DEVELOPMENT OF TOOL FOR MANAGING SEMANTIC TEXT CONTENT

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
ÇANKAYA UNIVERSITY

BY

SAMET KARAKAYNAK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

JANUARY 2009

Title of the Thesis: **Development of Tool for Managing Semantic Text Content**
Submitted by **Samet Karakaynak**

Approval of the Graduate School of Natural and Applied Sciences, Çankaya University

Prof. Dr. Yahya Kemal Baykal
Acting Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Prof. Dr. Mehmet R. Tolun
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Abdül Kadir Görür
Supervisor

Examination Date : 19.01.2009

Examining Committee Members

Asst. Prof. Dr. Abdül Kadir GÖRÜR (Çankaya Univ.)

Asst. Prof. Dr. Reza HASSANPOUR (Çankaya Univ.)

Asst. Prof. Dr. Tansel ÖZYER (TOBB ETU)

STATEMENT OF NON-PLAGIARISM

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name	: Samet KARAKAYNAK
Signature	: S. Karakaynak
Date	: 19.01.2009

ABSTRACT

CENTROID-BASED MULTI-DOCUMENT SUMMARIZATION
USING LATENT SEMANTIC ANALYSIS

Karakaynak, Samet
M.S.c., Department of Computer Engineering
Supervisor: Asst. Prof. Dr. Abdül Kadir Görür

January 2009, 54 Pages

The aim of this study is creating multi-document summaries using latent semantic analysis and centroid based approach. First, key-terms are extracted using latent semantic analysis (LSA). Key-terms are used to filter the redundant sentences before sentence extraction. Then summary sentences are extracted from the sentences containing the key-terms using latent semantic indexing (LSI) and centroid-based method with clustering consecutively.

Keywords: Multi-document summarization, Latent semantic analysis, Latent Semantic Indexing, Centroid Based Summarization

ÖZ

SAKLI ANLAMSAL ANALİZ KULLANARAK
ÇOKLU-DOKÜMANLARIN SANAL MERKEZE DAYALI ÖZETLENMESİ

Karakaynak, Samet
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Danışman: Asst. Prof. Dr. Abdül Kadir Görür

Ocak 2009, 54 Sayfa

Bu çalışma çoklu dokümanlardan saklı anlamsal analiz yöntemi kullanılarak sanal merkeze dayalı özet çıkarılması amacıyla gerçekleştirilmiştir. İlk olarak saklı anlamsal analiz yöntemi kullanılarak anahtar terimler çıkarılır. Anahtar terimler cümle çıkarmaya başlamadan önce anlama katkısı olmayan cümlelerin filtrelenmesi için kullanılır. Daha sonra özet cümleler, anahtar terimleri barındıran cümlelerden sırasıyla saklı anlam indeksleme ve kümeleme ile sanal merkeze dayalı yöntem kullanılarak çekilir.

Anahtar Kelimeler: Çoklu dokümanların özetlenmesi, Saklı anlamsal analiz, Saklı anlam indeksleme, Sanal merkeze dayalı özetleme

ACKNOWLEDGEMENTS

I would like to thank, first, my supervisor Assistant Professor Dr. Abdül Kadir Görür for his guidance and support throughout the completion of thesis.

I would also like to thank Gönenç Ercan for his informative comments and technical support throughout the thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

With the rapid growth of World Wide Web the tremendous amount of text documents is even increasing more and more. Hence conventional Information Retrieval methods become inadequate to retrieve the suitable information. The results returned by the conventional Information Retrieval systems have a great deal of redundant information. Summarization can be very beneficial when used as a complementary approach in Information Retrieval systems to overcome this redundancy problem. Additionally it is advantageous to give a summary of large amount and volume of text sources to the user instead of showing only the links. Hence great deals of works have been performed on this subject in recent years and the amount of studies is increasing daily.

A summary is a condensed representation of the content of its source [1]. From the definition of summary we can say that summarization is reduction of source text(s) to a shorter version, protecting its/their semantic content.

The goal of summarization is stated in [1].

> *The goal of automatic summarization is to take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's application's needs.*

Automated summarization tools called summarizers are used to reach an acceptable summary in a short time. A short definition of summarizer is given by Inderjeet Mani

[1]:"In brief, a summarizer is a system whose goal is to produce a condensed representation of the content of its input for human consumption"

Different summarization approaches are present: Generic vs. Query-Based, Extraction vs. Abstraction, Single-Document vs. Multi-Document.

Text summaries can be either query-based summaries or generic summaries. **Query-based summaries** give a result of content which is close to a search query. They reflect user's interest. This type of summaries is used to know whether the document is suitable for the user's interest, if suitable which part(s) of the document(s) is/are suitable. **Generic summaries** give the general idea of the documents' contents. These summaries reflect the author's point of view. The success of a generic summary can be understood from its coverage of the main topics of the original document(s) and keeping length of the summary and redundancy to a minimum.

A summary entirely consisting of fragments of the original source is **extract**. Extracts should be the most important parts of the original texts. A summary generated by paraphrasing/generating text from the original text source is **abstract**. Unlike extracts because of the nature of their production way there is no strict limit of reduction for abstracts while keeping the content of source texts. This means a shorter abstract may give more information from its source than a longer extract generated from the same source.

A summarization system taking a single document as input is **single-document summarization** system. A summarization system producing single summaries taking a set of documents as input is **multi-document summarization** system. Besides the challenges of single-document summarization, multi-document summarization has additional problems because of its nature. While summarizing a set of documents redundancy becomes a much bigger problem than redundancy in single-document summarization. Inconsistency may occur among different documents about the same topic or event. The time sequence of the events or the order of steps of a proceeding event/job may be confused. These additional problems make multi-document summarization more challenging.

CHAPTER 2

RELATED WORK

Different approaches have been used in the researches of text summarization since the 1950's. A major part of recent summarization systems use identification and extraction of salient sentences from document(s). Main methods of important sentence/clause identification are based on position in the text, cues, title/heading, term frequencies and cohesions among words/expressions.

2.1 Position-Based Method

Brandow, Mitze and Rau [2] found that important sentences occur at the **beginning** of the texts. But later according to a large scaled research of Lin and Hovy [3] on optimum position policy focus position changes with different text genres.

2.2 Cue-Based Method

Teufel [4] first used **cue phrases** on science articles. Cue phrases are grouped into two types: bonus phrases and stigma phrases. Phrases focusing the attention to the important sentences where they appear are **bonus phrases**. “Significantly”, “in conclusion”, “as a result” are some examples of bonus phrases. Phrases implying that their sentence is not important such as “hardly” and “impossible” are **stigma phrases**. Cue-phrase based method yielded the best result in scientific articles.

2.3 Title-Based Method

Edmundson [5] showed that the words in titles and headings occur mostly in semantically important sentences too. This heuristic is used as a complementary approach for other methods to increase the system performance.

2.4 Word Frequency Based Method

Luhn utilized word-frequency-based rules in the late 1950's to identify sentences for summaries [6]. According to Luhn important sentences contain frequently appearing words. But Edmundson [5] claimed that using word frequency is harmful for his system performance.

2.5 Cohesion Based Methods

Cohesion based methods look at the relations among words or expressions. According to the cohesion based methods important sentences/paragraphs are the entities having the tightest connections in cohesion models. Several approaches have been used to identify the connections among the words/expressions. The most famous approaches are based on term co-occurrence, coreference and lexical chains.

2.5.1 Term Co-occurrence Method

Salton, Mitra and Buckley [7] accepted documents as collections of paragraphs and generated intra-document links between paragraphs of a document. Based on the intra-document linkage pattern of a text, they characterized the structure of the text. They applied the knowledge of text structure to do automatic text summarization by paragraph extraction.

2.5.2 Coreference Method

According to Salience-Based Approach [8], the aim is to detect topic stamps which are important phrasal expressions representing the document's content. Local salience of candidate phrasal expressions, extracted from text using morphological analysis is defined by the sum of following parameters:

CNTX: 50 iff the expression is in the current discourse segment
SUBJ: 80 iff the expression is a subject
EXST: 70 iff the expression is an existential construction
ACC: 50 iff the expression is a direct object
HEAD: 80 iff the expression is not contained in another phrase
ARG: 50 iff the expression is not contained in an adjunct

By using the coreference links among candidate phrasal expressions coreference classes are identified. Salience of the coreference classes are defined by adding the salience factor values of the phrasal expressions in that class.

2.5.3 Lexical Chains – Based Method

A lexical chain is a list of related words, independent of the grammatical structure, in the text documents. Each word in a lexical chain has a distance relation to each other. Barzilay and Elhadad [9] created all possible lexical chains from text documents and created summaries focusing on strong chains.

CHAPTER 3

BACKGROUND WORK

3.1 Singular Value Decomposition

The singular value decomposition is used generally to solve unconstrained linear least squares problems, matrix rank estimation and canonical correlation analysis [10].

Having matrix A with dimensions m x n,
Where m ≥ n and rank (A) = r,
The Singular Value Decomposition of A "SVD (A)" is defined as:

$$A = U\Sigma V^T \tag{3.1}$$

Where $U^T U = V^T V = I_n$ and
$\Sigma = \text{diag}(\sigma_{1,\ldots,}\ \sigma_n)$,
$\sigma_i > 0$ for $1 \le i \le r$,
$\sigma_j = 0$ for $j \ge r + 1$

The first r columns of the orthogonal matrices U and V define the orthonormal eigenvectors associated with the r nonzero eigenvalues of $A A^T$ and $A^T A$.

- The columns of U are referred to as the left singular vectors,
- the columns of V are referred to as the right singular vectors,

- the singular values of A are the diagonal elements of Σ which are the nonnegative square roots of the n eigenvalues of AA^T [11].

We can show how SVD holds information of matrix structure with two theorems below:

Theorem 1.1.

Let,

SVD(A) is given in Equation (3.1),

$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{r+1} = \ldots = \sigma_n = 0$,

R(A) is range of A,

N(A) is null space of A

Then,

1. rank(A) = r

   N(A) ≡ span $\{v_{r+1}, \ldots, v_n\}$

   R(A) ≡ span $\{u_1, \ldots, u_r\}$

   where,

   U = $[u_1\ u_2 \ldots u_m]$

   V = $[v_1\ v_2 \ldots v_n]$

2. dyadic decomposition: $A = \sum_{i=1}^{r} u_i \cdot \sigma_i \cdot v_i^T$

3. norms: $\|A\|_F^2 = \sigma_1^2 + \cdots + \sigma_r^2$ and $\|A\|_2^2 = \sigma_1$

Theorem 1.2.

Let SVD(A) is given in Equation (3.1)

With r = rank(A) ≤ p = min(m,n) and define

$$A_k = \sum_{i=1}^{k} u_i \cdot \sigma_i \cdot v_i^T \tag{3.2}$$

Then

$$\min_{rank(B)=k} \|A - B\|_F^2 = \|A - A_k\|_F^2 = \sigma_{k+1}^2 + \cdots + \sigma_p^2$$

Constructed from the k largest singular triplets of A, A_k is the closest rank-k matrix to A [11]:

$$\min_{rank(B)=k} \|A - B\|_2 = \|A - A_k\|_2 = \sigma_{k+1} \qquad (3.3)$$

3.2 Latent Semantic Indexing

As stated in [11] a matrix of terms by documents is created. Cell values of this matrix are occurrences of each term in each document. Since each word does not appear in each document the matrix is usually sparse. We can denote this matrix as:

$$A = \lfloor a_{ij} \rfloor$$

where a_{ij} is the occurrence count of term i in document j.

To increase the importance of terms for each document local and global weightings are applied to the matrix.

$$a_{ij} = L(i, j) \times G(i)$$

where $L(i,j)$ is the local weighting of term i in document j, and $G(i)$ is the global weighting of term i.

The latent semantic structure model is derived by singular value decomposition (SVD) from the orthogonal matrix U containing left singular vectors, matrix V containing right singular vectors and the diagonal matrix Σ containing the singular values of A.

Table 1: Interpretation of SVD Components within LSI.

A_k = Best rank-k approximation to A	m = Number of terms
U = Term Vectors	n = Number of documents
Σ = Singular Values	k = Number of factors
V = Document Vectors	r = Rank of A

Using k–largest singular triplets means **approximation** of the original term-document matrix by A_k in Equation (3.2).

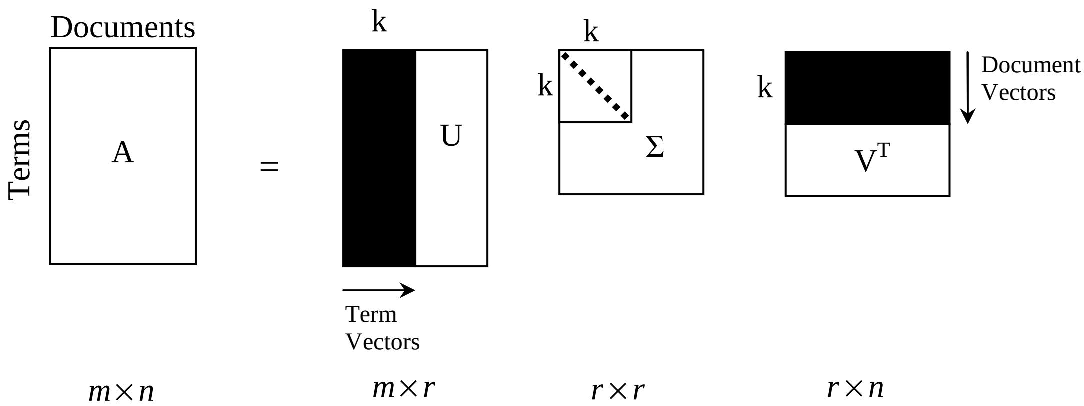

Figure 1: Mathematical Representation of the Matrix A_k.

As seen from Figure 1; U is the term vector, V is the document vector, and Σ represents the singular values. The shaded regions in U, V, and the diagonal line in Σ represent A_k from Equation (3.2).

The derived A_k matrix is not the reconstruction of the original term-document matrix A exactly. The truncated SVD captures most of the important underlying structure from the association of terms and documents and removes the noise from the word usage in documents. Because *k* is much smaller than the number of unique terms *m*, minor differences in terminology will be ignored. This means that terms not occurring in the same document but occurring in similar documents will be near to each other. Based on this point when we look at the document dimension; documents not sharing any words with a query may be near to that query in k-space.

3.3 Latent Semantic Analysis

The idea of using LSA in text summarization is published by Yihong Gong and Xin Liu in 2002 [12]. Inspired by the latent semantic indexing they applied the singular value decomposition (SVD) to generic text summarization.

The process starts with the creation of a terms-by-sentence matrix $A = [A_1 A_2 \cdots A_n]$. Each column vector A_i in this matrix represents the weighted term-frequency vector of sentence *i* in the document under consideration. If there are a total of m terms and n sentences in the document(s), then we will have an $m \times n$ matrix A for the document(s).

Applying SVD on matrix A, from the Equation (3.1) ($A = U\Sigma V^T$) we get:

- $U = [u_{ij}]$ is an $m \times n$ column-orthonormal matrix whose columns are called left-singular vectors
- $\Sigma = diag(\sigma_1, \sigma_2, \cdots, \sigma_n)$ is an $n \times n$ diagonal matrix, whose diagonal elements are non-negative singular values sorted in descending order.
- $V = [v_{ij}]$ is an $n \times n$ orthonormal matrix, whose columns are called right singular vectors.

If rank(A) = r [11], then Σ satisfies:

$$\sigma_1 \geq \sigma_2 \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0$$

The interpretation of applying the SVD to the terms-by-sentences matrix **A** can be made from two different viewpoints:

From transformation point of view, the SVD derives a mapping between the *m*-dimensional space spanned by the weighted term-frequency vectors and the *r*-dimensional singular vector space [12].

From semantic point of view, the SVD derives the latent semantic structure from the document represented by matrix **A**. This operation reflects a breakdown of the original document into r linearly-independent base vectors or concepts. Each term and sentence from the document is jointly indexed by these base vectors. Because SVD is capable of capturing and modeling interrelationships among terms, it can semantically cluster terms and sentences.

Consider the words *construction, building, architect, floor, plan, and design*. The words *construction* and *building* are synonyms, and *architect, floor, plan, design* are related concepts. The synonyms *construction* and *building* will occur in similar patterns holding common related words such as *architect, floor, plan, design* etc. Because of these similar patterns the words *construction* and *building* will have similar representations in r-dimensional singular vector space [12]. As declared in [11], if a word pattern is salient and recurring in the document(s), this pattern will be represented by one of the singular vectors. The importance of this pattern is shown by the magnitude of the related singular value. Any sentences containing this word combination pattern will be projected along this singular vector and the sentence that best represents this pattern will have the largest index value with this vector. As each particular word combination pattern describes a certain topic/concept in the document, the facts described above naturally lead to the hypothesis that each singular vector represents a salient topic/concept of the document, and the magnitude of its corresponding singular value represents the degree of importance of the salient topic/concept [12].

3.4 Centroid-based Summarization of Multiple Documents

3.4.1 What is Centroid

As declared in [13]:

"*A centroid is a set of words that are statistically important to a cluster of documents. As such, centroids could be used both to classify relevant documents and to identify salient sentences in a cluster*."

A centroid is a pseudo-document/sentence which consists of words which have average number of occurrence scores above a pre-defined threshold in the documents [13]. Centroid is used to find the sentences which represent the entire cluster the best.

3.4.2 Centroid-Based Summarization

Radev, Jing and Budzikowska [13] have developed a multi-document summarizer called MEAD which creates summaries using cluster centroids generated by a topic detection and tracking system and described two new techniques, based on cluster-based sentence utility and cross-sentence informational subsumption.

Cluster-based sentence utility is the degree of relevance of a sentence in the cluster to the general topic of the whole cluster. A degree of 0 means sentence is not relevant to the general topic, 10 means the sentence is essential for the topic of entire cluster.

Cross-sentence informational subsumption indicates that a sentence covers another sentence from information point of view. If the information content of the sentence S_1 is a subset of sentence S_2, then S_2 subsumes S_1 and S_1 is accepted as redundant from information perspective.

$$i(S_1) \subset i(S_2)$$

Equivalence classes consist of sentences subsuming each other. Sentences need not to exactly subsume each other to belong to the same equivalence class. An equivalence class may contain more than two sentences from the same or different articles.

A cluster centroid in the context of [13] is a pseudo-document which consist of words which have Count * IDF scores above a predefined threshold. Count is the average number of occurrences of a word in the whole cluster, IDF value is the ratio of the document number to the all occurrences of a word. According to the hypothesis in [13] sentences containing the words from the centroid are more representative of the topic of a cluster.

3.5 K-Means Clustering

K-Means [14] is an algorithm for clustering N data points into k disjoint subsets. The main point is defining k centroids, each belonging to a cluster. Each point in the data points is associated to the nearest one from k centroids until no point is pending. For the cluster set created in previous operation the new centroids are re-calculated. Points are re-associated to the nearest ones for the newly created centroids. These steps are repeated until centroids do not move any more. The algorithm tries to achieve to goal of minimizing an objective function: squared error function.

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2$$

Where there are k clusters S_i , $i=1,2,\ldots,k$ and μ_i is the centroid or mean point of all the points $x_j \in S_i$.

K-Means has drawback of results depending upon its two initial parameters: Cluster number k and initial center points. Firstly, inappropriate cluster number may give poor results. Secondly, the results change according to the initially selected cluster centers.

3.6 Cosine Similarity

Cosine Similarity is the cosine of the angle between two vectors of n dimensions. Given two vectors of attributes A and B, the cosine similarity θ using dot product and magnitude as:

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} \tag{3.4}$$

The result ranges from -1 to 1. -1 Means exactly opposite, 0 means independent, 1 means exactly the same [15].

Cosine similarity is often used for comparing documents in text mining. In text matching, the attribute vectors A and B are usually the TF.IDF vectors of the documents.

3.7 TF.IDF Weighting

TF.IDF (Term Frequency - Inverse Document Frequency) is a weighting scheme frequently used in information retrieval [16]. **Term Frequency** (TF) means how many times a term occurs in a document or document group. **Inverse Document Frequency** (IDF) shows the general importance of a term. To show the general importance IDF needs a large set of documents (corpus). According to IDF the importance of a term is inversely proportional with document number the term occurs in a corpus. We can denote TF.IDF with the following two formulas:

$$W_{d,t} = tf_{d,t} \bullet idf_t \tag{3.5}$$

$$idf_t = \log\left(\frac{D}{dft}\right)$$

Where;

- ***W*** is TF.IDF
- ***tf*** is the number of occurrences of a term in the document.
- ***D*** is the total number of documents in the whole document set (corpus)
- ***dft*** is the number of documents the term occurs in the corpus

Based on the definition above, $tf \cdot idf_{t,d}$ of term ***t*** and document ***d*** is;

- higher when
  - the term ***t*** occurs many times in smaller number of documents
- lower when
  - the term ***t*** occurs occasionally in a document
  - OR the term ***t*** occurs in many documents
- lowest when
  - the term ***t*** occurs virtually in all documents

CHAPTER 4

CENTROID-BASED MULTI-DOCUMENT SUMMARIZATION USING LATENT SEMANTIC ANALYSIS

4.1 Roadmap

Our method performs summarization in two major steps. First, **key-terms** are extracted using two main approaches: **Latent Semantic Analysis (LSA)** and choosing the terms with biggest TF.IDF values. Second, summary sentences are extracted from the sentences containing the key-terms from first step using **Latent Semantic Indexing (LSI)** and **centroid-based** approach with **K-Means clustering** consecutively. By using two steps we aim to bypass non-important sentences at the beginning. Our hypothesis here is that sentences containing key-terms are more important than the others.

In the first step we fetch sentences from documents using **sentence detector**. Then terms are fetched from sentences through two operations: **stemming** and **stop-words elimination**. Term Frequencies **(TF)** are found of each term for each document set then Term Frequency - Inverse Document Frequency **(TF.IDF)** values of each term for each document set are calculated multiplying term frequencies with Inverse Document Frequencies **(IDF)** prepared previously using the whole document corpus. Lastly, key-terms are extracted using two different methods. In first method, sentence-word matrix is created and filled with TF.IDF values and then key-terms are extracted using **LSA**. In second method, terms with biggest TF.IDF values are selected as key-terms. By using two different methods we aim to match the results of two methods and examine the performance of LSA in finding key-terms.

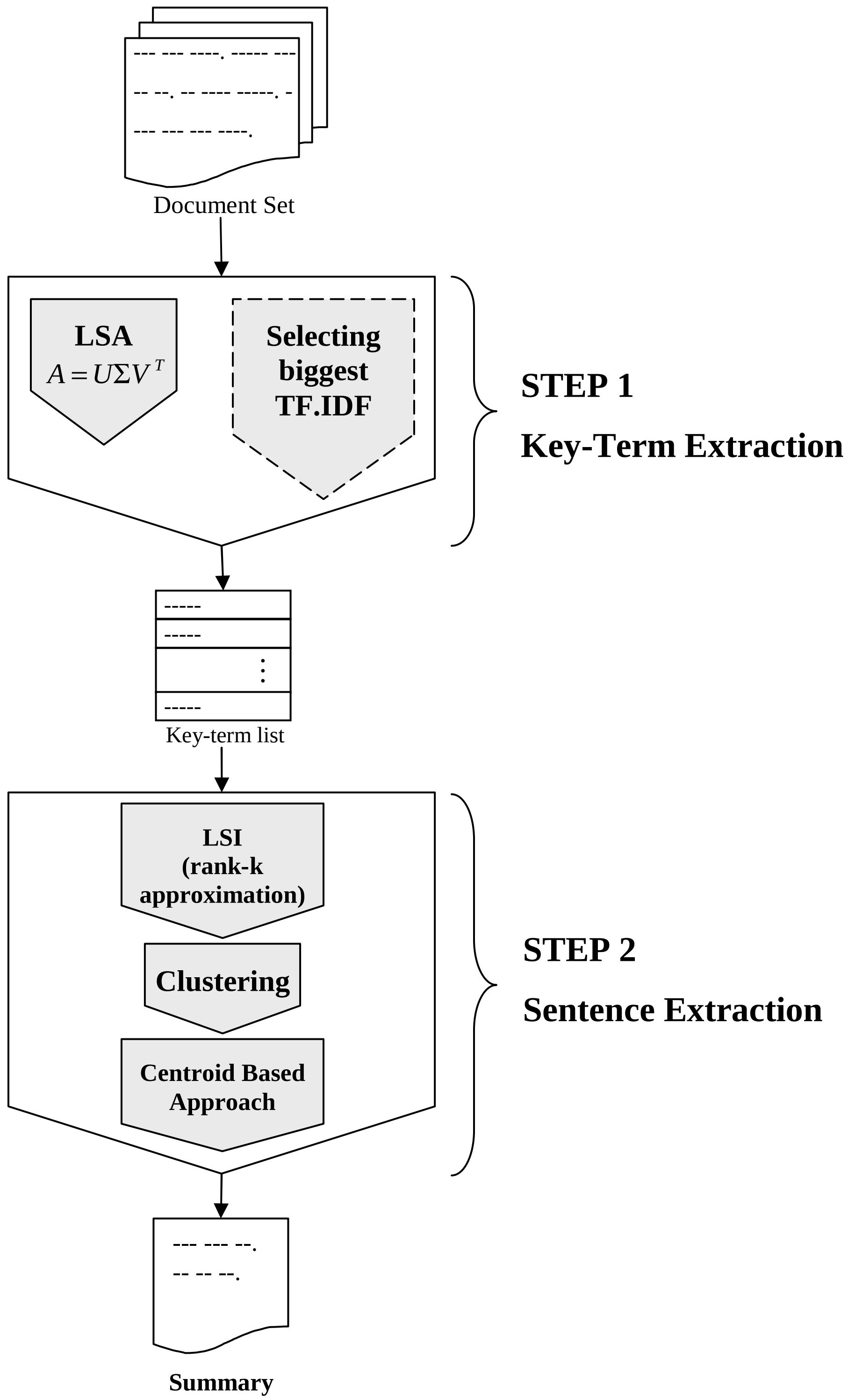

Figure 2: Roadmap

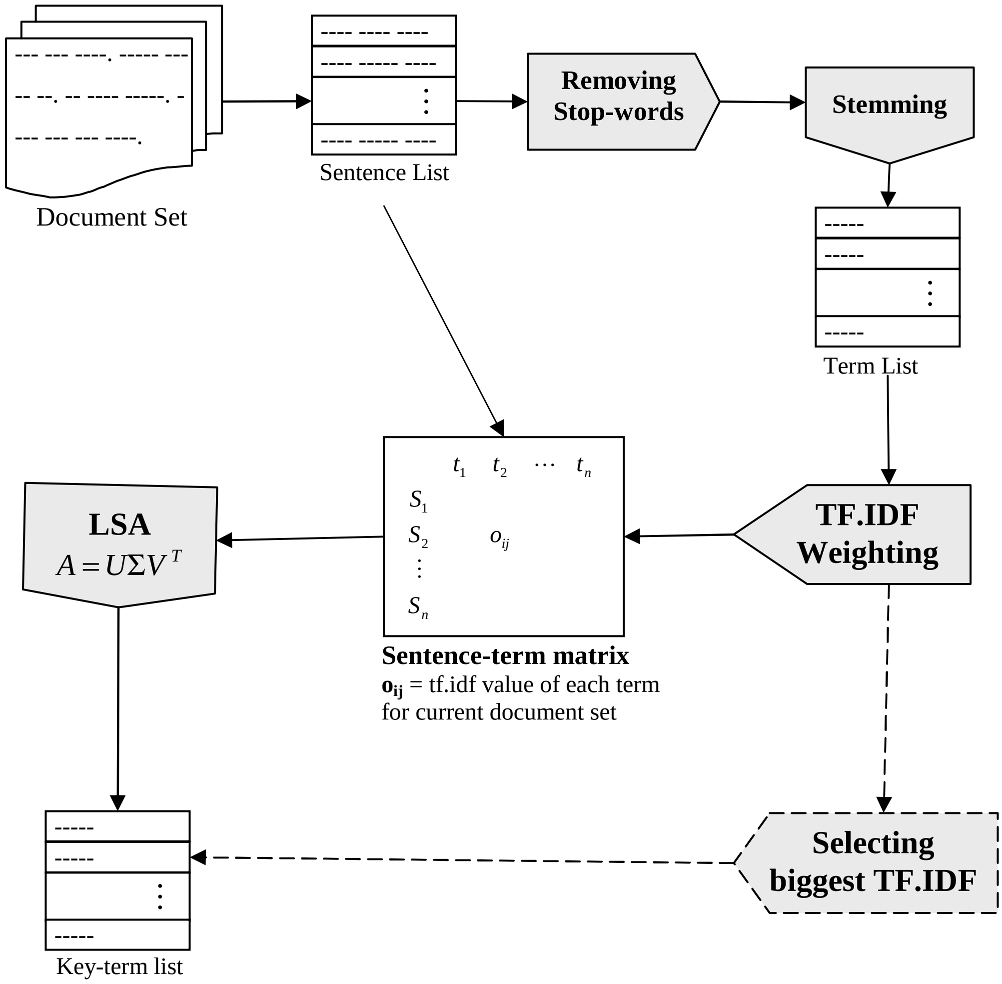

Figure 3: STEP 1: Key-Term Extraction

After extracting key-terms, sentences holding key-terms are detected and fetched from the whole sentence set. These are candidate sentences for our summary. Again after calculating the TF.IDF values of each key-term for each document set, "**key-term – candidate sentence**" matrix is created and filled with these TF.IDF values. Then dimension reduction is applied to the matrix using Latent Semantic Indexing **(LSI)** to eliminate the noise from the word usage in documents as stated in [11].

Each row, representing each candidate sentence, in the second matrix created in previous step is a vector of weighted key-terms. Based on this point of view similarity among candidate sentences is found calculating **cosine similarity** of all candidate sentences to each other and a sentence-sentence similarity matrix is created. Then, sentence clusters are extracted from the similarity matrix using **K-Means clustering** algorithm.

For each sentence cluster in the final level of our summarization method again sentence-term matrix is created and weighted with TF.IDF. Unlike previous levels average weighting of each key-term is calculated and a vector of average weightings called **centroid** is constructed in this level. For each cluster, sentences most similar to the **centroids** are detected using cosine similarity and added to the summary.

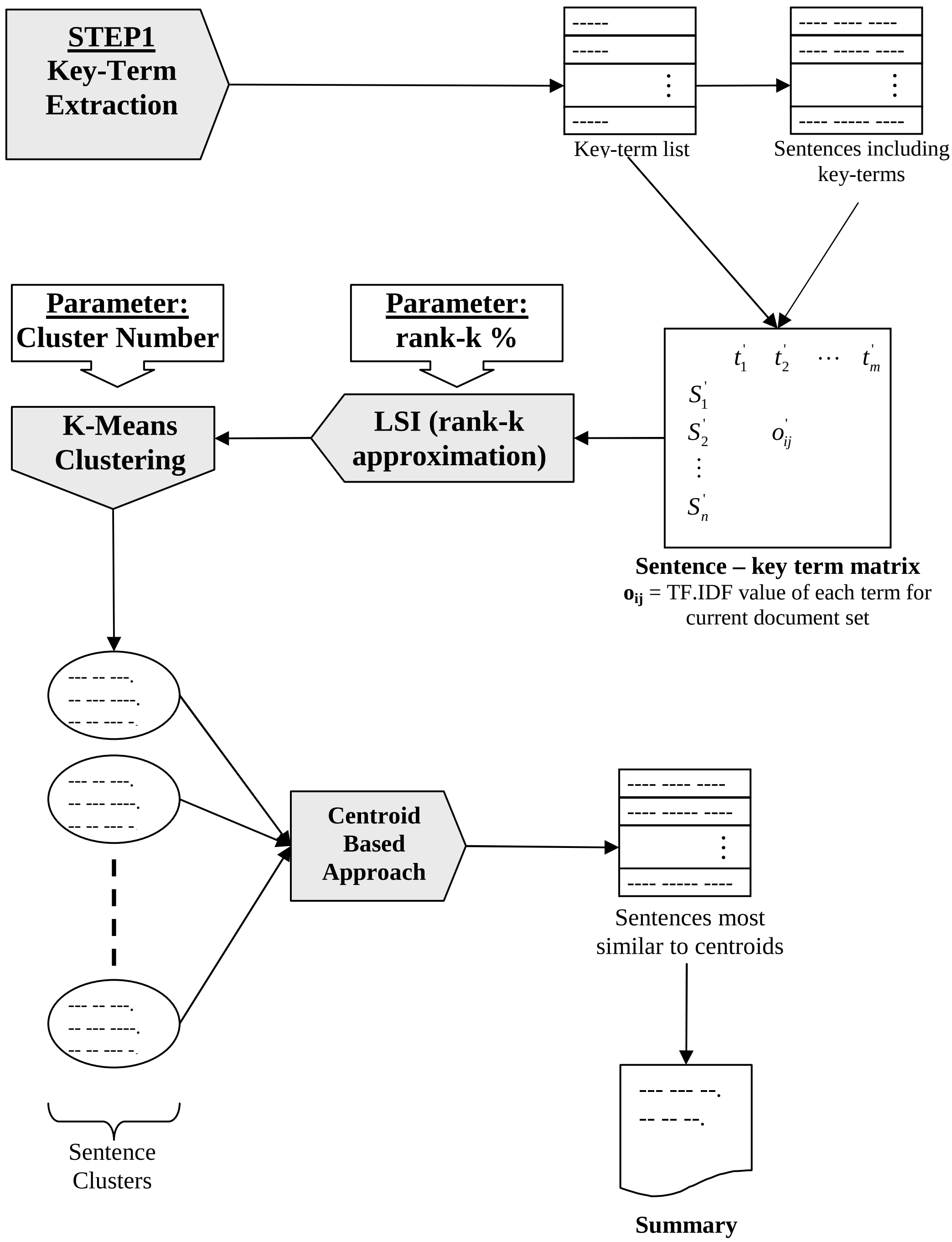

Figure 4: STEP 2: Sentence Extraction

4.2 Sentence Detector

Our sentence detector uses two heuristics to detect sentences [17]. First we use punctuations {., !, ?} to find sentence boundaries. But this native sentence boundary detection mechanism may work wrong when it encounters abbreviations. For example ‘*Dr. Smith works here.*’ can be detected as two separate sentences ‘*Dr.*’ and ‘*Smith works here*’. To overcome this problem the second heuristic of using the length of sentence to detect boundaries is used. If the number of letters in a sentence is less then a threshold value, first heuristic of punctuation is ignored and sentence boundary is detected. Our threshold value is six letters per sentence.

4.3 Stemming

Words existing in documents have many morphological variants. As morphological variants of words have similar semantic representations they can be considered as equivalent in summarization operations. Because of this situation a number of stemmers have been developed to reduce the words to their stems or root forms.

Stemming is a normalization process used to reduce words to their roots or stems. The stems do not have to be the morphological roots of the words. It is enough for a stem that semantically similar words can be reduced to the same stem, even if the stem is not a valid root. For example, the words "computes", "computation", and "computed" are considered as being from the same root and after stemming they will be considered as the same word.

The first published stemmer was written by Julie Beth Lovins in 1968 [18]. A new stemmer written by Martin Porter and published in the July 1980 [19] was very widely used and became the de-facto standard algorithm for English stemming. Martin Porter released an official free-software implementation of the algorithm around the year 2000 and implemented an improved English stemmer [20]. We have used Porter Stemmer for our stemming operation.

4.4 Removing Stop Words

Stop-words are insignificant words frequently appearing in documents. As stated in [21] the most frequent words are often the words with little meaning and stop word removal may affect substantially the document lengths which may deteriorate the effectiveness of weighting scheme.

There is no common list of stop words. Our stop word list is given in Appendix 1.

4.5 Extracting Key-Terms using Latent Semantic Analysis

Based on Latent Semantic Analysis Method described in Chapter 3.3 we focus on the patterns of sentence combinations in multi-documents. If a sentence pattern is salient and recurring in documents, this pattern will be captured and represented by one of the singular vectors. The magnitude of the corresponding singular value shows the importance degree of this pattern within the documents. Any words appearing in this sentence pattern will be projected along this singular vector, and the word that best represents this sentence pattern will have the largest index value with this vector. Because each particular sentence pattern describes a certain topic in the documents, we come up with a hypothesis that each singular vector represents a salient topic in the documents and the magnitude of its corresponding singular value represents the degree of importance of the salient topic.

Based on our discussion we propose the following SVD-based **key-term** extraction method.

1. Decompose the documents into individual sentences and set k = 1.
2. Construct the terms by sentences matrix A for the documents
3. Perform SVD on A to obtain the singular value matrix Σ, and the left singular vector matrix U. In the singular vector space, each sentence j is represented by the row vector $\varphi_j = \lfloor u_{1j} u_{2j} \cdots u_{rj} \rfloor$ of U.

4. Select the *k*'th left singular vector from matrix U.
5. Select the term which has the largest index value with the *k*'th left singular vector, and add it to the key-term list.
6. If *k* reaches the predefined number, terminate the operation; otherwise, increment *k* by one, and go to Step 4.

In Step 5 of the above operation, finding the term that has the largest index value with the *k*'th left singular vector is equivalent to finding the row vector φ_j whose *k*'th element u_{kj} is the largest. According to our hypothesis, this operation is equivalent to finding the most important term related the salient topic/concept represented by the *k*'th singular vector. Since the singular vectors are sorted in descending order of their corresponding singular values, the *k*'th singular vector represents the *k*'th important topic/concept. Because all the singular vectors are independent of each other, the words selected by this method have minimum semantic relation to each other.

4.5.1 Disadvantages

The two disadvantages declared for [12] in [22] are valid for our method too:

1. The higher is the number of dimensions of reduced space, the less significant topic we take into a summary.
2. A word with large index values but not the largest (it does not win in any dimension), will not be chosen although it is important enough to extract summary sentences.

4.6 LSI (Rank-k Approximation)

To eliminate the noise of word usage in documents the sentence – term matrix is approximated to rank-k as stated in chapter 3.2. Rank-k is found by multiplying the column number by **rank-k percentage (*k%*)** which is given as a parameter. Supposing that we have an $n \times m$ sentence-term matrix, rank-k (***k***) is found by the formula $k = rank(A) * k\%$. Our aim by using approximation percentage is to confine the parameter to 0 – 100 boundaries. Thus the approximation parameter (***k***) will be independent of the matrix rank which varies according to document set.

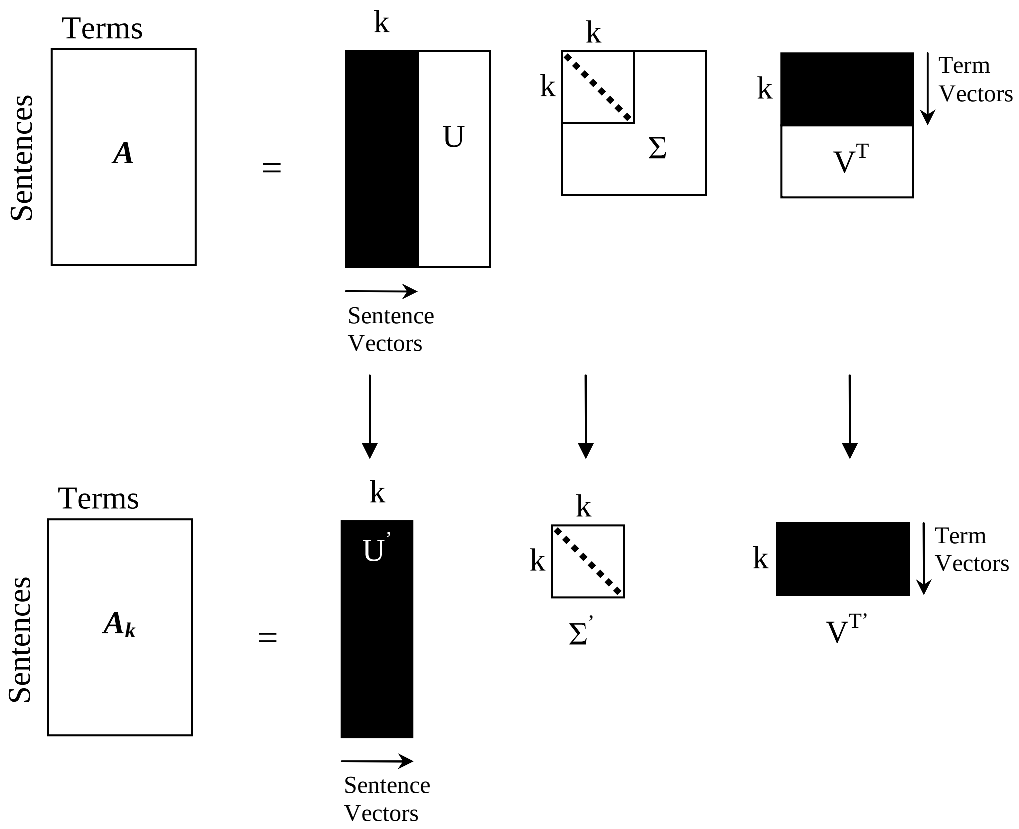

Figure 5: Rank-k Approximation

4.7 Clustering with K-Means

After rank-k approximation, sentence-term matrix is divided into clusters using K-Means algorithm. K-Means has two main problems stated in chapter 3.5. First problem is that the result is changed according to cluster number which should be predefined. Regarding to this problem ordered sets of cluster numbers are tried in an appropriate range intuitively.

Second problem is that the result changes according to selection of initial center sentence vectors. To get better results initial sentence vectors as far as possible from each other are selected. Our distance metric is **inverse cosine similarity** among the vectors. In other words sentences less similar are further and vice versa.

4.8 Sentence Extraction using Centroid-Based Approach

After sentences are partitioned into clusters, a sentence-term matrix weighted with TF.IDF is created for each cluster. Then average number of occurrences (frequency) of a term across the entire cluster is calculated by dividing the total occurrence number by total sentence number. This average occurrence number is multiplied by the IDF value of the term and average TF.IDF value of each term in each cluster is found. Then a vector of average TF.IDF values of all terms in the cluster is created. This pseudo sentence vector is called **centroid sentence vector**.

Having sentence-term occurrence matrix:

$$
\begin{array}{c|cccc}
 & t_1 & t_2 & \cdots & t_m \\
S_1 & & & & \\
S_2 & & o_{ij} & & \\
\vdots & & & & \\
S_n & & & &
\end{array}
$$

Figure 6: Sentence-Term Matrix in a Cluster

Where;

s = sentence

t = term

n = sentence number in the cluster

m = term number in the cluster

o_{ij} = TF.IDF value of j'th term in i'th sentence.

Centroid Value of each term is denoted by:

$$C_j = \frac{1}{n}\sum_{i=1}^{n} o_{ij} \tag{4.1}$$

Centroid Sentence is the vector denoted by:

$$S_{centroid} = \left[C_1 C_2 \cdots C_m\right] \tag{4.2}$$

After creating centroid vectors, cosine similarity of each sentence in the cluster is calculated and sentences are sorted according to their similarity to the centroid vector descending. In other words the sentence most similar to the centroid takes the first place; the one least similar to centroid takes the last place in the new sentence order. Additionally clusters are sorted according to their sentence number descending.

Starting from the biggest cluster the sentences most similar to the centroids are fetched from each cluster and added to the summary. This operation is repeated until the summary size reaches a predefined size limit.

4.9 Weighting

While constructing the **TF.IDF** weighting scheme we benefited from **DUC2004** documents explained in the next chapter. The IDF value of each term is calculated using 500 documents of DUC2004 as a corpus.

Unlike IDF, TF value depends on the working cluster. In the first (key-term extraction) step the clusters of DUC2004 each having 10 documents, in the second (sentence extraction) step clusters created by K-Means algorithm are used to calculate the TF values.

CHAPTER 5

EXPERIMENTS & EVALUATION

5.1 Experiments

We used **DUC2004** [23] conferences as an experiment area for our summaries. Task2 of DUC2004 conference is for multi document summarization [24]. DUC2004 experiment area includes 50 clusters each having its own topic and consisting of 10 documents. For each topic/cluster 4 model summaries written by humans exist. Addition to model summaries 35 system summaries exist in DUC2004 related with multi-document summarization branch (Task 2). There is a size restriction of not exceeding 665 characters for both model and system summaries.

Three sample documents, key-term lists and summaries created using both LSA and biggest TF.IDF are shown below.

Document Name: APW19981020.0241
Margaret Thatcher entertained former Children dictator Gen. Augusto Pinochet at her home two weeks before he was arrested in his bed in a London hospital, the ex-prime minister's office said Tuesday, amid growing diplomatic and domestic controversy over the move. Pinochet, who has vowed to fight attempts to extradite him to Spain on allegations of murder, genocide and torture, had drinks with Lady Thatcher and her husband, Denis, in their home in London's elite Belgravia district four days before he was hospitalized for back surgery performed Oct. 9. ``She regarded it as a private meeting,'' said Mark Worthington, spokesman for the Lady Thatcher, Conservative Party prime minister from 1979-90. The 82-year-old Pinochet was arrested Friday at a Spanish magistrate's request. In Conservative government days, Pinochet was welcomed on regular visits that included tea with the prime minister. He was the only Latin American leader to support Britain in its 1982 war against Argentina to reclaim the Falkland Islands. Pinochet and Lady Thatcher also implemented similar brands of right-wing economics. The current visit is Pincohet's first since Prime Minister Tony Blair's Labor Party administration was elected 18 months ago, ending 18 years of Conservative Party rule. Chile's ambassador delivered a formal protest to the Foreign Office on Monday, saying Britain has violated Pinochet's diplomatic immunity. He arrived last month on a diplomatic passport and is also a senator-for-life in Chile, which protects him from prosecution there. Pinochet's 17-year-rule was marked by torture and other human rights abuses against political opponents in which, the Chilean government has said, 4,299 people were killed or vanished. He remained Chilean army commander-in-chief until March. The magistrate broadened his charges Monday to include killings of Chileans as well as Spaniards, and genocide _ for which there is no diplomatic immunity. Chilean Ambassador Mario Artaza, himself an exile during Pinochet's rule, said Chile had a duty to protect a citizen with diplomatic immunity and senator status. ``We are not protecting the dictator of the '70s,'' Artaza said in a British Broadcasting Corp. radio interview Tuesday. ``What we are fighting for and discussing with the (British) government is the special situation of a senator in our transition who many people do not understand and many people don't like.'' ``We're not discussing his record during his period of dictatorship, that the present government does not support at all,'' added the ambassador. A Chilean specialist in international law was traveling to London for further meetings with British officials, Artaza said. Pinochet, expected to be hospitalized for perhaps two more weeks faces a long battle through British courts to avoid extradition, questioning by two Spanish judges who instigated the proceedings, and an appearance at London's Bow Street magistrate's court. British Conservative Party lawmakers accuse the Labor government of ``gesture'' politics and pandering to the party's left-wing.

Sample Document 1

Document Name: APW19981019.0098
Britain has defended its arrest of Gen. Augusto Pinochet, with one lawmaker saying that Chile's claim that the former Chilean dictator has diplomatic immunity is ridiculous. Chilean officials, meanwhile, issued strong protests and sent a delegation to London on Sunday to argue for Pinochet's release. The former strongman's son vowed to hire top attorneys to defend his 82-year-old father, who ruled Chile with an iron fist for 17 years. British police arrested Pinochet in his bed Friday at a private London hospital in response to a request from Spain, which wants to question Pinochet about allegations of murder during the decade after he seized power in 1973. Pinochet had gone to the hospital to have a back operation Oct. 9. ``The idea that such a brutal dictator as Pinochet should be claiming diplomatic immunity I think for most people in this country would be pretty gut-wrenching stuff," Trade Secretary Peter Mandelson said in a British Broadcasting Corp. television interview Sunday. Home Office Minister Alun Michael acknowledged Sunday that Pinochet entered Britain on a diplomatic passport, but said, ``That does not necessarily convey diplomatic immunity." The Foreign Office said only government officials visiting on official business and accredited diplomats have immunity. Pinochet has been a regular visitor to Britain, generally without publicity. His arrest this time appeared to reflect a tougher attitude toward right-wing dictators by Prime Minister Tony Blair's Labor Party government, which replaced a Conservative Party administration 18 months ago and promised an ``ethical" foreign policy. However, Michael Howard, a Conservative spokesman and former Cabinet minister, said he was concerned that Pinochet was arrested as a result of pressure from Labor lawmakers and lobby groups. Chilean President Eduardo Frei criticized the arrest, saying the Spanish magistrate's arrest order was tantamount to not recognizing Chile's institutions. ``Spain also lived under an authoritarian for 40 years and many of its present institutions are inherited from that regime," Frei said in Porto, Portugal, where he was attending the Ibero-American Summit. ``Would a Chilean court be allowed to start a trial for abuses that occurred under the Spanish authoritarian regime (of Francisco Franco)?" Frei asked. ``It is only for Chilean courts to try events that occurred in Chile." Franco's reign ended in 1975. Pinochet's family issued a statement Sunday calling the arrest ``an insult" and thanking the Chilean government, rightist politicians and the military for their support. In London, police guards were deployed Sunday outside the London Clinic, where Pinochet is believed to still be a patient. About 100 Chilean demonstrators pleased with the arrest gathered outside, chanting and waving placards bearing faded black and white portraits with the caption ``Disappeared in Chile." Across the Atlantic, the Chilean capital of Santiago was the scene of dueling demonstrations Sunday, reflecting the long-standing division of public opinion over Pinochet.

Sample Document 2 – Part 1

Document Name: APW19981019.0098 (cont.)
The rallies were mostly peaceful, although riot police used tear gas and water cannons on some pro-Pinochet protesters trying to break through police lines into the British embassy on Sunday evening. No arrests or injuries were reported. The envoy sent to London to argue for Pinochet's release, Santiago Benadava, would offer only diplomatic advice, said Chilean Foreign Minister Jose Miguel Insulza. Any legal defense would be up to Pinochet's family. Pinochet's son, Augusto, said the family would hire ``the best legal team available in London.'' Several right-wing Chilean politicians, including some who held posts in the Pinochet regime, also were flying to London to show their support to their former boss. Under extradition laws, Spain has 40 days from last Friday to formally apply for extradition. The final decision lies with British Home Secretary Jack Straw. There was no immediate word on when Pinochet would be questioned. But police sources, speaking on condition of anonymity, said questioning was not expected for a week or two. Pinochet has been widely accused of running a ruthless regime marked by disappearances and deaths of political opponents. His arrest was prompted by applications last week to question him by two Spanish judges investigating human rights violations. One of them, Baltasar Garzon, also wants to question Pinochet about the disappearances of Chilean dissidents in Argentina. The arrest warrant, however, referred only to questioning about allegations that he killed Spaniards in Chile between 1973 and 1983. In Chile, seven Spaniards have been identified as missing or dead under the Pinochet regime, including two Catholic priests and a U.N. official. According to a Chilean government report, a total of 4,299 political opponents died or disappeared during Pinochet's term. Pinochet, commander-in-chief of the Chilean army until March, has immunity from prosecution in Chile as a senator-for-life under a new constitution that his government crafted. He is also covered under an amnesty for crimes committed before 1978 _ when most of the human rights abuses took place.

Sample Document 2 - Part 2

Document Name: APW19981018.0423
Cuban President Fidel Castro said Sunday he disagreed with the arrest in London of former Chilean dictator Augusto Pinochet, calling it a case of ``international meddling.'' ``It seems to me that what has happened there (in London) is universal meddling,'' Castro told reporters covering the Ibero-American summit being held here Sunday. Castro had just finished breakfast with King Juan Carlos of Spain in a city hotel. He said the case seemed to be ``unprecedented and unusual.'' Pinochet, 82, was placed under arrest in London Friday by British police acting on a warrant issued by a Spanish judge. The judge is probing Pinochet's role in the death of Spaniards in Chile under his rule in the 1970s and 80s. The Chilean government has protested Pinochet's arrest, insisting that as a senator he was traveling on a diplomatic passport and had immunity from arrest. Castro, Latin America's only remaining authoritarian leader, said he lacked details on the case against Pinochet, but said he thought it placed the government of Chile and President Eduardo Frei in an uncomfortable position while Frei is attending the summit. Castro compared the action with the establishment in Rome in August of an International Criminal Court, a move Cuba has expressed reservations about. Castro said the court ought to be independent of the U.N. Security Council, because ``we already know who commands there,'' an apparent reference to the United States. The United States was one of only seven countries that voted against creating the court. ``The (Pinochet) case is serious ... the problem is delicate'' and the reactions of the Chilean Parliament and armed forces bear watching, Castro said. He expressed surprise that the British had arrested Pinochet, especially since he had provided support to England during its 1982 war with Argentina over the Falkland Islands. Although Chile maintained neutrality during the war, it was accused of providing military intelligence to the British. Castro joked that he would have thought police could have waited another 24 hours to avoid having the arrest of Pinochet overshadow the summit being held here. ``Now they are talking about the arrest of Pinochet instead of the summit,'' he said. Pinochet left government in 1990, but remained as army chief until March when he became a senator-for-life.

Sample Document 3

pinochet, chilean, pinochet', spanish, chile, london, extradit, british, augusto, immun, aznar, garzon, clinic, genocid, senat, warrant, castro, mundo, frei, regim, argentina, detent, diplomat, judici, geneva, espina, judg, thatcher, magistr, spaniard, chile', argu, madrid, abus, polic, trial, jose, artaza, mandelson, summit, wing, passport, legal, urinari, spain', exil, oct, oppon, detain, gen, lago, bertossa, protest, investig, santiago, terror, armi, blair, britain', privat, authoritarian, latin, magistrate', jack, stamp, deadlin, visit, seek, releas, 1990, sundai, polit, herniat, prosecut, pari, husband, hire, joaquin, compassion, prize, ladi, alun, ıbero, prime, meanwhil, issu, deleg, protect, newspap, formal, murder, room, 299, similar, underpin, regular, shout, case, delicate", 'ı, franco', hospit, pacemak, lawyer, cuba, command, life, eduardo, attempt, predica, public, european, appeal, defend, occas, 188, lawmak, reclaim, seem, avoid, enjoi, entitl, detail, europ, talk, interview, nonsens, appli, recov, question, ail, stir, resum, where, seven, radio, held, injuri, seriou, place, 1997, mr

Sample Key-Terms Extracted Using LSA

pinochet, chilean, spanish, pinochet', extradit, chile, spain, london, augusto, arrest, immun, garzon, aznar, 1973, british, dictat, clinic, genocid, warrant, magistr, britain, frei, mundo, disappear, castro, diplomat, senat, surgeri, baltasar, espina, spaniard, thatcher, chile', argentina, 82, regim, tortur, madrid, dictatorship, court, judici, el, geneva, detent, artaza, santer, garzon', falkland, swiss, judg, terror, general', switzerland, meddl, magistrate', conserv, spain', ladi, jose, request, maria, jack, instig, argu, prosecutor, pari, detain, hiriart, bertossa, mandelson, movoa, underpin, jaccard, wrench, blair', lucia, urinari, lago, alun, 299, pesl, santiago, legal, summit, straw, authoritarian, passport, 1982, abus, blair, gen, exil, oppon, trial, wing, ıbero, britain', diabet, widow, eduardo, polic, citizen, oct, ambassador, investig, toni, file, scotland, broaden, gut, stamp, husband, latin, rule, 1990, prime, parti, appeal, 17, labor, lawyer, london', rubber, compassion, human, crime, hospit, kidnap, infect, ail, rage, 1977, minist, right, protest, deleg, lawmak, armi, 90, kill, prize, deadlin, yard, releas, prosecut, recov, hire, bed, account, rightist, worthington, ınsulza, phalanx, token, luxemburg, porto, fernandez, nichol, dictator', oviedo, gesture", pander, herniat, galleri, 25th, benadava, accredit, jovino, margaret, julio, bingham, alberto, entangl, 83rd, spinal, vein, offshoot, coincident, lord, alpin, achil, tv13, franco, ef, perez, placard, joaquin, "thi, grounds", ethical", veronica, character", pacemak, pincohet', strongman', argentin, 'ı, impart, delicate", insult", belgravia, lakesid

Sample Key-Terms Extracted Using Biggest TF.IDF Method

Pinochet's 17-year-rule was marked by torture and other human rights abuses against political opponents in which, the Chilean government has said, 4,299 people were killed or vanished. Chilean officials, meanwhile, issued strong protests and sent a delegation to London on Sunday to argue for Pinochet's release. Cuban President Fidel Castro said Sunday he disagreed with the arrest in London of former Chilean dictator Augusto Pinochet, calling it a case of ``international meddling. The envoy sent to London to argue for Pinochet's release, Santiago Benadava, would offer only diplomatic advice, said Chilean Foreign Minister Jose Miguel Insulza.

Sample Summary Using Key-Terms from LSA

His lawyer, Clive Nicholls, said that if a bid to extradite the general succeeded, by the same token Queen Elizabeth II could be extradited to Argentina to face trial for the death of Argentine soldiers in the Falklands war in 1982. Cuban President Fidel Castro said Sunday he disagreed with the arrest in London of former Chilean dictator Augusto Pinochet, calling it a case of ``international meddling. " Pinochet, 82, was placed under arrest in London Friday by British police acting on a warrant issued by a Spanish judge. Castro had just finished breakfast with King Juan Carlos of Spain in a city hotel. Britain has defended its arrest of Gen.

Sample Summary Using Key-Terms from Biggest TF.IDF

5.2 Evaluation

We have evaluated our summaries with **ROUGE** (*Recall-Oriented Understudy for Gisting Evaluation*) [25, 26]. Rouge results are obtained according to N-Gram (Rouge 1/2/3/4), Longest Common Subsequence (Rouge L), Weighted Longest Common Subsequence (Rouge W 1.2) with F Measure (equal importance of recall and precision) and matched with other 35 systems for each scoring approaches.

The result of our system varies according to some parameters. First, **term percentage** is used to identify how many of the key-terms extracted in first step will be used in the second step. 10 levels of term percentages are used from 10% to 100%. Second, **rank-k approximation percentage** is used in matrix approximation operation during sentence extraction to find **rank-k** value for each document cluster of DUC2004. 10 levels of rank-k percentages are used as input starting from 10% to 100%. Third, **cluster number** is used in clustering by K-Means which needs cluster number as parameter from outside. From 1 to 8, eight cluster numbers are used as parameter for the clustering operation.

The number of configurations for all combinations of the parameters above is $10 \times 10 \times 8 = 800$. Summaries for these 800 combinations of parameters have been created and evaluated using ROUGE. The best eight configurations with their scores and order among other summarization systems are shown in Table 2. Same experiment is done with biggest TF.IDF to see the success of LSA approach in key-term extraction.

The best results for term percentage were achieved at 10% and the best results for rank-k percentage were obtained at 70%. High scores are observed at term percentage of 10% and rank-k percentage of %70 pair. Detailed score table for this parameter pair is given in Appendix 2. The best results for cluster number are obtained at 1, 2 and 3 clusters. The scores dropped with exceeding three clusters. The best result was obtained at term percentage of 10%, rank-k percentage of 70% and 3 clusters.

Table 2: System Configurations with Best ROUGE Results.

Configuration Parameters			ROUGE Scores & Orders					
Term %	Rank-k %	Cluster No	R1_AF	R2_AF	R3_AF	R4_AF	RL_AF	RW_12_AF
10	70	1	**27** 0.33053	**21** 0.06894	**16** 0.0232	**15** 0.01007	**20** 0.29541	**15** 0.13348
10	70	2	**23** 0.33504	**22** 0.06855	**20** 0.02164	**19** 0.00887	**16** 0.29938	**15** 0.13463
10	70	3	**19** 0.34066	**22** 0.0681	**21** 0.0212	**20** 0.00858	**<u>14</u>** 0.30349	**<u>13</u>** 0.13565
10	70	4	**21** 0.3392	**23** 0.06683	**20** 0.02224	**18** 0.00956	**<u>14</u>** 0.30305	**15** 0.13479
20	90	3	**19** 0.34015	**23** 0.06522	**22** 0.02017	**24** 0.00756	**<u>14</u>** 0.30257	**15** 0.13386
80	60	2	**23** 0.33517	**25** 0.06365	**24** 0.01992	**22** 0.00837	**16** 0.29921	**15** 0.13362
10	80	4	**21** 0.33906	**23** 0.0658	**20** 0.02198	**15** 0.00996	**<u>14</u>** 0.30316	**15** 0.13426
10	100	1	**27** 0.33053	**21** 0.06894	**16** 0.0232	**15** 0.01007	**20** 0.29541	**15** 0.13348

Term %:	term percentage to be used in STEP 2
Rank-k %:	rank-k approximation percentage
Cluster No:	cluster number
R1_AF:	ROUGE 1, F Measure
R2_AF:	ROUGE 2, F Measure
R3_AF:	ROUGE 3, F Measure
R4_AF:	ROUGE 4, F Measure
RL_AF:	ROUGE L, F Measure
RW_12_AF:	ROUGE W 1.2, F Measure

Figure 7: Meanings of Titles in Result Tables

Table 3: Best ROUGE Results for Biggest TF.IDF Method in Key-Term Extraction

Term %	Rank-k %	Cluster No	R1_AF	R2_AF	R3_AF	R4_AF	RL_AF	RW_12_AF
90	100	2	**26** 0.33151	**23** 0.06616	**21** 0.02105	**22** 0.00841	**19** 0.29677	**17** **0.133**
20	50	2	**27** 0.32996	**23** 0.06432	**23** 0.01997	**23** 0.0080	**20** 0.29549	**17** 0.13235
40	50	3	**24** 0.33383	**25** 0.06189	**25** 0.01868	**26** 0.00661	**19** 0.29717	**17** 0.13221
80	50	2	**27** 0.32837	**25** 0.0632	**24** 0.01896	**25** 0.00691	**21** 0.29485	**17** 0.13234
10	60	2	**26** 0.33238	**23** 0.06501	**22** 0.02038	**23** 0.00802	**20** 0.29632	**18** 0.13204
80	60	2	**27** 0.32932	**25** 0.06285	**25** 0.01863	**24** 0.00729	**21** 0.29471	**17** 0.13212
10	90	3	**25** 0.33349	**25** 0.06327	**24** 0.01911	**24** 0.00751	**19** 0.29758	**17** 0.13235
50	90	3	**27** 0.33072	**26** 0.06127	**25** 0.01852	**24** 0.00777	**20** 0.29608	**17** 0.13211

(Meanings of titles are shown in Figure 7)

CHAPTER 6

CONCLUSION AND FUTURE WORK

We performed summarization in two main steps. First, key-terms were extracted then important sentences were extracted using the key-terms through clustering and centroid based approach. Key-terms were extracted using two methods: LSA and biggest TF.IDF. The aim of using two methods was to observe the success of LSA in key-term extraction.

After matching the results of key-term extraction with LSA and biggest TF.IDF we can conclude that our hypothesis of using LSA in key-term extraction is successful. Additionally key-terms were ordered according to their importance in step 1. Getting the best results for key-term percentage generally at 10% shows us that LSA is useful in finding the importance of terms in documents.

Getting poorer results over 3 clusters we can conclude that cluster numbers higher than a threshold value (3 clusters here) is detrimental for the performance of summarization. Additionally using rank-k approximation using LSI before clustering increased our success rate.

Based on the scores and the order of our system in the ROUGE results we can say that the success of our 2-step summarization approach is acceptable.

Weighting approaches can be developed and new weighting schemes can be applied to the summarization system as a future work. Additionally a method for estimating the cluster number can be used before clustering or K-Means algorithm may be replaced with other clustering algorithms as a whole. Sentences can be ordered inside the summary after extracting sentences to keep the order of events as in the original documents and to make summaries more understandable.

REFERENCES

[1] **MANI, I.** (2001), Introduction, *Automatic Summarization*, John Benjamins Publishing Co., Amsterdam/Philadelphia, 1-5.

[2] **BRANDOW, R., MITZE, K., RAU, L. F.** (1995), Automatic Condensation of Electronic Publications by Sentence Selection., Information *Processing & Management*, 675–685, Vol 31(5).

[3] **LIN, C. Y., HOVY, E. H.** (1997), Identifying Topics by Position, *Applied Natural Language Processing Conference*, 283–290.

[4] **TEUFEL, S., MOENS, M.** (1997), Sentence Extraction as a Classification Task, *ACL/EACL97-WS*, Madrid, 58-65.

[5] **EDMUNDSON, H. P.** (1969), New Methods in Automatic Extracting, *Journal of tile Association for Computing Machinery*, 264–285, Vol 16(2).

[6] **LUHN, H. P.** (1958), The Automatic Creation of Literature Abstracts, *IBM Journal of Research and Development*, 159-165, Vol 2(2).

[7] **SALTON, G.** et. al. (1997), Automatic Text Structuring and Summarization, *Information Processing and Management*, Vol 33(2).

[8] **BOGURAEV, B., KENEDY, C.** (1997), Salience-Based Content Characterisation of Text Documents, *Advances in Automatic Text Summarization*, MIT Press, 2-9.

[9] **BARZILAY, R., ELHADAD, M.** (1997), Using Lexical Chains for Text Summarization, *In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, Madrid, 10-17.

[10] **BERRY, M. W.** (1992), Large Scale Singular Value Computations, *International Journal of Supercomputer Applications*, 13-49, Vol 6.

[11] **BERRY, M. W., Dumais, S. T., O'Brien, G.W.** (1995), Using Linear Algebra for Intelligent Information Retrieval, *SIAM: Review*, 573-595, Vol 37.

[12] **GONG, Y., LIU, X.** (2001), Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis, *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, Louisiana, 19-25.

[13] **RADEV, D. R., JING, H., BUDZIKOWSKA, M.** (2000), Centroid-Based Summarization of Multiple Documents: Sentence Extraction, Utility-Based Evaluation, and User Studies, *In ANLP/NAACL Workshop on Summarization*, Seattle, WA.

[14] **MACQUEEN, J. B.** (1967), Some Methods for Classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, 281-297.

[15] **MANI, I.** (2001), Morphological-Level Approaches, *Automatic Summarization*, John Benjamins Publishing Co., Amsterdam/Philadelphia, 181-182.

[16] **SALTON, G., MCGILL, M. J.** (1983), *Introduction to Modern Information Retrieval*, McGraw Hill Book Co., New York.

[17] **ERCAN, G. (2006),** *Automated Text Summarization and Keyphrase Extraction*, M. S. Thesis, Institude of Engineering and Science, Bilkent University, Ankara.

[18] **LOVINS, J. B.** (1968), Development of a Stemming Algorithm, *Mechanical Translation and Computational Linguistics* 11, 22-31.

[19] **PORTER, M. F.** (1980), An Algorithm for Suffix Stripping, *Program*, 14(3), 130–137.

[20] http://tartarus.org/~martin/PorterStemmer/

[21] **SCHAUBLE, Peter** (1997), Vocabularies for Text Indexing, *Multimedia Information Retrieval: Content-Based Information Retrieval from Large Text and Audio Databases*, Kluwer Academic Publishers, Norwell, MA, 54–56.

[22] **STEINBERGER, J., JEŽEK, K.** (2004), Using Latent Semantic Analysis in Text Summarization and Summary Evaluation, *In Proc. ISIM '04*, 93-100.

[23] http://duc.nist.gov/duc2004/

[24] **LIKOWSKY, K. C.** (2004), Summarization Experiments in DUC 2004, *In Proceedings of the HLT-NAACL Workshop on Automatic Summarization*, Boston.

[25] http://berouge.com

[26] **LIN, C.Y.** (2004), Looking for a Few Good Metrics: ROUGE and Its Evaluation. *In Proceedings of NTCIR Workshop 2004*, Tokyo.

APPENDIX A

STOP WORDS

A	Different	just	present	true
abaft	directly	k	probably	'twas
aboard	do	l	provided	'tween
about	does	large	providing	'twere
above	doesn't	last	public	'twill
across	doing	later	q	'twixt
afore	done	least	qua	two
aforesaid	don't	left	quite	'twould
after	dost	less	r	u
again	doth	lest	rather	under
against	down	let's	re	underneath
agin	during	like	real	unless
ago	durst	likewise	really	unlike
aint	e	little	respecting	until
albeit	each	living	right	unto
all	early	long	round	up
almost	either	m	s	upon
alone	em	many	same	us
along	english	may	sans	used
alongside	enough	mayn't	save	usually
already	ere	me	saving	v

also	even	mid	second	versus
although	ever	midst	several	very
always	every	might	shall	via
am	everybody	mightn't	shalt	vice
american	everyone	mine	shan't	vis-a-vis
amid	everything	minus	she	w
amidst	except	more	shed	wanna
among	excepting	most	shell	wanting
amongst	f	much	she's	was
an	failing	must	short	wasn't
and	far	mustn't	should	way
anent	few	my	shouldn't	we
another	first	myself	since	we'd
any	five	n	six	well
anybody	following	near	small	were
anyone	for	'neath	so	weren't
anything	four	need	some	wert
are	from	needed	somebody	we've
aren't	g	needing	someone	what
around	gonna	needn't	something	whatever
as	gotta	needs	sometimes	what'll
aslant	h	neither	soon	what's
astride	had	never	special	when
at	hadn't	nevertheless	still	whencesoever
athwart	hard	new	such	whenever
away	has	next	summat	when's
b	hasn't	nigh	supposing	whereas
back	hast	nigher	sure	where's
bar	hath	nighest	t	whether
barring	have	nisi	than	which
be	haven't	no	that	whichever

because	having	no-one	that'd	whichsoever
been	he	nobody	that'll	while
before	he'd	none	that's	whilst
behind	he'll	nor	the	who
being	her	not	thee	who'd
below	here	nothing	their	whoever
beneath	here's	notwithstanding	theirs	whole
beside	hers	now	their's	who'll
besides	herself	o	them	whom
best	he's	o'er	themselves	whore
better	high	of	then	who's
between	him	off	there	whose
betwixt	himself	often	there's	whoso
beyond	his	on	these	whosoever
both	home	once	they	will
but	how	one	they'd	with
by	howbeit	oneself	they'll	within
c	however	only	they're	without
can	how's	onto	they've	wont
cannot	i	open	thine	would
can't	id	or	this	wouldn't
certain	if	other	tho	wouldst
circa	ill	otherwise	those	x
close	i'm	ought	thou	y
concerning	immediately	oughtn't	though	ye
considering	important	our	three	yet
cos	in	ours	thro'	you
could	inside	ourselves	through	you'd
couldn't	instantly	out	throughout	you'll
couldst	into	outside	thru	your
d	is	over	thyself	you're

dare dared daren't dares daring despite did didn't	isn't it it'll it's its itself i've j	own p past pending per perhaps plus possible	till to today together too touching toward towards	yours yourself yourselves you've z

APPENDIX B

ROUGE SCORES

Table 4: ROUGE Results for Key-Term of 10% & Rank-k Approximation of 70%

Term %	Rank-k %	Cluster No	R1_AF	R2_AF	R3_AF	R4_AF	RL_AF	RW_12_AF
10	70	1	**27** 0.33053	**21** 0.06894	**16** 0.0232	**15** 0.01007	**20** 0.29541	**15** 0.13348
10	70	2	**23** 0.33504	**22** 0.06855	**20** 0.02164	**19** 0.00887	**16** 0.29938	**15** 0.13463
10	70	3	**19** 0.34066	**22** 0.0681	**21** 0.0212	**20** 0.00858	**14** 0.30349	**13** 0.13565
10	70	4	**21** 0.3392	**23** 0.06683	**20** 0.02224	**18** 0.00956	**14** 0.30305	**15** 0.13479
10	70	5	**27** 0.32906	**26** 0.05991	**25** 0.01876	**22** 0.00808	**23** 0.29253	**23** 0.12926
10	70	6	**27** 0.32415	**28** 0.05724	**26** 0.0172	**24** 0.0074	**23** 0.28873	**24** 0.1264
10	70	7	**30** 0.31253	**32** 0.05101	**26** 0.01461	**27** 0.00577	**26** 0.27977	**26** 0.1227
10	70	8	**29** 0.31448	**32** 0.05013	**28** 0.01406	**28** 0.00553	**26** 0.28147	**26** 0.12333
20	70	1	**27** 0.32419	**23** 0.06493	**21** 0.0208	**21** 0.00847	**23** 0.28999	**21** 0.13077
20	70	2	**25** 0.3325	**22** 0.06717	**21** 0.02095	**20** 0.00862	**19** 0.29657	**17** 0.13269
20	70	3	**23** 0.33597	**23** 0.06537	**21** 0.02066	**21** 0.00844	**17** 0.29875	**16** 0.13338
20	70	4	**23** 0.33637	**24** 0.06383	**22** 0.02011	**23** 0.00799	**19** 0.29711	**17** 0.13211

Term %	Rank-k %	Cluster No	R1_AF	R2_AF	R3_AF	R4_AF	RL_AF	RW_12_AF
20	70	5	**28** 0.32131	**29** 0.05443	**26** 0.01587	**26** 0.00638	**24** 0.28583	**24** 0.12679
20	70	6	**29** 0.31972	**31** 0.05197	**27** 0.01433	**27** 0.00563	**24** 0.28677	**24** 0.12655
20	70	7	**29** 0.31621	**29** 0.05387	**26** 0.01586	**26** 0.00669	**25** 0.28462	**25** 0.12602
20	70	8	**29** 0.31412	**32** 0.05115	**29** 0.01345	**27** 0.0056	**26** 0.28095	**26** 0.12379
30	70	1	**27** 0.32525	**23** 0.0644	**21** 0.02083	**20** 0.00859	**23** 0.29053	**21** 0.13081
30	70	2	**27** 0.32362	**26** 0.05912	**26** 0.01765	**24** 0.00737	**23** 0.2886	**23** 0.12954
30	70	3	**27** 0.32766	**26** 0.05939	**26** 0.01733	**26** 0.00671	**23** 0.29107	**23** 0.12961
30	70	4	**28** 0.32335	**28** 0.05712	**26** 0.01573	**27** 0.00579	**23** 0.28727	**23** 0.12734
30	70	5	**29** 0.31289	**32** 0.05073	**27** 0.01453	**28** 0.00547	**26** 0.27865	**26** 0.12319
30	70	6	**30** 0.3124	**32** 0.05036	**27** 0.0145	**26** 0.00587	**26** 0.2787	**26** 0.12241
30	70	7	**32** 0.3088	**33** 0.04628	**33** 0.01153	**31** 0.00392	**26** 0.27335	**29** 0.11989
30	70	8	**32** 0.30769	**33** 0.04739	**32** 0.01245	**28** 0.00491	**26** 0.27651	**27** 0.12151
40	70	1	**27** 0.32489	**23** 0.06477	**21** 0.02101	**19** 0.00868	**23** 0.29016	**20** 0.13085
40	70	2	**25** 0.3334	**25** 0.06239	**24** 0.01959	**24** 0.00773	**19** 0.29667	**17** 0.13294
40	70	3	**27** 0.32977	**25** 0.06204	**24** 0.01914	**24** 0.00763	**22** 0.29349	**23** 0.12995
40	70	4	**27** 0.32543	**28** 0.05722	**26** 0.01591	**26** 0.00639	**23** 0.29033	**23** 0.12894
40	70	5	**27** 0.32368	**27** 0.05808	**26** 0.01686	**25** 0.00688	**23** 0.28802	**23** 0.12763
40	70	6	**29** 0.31434	**32** 0.05058	**28** 0.01405	**26** 0.00601	**26** 0.2801	**26** 0.12288
40	70	7	**31** 0.3094	**32** 0.04973	**30** 0.01309	**27** 0.0057	**26** 0.27514	**28** 0.12125
40	70	8	**29** 0.31435	**31** 0.05183	**27** 0.01438	**26** 0.00604	**26** 0.2812	**26** 0.12324

Term %	Rank-k %	Cluster No	R1_AF	R2_AF	R3_AF	R4_AF	RL_AF	RW_12_AF
50	70	1	**28** 0.32225	**23** 0.06412	**21** 0.02092	**19** 0.00866	**23** 0.28751	**23** 0.12984
50	70	2	**27** 0.32759	**26** 0.06068	**25** 0.01843	**24** 0.00749	**23** 0.29052	**23** 0.12987
50	70	3	**27** 0.32531	**26** 0.05872	**26** 0.01747	**25** 0.0071	**23** 0.28903	**23** 0.12886
50	70	4	**27** 0.32855	**27** 0.05854	**26** 0.01578	**27** 0.00576	**23** 0.29254	**21** 0.13051
50	70	5	**29** 0.31734	**29** 0.05397	**30** 0.01323	**29** 0.00473	**26** 0.28172	**26** 0.12583
50	70	6	**29** 0.31812	**31** 0.05179	**29** 0.01338	**28** 0.00503	**26** 0.28264	**26** 0.12516
50	70	7	**29** 0.31454	**33** 0.04831	**31** 0.0125	**28** 0.00499	**26** 0.2805	**26** 0.12334
50	70	8	**32** 0.30852	**33** 0.04843	**33** 0.01129	**31** 0.00413	**26** 0.27455	**26** 0.12166
60	70	1	**28** 0.32178	**25** 0.06371	**21** 0.0207	**20** 0.00857	**24** 0.28697	**23** 0.12959
60	70	2	**26** 0.33187	**25** 0.06334	**24** 0.0195	**23** 0.0079	**21** 0.29503	**17** 0.13256
60	70	3	**29** 0.31858	**28** 0.05723	**26** 0.01756	**25** 0.00718	**24** 0.28521	**23** 0.12694
60	70	4	**29** 0.31282	**31** 0.05176	**26** 0.01466	**27** 0.00561	**26** 0.27855	**26** 0.1236
60	70	5	**31** 0.31209	**32** 0.05092	**26** 0.01499	**26** 0.00642	**26** 0.2793	**26** 0.12312
60	70	6	**31** 0.31161	**30** 0.05382	**26** 0.01644	**24** 0.00739	**26** 0.28091	**26** 0.12366
60	70	7	**32** 0.30723	**33** 0.04842	**27** 0.01427	**26** 0.00646	**26** 0.27542	**28** 0.12146
60	70	8	**32** 0.30565	**33** 0.04699	**33** 0.01199	**28** 0.00516	**26** 0.27387	**28** 0.12067
70	70	1	**27** 0.32377	**23** 0.06425	**21** 0.02089	**21** 0.00852	**23** 0.289	**21** 0.13033
70	70	2	**27** 0.33101	**23** 0.06421	**22** 0.02003	**23** 0.00796	**21** 0.29497	**17** 0.13254
70	70	3	**27** 0.32668	**27** 0.05775	**26** 0.01813	**24** 0.00778	**23** 0.2913	**23** 0.12967
70	70	4	**28** 0.32328	**29** 0.05489	**26** 0.01615	**26** 0.00667	**23** 0.2883	**23** 0.12804

Term %	Rank-k %	Cluster No	R1_AF	R2_AF	R3_AF	R4_AF	RL_AF	RW_12_AF
70	70	5	**29** 0.31598	**32** 0.05131	**30** 0.01277	**31** 0.00402	**26** 0.28214	**26** 0.12425
70	70	6	**29** 0.31361	**32** 0.05008	**32** 0.01234	**31** 0.00399	**26** 0.28042	**26** 0.12362
70	70	7	**32** 0.30844	**33** 0.04764	**33** 0.01144	**31** 0.00378	**26** 0.27596	**28** 0.12137
70	70	8	**32** 0.30894	**34** 0.04365	**34** 0.00934	**34** 0.00297	**26** 0.27503	**28** 0.12018
80	70	1	**27** 0.325	**23** 0.06422	**21** 0.02065	**21** 0.00846	**23** 0.2897	**21** 0.13064
80	70	2	**27** 0.33118	**25** 0.06226	**24** 0.01972	**23** 0.00799	**22** 0.29328	**20** 0.13118
80	70	3	**27** 0.32722	**28** 0.05718	**26** 0.01714	**24** 0.00726	**23** 0.29021	**23** 0.1289
80	70	4	**28** 0.32133	**28** 0.05659	**26** 0.01773	**24** 0.00777	**23** 0.28832	**23** 0.12771
80	70	5	**31** 0.31036	**33** 0.04594	**33** 0.0117	**28** 0.00495	**26** 0.27636	**26** 0.1225
80	70	6	**32** 0.30712	**33** 0.0484	**31** 0.0125	**28** 0.00549	**26** 0.2769	**26** 0.12224
80	70	7	**32** 0.30345	**33** 0.04488	**34** 0.01024	**30** 0.00432	**26** 0.26996	**30** 0.11882
80	70	8	**34** 0.29494	**34** 0.04023	**34** 0.00865	**33** 0.00341	**32** 0.26193	**33** 0.11566
90	70	1	**27** 0.32464	**23** 0.06407	**21** 0.02082	**21** 0.00845	**23** 0.28948	**21** 0.13048
90	70	2	**27** 0.3273	**25** 0.06278	**24** 0.01968	**24** 0.00736	**23** 0.29214	**20** 0.13093
90	70	3	**26** 0.33204	**28** 0.05741	**26** 0.01657	**26** 0.00631	**22** 0.29324	**20** 0.13122
90	70	4	**27** 0.32445	**30** 0.05323	**27** 0.01451	**27** 0.00575	**23** 0.28762	**23** 0.12691
90	70	5	**31** 0.31082	**33** 0.0479	**33** 0.0122	**28** 0.00488	**26** 0.27658	**26** 0.1219
90	70	6	**32** 0.30759	**33** 0.04749	**33** 0.01218	**28** 0.00495	**26** 0.27428	**28** 0.12059
90	70	7	**32** 0.30038	**34** 0.04233	**34** 0.01082	**30** 0.00423	**28** 0.26934	**30** 0.11866
90	70	8	**32** 0.30579	**33** 0.04589	**33** 0.01111	**30** 0.00443	**26** 0.27358	**28** 0.12063

Term %	Rank-k %	Cluster No	R1_AF	R2_AF	R3_AF	R4_AF	RL_AF	RW_12_AF
100	70	1	**27** 0.32497	**23** 0.06488	**21** 0.02104	**21** 0.00845	**23** 0.28961	**21** 0.13072
100	70	2	**27** 0.3238	**26** 0.06151	**24** 0.01963	**22** 0.00815	**23** 0.28869	**23** 0.1298
100	70	3	**27** 0.32461	**28** 0.05768	**25** 0.0184	**24** 0.00759	**23** 0.28924	**23** 0.12881
100	70	4	**28** 0.32297	**28** 0.05763	**26** 0.01659	**26** 0.00639	**23** 0.28749	**23** 0.12773
100	70	5	**29** 0.31496	**32** 0.05061	**29** 0.0134	**28** 0.00525	**26** 0.28184	**26** 0.12535
100	70	6	**31** 0.31105	**32** 0.05017	**33** 0.01193	**29** 0.00452	**26** 0.27791	**26** 0.12306
100	70	7	**31** 0.3102	**33** 0.04508	**34** 0.01029	**32** 0.0035	**26** 0.27604	**28** 0.12132
100	70	8	**32** 0.30688	**33** 0.04508	**34** 0.01017	**32** 0.00372	**26** 0.2744	**26** 0.12168

(Meanings of titles are shown in Figure 7)