

**Summarizing Text Articles with Dirichlet Distribution**

by

Noor Zalifah Mohamed

Dissertation submitted in partial fulfillment of

the requirements for the

Bachelor of Technology (Hons)

(Business Information System)

SEPTEMBER 2011

Universiti Teknologi PETRONAS

Bandar Seri Iskandar

31750 Tronoh

Perak Darul Ridzuan

CERTIFICATION OF APPROVAL

**Summarizing Text Articles with Dirichlet Distribution**

by

Noor Zalifah Mohamed

A project dissertation submitted to the  
Business and Information System Programme  
Universiti Teknologi PETRONAS  
in partial fulfillment of the requirements for the  
BACHELOR OF TECHNOLOGY (Hons)  
(BUSINESS INFORMATION SYSTEM)

Approved by,



Elaine Chen Yoke Yie

UNIVERSITI TEKNOLOGI PETRONAS

TRONOH, PERAK

MAY 2011

## CERTIFICATION OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this project, that the original work is my own except as specified in references and acknowledgements, and that the original work contained herein have not been undertaken or done by unspecified sources or persons.



---

Noor Zalifah Mohamed

## **ABSTRACT**

The Latent Dirichlet Allocation (LDA) is based on the hypothesis that a person writing a document has topics in mind. To write about a topic then means to pick a word with a certain probability from the pool of words of that topic. A document can then be represented as a mixture of various topics. LDA is a generative probabilistic model for a corpus of discrete data, such as the words in a set of documents. LDA models the words in the documents under “bag-of-words” assumption, which basically ignores the orders of the words in the documents. Following this “exchangeability”, the distribution of the words would be independent and identically distributed given conditioned on some parameters. This conditionally independence allows us to build a hierarchical Bayesian model for a corpus of documents and words. The objective is to develop a text summarization system base on the Latent Dirichlet Allocation (LDA) method. The system would be used to determine the accuracy level of the method. This is done by comparing the result produced by the text summarization system with an existing summary that is produced by a human.

## ACKNOWLEDGEMENT

In the name of ALLAH S.W.T, the most merciful and compassionate, praise to ALLAH, he is the almighty, eternal blessing and peace upon the Glory of the Universe, our beloved Prophet Muhammad (S.A.W), and his family and companions.

Upon completing one year of final year project, the author is greatly indebted to personnel namely below.

First and foremost **Elaine Chen Yoke Yie (Supervisor)** who found time in a very busy schedule to give author new functional scope tasks, monitors progression and answer questions. Author feels grateful for being placed under his supervision and also deeply grateful for his advices, encouragements and patience throughout the duration of the project. The support he has given to the author as a student in civil scope of work has been greatly appreciated too.

Author also would like to thank the Department Computer Information Science of UTP, for the chance to write this paper.

Last but not least, author's sincere appreciation also extends to all fellow colleagues and others who provided assistance at various occasions.

# TABLE OF CONTENTS

ABSTRACT.....	iii
ACKNOWLEDGEMENT .....	iv
CHAPTER 1 .....	1
INTRODUCTION .....	1
1.1. Background Study.....	1
1.2. Problem Statement .....	1
1.3. Objective and Scope of Study.....	2
CHAPTER 2 .....	3
LITERATURE REVIEW .....	3
2.1 How should a text summarization system proceed? .....	5
2.2 Approaches to text summarization.....	6
2.2.1 Classical Approaches.....	6
2.2.2 Corpus-based Approach .....	8
2.3 Method 1: Segmented topic model based on the two-parameter Poisson-Dirichlet process.....	9
2.3.1 Introduction .....	9
2.3.2 Segmented topic model .....	9
2.4 Method 2: Latent Dirichlet Allocation (LDA).....	13
CHAPTER 3 .....	17
METHODOLOGY .....	17
3.1 Research Methodology.....	17
3.2 Development Methodology.....	17
3.3 Tool Required.....	19
CHAPTER 4 .....	21
SYSTEM DEVELOPMENT .....	21
4.1 Design.....	21
4.2 Samples .....	26
CHAPTER 5 .....	28
RESULTS AND DISCUSSION .....	28
5.1 Psychology Review.....	29
CHAPTER 6 .....	33

CONCLUSION & RECOMMENDATIONS .....	33
6.1 Conclusion.....	33
6.2 Recommendations.....	33
REFERENCES .....	34

## LIST OF FIGURES

Figure 1: Graphical model representation of LDA .....	14
Figure 2: Waterfall prototyping methodology .....	18
Figure 3: Activity diagram of system .....	21
Figure 4: Use case diagram of the system .....	22
Figure 5: Proposed design.....	23
Figure 6: Current system.....	23
Figure 7: System displaying contents of text file.....	24
Figure 8: Shows the two files created by the system.....	25
Figure 9: Manipulating N - psychology article.....	30
Figure 10: Manipulating N - breast cancer article .....	32

## LIST OF TABLES

Table 1: Results by manipulating N – psychology article .....	29
Table 2: Manipulating N - breast cancer article.....	31



# **CHAPTER 1**

## **INTRODUCTION**

### **1.1. Background Study**

With the rapid growth of the World Wide Web and electronic information services, information is becoming available on-line at an incredible rate. No one has time to read everything, yet we often have to make critical decisions on what we are able to assimilate. The technology of automatic text summarization is becoming indispensable for dealing with this problem.

Automatic text summarization is the technique, where a computer summarizes a text. It retains the relevant points in context of the subject matter and in context of how the author of the document intended for us to consume it. This technique has its roots in the 60'. Today with the Internet and the WWW, the technique has become more important.

### **1.2. Problem Statement**

Due to the steadily growing amount of unstructured text on the web, it becomes more and more important to use methods for automatic text summarization, to get control over the information flood. The goal of such methods is to take one or more input texts and transform them into a shorter text. A summary should be informative and readable and should preserve the meaning of the original texts.

Simple methods to produce a summary – choose the first paragraph, count word frequencies, or look for cue words. More sophisticated methods use techniques from Natural Language Processing (e.g lexical chains, or the rhetorical structure theory), and utilize machine learning techniques (e.g Naïve Bayes, or decision trees).

This paper is used to determine the accuracy of using the Latent Dirichlet Allocation (LDA). Besides implementing this method, it would be used to compare with a summary generated by a human.

### **1.3. Objective and Scope of Study**

The objective is to develop a text summarization system base on the Latent Dirichlet Allocation (LDA) method. The system would be used to determine the accuracy level of the method. This is done by comparing the result produced by the text summarization system with an existing summary that is produced by a human.

## CHAPTER 2

### LITERATURE REVIEW

The subject of summarization has been investigated by the Natural Language Processing (NLP) community for nearly the last half century. Radev et al. (2002) define a summary as “a text that is produced from one or more texts, that convey important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that” (Das & Martins, 2007). This simple definition captures three important factors that characterize research on automatic summarization.

- Summaries may be produced from a single document or multiple documents
- Summaries should preserve important information
- Summaries should be short

We start by introducing some common terms in the summarization world:

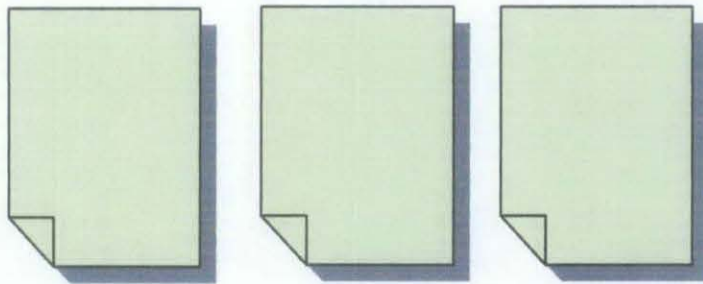
<b>Extraction</b>	Procedure of identifying important sections of the text and producing verbatim
<b>Abstraction</b>	Aims to produce important material in a new way
<b>Fusion</b>	Combines extracted parts coherently
<b>Compression</b>	Aims to throw out unimportant sections of the text

Earliest instances of research on summarizing scientific documents proposed paradigms for extracting salient sentences from text using features like word and phrase frequency, position in the text and key phrases. Various works published since then has concentrated on other domains, mostly on newswire data. Many

approaches addressed the problem by building systems depending on the type of required summary. While extractive summarization is mainly concerned with what the summary content should be, usually relying solely on extracting of sentences, abstractive summarization puts strong emphasis on the form, aiming to produce a grammatical summary. This usually requires advanced language generation techniques. In a paradigm more tuned to information retrieval (IR), one can also consider topic-driven summarization, that assumes that the summary content depends on the preference of the user and can be assess via query, making the final summary focused on a particular topic.

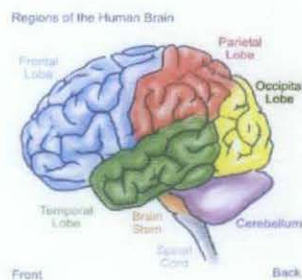
A crucial issue that will certainly drive future research on summarization is evaluation. During the last few years, many system evaluations have created sets of training material and have established baselines for performance levels. However, a universal strategy to evaluate summarization systems is still absent.

## 2.1 How should a text summarization system proceed?



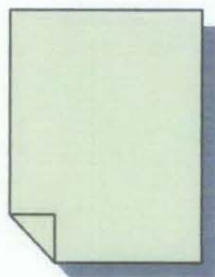
### Interpretation

Read the documents to obtain a text representation



### Transformation

Understand them and build a text representation into a summary representation



### Generation

Generate a summary from the summary representation

## 2.2 Approaches to text summarization

### 2.2.1 Classical Approaches

Many different approaches can be found with today's technology. Here we can look into a few different approaches.

#### *Surface level*

This approach inclines to represent information taking shallow features and then selectively combining them together in order to obtain a salience function that can be used to extract information. Among these features are:

- **Thematic features** rely on word (significant words) occurrence statistics. Thus, sentences containing words that occur frequently in a text have higher weight than the rest. That means that these sentences are the important ones and they are hence extracted. The term frequency technique is used to describe this. Before doing term frequency, a filtering task must be done using a stop-list words which contains words such as pronouns, prepositions and articles. This is the classical statistical approach. However, from a point of view of a corpus-based approach measure (commonly used in information retrieval) it is very useful to determine keywords in text.
- **Location** refers to the position in text, paragraph or any other particular section in the sense that they contain the target sentences to be included in the summary. This is usually genre-dependent, but there are two basic general methods, which are lead-method and title-based method. The first one consists of extracting only the first sentences, assuming that these are the most relevant ones. Whereas the second considers that words headings or titles are relevant to summarization.
- **Background** assumes that the importance of meaning units is determined by the presence of terms from the title or headings, initial part of the text or a user's query.

This means that words in the headings or initial parts of a text is considered important and is relevant to be inserted in a summary. This is also true for words from a user's query.

- **Cue words** and phrases such as “in conclusion”, “important”, “in this paper”, etc. can be very useful to determine signals of relevance or irrelevance. Words stated above can be considered relevant and words after it may be inserted in the summary.

### ***Entity Level***

This approach attempts to build a representation of the text, modeling text entities and their relationships. The objective is to help to determine what is salient. These relations between entities include:

- **Similarity** occurs when two words share a common stem. For example two words whose form is similar. This can be extended for phrases or paragraphs. Similarity can be calculated by vocabulary overlap or with linguistic techniques.
- **Proximity** refers to the distance between texts units. With that information it is possible to establish entity relations based on its distance.
- **Thesaural relationships among words** can be described as relationships like synonymy, hypernymy, part-of-relations (meronymy).
- **Coreference** is referring expressions can be linked so that, coreference chains can be built with coreferring expressions.
- **Logical relations** such as agreements, contradiction, entailment and consistency.
- **Syntactic relations** are based on parse trees.

- **Meaning representation-based relations**, establishing relations between entities in the text as for example, predicate-argument relations.

### *Discourse Level*

The target of discourse level approaches is to model the global structure of the text and its relations in order to achieve communicative goals. The information that can be exploited at this level is:

- **Format** of the document, such as hypertext markup or document outlines.
- **Threads of topics** as they are revealed in the text.
- **Rhetorical structure of text**, representing argumentative or narrative structure. The idea behind this deals with the possibility to build the coherence structure of a text, so that the ‘centrality’ of textual units will reflect their importance.

### **2.2.2 Corpus-based Approach**

A corpus-base approach is an approach whereby importance of different text features for any given summarization problem may be determined by counting the occurrences of features stated above in text corpora. A common use of a corpus is in calculating the importance of a word or phrase base on its frequency. Besides that, the *tf.idf* measure, a widely used measure in information retrieval as well as text summarization and is used to pick out words or phrases that distinguishes one document from another in a corpus.



## **2.3 Method 1: Segmented topic model based on the two-parameter Poisson-Dirichlet process**

### **2.3.1 Introduction**

The study of random probability measures has been around since the time of Bayes, its application to Bayesian non-parametric statistics proved burdensome and fairly intractable until a few years ago. A random probability measure was proposed, called a Dirichlet process, for treating Bayesian non-parametric problems. Ferguson defines the Dirichlet process by prescribing the joint distribution of this process applied to an arbitrary measurable partition of the measure space.

### **2.3.2 Segmented topic model**

A challenge in text analysis is the problem of understanding the document structure.

Given a collection of documents, each of which consists of a set of segments (e.g. sections, paragraphs, or sentences), each segment contains a group of words, we wish to explore the latent topic structure of each document by taking into account segments and their layout. In this method, it is believed that segments in a document not only have meaningful content but also provide preliminary structural information, which can aid in the analysis of the original text. The idea came about from the way people normally compose documents. When writing a document, we need to come up with the main ideas first, then organize segments around them and the ideas for segments could vary around the main ideas.

Take an essay as an example. Generally, an essay would have main ideas which indicate what the essay deals with. Then there are paragraphs, basic structural units in an essay which are organized around the main ideas. Besides that, one paragraph might have one or more ideas called sub-ideas in our work. These sub-ideas link to the main ideas. This means that they are not separated, but sub-ideas can

be more specific than main ideas, and generally be variations of them. The layout and progression of ideas give the meaningful structure of an essay.

For this method, the authors adopt probabilistic generative models called topic model. The idea is that each document is a random mixture over several latent topics, each of which is a distribution over words. Topic models specify a simple probabilistic process by which words can be generated.

A simple structure topic model using the two-parameter Poisson Dirichlet process (PDP) was developed based on recent theoretical results of the PDP for finite discrete cases. This allows a collapsed Gibbs sampler to be developed for the hierarchical structure model.

A Segmented Topic Model (STM) is a four level probabilistic generative topic model: two levels of proportions which consist of a level of topics and a level of words. Before specifying STM, here are the list of all notations and terminologies used.

- A word is the basic unit of our data, indexed by  $[1, \dots, W]$ .
- A segment is a sequence of  $L$  words. It can be a section, paragraph or even sentence. But in this method, we assume segments are paragraphs or sentences.
- A document is an assemblage of  $J$  segments
- A corpus is a collection of  $I$  documents

The basic idea of STM is to assume that each document  $i$  has a certain mixture of latent topics, denoted by probability  $\mu_i$ , and is composed of meaningful segments; each of these segments also has a mixture over the same space of latent topics as those for the document. And this is denoted by probability vector  $v_{i,j}$  for segment  $j$  of document  $i$ . Both the main ideas of a document and sub-ideas of its segments are modeled here by these distributions over topics. Sub-ideas are taken as variants of

the main ideas, and thus sub-ideas can be linked to the main ideas, giving correlations between a document and its segments.

How do the segment proportion  $v_{i,j}$  vary around the document proportions  $\mu_i$ ?

The use of PDP distribution as  $v_{i,j} \sim \text{PDP}(a, b, \mu_i)$  distribution is a key innovation. This equation is used instead of say  $v_{i,j} \sim \text{Dirichlet}(b \mu_i)$  where  $b$  plays the role of the “equivalent sample size”. However, such a distribution makes the prior non-conjugate to the likelihood so general MCMC sampling is required and parameter vectors such as  $\mu_i$  can no longer be integrated out to yield efficient collapsed Gibbs samplers. Thus, the following lemma is adopted from (Buntine and Hunter 2010):

*Lemma 1 The following approximations on distributions hold*

$$\text{PDP}(0, b, \text{discrete}(\Theta)) \sim \text{Dirichlet}(b\Theta),$$

$$\text{PDP}(a, 0, \text{discrete}(\Theta)) \sim \text{Dirichlet}(a\Theta) \rightarrow (\text{as } a \rightarrow 0),$$

<b>Notation</b>	<b>Description</b>
$K$	Number of topics
$I$	Number of documents
$J_i$	Number of segments in document $i$
$L_{i,j}$	Number of words in document $i$ , segment $j$
$W$	Number of words in dictionary
$\alpha$	Base distribution for document topic probabilities
$\mu_i$	Document topic probabilities for document $i$ , base distribution for segment topic probabilities
$v_{i,j}$	Segment topic probabilities for document $i$ and segment $j$
$\phi$	Word probability vectors as a $K \times W$ matrix

$\phi_k$	Word probability vector for topic $k$ , entries in $\phi$
$\gamma$	$W$ -dimensional vector for the Dirichlet prior for each $\phi_k$
$w_{i,j,l}$	Word in document $i$ , segment $j$ , at position $l$
$z_{i,j,l}$	Topic for word in document $i$ , segment $j$ , at position $l$

The PDP is a prior conjugate to the multinomial likelihoods, so allows Gibbs samplers of the kind used for Latent Dirichlet Allocation (LDA). Thus, conditioned on the model parameters  $\alpha$ ,  $\gamma$ ,  $\phi$  and PDP parameters  $a$ ,  $b$  (called discounts and strength respectively), STM assumes the following generative process for each document  $i$ :

1. Draw  $\mu_i \sim \text{Dirichlet}_k(\alpha)$
2. For each segments  $j \in \{1, \dots, J_i\}$ 
  - a) Draw  $v_{i,j} \sim \text{PDP}(a, b, \mu_i)$
  - b) For each  $w_{i,j,l}$  where  $l \in \{1, \dots, L_{i,j}\}$ 
    - i. Select a topic  $z_{i,j,l} \sim \text{discrete}_k(v_{i,j})$
    - ii. Generate a word  $w_{i,j,l} \sim \text{discrete}_w(\phi_{z_{i,j,l}})$

We have assumed the number of topics (i.e., the dimensionality of the Dirichlet distribution) is known and fixed, and the word probabilities are parameterized by a  $K \times W$  matrix  $\phi$ . The graphical representation of STM is shown in the figure 2.

Shaded nodes are observed random variables, non shaded nodes are latent random variables, and the plates indicate repeated sampling.

The goal of Gibbs sampling is to find estimates for the parameters of interest in order to determine how well the observable data fits the model of interest, and also

whether or not data independent of the observed data fits the model described by the observed data (Rouchka, 1997). Gibbs sampling requires a vector of parameters of interest that are initially unknown.

Gibbs sampling requires an initial starting point for the parameters as well. Then, one at a time, a value for each parameter of interest is sampled given values for the other parameters and data. Once all of the parameters of interest have been sampled, the nuisance parameters are sampled given the parameters of interest and the observed data. Then, the process is started over. The power of Gibbs sampling is that the joint distribution of the parameters will converge to the joint probability of the parameters given the observed data. (Rouchka, 1997)

#### **2.4 Method 2: Latent Dirichlet Allocation (LDA)**

Latent is something that is present or potential, but not evident or active (TheFreeDictionary) while allocation is to set apart for a special purpose or to designate (TheFreeDictionary).

The Latent Dirichlet Allocation (LDA) is based on the hypothesis that a person writing a document has topics in mind. To write about a topic then means to pick a word with a certain probability from the pool of words of that topic. A document can then be represented as a mixture of various topics.

LDA is a generative probabilistic model for a corpus of discrete data, such as the words in a set of documents. LDA models the words in the documents under “bag-of-words” assumption, which basically ignores the orders of the words in the documents. Following this “exchangeability”, the distribution of the words would be independent and identically distributed given conditioned on some parameters. This conditionally independence allows us to build a hierarchical Bayesian model for a corpus of documents and words. More specifically, the process of how LDA generates the words in a corpus can be illustrated by the graphical model representation below. (Chang & Yu)

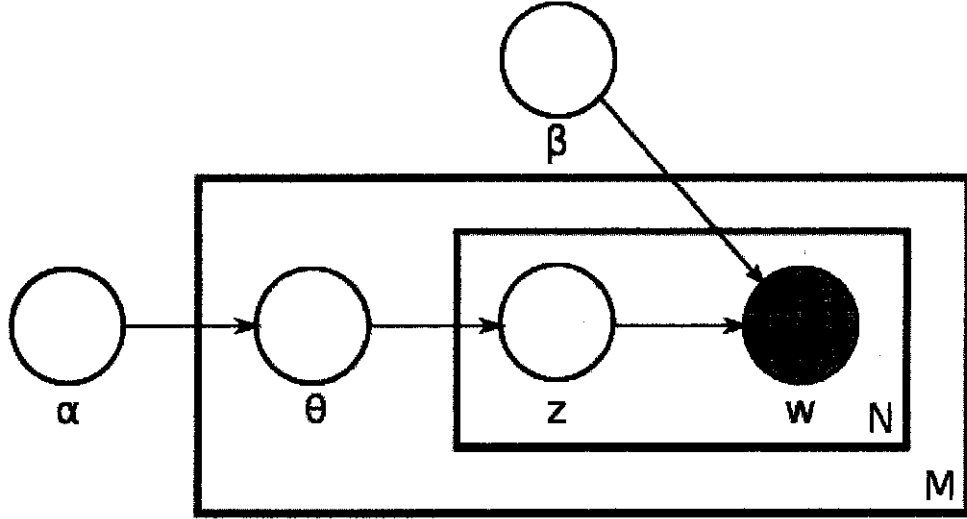


Figure 1: Graphical model representation of LDA

LDA helps explain the similarity of data by grouping features of this data into unobserved sets. A mixture of these sets then consists of the observable data. The method was first introduced by Blei et al and applied to solve various tasks including topic identification, entity resolution and Web spam classification (Krestel, Frankhauser, & Nejd, 2009).

The modeling process of LDA can be described as finding a mixture of topics for each resource. For example,  $P(z | d)$  with each topic described by terms following another probability distribution,  $P(t | z)$ . This can be formalized to be

$$P(t_i | d) = \sum_{j=1}^Z P(t_i | z_i = j) P(z_i = j | d), \quad (1)$$

Where  $P(t_i | d)$  is the probability of the  $i$ th term for a given document  $d$  and  $z_i$  is the latent topic.

$P(t_i | z_i = j)$  is the probability  $t_i$  within topic  $j$ .

$P(z_i = j | d)$  is the probability of picking a term from topic  $j$  in the document.

The number of latent topics  $Z$  has to be defined in advance and allows adjusting the degree of specialization of the latent topics. LDA then estimates the topic-term distribution  $P(t | z)$  and the document-topic distribution  $P(z | d)$  from an unlabeled corpus of documents using Dirichlet priors for the distributions and a fixed number of topics. Gibbs sampling is used to create the Dirichlet priors.

The goal of Gibbs sampling is to find estimates for the parameters of interest in order to determine how well the observable data fits the model of interest, and also whether or not data independent of the observed data fits the model described by the observed data (Rouchka, 1997). Gibbs sampling requires a vector of parameters of interest that are initially unknown.

Gibbs sampling requires an initial starting point for the parameters as well. Then, one at a time, a value for each parameter of interest is sampled given values for the other parameters and data. Once all of the parameters of interest have been sampled, the nuisance parameters are sampled given the parameters of interest and the observed data. Then, the process is started over. The power of Gibbs sampling is that the joint distribution of the parameters will converge to the joint probability of the parameters given the observed data. (Rouchka, 1997)

Gibbs sampling iterates multiple times over each term  $t_i$  in document  $d_i$ , and samples a new topic  $j$  for the term based on the probability  $P(z_i = j | t_i, d_i, z_{-i})$  based on Equation 2, until the LDA model parameters converge.

$$P(z_i = j | t_i, d_i, z_{-i}) \propto \frac{c_{t_i j}^{TZ} + \beta}{c_{t_j}^{TZ} + T\beta} \frac{c_{d_i j}^{DZ} + \alpha}{c_{d_i}^{DZ} + Z\alpha} \quad (2)$$

$C^{TZ}$  maintains a count of all topic-term assignments,  $C^{DZ}$  counts the document-topic assignments,  $z_{-i}$  represents all topic-term and document-topic assignments except the

current assignment  $z_i$  for term  $t_i$ , and  $\alpha$  and  $\beta$  are the hyper-parameters for the Dirichlet priors, serving as smoothing parameters for the counts. Based on the counts the posterior probabilities in Equation 1 can be estimated to be:

$$P(t_i | z_i = j) = \frac{c_{t_i j}^T Z + \beta}{\sum_t c_{d_i z}^D Z + T\beta} \quad (3)$$

$$P(z_i = j | d_i) = \frac{c_{d_i j}^D Z + \alpha}{\sum_z c_{d_i z}^D Z + Z\alpha} \quad (4)$$



## **CHAPTER 3**

### **METHODOLOGY**

#### **3.1 Research Methodology**

The study of automatic text summarizing is a vast field. A study that even after more than 50 years, has yet to find a system that is sufficient enough. Though this project focuses specifically on the method of segmented topic model based on the Latent Dirichlet Allocation method, to be able to implement this method, a thorough study regarding automatic text summarizing is needed.

Most important resources are located in the cyber world. Thus, for this project, majority of the resources were found online. Through the web, recent study regarding automatic text summarization can be found compared to other form of material.

Besides online materials, books regarding this topic and anything related to it can be found. However, the number is of limited value. Books are not as current compared to online resources. These are the two major resources used for the study of this project.

#### **3.2 Development Methodology**

The system will be developed using the Prototyping methodology. Prototyping methodology is one of the methods under the Rapid Application Development

(RAD) category of methodologies. RAD-based methodologies attempt to address both weaknesses of structured design methodology by adjusting the SDLC phases. This way, some part of the system will be developed quickly to better understand the system and continuously suggest revisions that bring system closer to what is needed.

A prototyping-based methodology performs the analysis, design and implementation phases concurrently. With all three phases performed repeatedly until the system is completed. The basic analysis and design are performed and work immediately begins on a system prototype with minimal features. The first prototype would consist of the first part of the system that is used. In this case, the prototype would consist of the input screen where users would be able to choose between pasting the text to be summarized or simply open a text document.

Prototype would be evaluated and commented on. Then the process of reanalyzing, redesigning and re-implementing would be done until a complete system have been developed.

Figure 1 below displays the development process using the Prototyping methodology.

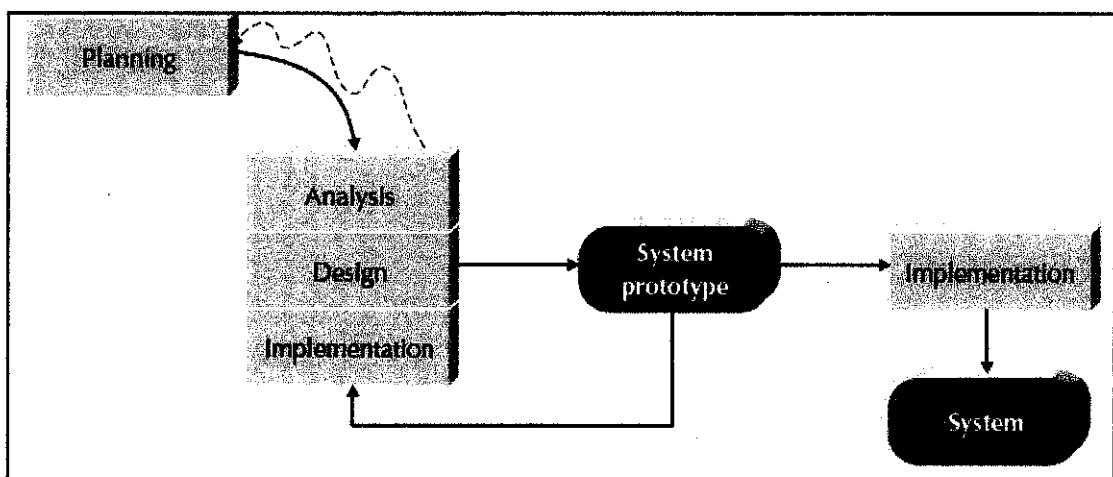


Figure 2: Waterfall prototyping methodology

### 3.3 Tool Required

To implement this method is to create an automatic text summarization system base on this method. This project would then focus heavily on the programming aspect of it. Method improvements would be done along the way.

Currently, there are many different programming languages that could be used for this particular system. Some of the more familiar languages would be C, C++ as well as Java. The full extent of the programming tool needed would be explored further as we go deeper into the project.

A database of documents would also be needed for this particular project. However, since text documents generally take up much storage space, no specific hardware is needed. For this project, a personal computer would suffice.

During the development of the system, three different platform were used; the Visual Studio, the C# platform, and last but not least the Matlab platform.

Each platform has its advantages and disadvantages. Both the Visual Studio and the C# platform enables one to easily create the graphical user interface, however, it is harder to code the statistical formula to extract the topic models from the document corpus. The challenge when it comes to coding the statistical formula is to actually understand the formula. Before any codes can be written, the formula to extract the topic models and to create a Gibbs sampling requires one to have a solid knowledge in statistical studies.

Ultimately the Matlab platform was chosen due to its capabilities as a high-level language and interactive environment that enables one to perform computationally

intensive tasks faster than with traditional programming languages such as C, C++ and Fortran. This is needed since to extract topic models from the document corpus, the system needs to process a mass amount of words and documents according to a statistical formula. MATLAB has been extended over the years to respond to the needs of various users. Hence, several toolboxes exist to add to the power of the original language.

Matlab was also chosen due to the fact that it contains a specific toolbox that would help facilitate the development of the system. The tool box that would be used in this particular project is called the Topic Modeling Toolbox 1.4. It is free for scientific use and was written by Mark Stevyers of University of California from the Department of Cognitive Sciences along with Tom Griffiths of University of California from the Department of Psychology (2011).

The toolbox helps identify the topic models which resides in a corpus documents. For now, its function to identify the most frequent words associated to a particular topic within a document corpus.

Topic models are based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words. A topic model is a generative model for documents: it specifies a simple probabilistic procedure by which documents can be generated. To make a new document, one chooses a distribution over topics. Then, for each word in that document, one chooses a topic at random according to this distribution, and draws a word from that topic. Standard statistical techniques can be used to invert this process, inferring the set of topics that were responsible for generating a collection of documents. (Stevyers & Griffiths)

## CHAPTER 4

### SYSTEM DEVELOPMENT

#### 4.1 Design

The system's design is fairly straightforward. The system will only interact with one user at a time and would refer to the corpus of whatever documents are available. Below is the activity diagram of the system:

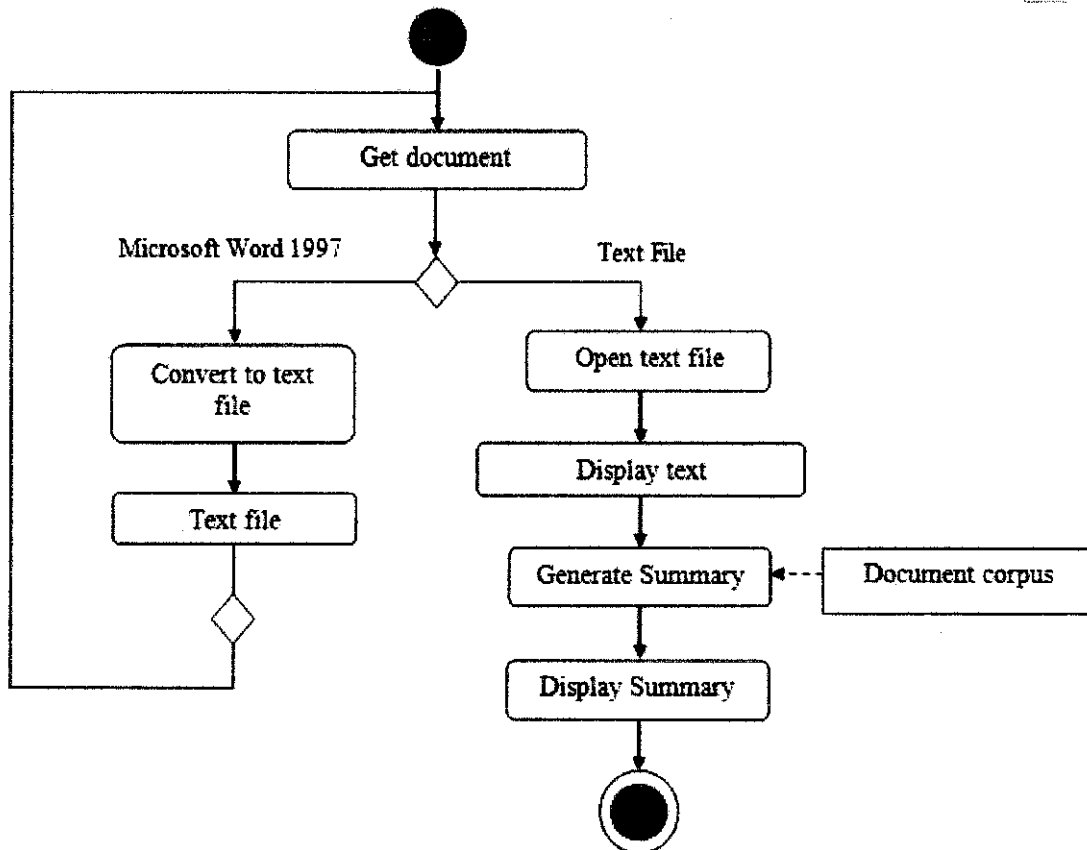


Figure 3: Activity diagram of system

Figure 1 shows the activity diagram of the system. It shows the processes that will be done by the system to accomplish its end task of summarizing a text article.

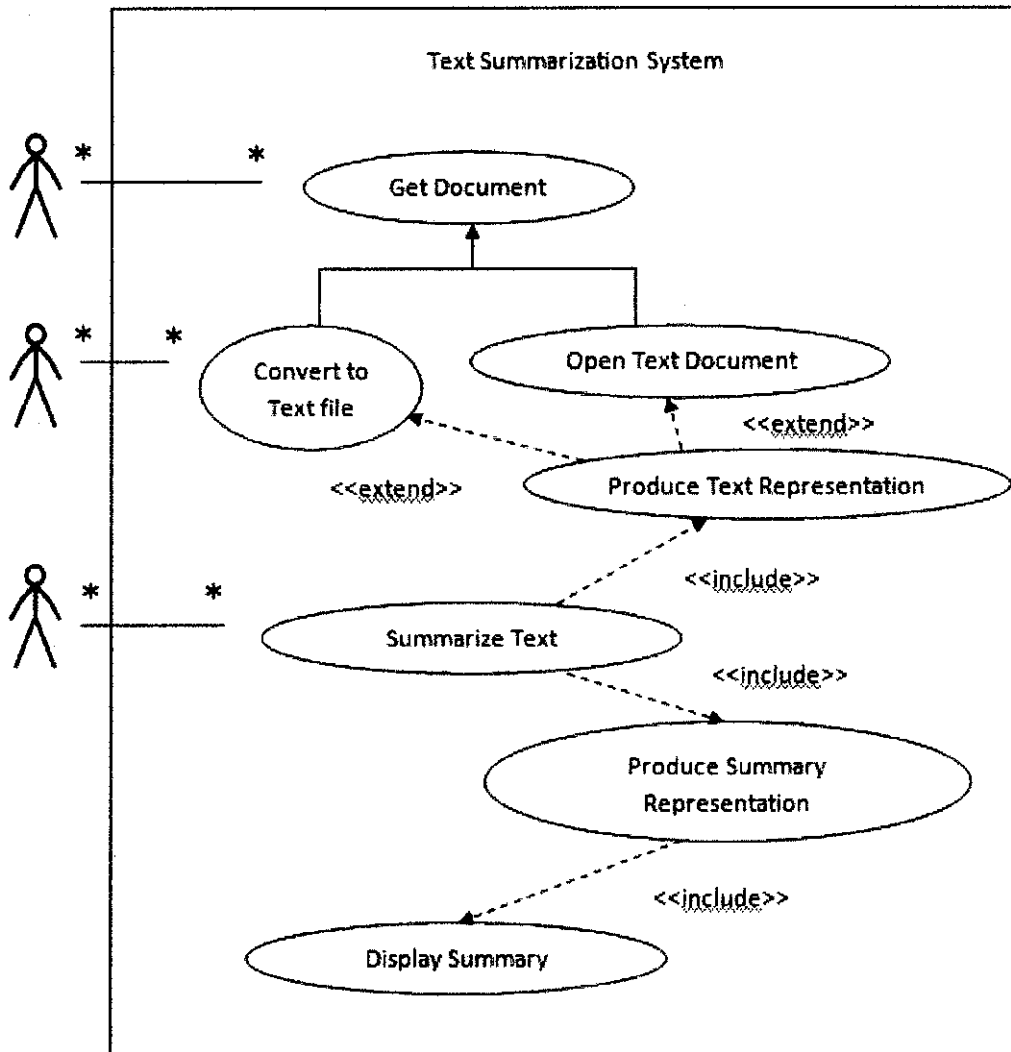


Figure 4: Use case diagram of the system

Figure 4 displays the use case diagram of the system. It tells readers how a user may interact with the system.

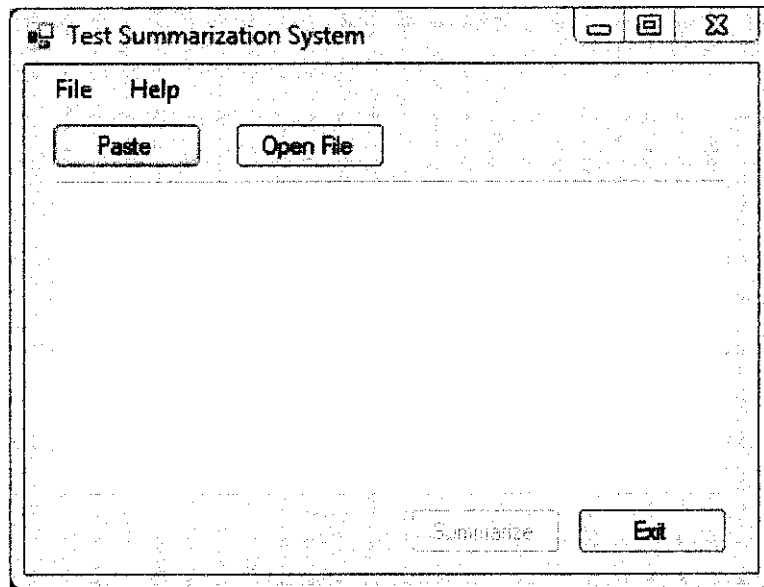


Figure 5: Proposed design

Figure 5 shows the proposed design of the system during development. However, after refinement of the system, it now looks as seen below.

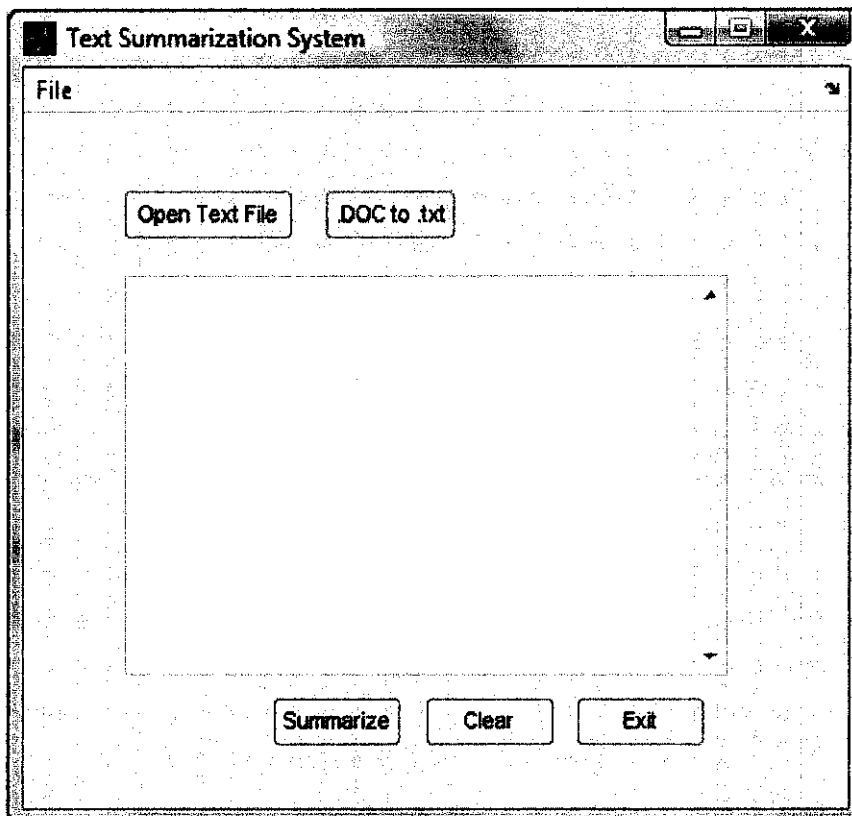


Figure 6: Current system

To start the system, user needs to click on the 'Open Text File' button to choose a text file to be summarized. The chosen file needs to be a text file with the '.txt' extension to it due to certain restrictions to the system. Once the user selects a text file, the contents of the text file would be displayed on the text box as seen below.

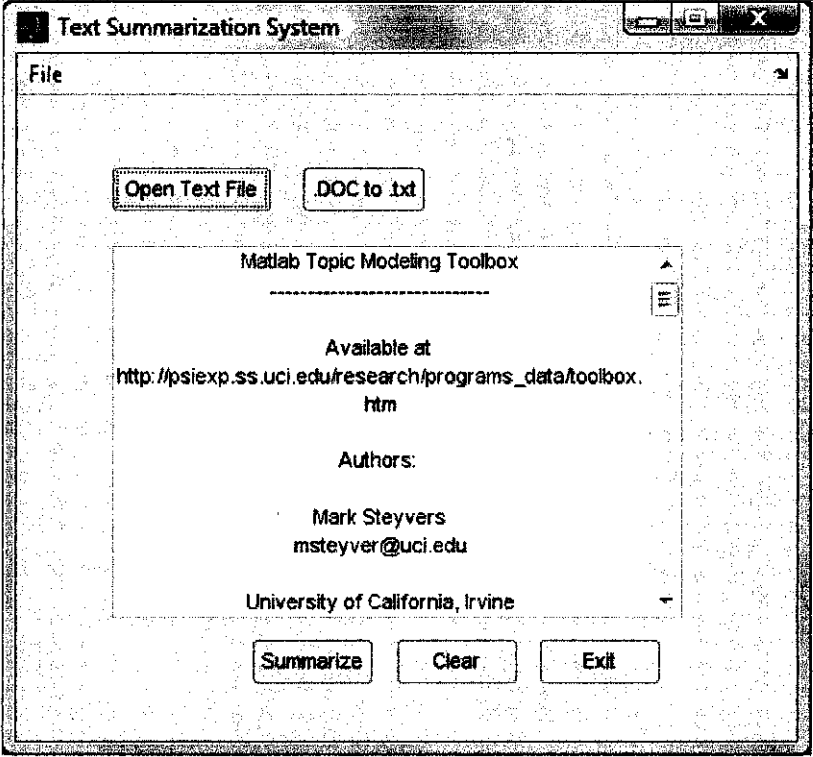


Figure 7: System displaying contents of text file

Once this is done, the user may summarize the contents by clicking on the 'Summarize' button.



Or user may click the 'Clear' button if the wrong text file was chosen at the earlier stage. This would then clear the text box and enables a user to open the correct text file.

Since currently the system only accepts text file documents as input, a user who has a file in the form of Microsoft Words 97 – 03 version can also convert the file into a text file format. This can be done by clicking on the '.DOC to .txt' button as can be seen in Figure 5 and Figure 6.

When the button is clicked, a new .txt file will be created in the same file path as the original document that was converted. This process will not however delete the original file. Thus, the end result would be that the user would now have two documents of the same name but in a different form.



Name	Type
Slides	File Folder
Survey	File Folder
 7. Comparison	Microsoft Office Word 97 - 2003 Document
 7. Comparison	Text Document

Figure 8: Shows the text file created by the system

## 4.2 Samples

For the purpose of this paper, the system would be tested to summarize documents from the medical industry. To further narrow it down, the study of Breast Cancer is chosen, simply because documents related to this study is abundant and easier to be obtained. Documents obtained are all in PDF format. Below is the list of documents to be used to test the system:

1. A Decade of Change An Institutional Experience
2. A Role for Estrogen Receptor Phosphorylation
3. Biological Characteristics and Medical Treatment
4. Bones, breasts, and bisphosphonates
5. Breast cancer and sexuality
6. Breast cancer in Singapore some perspectives
7. Breast cancer Malaysia
8. Cancer Multidisciplinary Team Meetings
9. Challenges in the development of future
10. Current and emerging treatment strategies
11. Diagnosis delay of breast cancer
12. Early Stage Breast Cancer and Its Association
13. Five Methods of Breast Volume Measurement
14. Global Health Inequalities and Breast Cancer
15. Help reduce your risk of breast cancer with vitamin D
16. Impact of Breast MRI on Surgical Treatment
17. In Search of Breast Cancer Culprits
18. Increased Circulating Level of the Survival Factor GP88
19. Lapatinib new opportunities for management
20. Malaysia And Breast Cancer

Before any of data can be used with the toolbox, it needs to go through another process whereby word frequencies are counted for every document. The information is then kept in a text file and is organized into three columns where each row contains the document index, the word index, and the word count. For example: 1 2

10, 1 3 4, 2 2 6 (where each comma represents a new line). This should be read as “word 2 occurs 10 times in document 1, word 3 occurs 4 times in document 1 and word 2 occurs 6 times in doc 2”.

Every document goes through the pre-processing where every word frequency is counted for. This word counts are then cross referred to words contained in the breast cancer vocabulary. The vocabulary is taken from the breastcancer.org website. Breastcancer.org is a nonprofit organization dedicated to providing the most reliable, complete and up-to-date information about breast cancer (breastcancer.org, 2011).

## CHAPTER 5

### RESULTS AND DISCUSSION

The summarization system select the most representative sentences in the input to form an extractive summary; whereby the selected sentences are strung together to form a summary without any modification of their original wording. In this kind of setting, information retrieval metrics of precision and recall are used. A person is asked to select sentences that seem to best convey the meaning of the text to be summarized and then the sentences selected automatically by the system are evaluated against the human selections.

Recall is the fraction of sentences chosen by the person that were also correctly identified by the system. It indicates what proportion of all the relevant sentences have been retrieved from the collection.

$$Recall = \frac{\text{system-human choice overlap}}{\text{sentences chosen by human}} \quad (5)$$

Precision is the fraction of system sentences that were correct. It indicates what proportion of the retrieved sentences is relevant.

$$Precision = \frac{\text{system-human choice overlap}}{\text{sentences chosen by system}} \quad (6)$$

F-score is a composite score that combines the precision and recall. It can be interpreted as a weighted average of the precision and recall. F-score reaches its best value at 1 and worst score at 0.

$$F - score = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (7)$$

For all the measurement, a higher value of each means that it the system is able to generate a summary that is relevant to the user. Thus, the higher value for each, the more accurate is the system.

## 5.1 Psychology Review

Testing is first done using the sample data provided by the toolbox. The samples used were words and vocabulary under the topic of psychology.

The full text is an article by Nassar-Mcmillan and Hakim-Larson titled ‘Counseling Considerations Among Arab Americans’ (2003). The summarized version was done by Mark H., a published author, ghostwriter, and editor. He is a free-lance published writer and editor who have experience in academic and environmental organizations. The summarizer also has a doctorate in English, masters in Professional Writing and a degree in English (Elance, 2011).

For this particular article, topic models are generated through the LDA method and cross referred to the original document. Here we are comparing the accuracy of the LDA method in extracting relevant sentences compared to the summary done by Mark H. Different results were received as shown in table below by using different numbers of topics being extracted.

Topics Extracted, N	Recall	Precision	F-score
20	0.52	0.12	0.195
30	0.39	0.12	0.184
40	0.09	0.06	0.072

Table 1: Results by manipulating N – psychology article

As can be seen, the more number of topics are extracted, the less accurate the system is able to extract relevant sentences from the original text.

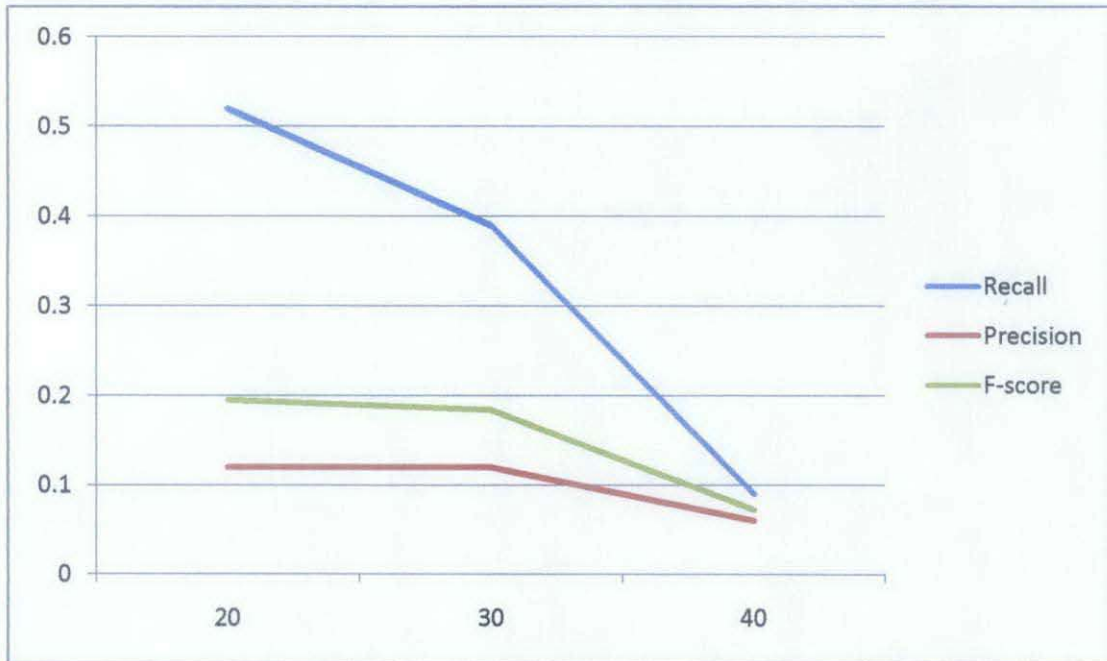


Figure 9: Manipulating N - psychology article

This is probably due to the small number of samples used in testing of the system. It is highly recommended that large values of data are used for the system to generate more populous topics.

The system needs to use as much sample possible to be more accurate in its extraction of topic models. The size of the sample data used is vital to have more topics that can be extracted. When the sample size is small, the number of topics that can be extracted becomes incredibly limited for the system to choose from. Hence, the above result was generated.

## 5.2 Breast Cancer

The next test is done using the bag of words retrieved using multiple journals on Breast Cancer. The vocabulary was taken from various sources, the main source coming from breastcancer.org.

The full text article is taken from the Natural Health Especially for Women website titled 'Breast Cancer: Alternative Treatment Hypothesis'. The summarized version is also taken from the same website.

The same process as in section 5.1 is repeated to get the table below.

<b>Topics Extracted, N</b>	<b>Recall</b>	<b>Precision</b>	<b>F-score</b>
20	0.11	0.14	0.12
30	0.05	0.60	0.09
40	0.02	0.31	0.03

Table 2: Manipulating N - breast cancer article

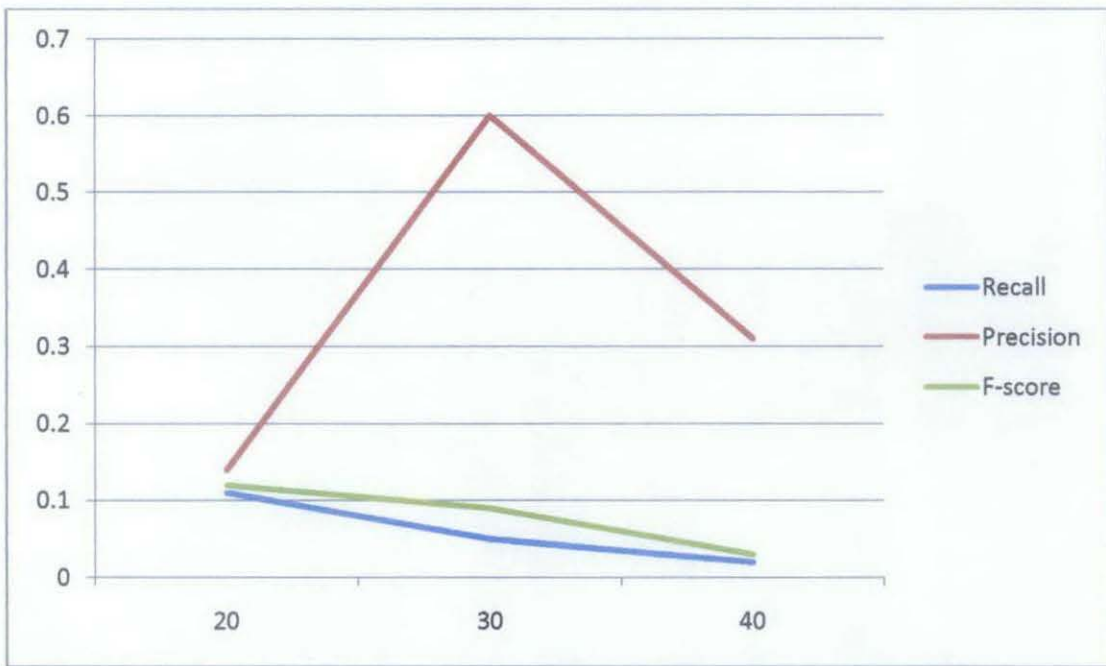


Figure 10: Manipulating N - breast cancer article

As observed, the same pattern of result is received when testing using the breast cancer article. The accuracy is higher when the number of topics being extracted using the LDA model is smaller.

This is again due to the small number of samples being used to extract the topic models. The number of sample used to test summarizing the breast cancer is smaller in number compared to the one used to summarize the psychology article. This factor influenced the performance of the text summarization system.



## **CHAPTER 6**

### **CONCLUSION & RECOMMENDATIONS**

#### **6.1 Conclusion**

The text summarization system using the Latent Dirichlet Allocation (LDA) is more accurate when using a larger number of samples as its bag of words. A larger sample would enable the system to extract more topic models more accurately. Currently the text summarization system extracts more accurately when the number of topics being extracted is small. However, a different conclusion might be found using the same LDA method if the number of sample is large enough.

#### **6.2 Recommendations**

- The system can be further refined by accepting more type of files as inputs to be summarized. Current system only accepts text files and has a feature to convert only Microsoft Word 97-03 documents.
- The system may add more samples to increase its accuracy when using Gibbs sampling.
- Current system only summarizes for Breast Cancer base texts. It can be further improved by increasing the samples and also increasing the vocabulary.

## REFERENCES

- (n.d.). Retrieved June 28, 2011, from TheFreeDictionary:  
<http://www.thefreedictionary.com/latent>
- (n.d.). Retrieved June 30, 2011, from TheFreeDictionary:  
<http://www.thefreedictionary.com/allocate>
- (n.d.). Retrieved June 18, 2011, from Carcinogenesis:  
<http://carcin.oxfordjournals.org/content/current>
- (n.d.). Retrieved June 18, 2011, from Internation Journal of Oncology:  
<http://www.spandidos-publications.com/ijo/index.jsp>
- (n.d.). Retrieved June 19, 2011, from Free Medical Journals:  
[http://www.freemedicaljournals.com/fmj/IP\\_ONCOL.HTM](http://www.freemedicaljournals.com/fmj/IP_ONCOL.HTM)
- (n.d.). Retrieved June 19, 2011, from Dovepress: <http://www.dovepress.com/cancer-management-and-research-journal>
- (n.d.). Retrieved June 19, 2011, from Asian Pacific Organization for Cancer Prevention:  
[http://www.apocp.org/journal\\_of\\_cancer\\_prevention\\_volume\\_10.php](http://www.apocp.org/journal_of_cancer_prevention_volume_10.php)
- breastcancer.org. (2011, July 25). Retrieved July 30, 2011, from  
[http://www.breastcancer.org/about\\_us/](http://www.breastcancer.org/about_us/)
- Chang, H.-A., & Yu, C.-H. (n.d.). Retrieved June 25, 2011, from  
<http://admis.fudan.edu.cn/seminars/ppt/lecture-lda.pdf>
- Das, D., & Martins, A. F. (2007, November 21). A Survey on Automatic Text Summarization. Retrieved July 23, 2011
- Dipanjana Das, A. F. (2007, November 21). *School of Computer Science, Carnegie Mellon*. Retrieved February 26, 2011, from  
[http://www.cs.cmu.edu/~afm/Home\\_files/Das\\_Martins\\_survey\\_summarization.pdf](http://www.cs.cmu.edu/~afm/Home_files/Das_Martins_survey_summarization.pdf)

- Elace. (2011, March 7). Retrieved August 4, 2011, from <http://www.elance.com/s/mthoyer/resume/>
- Ganapathiraju, M. K. (2002, November 26). Relevance of Cluster Size in MMR Based Summarizer: A Report.
- Hovy, E. (n.d.). Text Summarization.
- Inderjeet Mani, M. T. (1999). *Advances in Automatic Text Summarization*. London: The MIT Press.
- Krestel, R., Frankhauser, P., & Nejdil, W. (2009, October 23). Retrieved July 4, 2011, from <http://www.l3s.de/web/upload/documents/1/recSys09.pdf>
- Lan Du, W. B. (2010, July 23). A Segmented Topic Model Based on the Two-parameter Poisson-Dirichlet Process. Canberra, Australia.
- Lloret, E. (n.d.). Retrieved February 26, 2011, from University d'Alacant: <http://www.dlsi.ua.es/~elloret/publications/TextSummarization.pdf>
- Nassar-McMillan, S. C., & Hakim-Larson, J. (2003). Counseling Considerations Among Arab Americans. *Journal of Counseling & Development*, 150 - 159.
- Rouchka, E. C. (1997, May 20). A Brief Overview of Gibbs Sampling. 1.
- Stevyvers, M., & Griffiths, T. (n.d.). Retrieved June 18, 2011, from <http://psiexp.ss.uci.edu/research/papers/StevyversGriffithsLSABookFormatted.pdf>
- Stevyvers, M., & Griffiths, T. (2011, April 4). *Matlab Topic Modeling Toolbox 1.4*. Retrieved April 20, 2011, from [http://psiexp.ss.uci.edu/research/programs\\_data/toolbox.htm](http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm)