size of the training corpus of the LDA model.

1. Create a new model for every query. The number of topics of each model is determined by the length of the input text. (very specific topics, performance)
2. Model the topics from the searchable candidates in our database, which is a several hundred-megabyte corpus. (number of topics problem, drift, grows)
3. Model the topics from huge existing online corpus. For example, the English Wikipedia article dump, which is a 10-20 gigabyte corpus. (number of topics problem, time required)
Experiments

Three different approaches to the recommendation systems was implemented for comparison:

1. Filter out stopwords and stem the remaining words to create a bag of words for the text. Apply content-based filtering by extracting keywords from the input with TextRank. The keywords are then feed into the search engine implemented by Solr.

2. An LDA model is created for every entry to extract the topic of the input text from the bag of words of the input. The highest probability words in the topics extracted is treated as keywords and given to the search engine.

3. An LDA model is trained with the English Wikipedia article dump of articles of March 2017. 100 topics is generated by the model. The trained model is then saved to allow access when queries come in. The model is used to infer topic of the input. The highest probability words in the topics inferred is treated as keywords and given to the search engine.

All three approaches are written in Python and as a web server receiving POST requests. This allow us to easily perform unit testing without any GistNote implementation and allow us to quickly integrate the recommendation system to existing GistNote system. However, our implementation is not scalable, which after some consideration was deemed acceptable as only an experiment.

Results

Accuracy

Information retrieval accuracy is usually analysed by two properties, namely precision and recall, and also the F-score, shown in figure 4 and 5, the harmonic mean of the two. However, since our number of candidate is so large, and there is no ground truth in our recommendation, we are going to consider the average precision score of the top ten results that we considered relevant. (AP@10) Due to time consideration we only analysed with two sample text, one short, one medium length. (See appendix)

\[
F_1 = 2 \cdot \frac{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}}{\text{precision} \cdot \text{recall}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\]

Figure 4: Calculation of the F-score
Figure 5: Graphical representation of the concept of precision and recall
1st Approach
Using sample text 1:

Introduction to Automatic Text Summarization – Algorithm

- Figure 6: 1st approach, recommendation result for sample text 1

Number 1, 4, 9 shown in figure 6 are relevant.

AP@10 = (1/1 + 2/4 + 3/9) / 10 = 0.183
Using sample text 2:

**World War II - Wikipedia**

- In June 1941, the European Axis powers launched an armistice in the Soviet Union, opening the largest land theatre of war in history, which wrapped the entire periphery of the Axis' colonial forces into one entire front. In December 1941, Japan attacked the United States and European colonies in the Pacific Ocean, and quickly conquered much of the so-called "French Pacific".
- On 7 September 1939—2 September 1940 (or 8 years and 1 day later), London, Europe, Eastern, Southeast Asian, China, Middle East, Mediterranean, North Africa, and Massa of Africa, South and South America, Russia, Italy, and China, revealed the collapse of the Soviet Union in 1945, the disintegration of the United Nations, the emergence of the United States, and the Soviet Union in 1945, (or 8 years and 1 day later).

**A Brief History of the U.S. Army in World War II**

- In 1941, the United States entered the war, but its forces were initially delayed. In 1943, the United States and European forces were victorious over Germany, Italy, and Japan.
- A worldwide depression that began in 1929 destroyed the fragile democratic regime in Germany. In 1933, Adolf Hitler led to power the National Socialist German Workers (Nazi) Party, a regime responsible for mass murder and extermination.

**Allies of World War II - Wikipedia**

- On October 23, 1940, the first time an Axis power invaded a neutral country, an Axis force landed in Greece, and the Second World War had started. In 1943, Allied forces invaded Italy, and the war ended in 1945.

**BBC - History - World War Two: Summonary Outline of Key Events**

- A peace treaty was signed in 1945, ending the war in Europe. Japan surrendered unconditionally.
- The United Nations was created to prevent future wars.
- The United States and the United Kingdom were the major contributors to the war's end.

**War the Atomic Bombing of Japan Necessary? | By Robert Freeman | Common Dreams**

- Stalin agreed with Hitler's plan to use the atomic bomb in 1945, but he agreed to use it only if the United States would not use it in the war.
- The United Nations was created to prevent future wars.
- The atomic bomb had nothing to do with the end of the war, but it solved the problem of how to defeat the Soviet Union in Europe.

**The Decision to Drop the Bomb (polihistory.org)**

- A 10-million-ton bomb was dropped on Hiroshima, Japan, on 6 August 1945, and a 20,000-ton bomb was dropped on Nagasaki, Japan, on 9 August 1945.
- The United Nations was created to prevent future wars.
- Two days later, the Soviet Union declared war on Japan, and the war ended.

**Operation Tootable - Wikipedia**

- The goal of this operation was to capture the strategically important French town of Falaise and then the smaller towns of Ton and Chinon.

**World War II aftermath occupied Italy**

- The German invasion of Italy, which replaced the country's government, continued on the heels of the defeat of France, and the war ended.
- The United Nations was created to prevent future wars.
- The Nlea, gave the Italian soldiers the choice of fighting with the Germans or surrendering.

**CHAPTER 25: World War II: The War Against Japan**

- After the war, Japan was occupied by the United States and the United Kingdom.
- The United States and the United Kingdom occupied Japan, and the war ended.

**The History Place - This Month in History: August**

- On August 6, 1945, the first Atomic Bomb, developed by the U.S., was dropped on the Japanese city of Hiroshima.
- August 15, 1945: Following the two Atomic Bomb drops and the surrender of the emperor, the war ended.
- The United States and the United Kingdom occupied Japan, and the war ended.

Figure 7: 1st approach, recommendation result for sample text 2. Numbers 1-6, 8-10 shown in figure 7 are relevant.

AP@10 = (1/1+2/2+3/3+4/4+5/5+6/6+7/8+8/9+9/9+10)/10 = 0.866
2\textsuperscript{nd} Approach
Using sample text 1:

Figure 8: 2\textsuperscript{nd} approach, recommendation result for sample text 1

Number 1,2,3 shown in figure 8 are relevant.

AP@10 = (1/1+2/2+3/3)/10 = 0.3
Figure 9: 2nd approach, recommendation result for sample text 2

Number 1-10 shown in figure 9 are relevant.

AP@10 = 1
3rd Approach
Using sample text 1:

Data dictionary, Wikipedia
Highlights:
- A data dictionary system is used only by the designers, users, and administrators and not by the DBMS software; it is called a passive data dictionary.

Who issues SQL queries? (byom); Khan Academy
http://www.khanacademy.org/computer-programming/using-sql-to-query-a-database/5567777452735552
Highlights:
- SQL statements help process queries
- SQLite is a free software

Web Site
http://example.com
Highlights:
- In any system, the crucial thing is to be able to access the data in an unambiguous way and to have a way to update the data
- User interaction helps process a query
- SQLite is a free software

Not Invented Here -- Brave New Geek
http://bravenewgeek.com
Highlights:
- if it's a core business function—do it yourself, no matter what.

git.flipbox.org Git; flipbox.github.com · COPYING-LGPL-2.1
http://git.flipbox.org
Highlights:
- to make sure the software is free for all its users
- to make sure the software is free for all its users

What is the difference between dataset and database? - Stack Overflow
Highlights:
- In American English, "dataset" means "an organized collection of data.
- A database is usually under the control of a database management system, which in software first manages other things, manages access to the database.
- In American English, "dataset" usually refers to data selected and arranged in rows and columns for processing by statistical software.
- A dataset is the data—usually not a table or vector, such as XML or other types of data. However, in its data, it doesn't really do anything.

Crash Course: UI Design — HII Design — Medium
http://medium.com/
Highlights:
- Information architecture is concerned with how the information within an app is organized and how users cognitively process the information.
- Use flows may not be the specific purpose of user onboards or through the app to help solve their specific need. It is concerned with the most logical steps for your users to gain their needs.

How to build a Data Science portfolio that will get you a data scientist job?
http://medium.com/
Highlights:
- In Kaggle data science competitions data is cleaned and formatted and the candidates only have to focus on building the model whereas Declarative datasets focus on data science projects that begin with collecting data and navigating through the entire data science project lifecycle.

R programming language demands the right use case
http://medium.com/
Highlights:
- The growing popularity among programmers and organizations has attracted the attention of technology vendors, many of which now offer R-based products or support the language as their own software.

NotION vs SQL - 4 Reasons Why NotION is better for Big Data applications
http://example.com
Highlights:
- "NotION uses much less memory and has a smaller footprint. NotION uses fast and efficient algorithms to handle the big data operations.

Figure 10: 3rd approach, recommendation result for sample text 1
None shown in figure 10 are relevant.
AP@10 = 0
Using sample text 2:

**Figure 11:** 3rd approach, recommendation result for sample text 2

Number 10 shown in figure 11 is relevant.

\[ \text{AP@10} = 0.01 \]
From the results, we can see that the 3rd approach is the worst performing one, and the 2nd is the best. We also found that the 3rd approach produced more general higher level topics instead of good keywords, so it may still have it use when users want to explore similar resources that is less directly related.

**Speed and Scalability**
Time of the execution is measured as the time between the input text arrival to the server and passing the keywords to Solr. The time required to train the Wikipedia model was measured separately.

**1st Approach**
Using sample text 1:
Mean of 5 measurements: 0.0133s

Using sample text 2:
Mean of 5 measurements: 0.114s

**2nd Approach**
Using sample text 1:
Mean of 5 measurements: 0.0732s

Using sample text 2:
Mean of 5 measurements: 0.629s

**3rd Approach**
Using sample text 1:
Mean of 5 measurements: 0.926s

Using sample text 2:
Mean of 5 measurements: 1.15s

Time required to train the model using the multicore implementation of LDA in the Python gensim package: 4 hours 31 minutes (Intel Core i7 4770k, using 3 worker threads)
<table>
<thead>
<tr>
<th>Approach</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample text 1</td>
<td>0.0133s</td>
<td>0.0732s</td>
<td>0.0926s</td>
</tr>
<tr>
<td>Sample text 2</td>
<td>0.114s</td>
<td>0.629s</td>
<td>1.15s</td>
</tr>
</tbody>
</table>

Table 4: Mean of time measured

While the 3rd approach is again the worst, we can observe that the time growth is only at the level of the 1st approach. Considering the percentage growth, the 3rd approach grows the least. 2nd produced the most accurate results, but its scalability is the worst.

**Difficulties and Limitations**

The biggest problem we faced during development and testing is the specification required to quickly train an LDA model of the entire Wikipedia corpus, especially disk space and memory. The disk space available for the project virtual machine is fairly sized. However, it is not enough to contain the Solr search engine with our test data, English stopwords data, English stemming data, the entire English Wikipedia corpus, TF-IDF vectors for the Wikipedia corpus and the trained LDA model. Not even my personal computer had enough space for the process. Similarly, the memory required to train the model is also huge. We have encountered at least 3 memory problems trying to train the model. At the end, we limited our worker threads to 3 to reduce the simultaneous memory requirement. Disk space was cleared out from my personal computer by getting an external hard drive, and we trained the model on my personal computer, later transferring the trained model to the project virtual machine. We still did not manage to train the model with the intended number of topics due to memory constraints. The final number of topics is 100, which we found unsatisfactory. However, the remaining time did not allow us to retrain a new model, debug and rerun the testing.

We would also like to have a 4th approach, with another LDA model that is trained with our own data. However, it would require us to develop a new preprocessing that is different from the Wikipedia corpus preprocessing, which took 15 hours 52 minutes. The project virtual machine’s remaining disk space also may not be enough to train and maintain another model, which needs constant updating when more data from GistNote arrive.
Concerns

Security
Currently, both the Solr search engine server and our own NLP server APIs are hosted with HTTP, i.e. without encryption. The APIs also do not require any password authentication or API keys for access. This is absolutely not the proper way to implement an application’s security. However, considering that this setup is only for testing and demonstration purposes, we relaxed our security requirement. Moreover, our server does not handle concurrent queries, and Solr handles overloading very well, so we are less concerned about the possibility of crashing the server machine.

Legal
Since in our testing we are using the highlight data from GistNote. We have been paying a close attention to data privacy issues. In order to avoid violating the data privacy, we have contacted GistNote to confirm that our use is not in violation of the data privacy agreement of the application. Furthermore, we have stripped the user identifiers off the data. Any entry of highlight data in our system cannot be directly traced back to any entry in the GistNote system.

Conclusion and Future Works
From the results of the testing, one can observer that the improvement of recommendation from a topic model is not dramatic over simple content-based filtering using keywords extraction algorithms for our specific scenario. One likely explanation is that our information retrieval system is not optimized for document relevance to topics generated from our LDA model. The fact that the most probable inputs to the system are short text with incomplete sentences also does not help our LDA model result. Here we propose that another possible approach to be to index our data by topics weighting from an LDA model that contain at least several times more topics instead of TF-IDF vectors. Information retrieval is then based on the cosine similarity between the query and the documents in the system. While a similar research on this specific topic had been completed recently, with their implementation providing marginal improvement in certain scenario and number of topics, we feel that the performance on a large corpus warrant further research and improvement. [5]
References


List of Figures

Figure 1. Representation of the Google PageRank algorithm Page 9
Figure 2. Example of a co-occurrence graph of a body of text Page 9
Figure 3. Calculation of term weight using TF-IDF Page 10
Figure 4. Calculation of the F-score Page 13
Figure 5. Graphical representation of the concept of precision and recall Page 14
Figure 6. 1st approach, recommendation result for sample text 1 Page 15
Figure 7. 1st approach, recommendation result for sample text 2 Page 16
Figure 8. 2nd approach, recommendation result for sample text 1 Page 17
Figure 9. 2nd approach, recommendation result for sample text 2 Page 18
Figure 10. 3rd approach, recommendation result for sample text 1 Page 19
Figure 11. 3rd approach, recommendation result for sample text 2 Page 20

List of Tables

Table 1. Summary of current academic research tools Page 6
Table 2. Comparison of project deliverable and other research tools Page 7
Table 3. Average precision@10 Page 21
Table 4. Mean of time measured Page 22
Appendix

Sample Text 1
Automatic summarization is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document. As the problem of information overload has grown, and as the quantity of data has increased, so has interest in automatic summarization. Technologies that can make a coherent summary take into account variables such as length, writing style and syntax. An example of the use of summarization technology is search engines such as Google. Document summarization is another.

Sample Text 2
World War II (often abbreviated to WWII or WW2), also known as the Second World War, was a global war that lasted from 1939 to 1945, although related conflicts began earlier. It involved the vast majority of the world's nations—including all of the great powers—eventually forming two opposing military alliances: the Allies and the Axis. It was the most widespread war in history, and directly involved more than 100 million people from over 30 countries. In a state of "total war", the major participants threw their entire economic, industrial, and scientific capabilities behind the war effort, erasing the distinction between civilian and military resources. Marked by mass deaths of civilians, including the Holocaust (in which approximately 11 million people were killed) and the strategic bombing of industrial and population centres (in which approximately one million were killed, and which included the atomic bombings of Hiroshima and Nagasaki), it resulted in an estimated 50 million to 85 million fatalities. These made World War II the deadliest conflict in human history.

The Empire of Japan aimed to dominate Asia and the Pacific and was already at war with the Republic of China in 1937, but the world war is generally said to have begun on 1 September 1939 with the invasion of Poland by Nazi Germany and subsequent declarations of war on Germany by France and the United Kingdom. From late 1939 to early 1941, in a series of campaigns and treaties, Germany conquered or controlled much of continental Europe, and formed the Axis alliance with Italy and Japan. Under the Molotov–Ribbentrop Pact of August 1939, Germany and the Soviet Union partitioned and annexed territories of their European neighbours, Poland, Finland, Romania and the Baltic states. The war continued primarily between the European Axis powers and the coalition of the United Kingdom and the British Commonwealth, with campaigns including the North Africa and East Africa campaigns, the aerial Battle of Britain, the Blitz bombing campaign, the Balkan Campaign as well as the long-running Battle of the Atlantic. In June 1941, the European Axis powers launched an invasion of the Soviet Union, opening the largest land theatre of war in history, which trapped the major part of the Axis' military forces into a war of attrition. In December 1941, Japan attacked the United States and European territories in the Pacific Ocean, and quickly conquered much of the Western Pacific.

The Axis advance halted in 1942 when Japan lost the critical Battle of Midway, near Hawaii, and Germany was defeated in North Africa and then, decisively, at Stalingrad in the Soviet Union. In 1943, with a series of German defeats on the Eastern Front, the Allied invasion of Sicily and the Allied invasion of Italy which brought about Italian surrender, and Allied victories in the Pacific, the Axis lost the initiative and undertook strategic retreat on all fronts. In 1944, the Western Allies invaded German-occupied France, while the Soviet Union regained all of its territorial losses and invaded Germany and its allies. During 1944 and 1945 the Japanese suffered major reverses in mainland Asia in South Central China and Burma, while the Allies crippled the Japanese Navy and captured key Western Pacific islands.

The war in Europe concluded with an invasion of Germany by the Western Allies and the
Soviet Union, culminating in the capture of Berlin by Soviet and Polish troops and the
subsequent German unconditional surrender on 8 May 1945. Following the Potsdam
Declaration by the Allies on 26 July 1945 and the refusal of Japan to surrender under its terms,
the United States dropped atomic bombs on the Japanese cities of Hiroshima and Nagasaki on
6 August and 9 August respectively. With an invasion of the Japanese archipelago imminent,
the possibility of additional atomic bombings, and the Soviet Union's declaration of war on
Japan and invasion of Manchuria, Japan surrendered on 15 August 1945. Thus ended the war
in Asia, cementing the total victory of the Allies.
World War II altered the political alignment and social structure of the world. The United
Nations (UN) was established to foster international co-operation and prevent future conflicts.
The victorious great powers—the United States, the Soviet Union, China, the United
Kingdom, and France—became the permanent members of the United Nations Security
Council. The Soviet Union and the United States emerged as rival superpowers, setting the
stage for the Cold War, which lasted for the next 46 years. Meanwhile, the influence of
European great powers waned, while the decolonisation of Asia and Africa began. Most
countries whose industries had been damaged moved towards economic recovery. Political
integration, especially in Europe, emerged as an effort to end pre-war enmities and to create a
common identity.