

LANGUAGE INDEPENDENT MULTI DOCUMENT SUMMARIZATION
USING LATENT SEMANTIC INDEXING/CLUSTERING TECHNIQUES

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

BY

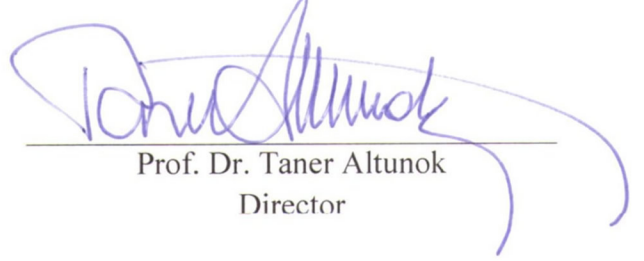
SUAT ALİM

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

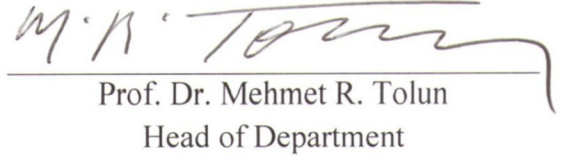
DECEMBER 2009

Title of Thesis: **Language Independent Multi Document Summarization Using
Latent Semantic Indexing/Clustering Techniques**
Submitted by **Suat Alim**

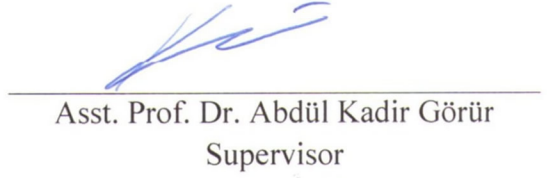
Approval of the Graduate School of Natural and Applied Sciences, Çankaya
University


Prof. Dr. Taner Altunok
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

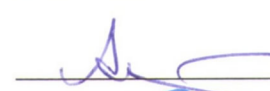
This is to certify that we have read this thesis and that in our opinion it is fully
adequate, in scope and quality, as a thesis for the degree of Master of Science.



Asst. Prof. Dr. Abdül Kadir Görür
Supervisor

Examination Date: 03.12.2009

Examining Committee Members

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

Dr. Ali Rıza AŞKUN (Çankaya Univ.) 

Prof. Dr. Hayri SEVER (Hacettepe Univ.) 

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 : Suat ALİM

Signature :



Date

: 03.12.2009

ABSTRACT

LANGUAGE INDEPENDENT MULTI DOCUMENT SUMMARIZATION USING LATENT SEMANTIC INDEXING/CLUSTERING TECHNIQUES

Alim, Suat

M.S.c., Department of Computer Engineering

Supervisor: Asst. Prof. Dr. Abdül Kadir Görür

December 2009, 93 Pages

This thesis discusses our research on language independent multi-document summarization. We used latent semantic indexing and centroid based clustering methods in our summarization process. Firstly, our algorithm uses latent semantic analysis (LSA) to extract key-terms. Secondly, important sentences holding these key-terms are extracted by applying latent semantic indexing (LSI) and centroid-based clustering methods. Our experiments show that LSA improve key-term extraction. Also, our summarization system has achieved good results, compared to some other multi-document summarization systems.

Keywords: Language Independent Multi-document Summarization, Latent Semantic Analysis, Latent Semantic Indexing, Centroid Based Summarization

ÖZ

SAKLI ANLAM İNDEKSLEME VE KÜMELEME TEKNİKLERİ İLE DİLDEN BAĞIMSIZ ÇOKLU DOKÜMAN ÖZETLEME

Alim, Suat

Yüksek Lisans, Bilgisayar Mühendisliği Bölümü

Danışman: Asst. Prof. Dr. Abdül Kadir Görür

Aralık 2009, 93 Sayfa

Bu tez dilden bağımsız olarak çoklu dokümanlardan özet çıkarılması üzerine yaptığımız araştırmayı içermektedir. Özetleme işlemimizde saklı anlamsal indeksleme ve sanal merkeze dayalı kümeleme yöntemlerinden yararlandık. Sistemimizde ilk olarak saklı anlamsal analiz yöntemi kullanılarak anahtar terimler çıkarılır. Daha sonra anahtar terimleri içeren özet cümleler saklı anlam indeksleme ve sanal merkeze dayalı kümeleme yöntemleri kullanılarak çıkarılır. Yaptığımız deneyler saklı anlamsal analiz yönteminin anahtar kelimelerin çıkarılmasında başarılı olduğunu ortaya koymaktadır. Ayrıca, özet çıkarma sistemimiz diğer çoklu doküman özetleme sistemleri ile karşılaştırılınca iyi sonuçlar elde etmiştir.

Anahtar Kelimeler: Dilden Bağımsız Olarak Çoklu Dokümanların Özetlenmesi, Saklı Anlamsal Analiz, Saklı Anlam İndeksleme, 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 Samet Karakaynak for his informative comments and technical support throughout the thesis.

TABLE OF CONTENTS

STATEMENT OF NON-PLAGIARISM	iii
ABSTRACT	iv
ÖZ	v
ACKNOWLEDGEMENTS	vi
TABLE OF CONTENTS	vii
LIST OF TABLES	x
LIST OF FIGURES	xi
CHAPTERS:	
1. INTRODUCTION	1
1.1 Summarization	1
1.2 Thesis Outline	3
2. RELATED WORK	5
2.1 Content Selection and Importance Identification	5
2.2 Text Generation, Text Compression and Smoothing	7
3. BACKGROUND WORK	8
3.1 Singular Value Decomposition	8
3.1.1 Mathematical Definition of SVD	8
3.1.2 Computing SVD of a Matrix	10
3.2 Latent Semantic Indexing	12
3.2.1 LSI Working Structure	12
3.3 Latent Semantic Analysis	14
3.3.1 LSA Working Structure	15
3.4 Centroid-Based Summarization of Multiple Documents	16
3.4.1 What is Centroid	16

3.4.2 Centroid-Based Summarization	16
3.4.3 Centroid-Based Clustering	16
3.5 Clustering Methods	17
3.5.1 K-Means Clustering	17
3.5.2 QT (Quality Threshold) Clustering	17
3.5.3 Agglomerative Hierarchical Clustering	18
3.6 Cosine Similarity	18
3.7 TF.IDF Weighting.....	19
4. LANGUAGE INDEPENDENT MULTI DOCUMENT SUMMARIZATION USING LATENT SEMANTIC INDEXING/CLUSTERING TECHNIQUES	21
4.1 Roadmap.....	21
4.1.1 Step 1	23
4.1.2 Step 2	25
4.2 Sentence Detector	27
4.3 Removing Stop Words	27
4.4 Stemming.....	27
4.4.1 English Stemmer	28
4.4.2 Turkish Stemmer	28
4.5 Extracting Key-Terms using Latent Semantic Analysis	28
4.6 LSI (Rank-k Approximation)	30
4.7 Clustering	31
4.8 Sentence Extraction using Centroid-Based Approach	31
4.9 Weighting	32
5. EXPERIMENTS & EVALUATION.....	34
5.1 Experiments	34
5.1.1 English Documents for Summarization.....	34
5.1.2 Turkish Documents for Summarization.....	41
5.2 Evaluation.....	53
6. CONCLUSION AND FUTURE WORK	61
6.1 Future Work.....	62
REFERENCES	R1

APPENDICIES:

A. STOP WORDS	A1
English Stop Words	A1
English Stop Words (cont.)	A2
English Stop Words (cont.)	A3
Turkish Stop Words	A4
B. ROUGE SCORES.....	A5
Top ROUGE Results with K-Means Clustering	A5
Top ROUGE Results with K-Means Clustering (cont.)	A6
Top ROUGE Results with K-Means Clustering (cont.)	A7
Top ROUGE Results with K-Means Clustering (cont.)	A8
Top ROUGE Results with K-Means Clustering (cont.)	A9
Top ROUGE Results with QT Clustering.....	A10
Top ROUGE Results with QT Clustering (cont.).....	A11
Top ROUGE Results with QT Clustering (cont.).....	A12
Top ROUGE Results with Agglomerative Hierarchical Clustering.....	A13
Top ROUGE Results with Agglomerative Hierarchical Clustering (cont.).....	A14
C. CURRICULUM VITAE	A15

LIST OF TABLES

TABLES	PAGE
Table 3.1: Interpretation of SVD Components within LSI.....	13
Table 5.1: Best ROUGE Results with K-Means Clustering.....	54
Table 5.2: Best ROUGE Results with QT Clustering	55
Table 5.3: Best ROUGE Results with Agglomerative Hierarchical Clustering ..	56
Table 5.4: Best ROUGE Results for Biggest TF.IDF Method in Key-Term Extraction	59
Table 5.5: Best ROUGE Results for Random Sentence Selection	60

LIST OF FIGURES

FIGURES	PAGE
Figure 3.1: Mathematical Representation of the Matrix A_k	13
Figure 4.1: Roadmap	22
Figure 4.2: STEP 1: Key-Term Extraction	24
Figure 4.3: STEP 2: Sentence Extraction	26
Figure 4.4: Rank-k Approximation	30
Figure 4.5: Sentence-Term Matrix in a Cluster	31
Figure 5.1: Sample Document for English 1	35
Figure 5.2: Sample Document for English 2 – Part 1	36
Figure 5.3: Sample Document for English 2 - Part 2	37
Figure 5.4: Sample Document for English 3	38
Figure 5.5: Sample Key-Terms Extracted Using LSA for English.....	39
Figure 5.6: Sample Summary Using Key-Terms from LSA for English.....	39
Figure 5.7: Sample Key-Terms Extracted Using Biggest TF.IDF Method for English.....	40
Figure 5.8: Sample Summary Using Key-Terms from Biggest TF.IDF for English.....	40
Figure 5.9: Sample Document 1 for Turkish Set 1	42
Figure 5.10: Part 1 of Sample Document 2 for Turkish Set 1	43
Figure 5.11: Part 2 of Sample Document 2 for Turkish Set 1	44
Figure 5.12: Part 1 of Sample Document 3 for Turkish Set 1	45
Figure 5.13: Part 2 of Sample Document 3 for Turkish Set 1	46
Figure 5.14: Sample Key-Terms Extracted Using LSA for Turkish Set 1.....	47
Figure 5.15: Sample Summary Using Key-Terms from LSA for Turkish Set 1 .47	

Figure 5.16: Sample Document 1 for Turkish Set 2	48
Figure 5.17: Sample Document 2 for Turkish Set 2	49
Figure 5.18: Part 1 of Sample Document 3 for Turkish Set 2	50
Figure 5.19: Part 2 of Sample Document 3 for Turkish Set 2	51
Figure 5.20: Sample Key-Terms Extracted Using LSA for Turkish Set 2.....	52
Figure 5.21: Sample Summary Using Key-Terms from LSA for Turkish Set 2.	52
Figure 5.22: Meanings of Titles in Result Tables	55
Figure 5.23: Number of Best Results for Each Term Percentage	57
Figure 5.24: Number of Best Results for Each Rank-k Percentage	57
Figure 5.25: Number of Best Results for Each Cluster Number	58
Figure 5.26: Number of Best Results for Each Threshold Value	58

CHAPTER 1

INTRODUCTION

1.1 Summarization

As the number of electronic documents increase rapidly depending on the growth of internet access, the need for faster techniques to retrieve the suitable information becomes important. Also duplication of many documents with the same or similar topics is another problem. This kind of data duplication problem increases the necessity for effective document summarization. Since traditional Information Retrieval systems returns redundant information, this problem can be eliminated by using summarization as a complementary approach in Information Retrieval systems. Thus, creating a summary by extracting important sentences from the original text is a common technique in automated text summarization.

In addition, showing a summary of text sources can be better than showing only the links to the user. Hence number of automatic text summarization researches increase every day.

A summary is defined as a condensed representation of the underlying text [2]. An ideal summary must contain full meaning of the document. It should provide the most important information in the document. A summary should be non-repetitive and as brief as possible. From the definition of summary we can say that

summarization is reduction of source text(s) to a shorter version without losing semantic content.

Goal of the summarization is defined in [1] as: "The goal of text summarization is to present the most important information in a shorter version of the original text while keeping its main content and helps the user to quickly understand large volumes of information."

Inderjeet Mani defines summarizer in [2] as: "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 systems are used to create summaries. They differ depending on the text extraction method and the number of documents used as input.

Text summaries created by **Query based summaries** are based on a given search query. Query based summarization is very similar to question answering. The generated summary is shaped by the user's interest.

However **Generic summarization** is the process of separating the most important information from a source to produce a summary. A generic summary presents a general meaning of the documents' contents.

A summary generated by selecting fragments of the original text source is **extract**. In an extract, selected fragments should be representative and the most important parts of the original text(s). Text generation is not needed for extracts, since it is formed from text selected from the original text(s).

Abstract is a special form of summary that is generated/paraphrased text from the original text source. Programs for abstraction are harder to develop, since they use their own words to create summary. However a shorter abstract can express more meaning than a longer extract for a text.

A **single-document summarization** system uses a single document as input to produce a single summary. A **multi-document summarization** system uses a set of documents as input to produce a single summary. Multi-document summarization has additional problems compared to single-document summarization such as redundancy, inconsistency problems. Also meaning of a summary created by multi-document summarization can be weak as a result of confusion on time sequence of the events. Because of these problems multi-document summarization becomes more challenging.

In our thesis, we created a multi-document summarization system that uses key-terms and sentence extraction. We focus on important sentence and key-term extraction. All key-terms and sentences are evaluated by their importance. Extracting sentences and key-terms are similar problems. Hence our system uses LSA, LSI and centroid based summarization methods to overcome these problems.

1.2 Thesis Outline

In Chapter 2, related work in summarization research is outlined. Different summarization methods developed over the years are briefly introduced.

Techniques and algorithms used in our summarization system are described as background work in chapter 3. These terms are explained in computational perspective.

Chapter 4 defines our multiple document summarization algorithm based on latent semantic indexing and clustering algorithms. This chapter gives the details of our implementation step by step.

Samples for experiment sets used by our summarization system and some sample summaries created by our system are given in chapter 5. Also results obtained from our summarization method are evaluated. Overall performance of our algorithm and comparison of our results to other algorithms results are provided in this chapter.

As a conclusion, in Chapter 6, possible improvements and possible applications for the work on this thesis are discussed.

CHAPTER 2

RELATED WORK

Summarization has been an active research area for last 50 years. Content selection/importance identification and text generation/smoothing extracts are main phases of summarization task. The most important part of recent summarization systems use identifications and extractions of significant sentences from document(s).

2.1 Content Selection and Importance Identification

A summarization system tries to identify significant information that is important enough to be in the summary. For sentence/clause identification different methods have been used. Most widely used methods based on positions in the text, cues, titles and headings, term frequencies and cohesions among words and expressions.

According to Brandow, Mitze and Rau [3] important sentences occur at the **beginning** (and/or end) of texts. Many experiments show that, this simple technique gives the best results in news articles and scientific reports. An algorithm extracts the first sentences from a document has been one of the best scoring algorithm when the summary is limited to 75 characters in DUC 2004 [30] Conference. But according to experiments given below;

- In 85% of 200 individual paragraphs the topic sentences occurred in initial position and in 7% in final position [4].

- Only 13% of the paragraphs of contemporary writers start with topic sentences [5].

Also, according to a large scaled research of Lin and Hovy [6] on optimum position policy focus position changes with different text genres.

Cue phrases were firstly used by Teufel [7] on science articles. This method yielded the best result in scientific articles. Two types of cue phrases are defined as follows:

- **Bonus Phrases** attracts attention to the important sentences where they appear. “Significantly”, “in conclusion”, “as a result” are some examples of bonus phrases.
- **Stigma Phrases** indicates sentences where they appear as not important. “Hardly” and “impossible” are some examples of stigma phrases.

According to Edmundson [8] the words in titles and headings occur mostly in semantically important sentences too. Edmundson showed that this method statistically valid at 99% level of significance. Also using the formatted features like bold words could improve the summarization performance. This method is used by other approaches as a complementary approach to increase the system performance.

Luhn [9] claims that important sentences contain unusually frequent words in the text. But Edmundson [8] claimed that using word frequency is harmful for his system performance. Luhn applied word frequency rules to identify sentences to create summaries. This method increases sentence score for each frequent word.

Cohesion based methods research the relations among words or expressions. According to the cohesion based methods the entities having the tightest connections in cohesion models are important sentences or paragraphs. To identify the connections among the words or expressions several approaches have been used and most commons are based on term co-occurrence [10], co-reference [11] and lexical chains [12].

A lexical chain is defined as a list of related words in the text documents. A lexical chain is independent of the grammatical structure. All words in a lexical chain have a distance relation with each others. Barzilay and Elhadad [12] created all possible lexical chains from text documents. Then summaries are created by focusing on strong chains.

2.2 Text Generation, Text Compression and Smoothing

Ideally, a summarization system should interpret the text, transform it into a semantic representation and generate the summary from the semantic representation. Interpreting the text is a hard problem. Extensive domain knowledge is required for interpretation.

Some researchers tried to fill some predefined templates to create summaries, by treating summarization as information extraction problem. Paraphrasing or reducing the sentences extracted by extractive summarization systems could provide more coherent and shorter summaries. A text compression algorithm to reduce sentences to shorter ones could be used and multiple sentences can be reduced into one. A summary revision system could be used which takes an extract and produces a shorter and more readable version for it.

CHAPTER 3

BACKGROUND WORK

3.1 Singular Value Decomposition

Singular value decomposition is a factorization of a rectangular real or complex matrix. That method is generally used to solve unconstrained linear least squares problems, matrix rank estimation and canonical correlation analysis [13].

3.1.1 Mathematical Definition of SVD

Let,

A denote an $m \times n$ matrix of a real valued data,

U and V are the orthogonal matrices and their first r columns identify the orthonormal eigenvectors with the r nonzero eigenvalues of $A^T A$ and $A A^T$.

- Right singular vectors is associated with the columns of V matrices,
- Left singular vectors is associated with the columns of U matrices,
- Diagonal elements of Σ which are the nonnegative square roots of the n eigenvalues of AA^T , refers the singular values of A [13].

Where,

$$m \geq n,$$

$$r \leq n,$$

$$\text{rank}(A) = r,$$

$$V^T V = U^T U = I_n,$$

$$\begin{aligned}\Sigma &= \text{diag}(\sigma_1, \dots, \sigma_n), \\ \sigma_i &> 0 \text{ for } 1 \leq i \leq r, \\ \sigma_j &= 0 \text{ for } j \geq r + 1\end{aligned}$$

Then the Singular Value Decomposition equation of matrix A “SVD (A)” is the following:

$$A = U\Sigma V^T \quad (3.1)$$

The following two theorems given in [13] illustrate the ways how SVD show important information about the structure of a matrix.

Theorem 1:

Let,

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

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

R(A) is range of A,

N(A) is null space of A.

Then,

1. $\text{rank}(A) = r$

$N(A) \equiv \text{span} \{v_{r+1}, \dots, v_n\}$

$R(A) \equiv \text{span} \{u_1, \dots, u_r\}$

where,

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

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

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

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

Theorem 2

Let,

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

with $r = \text{rank}(A) \leq p = \min(m, n)$,

$k < r$ and define

$$A_k = \sum_{i=1}^k u^i \cdot \sigma^i \cdot v_i^T \quad (3.2)$$

Then,

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

Constructed from the k largest singular triplets of A , A_k is the best rank- k approximation matrix to the matrix A [14]:

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

3.1.2 Computing SVD of a Matrix

Following example provides instructions for decomposing the matrix A using the singular value decomposition (SVD) algorithm. It covers singular values, right and left eigenvectors and a shortcut for computing the full SVD of the matrix.

$$\text{Let, } A = \begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix}$$

$$\text{Then, } A^T = \begin{bmatrix} 4 & 3 \\ 0 & -5 \end{bmatrix}$$

Now by using these matrices we construct a new matrix by multiplying them $A^T \cdot A$.

$$A^T \cdot A = \begin{bmatrix} 4 & 3 \\ 0 & -5 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix} = \begin{bmatrix} 25 & -15 \\ -15 & 25 \end{bmatrix}$$

Then we will find the eigenvalues of $A^T \cdot A$ and sort them in descending order. Eigenvalues of a matrix are found by characteristic equation.

$$\det \begin{bmatrix} 25 - \lambda & -15 \\ -15 & 25 - \lambda \end{bmatrix} = 0$$

$$(25 - \lambda)(25 - \lambda) - (-15)(-15) = 0$$

$$\lambda^2 - 50\lambda + 400 = 0$$

From the equation eigenvalues will be $\lambda_1 = 40$ and $\lambda_2 = 10$.

These eigenvalues are used to find eigenvectors which are the columns of V .

for $\lambda_1 = 40$

$$\begin{bmatrix} 25 & -15 \\ -15 & 25 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 40 \begin{bmatrix} x \\ y \end{bmatrix}$$

$$25x - 15y = 40x \text{ and } -15x + 25y = 40y$$

Then $x = -y$

$$\text{Length}(V_1) = \sqrt{x^2 + y^2} = \sqrt{2x^2} = x\sqrt{2}$$

$$V_1 = \begin{bmatrix} x/x\sqrt{2} \\ -x/x\sqrt{2} \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$$

$$V_1 = \begin{bmatrix} 0,7071 \\ -0,7071 \end{bmatrix}$$

for $\lambda_2 = 10$

$$\begin{bmatrix} 25 & -15 \\ -15 & 25 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 10 \begin{bmatrix} x \\ y \end{bmatrix}$$

$$25x - 15y = 10x \text{ and } -15x + 25y = 10y$$

Then $x = y$

$$\text{Length}(V_2) = \sqrt{x^2 + y^2} = \sqrt{2x^2} = x\sqrt{2}$$

$$V_2 = \begin{bmatrix} x/x\sqrt{2} \\ x/x\sqrt{2} \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

$$V_2 = \begin{bmatrix} 0,7071 \\ 0,7071 \end{bmatrix}$$

$$V = [V_1 \quad V_2] = \begin{bmatrix} 0,7071 & 0,7071 \\ -0,7071 & 0,7071 \end{bmatrix}$$

$$V^T = \begin{bmatrix} 0,7071 & -0,7071 \\ 0,7071 & 0,7071 \end{bmatrix}$$

Then singular values of these eigenvalues will be calculated by taking square root of them. So singular values are found as $\Sigma_1 = \sqrt{40} = 6,3245$ and $\Sigma_2 = \sqrt{10} = 3,1622$.

These singular values are placed in a new matrix Σ in descending order along its diagonal, and its inverse Σ^{-1} is calculated.

$$\Sigma = \begin{bmatrix} 6,3245 & 0 \\ 0 & 3,1622 \end{bmatrix}$$

$$\Sigma^{-1} = \begin{bmatrix} 6,3245 & 0 \\ 0 & 3,1622 \end{bmatrix}^{-1} = 1/(6,3245 \cdot 3,1622) \begin{bmatrix} 3,1622 & 0 \\ 0 & 6,3245 \end{bmatrix}$$

$$\Sigma^{-1} = \begin{bmatrix} 0,1581 & 0 \\ 0 & 0,3162 \end{bmatrix}$$

Finally we take the U from the equation of SVD as $U = AV\Sigma^{-1}$

$$U = AV\Sigma^{-1} = \begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix} \begin{bmatrix} 0,7071 & 0,7071 \\ -0,7071 & 0,7071 \end{bmatrix} \begin{bmatrix} 0,1581 & 0 \\ 0 & 0,3162 \end{bmatrix}$$

$$U = AV\Sigma^{-1} = \begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix} \begin{bmatrix} 0,1118 & 0,2236 \\ -0,1118 & 0,2236 \end{bmatrix}$$

$$U = AV\Sigma^{-1} = \begin{bmatrix} 0,4472 & 0,8944 \\ 0,8944 & -0,4472 \end{bmatrix}$$

By multiplying found results of matrices we can proof that operation is correct.

$$A = U\Sigma V^T = \begin{bmatrix} 0,4472 & 0,8944 \\ 0,8944 & -0,4472 \end{bmatrix} \begin{bmatrix} 6,3245 & 0 \\ 0 & 3,1622 \end{bmatrix} \begin{bmatrix} 0,7071 & -0,7071 \\ 0,7071 & 0,7071 \end{bmatrix}$$

$$A = U\Sigma V^T = \begin{bmatrix} 0,4472 & 0,8944 \\ 0,8944 & -0,4472 \end{bmatrix} \begin{bmatrix} 4,4721 & -4,4721 \\ 2,2360 & 2,2360 \end{bmatrix}$$

$$A = U\Sigma V^T = \begin{bmatrix} 3,9998 & 0 \\ 2,9999 & -4,9997 \end{bmatrix} \approx \begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix}$$

3.2 Latent Semantic Indexing

Latent Semantic Indexing creates a result set by looking through each document if certain keywords exist or not and separates the documents which do not contain them. LSI considers the document collection as a whole to determine which other documents contain some of those same words. Then it defines documents that have many words in common to be semantically close, and ones with few words in common to be semantically distant. Although the LSI doesn't understand meanings of the words, by noticing the patterns it seems amazingly intelligent.

In addition, when you search an LSI-indexed database, the search engine returns the documents that it thinks best fit to your query by looking at the calculated similarity values of each content word. Then a returned result may contain two documents they do not share a particular keyword. Since two documents may be semantically very close even if they do not share a particular keyword. LSI does not require an exact match to return useful results.

3.2.1 LSI Working Structure

In order to implement LSI, a term by document matrix is constructed as stated in [14]. Each value a_{ij} in this matrix represents the occurrence of term i in document j .

$$A = [a_{ij}] \quad (3.4)$$

Since every word does not appear in each document, the matrix is usually sparse (populated primarily with zeros). Then, each element of that matrix is factorized as:

$$a_{ij} = L(i, j) \times G(i) \quad (3.5)$$

where $L(i, j)$ is the local weighting of term i in document j and $G(i)$ is the global weighting of term i . This factorization used to increase/decrease the importance of terms for each document.

As defined before, singular value decomposition equations is derived from the orthogonal matrix U , which contains left singular vectors, matrix V , which contains right singular vectors and diagonal matrix Σ , which contains the singular values of A . Also, latent semantic structure model is derived from the SVD components [14]. The SVD can be viewed as a technique for deriving a set of uncorrelated indexing variables or factors. The use of k -largest singular triplets is equivalent to approximating the original term-document matrix A by A_k in Equation (3.2).

Table 3.1: 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

The following figure represents the Matrix A_k , mathematically. In the figure U represents the term vector, V represents the document vector and Σ represents the singular values. The black regions in U , V and the diagonal line in Σ represent A_k from equation (3.2).

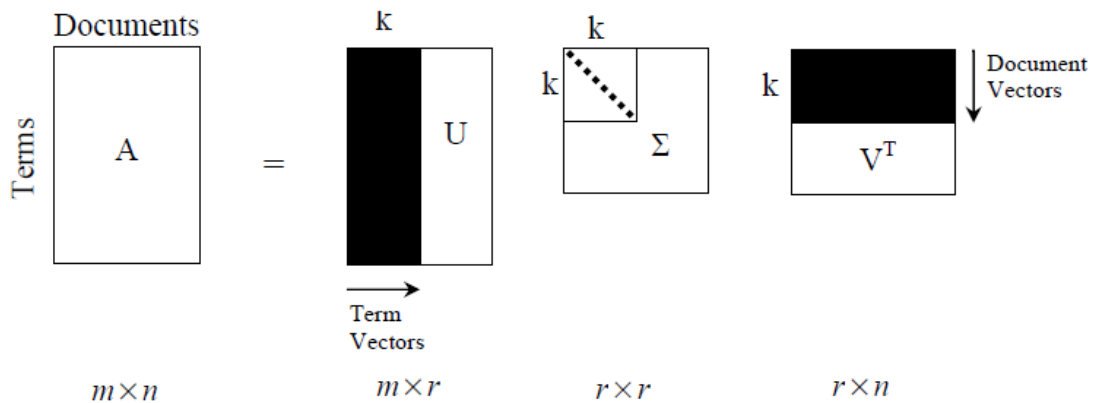


Figure 3.1: Mathematical Representation of the Matrix A_k

Applying LSI method on matrix A generates the derived A_k matrix which must be different than A . The SVD captures most of the important underlying structure in the association of terms and documents. Also it provides another profit by removing the noise and variability in word usage. Since the number of dimensions k , is much smaller than the number of unique terms m , minor differences in terminology will be ignored. Terms which occur in similar documents, will be near each other in the k -dimensional factor space even if they never co-occur in the same document. This means that some documents which do not share any words may be near in k -space. This derived representation which captures term-term associations is used for retrieval [14].

3.3 Latent Semantic Analysis

As stated in [15]:

“Latent semantic analysis (LSA) is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text.”

In 2002, Yihong Gong and Xin LUI bring the idea of using LSA in text summarization [16]. LSA is used for generic text summarization by applying SVD.

The meaning of applying the SVD to the matrix A can be explained by following two viewpoints. From semantic point of view, the latent semantic structure is derived from the SVD from the document represented by matrix A . So, a breakdown of the original document is reflected to the r linearly-independent base vectors. Since, SVD can capture and model an interrelationship among terms, terms and sentences can be clustered by SVD semantically. So, each sentence and term in the documents is indexed by these base vectors. Transformation point of view declares that, the SVD derives a mapping between the m dimensional space spanned by the weighted term-frequency vectors and the r dimensional singular vector space [16].

3.3.1 LSA Working Structure

The LSA process starts with creating a term by sentences matrix $A = [A_1 A_2 \cdots A_n]$ with each column vector A_i representing the weighted term-frequency of sentence i in the document. 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). Since every word does not normally appear in each sentence, the matrix A is usually sparse.

By applying SVD on matrix A from the equation 3.1 ($A = U \Sigma V^T$) we get $n \times n$ diagonal matrix Σ , $n \times n$ orthonormal matrix V and $m \times n$ column-orthonormal matrix U . From the [14], when $\text{rank}(A)=r$ then Σ satisfies:

$$\sigma_1 \geq \sigma_2 \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0 \quad (3.6)$$

Consider the words *machine*, *device*, *engineer*, *plan*, and *production*. The words *machine* and *device* are synonyms, and *engineer*, *plan*, and *production* are related concepts. The synonyms *machine* and *device* will occur in similar patterns holding common related words such as *engineer*, *plan*, and *production* etc. Because of these similar patterns the words *machine* and *device* will have similar representations in r -dimensional singular vector space [16]. As declared in [14], 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 [16].

3.4 Centroid-Based Summarization of Multiple Documents

3.4.1 What is Centroid

As described in [17]:

“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.”

3.4.2 Centroid-Based Summarization

In 2000, Radev, Jing and Budzikowka developed a multi-document summarizer named MEAD [16]. They used cluster centroids produced by topic detection and tracking system to create summaries. Also they described two new techniques as cluster-based relative utility (CBRU) and cross-sentence informational subsumption (CSIS) [17].

CBRU technique gives a relevant degree of a sentence to the general topic of the entire cluster between the 0 and 10. The 0 degree means; sentence is not relevant to the cluster and 10 degree means that sentence is the main sentence for the topic of the cluster.

CSIS technique compares sentences to each other. After comparison, technique decides that a sentence subsumes another sentence according to the information they have. Such, if the information of sentence A covers the information of sentence B, then B becomes redundant because of the information it has. Then, sentence A and B analysis in the same equivalent class and B should be omitted during the summarization.

3.4.3 Centroid-Based Clustering

Relative documents are grouped together into clusters by a clustering algorithm. Each sentence is represented as a weighted vector of TF.IDF. A centroid is generated by using only the first sentence in the cluster. As new sentences are

processed, their TF.IDF values are compared with the centroid. If the similarity measure is within a threshold, the new sentence is included in the cluster.

3.5 Clustering Methods

Clustering is the assignment of a set of objects into subsets (called clusters) so that objects in the same cluster are similar in some sense.

3.5.1 K-Means Clustering

The term "k-means" was first used by James MacQueen in 1967 and is used to break N terms, into k sets [18]. Each set (cluster) contains its own center point (centroid). K-means is preferred because of its simplicity and speed in large datasets.

K-means algorithm uses a number of clusters "k" settable by the user to cluster points. Then randomly k points selected as cluster centers. Once k centers selected, remaining points are assigned to nearest cluster center. After assigning all points to the clusters, cluster centers are recalculated. This algorithm is repeated recursively until no change is occurred on the selected centers.

The center point of a cluster is found by calculating the average of all points in the cluster.

3.5.2 QT (Quality Threshold) Clustering

As stated in [19] QT (quality threshold) clustering is an alternative method of partitioning data, invented for gene clustering. It requires more computing power than k-means, but does not require specifying the number of clusters a priori, and always returns the same result when run several times.

The Quality Threshold (QT) algorithm uses a maximum diameter settable by the user to cluster points. The first cluster is built with the first point in the collection. As long as other points are close enough to be within the diameter, they are added

to the cluster. Once all points are read, the points that have been added to the cluster are set aside and the algorithm repeated recursively on the rest of the point collection. The program stops when there are no more points.

The distance between a point and a cluster is computed using Complete Linkage Distance. Complete Linkage Distance is the distance from the point and the furthest point in the cluster.

3.5.3 Agglomerative Hierarchical Clustering

In agglomerative clustering, we create clusters in a bottom-up manner. That is, starting with n points in separate clusters, we repeatedly merge the closest pair of clusters until all points are members of the same cluster.

The agglomerative clustering algorithm uses a number of clusters “ k ” settable by the user to cluster points. Initially, all points are assigned to a separate cluster. In each iteration, the two closest pair of clusters C_1 and C_2 are merged into new cluster. Then the old clusters C_1 and C_2 are removed from cluster set and newly created cluster added to set. The algorithm repeated recursively until number of clusters in the set reaches to the “ k ”.

The main step in the algorithm is how to define the closest pair of clusters. Several distance measures, such as single link, complete link, average link, and so on, can be used to compute the distance between any two clusters. We used complete link distance to find the closest pair.

3.6 Cosine Similarity

Cosine similarity calculates the similarity between two vectors with n dimension, by evaluating the cosine of the angle between them.

Consider A and B are two vectors with n dimension, which are generally TF.IDF vectors of the documents, cosine similarity for that vectors represented by:

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} \quad (3.7)$$

When two vectors are

- same, the angle between them is 0° and their similarity is 1.
- opposite, the angle between them is 180° and their similarity is -1.

So similarity between two vectors ranges from -1 to 1. Such, -1 means exactly opposite, 1 means exactly same, 0 means independence and in-between values indicating intermediate similarity or dissimilarity. [20]. But in information retrieval, the cosine similarity of two documents will range from 0 to 1, since the term frequencies (TF.IDF weights) cannot be negative. The angle between two term frequency vectors cannot be greater than 90° .

3.7 TF.IDF Weighting

TF.IDF (Term Frequency - Inverse Document Frequency) is an often used weighting scheme in information retrieval [21]. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. This importance is higher when a word occurs many times in smaller number of documents, lower when a word occurs occasionally in a document or occurs in many documents and lowest when word occurs rarely in all documents.

Term Frequency (TF) is defined as number of occurrence of a term in a document or document set, obtained from the ratio of number of occurrences of the term t in document d ($n_{t,d}$) to the sum of number of occurrences of all term in document d ($\sum_{i=1}^k n_{i,d}$).

$$tf_{d,t} = \frac{n_{t,d}}{\sum_{i=1}^k n_{i,d}} \quad (3.8)$$

Inverse Document Frequency (IDF) is a measure of the general importance of the term, obtained by taking the logarithm of the ratio of the number of all documents (D) to the number of documents containing the term (d_{ft}). From this formula, the

importance of a term is inversely proportional with document number the term occurs in a corpus.

$$idf_t = \log\left(\frac{D}{df_t}\right) \quad (3.9)$$

Then TF.IDF of term t in document d is defined as:

$$W_{d,t} = tf_{d,t} \cdot idf_t \quad (3.10)$$

CHAPTER 4

LANGUAGE INDEPENDENT MULTI DOCUMENT SUMMARIZATION USING LATENT SEMANTIC INDEXING/CLUSTERING TECHNIQUES

4.1 Roadmap

Summarization process in our method consists of two main steps. By these two steps we aim to eliminate non-important sentences at the beginning. Figure 4.1 shows the main steps of our summarization method.

At first step, we aim to extract key-terms from document corpus. Firstly, terms are extracted from documents. Then these terms are weighted with TF.IDF weighting. At the end of first step Latent Semantic Analysis (LSA) is used to extract key-terms.

At second step, we aim to find sentences to create summary. Firstly, we found sentences holding these key-terms based on our assumption that sentences containing key-terms are more important than the others. Then a sentence key-term matrix is created with TF.IDF values, and Latent Semantic Indexing (LSI) is applied on that matrix. Finally a centroid-based clustering algorithm is applied to find sentences most similar to centroid, and a summary is created from these sentences.

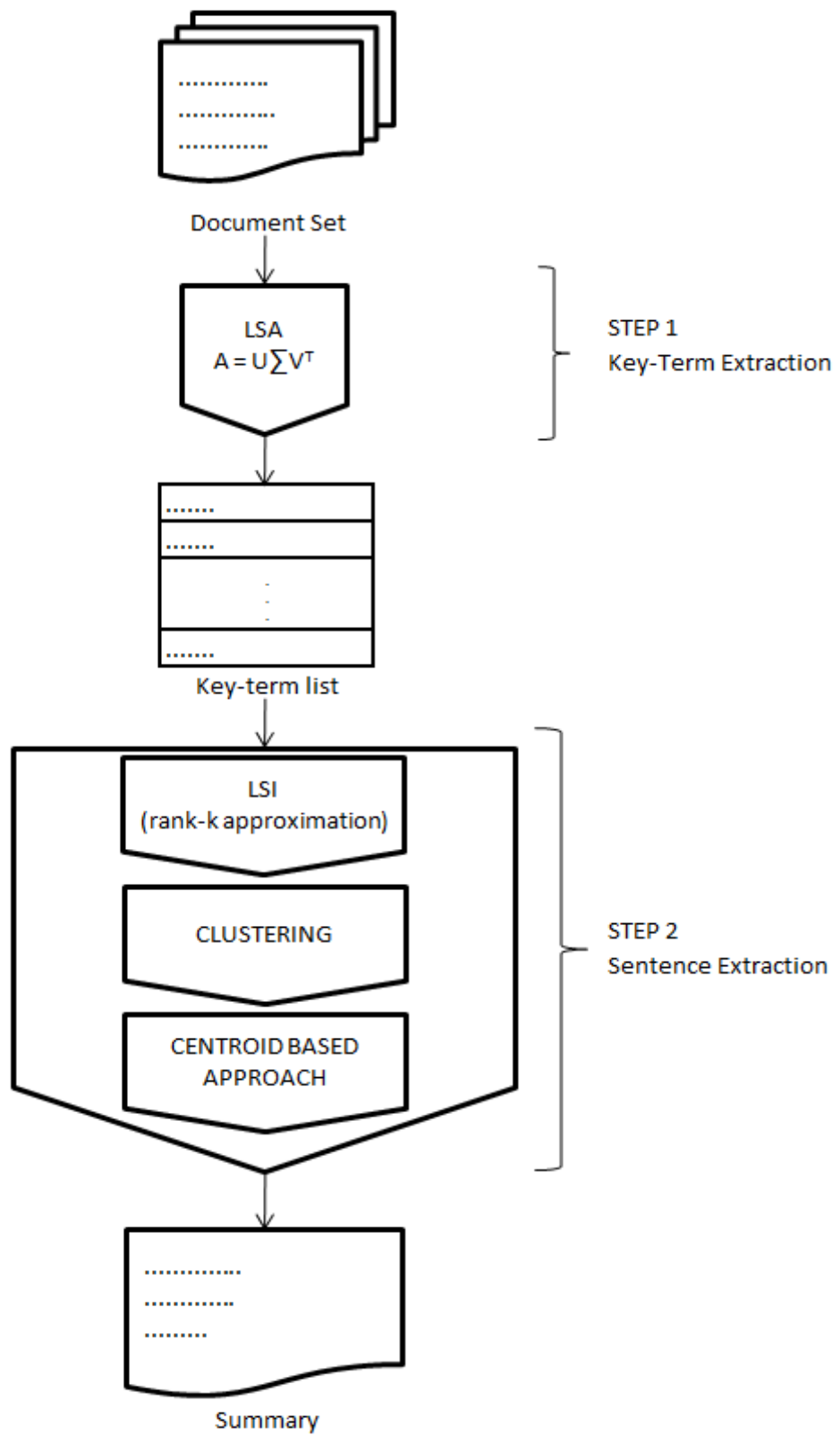


Figure 4.1: Roadmap

4.1.1 Step 1

Firstly, we read content of documents consecutively. While reading documents, we extract sentences by using sentence detector. Then stop-words elimination process is applied on these sentences to eliminate non-important words. Finally we apply a stemming algorithm to extract term list consisting of all root words.

After obtaining term list, Term Frequencies (**TF**) of each term for each document set are calculated. Also, we prepared Inverse Document Frequencies (**IDF**) using the whole document corpus. Then Term Frequency - Inverse Document Frequency (**TF.IDF**) values of each term for each document set are calculated multiplying Term Frequencies (TF) with Inverse Document Frequencies (IDF).

Finally, we create a sentence-term matrix consisting of rows with sentences and columns with words. Elements of this matrix are filled with TF.IDF value of corresponding word. Then we applied Latent Semantic Analysis (**LSA**) on that matrix to create key-terms list. Also we are able to extract key-terms from terms with biggest TF.IDF values. But as mentioned in [22] LSA approach is more successful than biggest TF.IDF method in key-terms extraction. Diagram of step-1 is given in Figure 4.2.

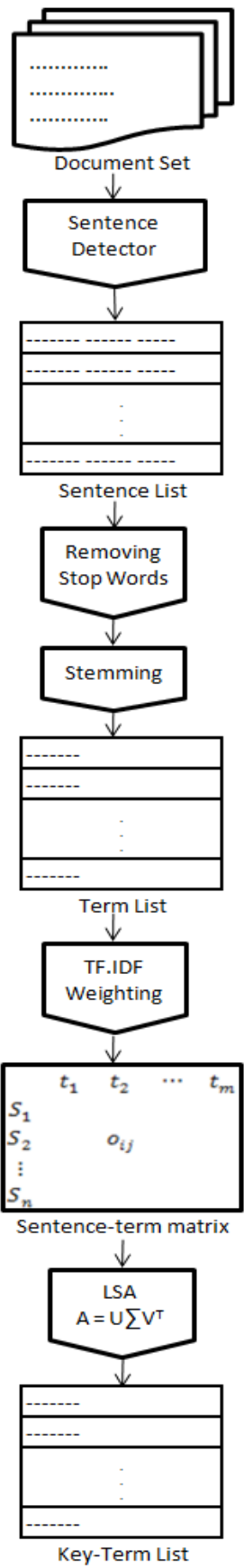


Figure 4.2: STEP 1: Key-Term Extraction

4.1.2 Step 2

At the end of the first step we have key-terms as output, and will be used as input to the second step.

Firstly, sentences holding key-terms are detected and fetched from the whole sentence set. These are candidate sentences for our summary. Then, these sentences and key-terms are used create “**key-term – candidate sentence**” matrix consisting of rows with sentences and columns with words. Elements of this matrix are filled with TF.IDF values of each key-term for each document set. Each row represents a vector of weighted key-terms of corresponding sentence.

Secondly, Latent Semantic Indexing (**LSI**) is applied to the matrix for dimension reduction. By using LSI we aim to eliminate the noise from the word usage in documents as stated in [14]. Result of LSI is a subset of initial matrix. The size of subset is defined by given parameter. Then the similarities between vectors of row elements (candidate sentences) of this reduced matrix are calculated by using **cosine similarity**. By using these similarity values a sentence-sentence matrix is created that holds similarity values of sentences to each other. Finally a clustering algorithm (**K-Means, QT (Quality Threshold) or Agglomerative Hierarchical**) is applied on that matrix to extract sentence clusters.

Finally, a sentence-term matrix is created for each sentence cluster. Elements of this sentence-term matrix are weighted with TF.IDF. On the other hand a vector consisting of calculated average weighting of each term is created. This is the **centroid** vector of terms. After that similarities between these vectors are calculated using cosine similarity. At the end, sentences most similar to the centroids are detected. These detected sentences are added to the summary until size limit reached.

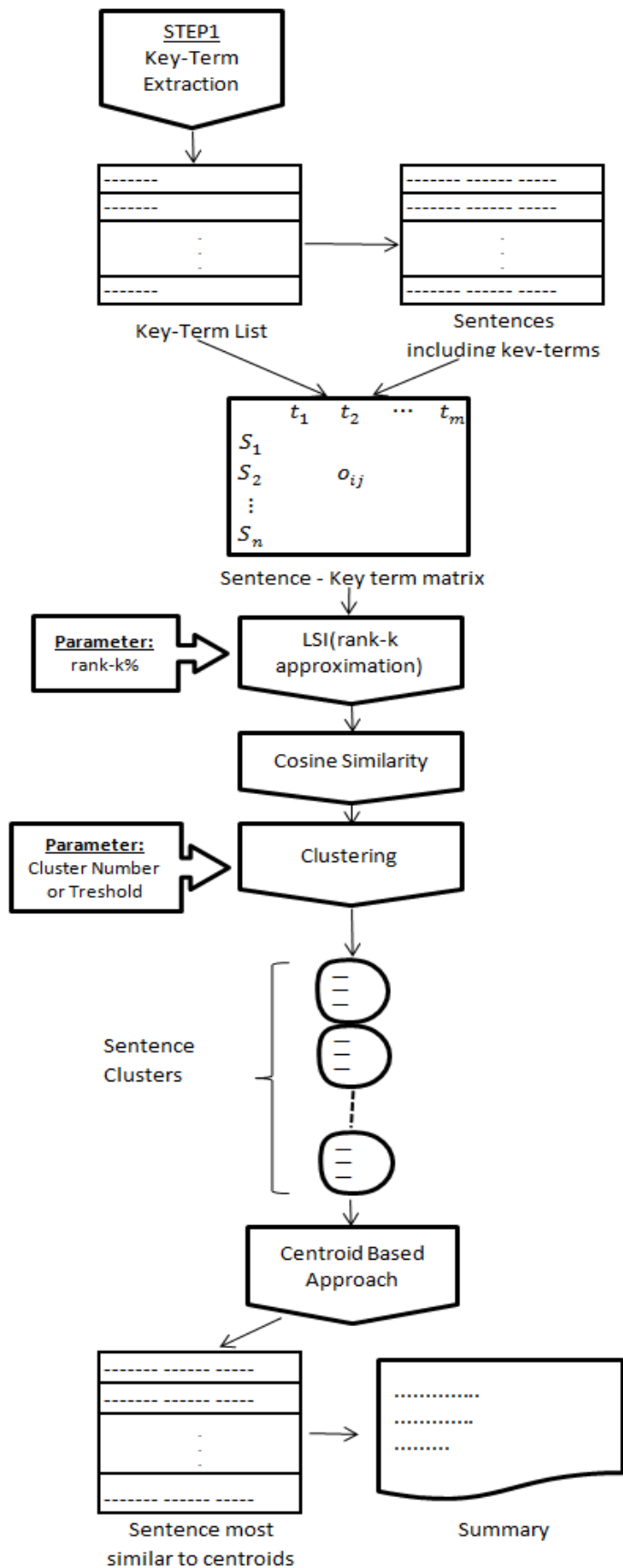


Figure 4.3: STEP 2: Sentence Extraction

4.2 Sentence Detector

Sentence detector is a mechanism used to find the boundaries of sentences in documents to extract them. Our sentence detector uses two heuristics to detect sentences [23]. First we use punctuations {., !, ?} to find sentence boundaries. But this boundary detection mechanism may work wrong when it encounters abbreviations. For example ‘Mr. John has married.’ can be extracted as two separate sentences as ‘Mr.’ and ‘John has married. To overcome this problem we use length of sentence as second heuristic to detect boundaries. If the number of letters in a sentence is less than a threshold value, first heuristic is ignored and sentence boundary is detected. Our threshold value is six letters per sentence.

4.3 Removing Stop Words

It is possible to have tens of thousands of different words occurring in a small set of documents. Many of them are not important, and their usage can reduce the performance. As stated in [28] the most frequent words are often the words with little meaning. Removing stop words may reduce the size of the documents which can not be ignored. On the other hand, not removing stop words may reduce the effectiveness of weighting scheme since we used TF.IDF weighting.

In our approach all occurrences of words that are considered to be useless are removed. Since there is no common list of stop words that is universally used and vary from language to language, our stop word lists in our summarization process used for English and Turkish languages are given in Appendix A.

4.4 Stemming

Words in documents have many morphological variants having similar semantic representations. They can be considered as equivalent in summarization operations. Because of this situation a number of stemmers have been developed. First paper about the stemmers was published in 1968 [24]. A very widely used English stemmer was written by Martin Porter and published in the July 1980 [25]. Stemmers are used to reduce the words to their stems or root forms. The stems do

not have to be the morphological roots of the words. It is sufficient to map semantically similar words to the same stem, even if the stem is not a valid root. For example, the words "stemmer", "stemming", and "stemmed" are considered as being from the same root and after stemming they will be considered as the same word.

4.4.1 English Stemmer

We have used Porter Stemmer for our stemming operation in summarization of English documents. This stemmer was released by Martin Porter around the year 2000 [26].

4.4.2 Turkish Stemmer

We have used Zemberek for our stemming operation in summarization of Turkish documents. Zemberek is an open source project intends to provide library and applications for solving Turkish Natural Language Processing (NLP) related computational problems [27]. Zemberek is one of the very interesting projects oriented around the Turkish language.

4.5 Extracting Key-Terms using Latent Semantic Analysis

Based on Latent Semantic Analysis we focus on the patterns of sentence combinations in multi-documents. A sentence pattern is represented by one of the singular vectors when it is salient and recurring in documents. Words appearing in this pattern will be projected along this singular vector. The word that best represents this pattern will have the largest index value with this vector. The importance degree of this pattern within the document is represented by the magnitude of the corresponding singular value. Since each pattern describes a certain topic in the documents, we can say that each singular vector represents a salient topic in the document. Then the magnitude of its corresponding singular value also represents the degree of importance of the salient topic.

Based on our hypothesis we suggest 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 $\varphi_j = [u_{1j} u_{2j} \cdots u_{rj}]$ 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.

The two disadvantages declared for LSA in [29] are valid for our method too. First, it is possible to use the same number of dimensions as the number of sentences in a summary. However, increasing the number of dimensions causes taking the less significant topics into a summary. Second, a sentence with large index values, but not the largest (it doesn't win in any dimension), will not be chosen although its content is possibly suitable for the summary.

4.6 LSI (Rank-k Approximation)

LSI is used to eliminate the noise of word usage in documents. Thus the sentence – term matrix is approximated to rank-k as stated in chapter 3.2. Rank-k is defined as multiplication of the column number and **rank-k percentage (k%)** which is given as a parameter. Suppose that we have a sentence-term matrix A with dimensions $n \times m$, **rank-k (k)** is found by the formula $k = \text{rank}(A) * k\%$. By using approximation percentage we aim 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.

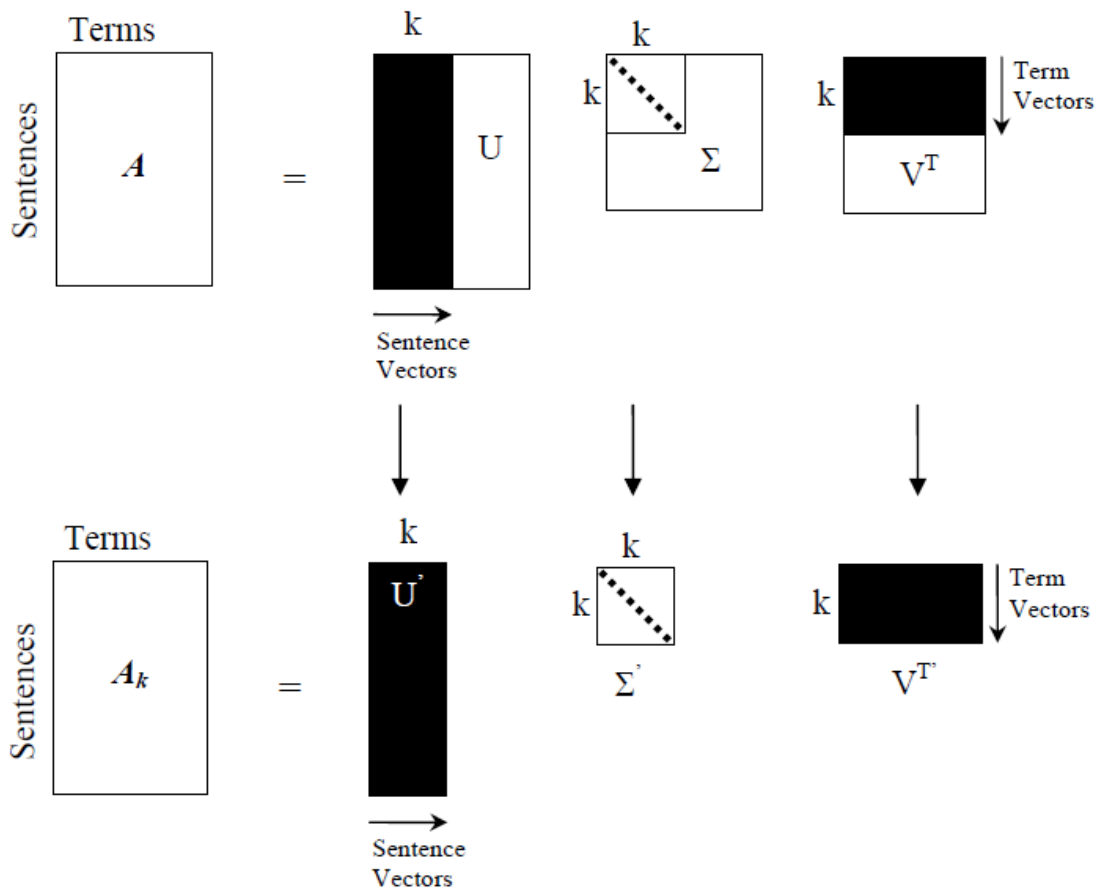


Figure 4.4: Rank-k Approximation

4.7 Clustering

After rank-k approximation, sentence-term matrix is divided into clusters using different clustering algorithms. We use K-Means, QT (Quality Threshold) and Agglomerative Hierarchical Clustering algorithms.

K-means clustering has two main problems effect the result. First, the result is changed according to cluster number which should be predefined. So that ordered sets of cluster numbers are tried in an appropriate range intuitively. Second, the result changes according to selected initial center sentence vectors. To get consistent results initial sentence vectors are selected as far as possible from each others. Since we use inverse cosine similarity as a distance metric among the vectors, words sentences less similar are further and vice versa.

QT clustering has a problem of defining the threshold value. Since our distance metric is cosine similarity, threshold value can be between 0 and 1.

In agglomerative hierarchical clustering the result is changed according to cluster number, so it is predefined like in K-Means. So that ordered sets of cluster numbers are tried in an appropriate range intuitively.

4.8 Sentence Extraction using Centroid-Based Approach

After sentences are partitioned into clusters, for each cluster a sentence-term matrix is created weighted with TF.IDF.

	t_1	t_2	\dots	t_m
S_1				
S_2		o_{ij}		
\vdots				
S_n				

Figure 4.5: Sentence-Term Matrix in a Cluster

Where; s indicates sentences, t indicates terms, n is the number of sentences in the cluster, m is the number of terms in the cluster and o_{ij} is the TF.IDF value of j 'th term in i 'th sentence.

Then frequency of a term (average number of occurrences across the entire cluster) is calculated by dividing the total occurrence number by total sentence number. Then average TF.IDF value of each term in each cluster is found by multiplying the IDF value of the term and frequency of a term.

$$C_j = \frac{1}{n} \sum_{i=1}^n o_{ij} \quad (4.1)$$

Then a vector called **centroid sentence vector** of average TF.IDF values of all terms in the cluster is created.

$$S_{centroid} = [C_1 C_2 \cdots C_m] \quad (4.2)$$

After creating centroid vectors, cosine similarity of each sentence in the cluster is calculated. Then sentences are sorted according to their similarity to the centroid vector descending. By sorting 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. Finally, clusters are sorted according to their sentence number descending.

Sentences most similar to the centroids are fetched from clusters starting from the biggest (first) cluster. Then selected sentence 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 documents of DUC2004 as a corpus.

Since TF values depend on the working clusters, in each step TF values are calculated around their corresponding clusters. In the first (key-term extraction) step

the clusters of DUC2004 each have 10 documents are used to calculate the TF values. In the second (sentence extraction) step clusters created by clustering algorithms are used to calculate the TF values.

CHAPTER 5

EXPERIMENTS & EVALUATION

5.1 Experiments

We used DUC2004 [30] conferences as an experiment area for our summaries. Task2 of DUC2004 conference is for multi document summarization [31].

5.1.1 English Documents for Summarization

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. These summaries are restricted with the max size of 665 characters. So our summarization system also creates summaries within this restriction.

Three sample documents from DUC2004 are given below. Also key-term lists extracted from these document using both LSA and biggest TF.IDF methods given below. Finally summaries created from these key-terms using LSI and K-Means clustering are also given.

Document Name: APW19981027.0241

Honduras braced for potential catastrophe Tuesday as Hurricane Mitch roared through the northwest Caribbean, churning up high waves and intense rain that sent coastal residents scurrying for safer ground. President Carlos Flores Facusse declared a state of maximum alert and the Honduran military sent planes to pluck residents from their homes on islands near the coast. At 0900 GMT Tuesday, Mitch was 95 miles (152 kilometers) north of Honduras, near the Swan Islands. With winds near 180 mph (289 kph), and even higher gusts, it was a Category 5 monster _ the highest, most dangerous rating for a storm. The 350-mile (560-kilometer) wide hurricane was moving west at 8 mph (12 kph). "Mitch is closing in," said Monterrey Cardenas, mayor of Utila, an island 20 miles (32 kilometers) off the Honduran coast. "And God help us." Mitch posed no immediate threat to the United States, forecasters said, but was expected to remain in the northwest Caribbean for five days. The U.S. National Weather Service in Miami said Mitch could weaken somewhat, but warned it would still remain "a very dangerous hurricane capable of causing catastrophic damage." The entire coast of Honduras was under a hurricane warning and up to 15 inches (38 centimeters) of rain was forecast in mountain areas. The Honduran president closed schools and public offices on the coast Monday and ordered all air force planes and helicopters to evacuate people from the Islas de la Bahia, a string of small islands off the country's central coast. The head of the Honduran armed forces, Gen. Mario Hung Pacheco, said 5,000 soldiers were standing by to help victims of the storm, but he warned the military could not reach everyone. "For that humanitarian work, we would need more than 300 Hercules C-137 planes," he said. "Honduras doesn't have them." A hurricane warning was also in effect for the Caribbean coast of Guatemala. In Belize, a hurricane watch was in place and the government also closed schools and sent workers home early Monday. Panic buying stripped bread from the shelves of some stores and some gasoline stations ran dry. Coastal Belize City was hit so hard by Hurricane Hattie in 1961 that the country built a new capital inland at Belmopan. Mexico mobilized troops and emergency workers Monday on the east coast of the Yucatan peninsula, which was also under a hurricane watch, and Cuba said it had evacuated 600 vacationers from the Island of Youth. Jerry Jarrell, the weather center director, said Mitch was the strongest hurricane to strike the Caribbean since 1988, when Gilbert killed more than 300 people. In La Ceiba, on Honduras' northern coast, people stood in long lines at filling stations Monday to buy gasoline under a steady rain. Maria Gonzalez said she needed the gas to cook with when her firewood gets wet. Still, she bought only 37 cents worth _ all she could afford. "I have six children, and we live in a riverbed," she said. "If it gets really bad, we'll go to the church and see what the architect of the world has in store for us." Swinwick Jackson, a fisherman on Utila, had tied up his boats and was taking his family to stay with a relative on higher ground. National police spokesman Ivan Mejia said the Coco, Segovia and Cruta rivers all overflowed their banks Monday along Honduras' eastern coast. "Frightened people are moving into the mountains to search for shelter," he said. In El Progreso, 100 miles (160 kilometers) north of the Honduran capital of Tegucigalpa, the army evacuated more than 5,000 people who live in low-lying banana plantations along the Ulua River, said Nolly Soliman, a resident. Before bearing down on Honduras, Mitch swept past Jamaica and the Cayman Islands. Rain squalls flooded streets in the Jamaican capital, Kingston, and government offices and schools closed in the Caymans, a British colony of 28,000 people. The strongest hurricane to hit Honduras in recent memory was Fifi in 1974, which ravaged Honduras' Caribbean coast, killing at least 2,000 people.

Figure 5.1: Sample Document for English 1

Document Name: APW19981029.0570

Hurricane Mitch cut through the Honduran coast like a rip saw Thursday, its devastating winds whirling for a third day through resort islands and mainland communities. At least 32 people were killed and widespread flooding prompted more than 150,000 to seek higher ground. Mitch, once among the century's most powerful hurricanes, weakened today as it blasted this Central American nation, bringing downpours that flooded at least 50 rivers. It also kicked up huge waves that pounded seaside communities. The storm's power was easing and by 1200 GMT, it had sustained winds of 80 mph (130 kph), down from 100 mph (160 kph) around midnight and well below its 180 mph (290 kph) peak of early Tuesday. After remaining virtually stationary for more than a day, the U.S. National Hurricane Center said Thursday the center of the 350-mile-wide (560-kilometer-wide) storm had moved slightly to the south but remained just off the Honduran coast. Hurricane-force winds whirled up to 30 miles (50 kilometers) from the center, with rain-laden tropical storm winds extending well beyond that. Caught near the heart of the storm were the Bay Islands, about 25 miles (40 kilometers) off Honduras' coast and popular with divers and beachcombers. "The hurricane has destroyed almost everything," said Mike Brown, a resident of Guanaja Island, 20 miles (32 kilometers) off the coast. "Few houses have remained standing." Honduran officials said 14 people had died on that small island alone, and at least nine had died elsewhere in the country. More than 72,000 people had been evacuated to shelters. Nine other deaths had been reported elsewhere in the region by early Thursday _ more than a day after Mitch drifted to just off the coast and seemed to park there. An American was thrown from his boat south of Cancun, Mexico, on Monday and was presumed dead. Eight others died in Nicaragua in flooding. Honduran officials said more than 200 towns and villages had been isolated by the storm, left without power, telephones or clean drinking water. Agriculture Minister Pedro Arturo Sevilla said crucial grain, citrus and banana crops had been damaged "and the economic future of Honduras is uncertain." Rain-swollen rivers knocked out bridges and roads, isolating La Ceiba, a coastal city of 40,000 people located 80 miles (128 kilometers) from the Bay Islands. About 10,000 residents fled to crowded shelters in schools, churches and firehouses. While supplies of food and gasoline seemed to hold up, drivers worried about the coming days formed long lines to fill their tanks at gas stations and some supermarkets took measures to limit panic buying. La Ceiba officials appealed for pure water for those in shelters and some residents set out plastic buckets to collect rainwater. Only a few hotels and offices with their own generators had electricity. Wind-whipped waves almost buried some houses near the shore. People evacuated low-lying houses by wading through chest-deep water with sodden bags of belongings on their heads. In neighboring Belize, most of the 75,000 residents of coastal Belize City had left by Wednesday, turning the country's largest city into a ghost town.

Figure 5.2: Sample Document for English 2 – Part 1

Document Name: APW19981029.0570 (cont.)

Police and soldiers patrolled the streets, and a few people wandered amid the boarded-up houses. The cable television company was broadcasting only The Weather Channel. With the storm seemingly anchored off Honduras, officials in Mexico to the north eased emergency measures on the Caribbean coast of the Yucatan Peninsula, where hundreds of people remained in shelters as a precaution Wednesday night. More than 20,000 tourists had abandoned Cancun and nearby resort areas, leaving hotels at about 20 percent of capacity. Houston accountant Kathy Montgomery said that she and her friend Nina Devries had tried to leave Cancun but found all the flights full. "It's been horrible," said Montgomery, as she and her friend drank cocktails at an outdoor restaurant. "We couldn't go out on a boat, we couldn't go snorkeling. Even Carlos' N Charlie's and Senor Frog's are closed," she said dejectedly, referring to two restaurants. "Some vacation." The U.S. Agency for International Development sent two helicopters each to Belize and Honduras to help in search, rescue and relief efforts. At its peak, Mitch was the fourth-strongest Caribbean hurricane in this century, behind Gilbert in 1988, Allen in 1980 and the Labor Day hurricane of 1935.

Figure 5.3: Sample Document for English 2 - Part 2

Document Name: APW19981106.0869

Aid workers struggled Friday to reach survivors of Hurricane Mitch, who are in danger of dying from starvation and disease in the wake of the storm that officials estimate killed more than 10,000 people. Foreign aid and pledges of assistance poured into Central America, but damage to roads and bridges reduced the amount of supplies reaching hundreds of isolated communities to a trickle: only as much as could be dropped from a helicopter, when the aircraft can get through. In the Aguan River Valley in northern Honduras, floodwaters have receded, leaving a carpet of mud over hundreds of acres (hectares). In many nearby villages, residents have gone days without potable water or food. A 7-month-old baby died in the village of Olvido after three days without food. Residents feared more children would die. "The worst thing, the saddest thing, are the children. The children are suffering, even dying," said the Rev. Cecilio Escobar Gallindo, the parish priest. A score of cargo aircraft landed Thursday at the normally quiet Toncontin airport in the Honduran capital of Tegucigalpa, delivering aid from Mexico, the United States, Japan and Argentina. Former U.S. President Jimmy Carter and his wife, Rosalynn, intended to visit Nicaragua on Friday to learn more about the hurricane's impact, The Carter Center in Atlanta announced. "We hope this visit will help call attention to the suffering and humanitarian need this disaster has created," Carter said in a statement. U.S. President Bill Clinton requested a "global relief effort" to help Central America and boosted U.S. emergency aid to dlr 70 million. Clinton is dispatching a delegation next week led by Tipper Gore, wife of Vice President Al Gore, to deliver some of the supplies destined for Honduras, Nicaragua, El Salvador and Guatemala. First lady Hillary Rodham Clinton added Nicaragua and Honduras to a trip she plans to the region beginning Nov. 16. Taiwan said today it will donate dlr 2.6 million in relief to Honduras, Nicaragua, El Salvador and Guatemala. The four countries are among a dwindling number of nations that recognize Taiwan, which China claims is a breakaway province. Two British ships that were in the area on an exercise were on their way to Honduras to join relief efforts, the Defense Ministry said Friday. "It's a coincidence that the ships are there but they've got men and equipment that can be put to work in an organized way," said International Development Secretary Clare Short. Nicaragua said Friday it will accept Cuba's offer to send doctors as long as the communist nation flies them in on its own helicopters and with their own supplies. Nicaraguan leaders previously had refused Cuba's offer of medical help, saying it did not have the means to transport or support the doctors. Nicaragua's leftist Sandinistas, who maintained close relations with Fidel Castro during their 1979-90 rule, had criticized the refusal by President Arnaldo Aleman's administration.

Figure 5.4: Sample Document for English 3

hurricane, mitch, honduran, coast, caribbean, kph, honduras', guatemala, food, shelter, tegucigalpa, rain, wind, beliz, 000, mexico, flood, helicopter, kilomet, children, island, river, resid, northwest, destroy, buri, bridge, region, whirl, gore, di, urbizo, search, island, coastal, km, dlr, gilbert, mitch', america, swept, gasoline, boat, devast, estim, aircraft, ravag, death, flore, ship, montgomeri, north, medicin, humanitarian, yucatan, bolano, vaccin, 7, hous, cacer, banana, restaur, weather, cayman, clinton, taiwan, disast, aleman', bai, affect, outbreak, mario, dead, dy, 6, we'll, 32, maximum, firewood, cecilio, help, bread, cuba', 8, clare, rescu, higher, celaya, dog, snorkel, villag, 1935, 194, been", provinc, ly, seasid, overwhelm, downriv, channel, audienc, cosmonaut, row, donat, mountain, afford, potabl, saddest, wander, riverb, reliabl, vari, drift, die, hotel, 480, ladi, visit, brussel, gen, 076, vacat, will", basic

Figure 5.5: Sample Key-Terms Extracted Using LSA for English

" Mitch posed no immediate threat to the United States, forecasters said, but was expected to remain in the northwest Caribbean for five days.

" A hurricane warning was also in effect for the Caribbean coast of Guatemala. Honduras braced for potential catastrophe Tuesday as Hurricane Mitch roared through the northwest Caribbean, churning up high waves and intense rain that sent coastal residents scurrying for safer ground.

Hurricane Mitch cut through the Honduran coast like a rip saw Thursday, its devastating winds whirling for a third day through resort islands and mainland communities.

The head of the Honduran armed forces, Gen.

Figure 5.6: Sample Summary Using Key-Terms from LSA for English

hurricane, honduras, mitch, honduran, nicaragua, caribbean, coast, guatemala, kph, honduras', mph, storm, shelter, tegucigalpa, beliz, el, food, wind, salvador, relief, flood, mexico, rain, coastal, km, helicopt, island, whirl, cancan, urbizo, storm', ceiba, evacu, mile, kilomet, river, buri, island, bridg, home, carter, northwest, cargo, resid, central, children, destroi, region, america, water, 000, gore, swept, 180, gasolin, search, medicin, aircraft, eu, la, nicaraguan, 560, vaccin, volcano, gilbert, mitch', cuba', utila, cacer, tipper, montgomeri, bolano, ecu, cayman, flore, yucatan, homeless, guanaja, 290, 231, ravag, boat, tropic, 32, di, humanitarian, aid, bai, tourist, devast, dlr, mexican, 350, banana, estim, suppli, strongest, isol, damag, plane, weather, emerg, equip, taiwan, restaur, rodham, pope, dy, 130, kill, disast, north, dead, 5, 600, ship, row, hillari, peak, stricken, outbreak, avalanch, clinton, death, doctor, affect, panic, peninsula, infect, mike, maximum, ly, load, rescu, higher, 7, 6, 8, commiss, toll, paul, mario, catastroph, forecast, 160, grave, donat, highwai, neighborhood, dog, mayor, area, weaken, carlo, we'll, overwhelm, drift, ton, squall, pacheco, rip saw, charlie', atlantida, nicaragua', dejectedli, carlos', devri, coco, riverb, nahum, 580, ulua, overflow, blackhawk, olvido, will", diver, saddest, sevilla

Figure 5.7: Sample Key-Terms Extracted Using Biggest TF.IDF Method for English

Honduras braced for potential catastrophe Tuesday as Hurricane Mitch roared through the northwest Caribbean, churning up high waves and intense rain that sent coastal residents scurrying for safer ground. President Carlos Flores Facusse declared a state of maximum alert and the Honduran military sent planes to pluck residents from their homes on islands near the coast. At 0900 GMT Tuesday, Mitch was 95 miles (152 kilometers) north of Honduras, near the Swan Islands. With winds near 180 mph (289 kph), and even higher gusts, it was a Category 5 monster _ the highest, most dangerous rating for a storm. The head of the Honduran armed forces, Gen.

Figure 5.8: Sample Summary Using Key-Terms from Biggest TF.IDF for English

5.1.2 Turkish Documents for Summarization

Since there is no Turkish document set available in DUC2004, we aimed to create a DUC2004 like document set for Turkish.

Firstly, we created a small news gathering system that uses RSS feeds of news portals and stored them in a database. Then we aimed to group news with the same topic to create document clusters. After creating a cluster, summaries for that cluster have been created by willing people with a size restriction of not exceeding 665 characters.

Two sets of 3documents generated as Turkish corpus are given below. Also key-term lists created by LSA method for both sets are given below. Finally two summaries created from both key-term lists using LSI and K-Means clustering are also given.

Document Name: TD01.0001

ABD'de ocak ayında göreve gelen Obama yönetiminin, eski Başkan George Bush döneminde terör zanlılarının işkence ve kötü muamele gördüğü iddialarının soruşturulması talimatı tartışma yarattı.

Bu çerçevede Adalet Bakanı Eric Holder, savcı John Durham'ı, ABD Merkezi Haberalma Örgütü (CIA) elemanlarının zanlılara kötü muamele ve işkence yaparak yasaları ihlal edip etmediklerini incelemek üzere görevlendirdi. New York Times gazetesi, CIA çalışanlarının işkence yaptığı iddiaları konusunda Adalet Bakanı Holder'in soruşturma için savcı görevlendirme kararının Kongre'de tartışmalara yol açtığını belirtti.

Demokrat Parti Oregon Senatörü ve Senato İstihbarat Komitesi üyesi Ron Wyden, sorgucuların odak noktasına alınmasını eleştirirken, Cumhuriyetçi Parti'den Peter Hoekstra da soruşturmanın terörle mücadele operasyonlarını riske attığı uyarısında bulundu.

Newsweek dergisi de CIA çalışanlarının incelenmesi için Durham'ın atanmasının, istihbarat alanında kızgınlığa yol açtığını, özellikle Cumhuriyetçi kanadın öfkelenildiğini aktardı.

Haberde, adı açıklanmayan bir üst düzey yetkilinin, CIA çalışanlarının kötü muamelede bulunduğu iddialarına ilişkin bir düzine davadan daha azının gözden geçirileceği iddiasına da yer verildi.

Adalet Bakanlığı yetkilileri de Durham'ın yapacağı teftişin herhangi bir kovuşturmayla sonuçlanmayabileceğine dikkati çekti. Yetkililer, bu iddiaların tamamının daha önce özel bir Adalet Bakanlığı birimince değerlendirildiğini, tanık ve kanıt olmadığı için kovuşturmaya gerek olmadığı kararına varıldığını hatırlatıyor.

Merkezi New York'ta bulunan İnsan Hakları İzleme Merkezi'nden Tom Malinowski ise bu sürecin CIA ve Adalet Bakanlığı'ndaki kişilere dokunulmadan, Adalet Bakanlığı'nın kırmızı çizgilerini geçen birkaç alt rütbelinin kovuşturulmasıyla sona ermesi halinde durumun daha kötü olacağına işaret ediyor.

Adalet Bakanlığı, Holder'ın bakanlığın etik dairesinin tavsiyesiyle Durham'ı atmasına paralel olarak CIA'in işkence ve kötü muameleyi içeren ayrıntılı sorgulama tekniklerini kapsayan 5 yıllık bir raporunu da yayımladı.

CIA'in "Terörle Mücadele ve Sorgulama Faaliyetleri Eylül 2001-Ekim 2003" başlıklı ve Mayıs 2004 tarihini taşıyan 109 sayfalık çok gizli raporunun, konu başlıkları dahil önemli bir kısmının karartılarak ağır bir sansürle yayımlandığı görülüyor.

New York Times'in haberine göre, sansüre rağmen, raporda CIA'in deniz aşırı ülkelerdeki cezaevlerinde uygulanan bir dizi kötü muameleye ilişkin ayrıntılar yer alıyor. Bunlar arasında gözaltındaki bir zanlının aile üyelerine cinsel saldırı tehdidi, sahte infaz, silah ya da matkapla korkutma, zanlıyı kusturuncaya kadar yüzüne sigara dumanı üfleme gibi unsurlar bulunuyor.

ABD'nin eski başkan yardımcısı Dick Cheney, CIA hakkında soruşturma başlatılması kararının, Amerikalıların Obama yönetiminin ülkenin güvenliğini koruma konusundaki şüphelerini bir kez daha hatırlattığını öne sürdü.

Figure 5.9: Sample Document 1 for Turkish Set 1

Document Name: TD02.0001

Obama, CIA işkencelerine dair yeni bilgiler yüzünden soruşturma için düğmeye basmak zorunda kaldı. Adalet Bakanı Holder, işkence soruşturmasının başına savcı Durham'ı atadı. Cumhuriyetçiler köpürürken, işkenceli sorguların mimarı Cheney, Obama'yı 'ABD'yi tehlikeye atmakla' suçladı.

ABD'de CIA'nin esirlere işkencelerine dair gizli rapor mahkeme kararıyla açıklanınca, Obama yönetimine örgütün terör zanlılarına yönelik politikasıyla ilgili soruşturma başlatmaktan başka yol kalmadı. Obama "Geleceğe bakıyoruz" diyerek selefi George W. Bush döneminin yetkilileri hakkında dava açılmayacağı güvencesi vermiş olsa da, Adalet Bakanı Eric Holder, işkence tekniklerini ortaya serip esirlerin çocuklarını öldürme ve yakınlarına tecavüz tehdidinde varan olayları sergileyen CIA başmüfettişi John Helgerson'un raporu karşısında düğmeye basmak zorunda kaldı. Holder, işkencelerle ilgili soruşturmayı yürütmek üzere savcı John Durham'ı görevlendirirken, Cumhuriyetçiler ayaklandı, Demokrat cepheden de tepkiler geldi.

Irak, Afganistan ve Pakistan gibi ülkelerden toplanıp Guantanamo'ya tıklan esirlere Beyaz Saray'ın izniyle işkence yapmış CIA ile ilgili yönetimin tavrında dramatik bir değişim yaşandı. Önce Adalet Bakanlığı'nın etik grubunun elit sorgu ekibi kurulması dahil bir dizi tavsiyesinin yer aldığı raporu açıklandı. Paralel olarak mahkeme kararıyla Helgerson'un 2004'de kaleme aldığı 109 sayfalık rapor sansürlü de olsa günyüzüne çıktı. Raporun daha önce açıklanan bölümleriyle dünya CIA'in uykusuz bırakma, zor pozisyonlarda tutma, beton zeminde bekletme, aşırı soğuk ve sıcak verme, sanal duvara çarpma, suda boğulma hissi yaratma (waterboarding), böceklerle aynı odada tutma ve aç-susuz bırakma gibi tekniklerini öğrenmişti. Raporun önceki gün yayımlanan kalan bölümleri bunlara yeni yöntemler ekledi. Rapora göre sorgu memurları, 11 Eylül sanığı Halid Şeyh Muhammed'e ABD'de yeni saldırı düzenlenmesi halinde çocuklarını öldürecekleri tehdidi savurdu. O sırada Muhammed'in oğulları Pakistan'da gözaltındaydı. Memurlar, 2000'de Yemen'de USS Cole gemisine düzenlenen saldırının zanlısı Abdülrahim Naşiri'ye bilgi vermezse annesi ve ailesini getirip gözleri önünde tecavüz etme tehdidi savurdu. Naşiri'nin başına boş silah dayayıp tetik çekildi ve vücuduna yakın mesafede matkap çalıştırıldı.

Ayrıca üç kez üst üste esirin şah damarını bayılıncaya kadar sıkma, sahte infaz, esiri kusturuncaya dek yüzüne sigara dumanı üfleme gibi taktikler uygulandı. Raporda CIA ajanları elde ettikleri bilgiler sayesinde bazı saldırıların önlendiğini anlatılıyor. Bir CIA yetkilisinin gelecekte yargılanma riskine atfen "10 yıl sonra pişman olacağız" dediğini de aktaran rapor, CIA'in 'yetkisiz, uyduruk ve insanlık dışı taktikler uyguladığı' sonucuna varıyor.

Şimdi savcı Durham, bu işe karışan CIA ajanları hakkında dava açılıp açılmamasına karar verecek. Ancak bu adım tartışma kopardı. Senato İstihbarat Komitesi'nin Demokrat üyesi Ron Wyden, sorgu memurlarının odak noktasına alınmasını eleştirirken, Cumhuriyetçi Senatör Peter Hoekstra terörle mücadele operasyonlarının riske atıldığını savundu.

Figure 5.10: Part 1 of Sample Document 2 for Turkish Set 1

Document Name: TD02.0001 (cont.)

Tartışmalı politikanın mimarı eski Başkan Yardımcısı Dick Cheney ise Weekly Standard dergisine “Yayımlanan belgeler, Kaide hakkında elde ettiğimiz bilgileri bize ileri sorgulama teknikleriyle sorgulanan kişilerin verdiğini açıkça göstermektedir.

Bu bilgiler hayatlar kurtardı ve terörist saldırıları önledi. Bu kişiler, 2002’den beri ele geçirilen Kaide üyelerinin neredeyse tamamının yakalanmasında rol oynadı. Bu bilgileri elde edenler, haklarında siyasi soruşturma ya da kovuşturma yapılmasını değil minnettarlığımızı hak ediyor” dedi. Cheney, Obama’nın bundan sonra sorgu işini CIA değil FBI’da üslenmiş elit bir ekiple yürütme ve bunların denetimini Ulusal Güvenlik Konseyi’ne devretme kararını da “Bu hükümetin ulusal güvenliğimizi koruma yeteneği konusundaki şüphelerini hatırlatmaktadır” diye eleştirdi.

Newsweek dergisi de ajanlar hakkındaki incelemenin CIA içinde kızgınlığa yol açtığını yazdı. Fakat soruşturmadan netice çıkıp çıkmayacağı meçhul. Zira Adalet Bakanlığı yetkilileri, Durham’ın yapacağı incelemenin herhangi bir kovuşturmayla sonuçlanmayabileceğine dikkati çekti.

CIA Başkanı Leon Panetta çalışanlarına e-posta ile kurallara uymaları ve kötü muameleyi savunmaktan dikkatle kaçınmalarını isterken, Beyaz Saray ise Obama’nın ‘iyi niyetle işini yapanlar ve yasal rehberi takip etmiş olanların suçlama ile yüzleşmeyeceğini söylediğini’ hatırlattı.

Figure 5.11: Part 2 of Sample Document 2 for Turkish Set 1

Document Name: TD03.0001

2004 tarihli CIA müfettiş raporuyla ilgili yeni detayların açığa çıkması sadece Amerika Birleşik Devletleri kamuoyunda değil dünya çapında büyük tartışma yarattı. CIA Başkanı Leon Panetta raporda adı geçen çalışanları koruyacağını açıklarken Uluslararası Af Örgütü İrlanda Şubesi direktörü İrlanda hükümetinin bu süreçteki payının araştırılması çağrısında bulundu.

Amerika Birleşik Devletleri'nde Obama hükümeti Afganistan savaşı, Ortadoğu barış süreci, gittikçe büyümekte olan bütçe açığı ve sağlık sektörü reformlarıyla uğraşırken geçtiğimiz günlerde ortaya çıkan 2004 tarihli CIA raporu ülkede ortalığı karıştırdı. 11 Eylül sonrası süreçte terör zanlılarının sorgularında kullanılan ağırlaştırılmış tekniklerin işkence düzeyine ulaştığı yönündeki iddiaların araştırılması için Başsavcı Eric Holder, federal savcı John Durham'ı görevlendirdi.

Raporun büyük bir kısmı karartılmış olsa da piyasaya çıkan kısımlarından elde edilen bilgiler ışığında CIA görevlilerinin sorgu sırasında suyla ıslatma, fırçalama, yüzüne duman üfleme, vücuttaki hayati noktalara basınç uygulama, soğukta bırakma gibi fiziksel metotların yanı sıra tutukluların ailelerini tehdit etme, sahte infaz ve bebek bezi bağlama gibi psikolojik metotlar da uyguladığı bildirildi.

ABD yasalarına göre bir tutukluyu ölümlle tehdit etmek suç sayılıyor. Holder'ın savcı Durham'ı görevlendirmesinin CIA'e ciddi bir darbe vuracağı belirtilirken New York Times'in haberine göre Holder başka seçeneğinin olmadığını ifade etti. "Benim görevim gerçekleri araştırıp hukuku uygulamaktır" diyen Holder CIA'in işlerine engel olduğu için eleştirileceğini bildiğini fakat başka seçeneğinin olmadığını belirtti.

Holder "Ben de Başkan Obama'nın Başkan Bush'un politikalarıyla ilgili tartışmalara girmeme kararını destekliyorum ama CIA raporlarını incelediğimde sorumluklarımın bunu gerektirdiğine karar verdim," dedi.

Diğer yandan Los Angeles Times'in bildirdiğine göre CIA yetkilileri raporun detaylarıyla ilgili açıklama yapmayı reddederken, kurum sözcüsü Paul Gimigliano raporun 2004 yılından beri Adalet Bakanlığı'nın elinde olduğunu ve savcıların denetiminden geçirildiğini belirtti.

Gimigliano "CIA'in yaptıkları kesinlikle suç teşkil etmemektedir. Adalet Bakanlığı yetkilileri dosyayı gözden geçirip dava ile ilgili kararlarını zaten vermişlerdir," şeklinde konuştu.

Diğer yandan CIA Başkanı Leon Panetta kurum çalışanlarına gönderdiği bir e-mail'le raporda adı geçen çalışanları koruyacağını belirtirken, raporu "eski bir hikaye" diye nitelendirdi. Öncelikli amacının kendilerine verilen yasal çerçeve içinde ülkelere hizmet etmeye çalışan görevlileri korumak olduğunu belirten Panetta "Başkan'ın pozisyonu da bu yönde," diye ekledi.

Raporda açıklanan ağırlaştırılmış sorgu metotlarının zaten herkesçe bilindiğine dikkat çeken Panetta, "bu eski bir hikaye" yorumunu yaptı.

Figure 5.12: Part 1 of Sample Document 3 for Turkish Set 1

Document Name: TD03.0001 (cont.)

New York Üniversitesi İnsan Hakları ve Küresel Adalet Merkezi'nden araştırma sorumlusu Jayne Huckerby'a göre ise Holder araştırmanın sınırlarını yeterince genişletmiş değil. El-Cezire televizyonuna konuşan Huckerby "Başsavcı bunun deniz aşırı hapisanelerdeki tutukluların sorgusu sırasında yasaların çiğnenip çiğnenmediğini incelemek için yapılan öncül bir inceleme olacağını belirtti. Bu açıdan bakılırsa dinlenecek tanıkların ve incelenecek belgelerin çok kısıtlı olduğunu söyleyebiliriz," dedi.

Öte yandan Başsavcı'nın kararına Cumhuriyetçi Parti'den ağır eleştiriler geldi. Fox News televizyonunda Greta Van Susteren'in konuğu olan Cumhuriyetçi Senatör Pete Hoekstra Başsavcıyı kendi başına hareket etmekle suçlarken Başkan'ı liderliğini göstermeye davet etti. "Başkan uzun zamandır geriye değil önümüze bakmamız gerektiğini ifade ediyor. Buna rağmen Holder eski defterleri tekrar açmaya çalışıyor. Bu iddialar yeni şeyler değil. Ordularımız Afganistan'da savaşıyor ve işler iyi gitmiyor. Bu zaman eski defterler açma zamanı değildir," dedi.

Hoekstra, Van Susteren'in bugüne kadar açığa çıkan belgelerin hep karartılmış olduğu, kamuoyunun bu konuda hala bilgisiz olduğu yönündeki sorusuna istihbarat kurumlarının ellerindeki bütün belgelerin açığa çıkarılmasının doğru olmadığı, bunun ülkenin güvenliğine bir tehdit oluşturacağı cevabını verdi. Wall Street Journal'dan Bret Stephens da köşesinde bu konuya yer verirken hükümetin içindeki ve dışarıdaki liberallerin tutarlılıktan uzak olduğunu ve CIA operasyonlarına ihanet ettiklerini yazdı.

Konuyla ilgili bir başka açıklama da Uluslar arası Af Örgütü'nün İrlanda Şubesi'nden geldi. Şube direktörü Colm O'Gorman ortaya çıkanlardan duydukları rahatsızlığı belirtirken İrlanda hükümetinin CIA operasyonlarındaki rolünün araştırılması için çağrıda bulundu.

CIA uçaklarının İrlanda hava sahasını kullandıklarını belirten O'Gorman geçtiğimiz yıl kurulan araştırma komitesinin görevini yapmasını istedi.

Figure 5.13: Part 2 of Sample Document 3 for Turkish Set 1

cia, rapor, iskence, savci, bilgi, yen, bakani, bush, bakanligi, holder, 2004, sorusturma, cocuklarini, anne, nin, orta, orgutu, panetta, teknik, iliskin, eric, bas, ihlal, cin, calisanlarinin, dav, zanlilarinin, federal, matkap, icinde, durham, yeni, sirada, istihbarat, yil, mahk, ırlanda, cumhuriyetci, suclanan, muhammed, kaide, esir, kar, yonetiminin, acilmasi, holderin, ek, hoekstra, 2002, karartilmis, mufettis, diz, siyasi, ragmen, is, ulu, paralel, risk, istihbarat, mucadele, sek, belge, zanlilari, leon, kararın, yapacağı, si, kural, suphelilerinin, gozalti, onunde, iddialari, ciddi, hukümetin, goruluyor, bazi, yazdi, iskenceye, holdera, tartisma, bugün, suc, birkac, olumle, baslatilmasi, hak, cevadin, calisanlarına, sucsuz, zanliyi, cikti, sorusturmanın, ac, hayat, bakmamız, defter, bak, ti, cercevesinde, herkes, dunya, dogrudan, aciklanirken, yegenim, 12, sayfalik, 117, 18, 2005, abdnin, abd, 4, aciklanmayan, 2004e, ancak, arasi, arindan, basinc, ayri, an,

Figure 5.14: Sample Key-Terms Extracted Using LSA for Turkish Set 1

ABDde ocak ayında goreve gelen Obama yonetiminin, eski Baskan George Bush doneminde teror zanlilarinin iskence ve kotu muamele gordugu iddialarının sorusturulmasi talimati tartisma yaratti. Bu cercevede Adalet Bakani Eric Holder, savci John Durhami, ABD Merkezi Haberalma Orgutu (CIA) elemanlarının zanlilara kotu muamele ve iskence yaparak yasaları ihlal edip etmediklerini incelemek uzere gorevlendirdi. New York Times gazetesi, CIA calisanlarının iskence yaptigi iddialari konusunda Adalet Bakani Holderin sorusturma icin savci gorevlendirme kararının Kongrede tartismalara yol actigini belirtti. Herkese iskence yaptılar. Bu iddialar yeni seyler degil.

Figure 5.15: Sample Summary Using Key-Terms from LSA for Turkish Set 1

Document Name: TD01.0002

ABDnin Fort Hood eyaletinde 12 kisinin olumu 31 kisinin de yaralanmasi ile sonuclanan kanli saldirinin yankilari suruyor.

Amerikan medyasinda cikan haberlere gore, oldurulen saldirgan Iraka gonderilmek uzere olan Malik Hasan Nidal adli bir binbasi. Onceleri 3 saldirgan oldugu iddia edilse de sonradan olayin tek failinin 39 yasindaki orduda psikiyatrist olarak calisan Nidal Malik Hasan oldugu aciklandi. Amerikan medyasi askeri usse saldiri haberlerini, Malik Hasan Nidalin Musluman oldugunu one cikararak veriyor. Nedeni henuz bilinmeyen olayin ardindan usse giris cikislar yasaklandi. Olen 12 kisten birinin polis digerlerinin de asker oldugu belirtildi.

1942 yilinda insan edilen Fort Hood, dunyadaki en buyuk Amerikan ussu. Uste yaklasik 50 bin asker bulunuyor. Us, ayrica Iraka en cok asker gonderen Amerikan uslerinden biri olarak da taniniyor. Irakta su anda Fort Hood ussunden 15 bine yakin asker bulunuyor. Us komutani Tuggeneral Cone, saldirinin goreve gidecek askerlerin son saglik kontrollerinden gecirildigi bir merkezde meydana geldigini soyledi.

Us komutani, saglik merkezinin yakinlarindaki bir binada da olayla baglantisi oldugundan suphelenilen iki kisinin gozaltina alindigini belirtti. Cone, gorgu taniklari ifadelerine gore, olayda birden fazla ates eden kisi olabilecegine degindi. Uste, savas sonrasinda psikolojik sorunlar yasayan askerler icin bir rehabilitasyon merkezi de bulunuyor.

Fort Hood askeri ussune yapilan saldiri ile ilgili ABD Baskani Barack Obama da, Washingtonda Amerikan kamuoyuna bir aciklama yapti. Olayi korkunc bir siddet patlamasi olarak niteleyen Obama, Ulke disinda cesur erkeklerimizi, kadinlarimizi kaybetmek zaten zor. Ama askerlerimizin Amerikan topraklarindaki bir uste saldiriya ugramalari korkunc dedi. Obama, olenlerin yakinlarına bassagligi dilerken, olayin ardindaki her seyi aciga cikaracaklarini soyledi.

Barack Obama, ussun guvenligini garanti altina almak icin Beyaz Sarayin, Pentagon, FBI ve Ic Guvenlik Bakanligiyla birlikte calistigini kaydetti.

Figure 5.16: Sample Document 1 for Turkish Set 2

Document Name: TD02.0002

Amerika Birlesik Devletlerinin Teksas eyaletindeki Fort Hood Askeri Usssunde psikayatr olarak gorev yapan Nidal Malik Hasan adli bir binbasi, 13 kisiyi oldurdu.

Yegeninin inancly bir Musluman oldugunu ifade eden Hasan, Nidal 11 Eylul'den bu yana inanciyla ilgili her turlu tacize ve alaya direnmekle birlikte yillardir ordudan terhisini istemekteydi" dedi.

30 kisiyi de yaralayan saldirgan yarali olarak ele gecirildi. Daha once saldirganin oldugu aciklanmisti. Olenlerden biri polis digerleri asker. Olayin ardindan usse giris cikislar yasaklandi. Binbasi Hasanin Iraka gitmek istemedigi bildiriliyor. Subay arkadaslari, Amerikan televizyonlarindaki mulakatlarinda, Hasanin, Amerikan ordusunun Irak ve Afganistandaki operasyonlarindan rahatsizlik duydugunu aktardilar.

1942 yilinda insa edilen Fort Hood, dunyadaki en buyuk Amerikan ussu. Uste yaklasik 50 bin asker bulunuyor. Korkunc bir siddet patlamasi ABD Baskani Barack Obama, olayi korkunc bir siddet patlamasi olarak niteledi.

Washingtonda basin toplantisinde konusan Obama, Ulke disinda cesur erkeklerimizi, kadinlarimizi kaybetmek zaten zor. Ama askerlerimizin Amerikan topraklarindaki bir uste saldiriya ugramalari korkunc dedi. Obama, olenlerin yakinlarina bassagligi dilerken, olayin ardindaki herseyi aciga cikaracaklarini soyledi. Barack Obama, ussun guvenligini garanti altina almak icin Beyaz Sarayin, Pentagon, FBI ve Ic Guvenlik Bakanligiyla birlikte calistigini kaydetti.

Us komutani Tuggeneral Bob Cone, saldirinin, goreve gidecek askerlerin son saglik kontrollerinden gecirildigi bir merkezde meydana geldigini soyledi.

Conea gore, saldirgan elindeki iki silahla askerlere ates acti ve bir polis gorevlisi tarafından vuruldu. Us komutanina gore yetkililer, saldirida baska kimsenin rolu oldugunu dusunmuyor. BBC Washington muhabiri Adam Brooks, Fort Hooda Irak ve Afganistana asker gonderen birlikler bulundugunu, uste bolgeden donen askerler oldugunu soyluyor. Uste, savas sonrasinda psikolojik sorunlar yasayan askerler icin bir rehabilitasyon merkezi de bulunuyor.

Figure 5.17: Sample Document 2 for Turkish Set 2

Document Name: TD03.0002

ABDnin Teksas eyaletindeki Fort Hood Askeri Ussune duzenlenen silahlı saldirida 12 kisinin olduđu 31 kisinin de yaralandıđı bildirildi. Olenlerin 10unun ABD askeri, birinin de uste sozlesmeli olarak calisan sivil bir guvenlik gorevlisi olduđu belirtildi.

Yerel saatle 13.30 gercekleşen saldiri uste gorevli 3 asker tarafından yapıldı. Saldirinin ABDnin tarihinde kendi sinirlari icindeki bir askeri usse yapılan en buyuk saldiri olduđu aciklandı.

The Los Angeles Times gazetesinin ifadelerine yer verdiği ordu yetkilileri saldirida kullanılan silahların orduya ait mi yoksa şahsi silahlar mı olduğunu bilmediklerini kaydetti.

Normal kosullar altında ABDdeki askeri birliklerde silah tasima yetkisi sadece guvenliğin sağlanmasından sorumlu askeri polislere veriliyor. Onların dışındaki askerlerin silahları sürekli olarak kontrol altında tutuluyor ve atış talimleri ile bakım yapılması durumları dışında askerlere verilmiyor.

Şahsi silahların da us yetkililerinin denetiminde kilit altında tutulması gerekiyor. Butun bunların kayıtlarını askeri polis elinde tutuyor.

ABD ordusunda psikiyatrist olarak görev yapan binbasi Nadal Malik Hasan, ABDnin Virginia eyaletinde doğdu. 39 yaşındaki Malik Hasanın inançlı bir Müslüman olduğu ve askeri uniformasıyla Marylanddaki camiide namaz kılariken defalarca görüldüğü iddia edildi. Camii imami Faisal Han, Malik Hasanın evlenmek için kendisinden yardım istediğini söyledi. Malik Hasan camideki bir toplantıda kendisini Filistin kökenli olarak tanıttı.

Virginia Tech Üniversitesinden mezun olan Hasan, Fort Hood Darnall Ordu Tip Merkezinde görevliydi. Askeri ve mesleki kayıtlara göre Hasan daha önce de Walter Reed Askeri Tip Merkezinde çalışıyordu.

Bekar ve çocuksuz olduğu bildirilen Hasan Fort Hooda Temmuz ayında transfer edilmisti. Hasanın eski ussundeki ustlerinin performansını zayıf olarak değerlendirdikleri belirtildi.

Saldiri, askerler için tıbbi gözlemler yapılan Askeri Hazırlık Merkezinde meydana geldi. olay bir askerin elindeki tabancaları ateşlemesiyle başladı.

Saldiri anında üstteki askerlere Irak veya Afganistana gitmesi için form dolduruluyordu. Morali bozuk olduğu söylenen saldırgan da İraka gidecek askerler arasındaydı. Teksastaki uste yaklaşık 50 bin asker bulunuyor.

Saldirinin ardından ülke genelinde ve yurt dışındaki askeri uslerde en üst düzeyde güvenlik önlemleri alındı. Saldirinin meydana geldiği yerde kapılar kapatıldı.

Teksasda meydan gelen olay ABDde daha öncede benzer şekilde yaşanan saldirıları hatırlattı. Geçen Martta New Yorktaki bir göçmen bürosunda yaşanan silahlı saldirıda 13 kişi hayatını kaybederken, yine Mart ayında Alabamada ofke patlaması sonucu meydana gelen silahlı saldirıda 10 kişi yaşamını yitirmisti. Bu tip olayların en kanlı bilanço ile sonuçlanani ise 2007 yılında 32 kişinin olduğu Virjiniada yaşanmisti.

Figure 5.18: Part 1 of Sample Document 3 for Turkish Set 2

Document Name: TD03.0002 (cont.)

Diger yandan failin ailesinden de yasanan saldiriyla ilgili aciklama geldi. Hasan'in the Washington Post gazetesine konusan halasi Noel Hasan, askerin ordudan terhisini istemesine karsin cepheye gonderilmek uzere oldugunu ifade etti. Hasanin, ABDnin Irak ve Afganistanda surdurdugu savaflara muhalif oldugu belirtildi.

Yegeninin inancly bir Musluman oldugunu ifade eden Hasan, Nidal 11 Eylul'den bu yana inanciyla ilgili her turlu tacize ve alaya direnmekle birlikte yillardir ordudan terhisini istemekteydi" dedi.

Ancak ordu sozcusu George Wright, Hasan'in terhisini istedigini teyit edemeyecegini ifade etti.

The Associated Press haber ajansi Hasan'in yakin zamanda NidalHasan" nickname'i ile internette yazdigi bir yazida Islami intihar bombacilarini Japon kamikaze pilotlariyla karsilastirdigi gerekcesiyle icra makamlarinin harekete gectigini duyurdu.

Hasan'in yazisinda Bu askerin intihar ettigini soylemek uygun degildir.

Soylenmesi asil uygun olan sey onun daha soylu bir amac icin hayatini feda eden cesur bir kahraman oldugudur" dedigi belirtildi.

Figure 5.19: Part 2 of Sample Document 3 for Turkish Set 2

hasan, malik, us, meydan, fort, musluman, korkunc, olayin, saldirgan, 13, saldirida, saldirganin, iraka, patlamasi, dehset, ulke, tip, binbasi, usse, cone, 30, ussunde, disindaki, kuzen, amerikalı, 31, bakanligıyla, kokenli, yakalandı, altına, cesur, hooper, virginia, denizasiri, anında, kontrol, disında, yaralandı, psikiyatrist, gonderilmek, dunyadaki, katildigi, dil, inancli, komutani, serbest, duy, cami, gorevli, saldiridan, afganistandaki, istemedigi, istedigini, 2, rehabilitasyon, saldiriyi, sokta, teksastaki, yerel, bayrak, yansimasi, tabanca, top, degerlendirdikleri, digerlerinin, oldurulen, saldirisi, gidecegi, 45, nadal, dua, muhalif, yer, boyle, randevulari, atesle, digerleri, neden, aciklanmisti, tabancalari, cikislar, karsi, kayitlarini, texas, aciklamasini, insa, seyi, insan, trajik, basin, 2007, aciklama, bilanço, aciklamalarda, ajansi, basladigini, arkadaslariyla, bekar, bin, arkadaslarina, ayrica, bir, amerikan, ancak, bbc

Figure 5.20: Sample Key-Terms Extracted Using LSA for Turkish Set 2

ABDnin Fort Hood eyaletinde 12 kisinin olumu 31 kisinin de yaralanmasi ile sonuclanan kanli saldirinin yankilari suruyor. Amerikan medyasında cikan haberlere gore, oldurulen saldirgan Iraka gonderilmek uzere olan Malik Hasan Nidal adli bir binbasi. Onceleri 3 saldirgan oldugu iddia edilse de sonradan olayin tek failinin 39 yasındaki orduda psikiyatrist olarak calisan Nidal Malik Hasan olduğu aciklandı. Amerikan medyasi askeri usse saldiri haberlerini, Malik Hasan Nidalin Musluman oldugunu one cikararak veriyor. Nedeni henüz bilinmeyen olayin ardından usse giris cikislar yasaklandı. Irakta su anda Fort Hood ussunden 15 bine yakin asker bulunuyor.

Figure 5.21: Sample Summary Using Key-Terms from LSA for Turkish Set 2

5.2 Evaluation

Evaluation of automatic summarization in a standard and inexpensive way is a difficult task. Evaluation is as important as summarization process. Since this evaluation shows how successful is our summarization process. That's why many evaluation approaches were developed.

We have used **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) [32], [33] to evaluate our summaries. Rouge defines the quality of a summary by comparing it to other summaries created by humans. It uses different metrics to create results such as 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). Then these results are matched with other 35 systems for each scoring approaches.

The result of our summarization system varies according to some parameters.

First, **term percentage** is used to identify to number 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 to find **rank-k** value for each document cluster at matrix approximation operation during sentence extraction. 10 levels of rank-k percentages are used as input starting from 10% to 100%.

Third, depending on the used clustering technique a parameter is needed from outside. For K-Means and Agglomerative Hierarchical Clustering **cluster number** is used as parameter. From 1 to 8, eight cluster numbers are used as parameter. For QT Clustering **threshold value** is used as a parameter. From 0.1 to 1, ten threshold values are used as parameter.

The number of configurations for all combinations of the parameters above is $10 \times 10 \times (8 + 8 + 10) = 2600$. Summaries for these 2600 combinations of parameters

have been created and evaluated using ROUGE. The best 10 configurations with their scores and order among other summarization systems for each clustering methods are shown in Table 2, 3 and 4.

Table 5.1: Best ROUGE Results with K-Means Clustering

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	3	18 0.3424	21 0.06929	20 0.02205	18 0.00921	<u>14</u> <u>0.30399</u>	<u>13</u> <u>0.13625</u>
20	40	2	21 0.33826	21 0.06909	19 0.02239	19 0.00894	15 0.30239	<u>13</u> <u>0.13601</u>
10	30	3	17 0.34433	18 0.0706	18 0.02279	19 0.0091	<u>14</u> <u>0.30535</u>	<u>13</u> <u>0.13599</u>
30	20	2	21 0.33892	22 0.06845	20 0.02222	19 0.00888	15 0.30164	<u>13</u> <u>0.13592</u>
10	70	2	23 0.33724	21 0.06927	19 0.0224	19 0.00914	16 0.29989	14 0.13535
20	70	3	21 0.33919	22 0.06769	20 0.02142	19 0.0088	16 0.30131	14 0.13534
10	80	2	21 0.33871	20 0.07016	18 0.02283	19 0.00911	16 0.30081	14 0.13533
10	80	3	19 0.34021	20 0.0704	16 0.0234	17 0.0098	15 0.30206	14 0.13514
20	70	2	21 0.33853	20 0.06981	19 0.02253	18 0.00954	16 0.30096	15 0.13499
10	30	2	19 0.34117	22 0.06807	20 0.02154	22 0.00815	16 0.30065	15 0.13497

(Meanings of titles are shown in Figure 5.22)

Term %:	term percentage to be used in STEP 2
Rank-k %:	rank-k approximation percentage
Cluster No:	cluster number
Threshold Value:	QT Clustering Threshold Value
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 5.22: Meanings of Titles in Result Tables

Table 5.2: Best ROUGE Results with QT Clustering

Configuration Parameters			ROUGE Scores & Orders					
Term %	Rank-k %	Threshold Value	R1_AF	R2_AF	R3_AF	R4_AF	RL_AF	RW_12_AF
60	40	0.5	21 0.33835	23 0.06676	21 0.02139	19 0.00912	<u>14</u> 0.30359	<u>13</u> 0.13579
20	40	0.5	19 0.34183	23 0.0666	24 0.0192	24 0.00731	15 0.30224	<u>14</u> 0.13509
20	50	0.5	19 0.34049	25 0.06369	24 0.01993	20 0.00862	<u>14</u> 0.30343	<u>14</u> 0.13502
10	30	0.3	23 0.33715	25 0.0624	25 0.01874	24 0.0073	16 0.30033	15 0.13424
10	20	0.5	19 0.34138	25 0.06261	24 0.01908	25 0.00719	16 0.30119	15 0.1341
100	40	0.5	24 0.33451	23 0.06572	17 0.02291	15 0.01007	17 0.29869	15 0.13374
10	100	0.8	26 0.33216	23 0.06677	20 0.02206	15 0.00993	21 0.2946	17 0.13251
10	100	0.9	24 0.33374	22 0.06711	20 0.02213	15 0.00993	20 0.29593	17 0.133
10	30	0.4	19 0.34071	25 0.06368	24 0.01961	22 0.00806	16 0.29952	16 0.13321
40	50	0.4	21 0.33946	22 0.06885	20 0.0218	19 0.00912	16 0.29962	17 0.13283

(Meanings of titles are shown in Figure 5.22)

Table 5.3: Best ROUGE Results with Agglomerative Hierarchical Clustering

Configuration Parameters			ROUGE Scores & Orders					
Term %	Rank-k %	Cluster No	R1_AF	R2_AF	R3_AF	R4_AF	RL_AF	RW_12_AF
10	10	1	23 0.33507	23 0.06628	20 0.02164	19 0.00883	18 0.2983	<u>15</u> <u>0.13373</u>
30	100	1	23 0.3364	25 0.06361	25 0.01838	24 0.00727	19 0.29666	17 0.13268
20	30	1	24 0.33408	26 0.06099	25 0.01834	25 0.00688	20 0.29581	18 0.13206
30	50	1	23 0.33555	25 0.06236	25 0.01842	25 0.00705	20 0.29601	18 0.13205
40	60	1	24 0.33433	25 0.06235	25 0.01854	25 0.00705	20 0.29545	19 0.13174
30	80	1	24 0.33434	26 0.06099	26 0.01721	26 0.00681	21 0.29526	19 0.13171
30	20	1	24 0.33459	26 0.05965	26 0.01749	26 0.00676	19 0.29666	19 0.13167
10	30	2	27 0.3313	23 0.06558	21 0.02098	22 0.00837	21 0.29487	19 0.13155
30	10	1	25 0.33361	26 0.05993	26 0.01788	24 0.00737	21 0.29496	19 0.13153
30	70	1	25 0.3336	26 0.06036	26 0.01759	25 0.00685	21 0.29442	19 0.1315

(Meanings of titles are shown in Figure 5.22)

The best results for term percentage were generally obtained at 10% and 20%. This shows us that term percentage is useful in finding the importance of terms in documents. A graph that shows the number of obtained best results for each term-percentage is shown in Figure 5.23.

The best results for rank-k percentage were generally when less than 50%. But we reach the highest score with rank-k of 70%. So there is not a range that can be specified to limit rank-k approximation. Although, using LSI (rank-k approximation) before clustering increased our success rate. A graph that shows the number of obtained best results for each rank-k percentage is shown in Figure 5.24.

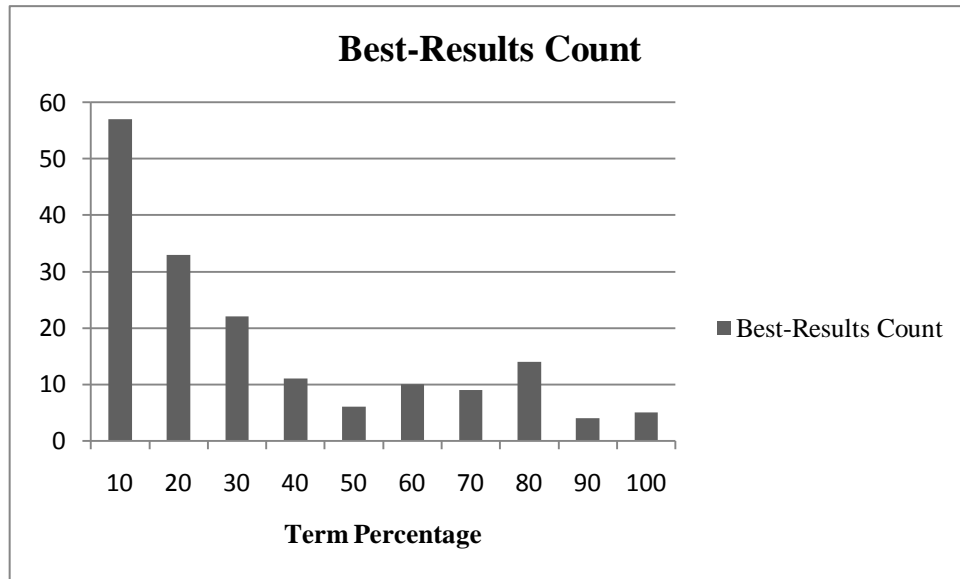


Figure 5.23: Number of Best Results for Each Term Percentage

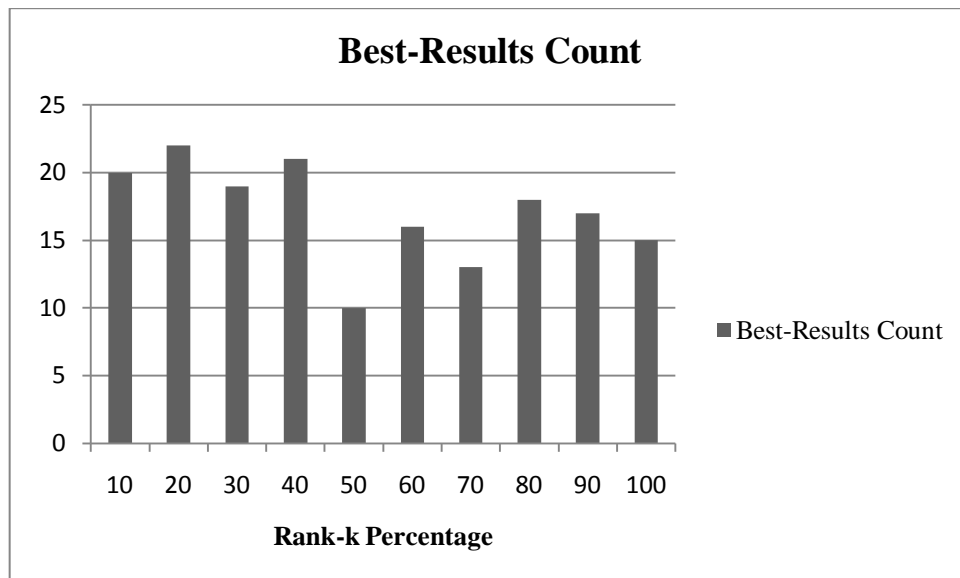


Figure 5.24: Number of Best Results for Each Rank-k Percentage

The best results for both K-Means and agglomerative hierarchical clustering algorithms were obtained with cluster numbers for 1, 2 and 3. When cluster number is greater than 3, scores dropped. A graph that shows the number of obtained best results for each cluster number is shown in Figure 5.25.

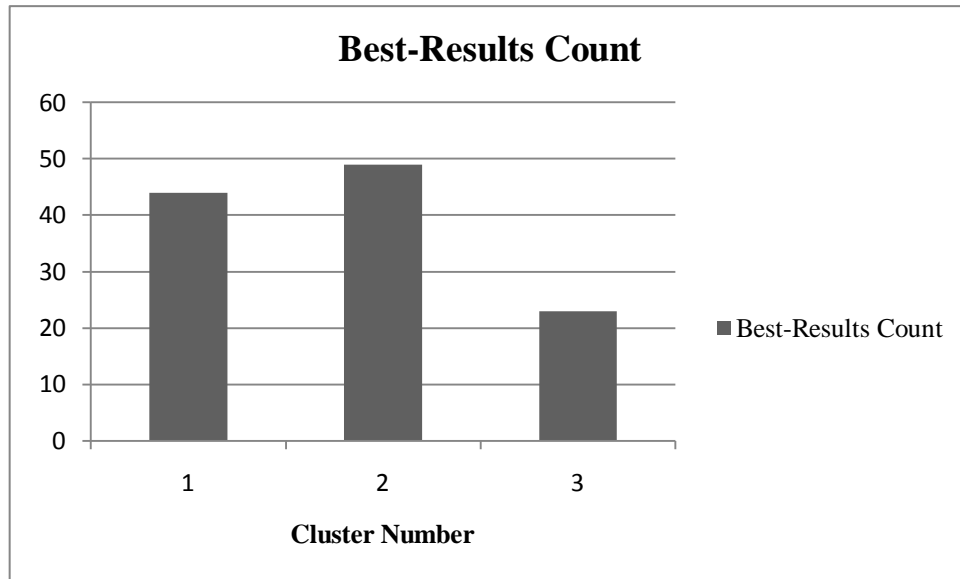


Figure 5.25: Number of Best Results for Each Cluster Number

The best results for QT clustering were obtained with threshold value around to 0.5. This is the average of min and max values of our distance metric. A graph that shows the number of obtained best results for each threshold value is shown in Figure 5.26.

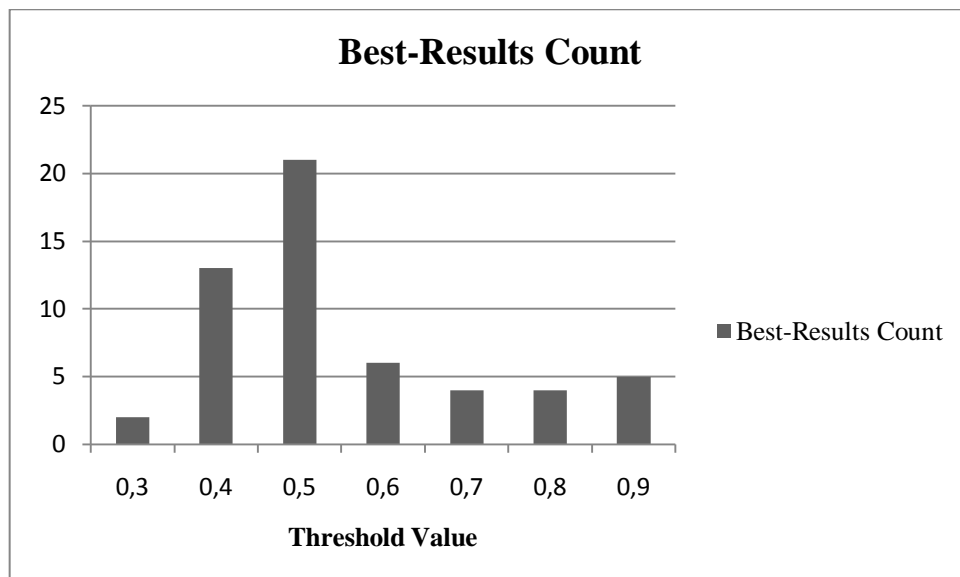


Figure 5.26: Number of Best Results for Each Threshold Value

Highest score is obtained by K-Means clustering at term percentage of 10%, rank-k percentage of %70 and cluster number of 3.

Detailed score tables for each clustering methods with different parameter values are given in Appendix B.

Same experiment is done in [22] with biggest TF.IDF to see the success of LSA approach in key-term extraction. Best ROUGE results obtained by biggest TF.IDF values used as Key-Term Extraction method and K-Means as clustering algorithm are given in Table 5.

Table 5.4: Best ROUGE Results for Biggest TF.IDF Method in Key-Term Extraction

Configuration Parameters			ROUGE Scores & Orders					
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 <u>0.133</u>
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 5.22)

We also create a random sentence selection algorithm to create summaries. In this method we fetch sentences from documents randomly until summary size reaches a predefined size limit. Then we calculated ROUGE results for this method. Best 10 ROUGE results obtained by random sentence selection are given in Table 6. By comparing the results obtained from random sentence selection and our summarization algorithm, we can see the success of our system.

Table 5.5: Best ROUGE Results for Random Sentence Selection

R1_AF	R2_AF	R3_AF	R4_AF	RL_AF	RW_12_AF
29 0.31465	33 0.04888	32 0.01229	29 0.00461	26 0.28243	26 0.12441
29 0.31544	33 0.0473	34 0.0101	34 0.0029	26 0.28064	26 0.12342
30 0.31216	33 0.04797	33 0.01148	30 0.00416	26 0.27997	26 0.12341
30 0.3122	32 0.05123	27 0.0146	26 0.00626	26 0.27768	26 0.12327
31 0.31028	33 0.04538	34 0.01068	30 0.00424	26 0.27771	26 0.12238
32 0.30661	33 0.04632	33 0.01188	29 0.00453	26 0.2748	26 0.12213
31 0.31017	33 0.04804	33 0.01208	28 0.00505	26 0.27818	26 0.12201
31 0.31043	33 0.04778	33 0.01197	31 0.00407	26 0.27627	26 0.12201
32 0.30811	33 0.04521	33 0.0116	30 0.00443	26 0.27567	26 0.12198
31 0.31032	33 0.04653	34 0.01068	29 0.00454	26 0.27662	26 0.12175

(Meanings of titles are shown in Figure 5.22)

CHAPTER 6

CONCLUSION AND FUTURE WORK

We created a summarization approach consisting of two main steps. At first step, key-terms were extracted using LSA. Then key-terms are used to extract important sentences through clustering and centroid based approach.

At the end of first step, key terms were ordered according to their importance. Then we used key-term percentage in order to find best results. Best results are generally obtained at lower percentages such as 10% and 20%. This shows that LSA is useful to finding the importance of terms in documents.

In second step we used clustering. Results with cluster number greater than 3 gives poorer results. We can say that cluster numbers higher than a value (3 clusters) is harmful for the performance of our summarization system. Additionally success rate of our summarization system is increased when applying rank-k approximation using LSI before clustering.

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

Additionally we are able to create summaries of Turkish documents by only changing stemmer and stop word list. So we can create summaries of documents of

other languages after providing stemmer and stop word list for corresponding language.

6.1 Future Work

As a future work new weighting schemes can be applied to the summarization system. Additionally a method for estimating the cluster number or diameter threshold can be used before clustering. In order to make summaries more understandable, we can order sentences in the summary to keep the order of events.

In order to improve Turkish summaries other another stemmer algorithm and stop word list can be used. Also documents of other languages can be summarized after providing stemmer and stop word list for that language.

Finally, to make our language independent summarization system more flexible, automatic language detection algorithms can be used to detect language of document set. After detecting the language, stemmer and stop word list for that language can be used in summarization operation.

REFERENCES

- [1] **Suanmali, L., Salim, N., Binwahlan, M.S.** (2009), Introduction, *Fuzzy Logic Based Method for Improving Text Summarization*, International Journal of Computer Science and Information Security, Vol 2, No 1.
- [2] **MANI, I.** (2001), Introduction, *Automatic Summarization*, John Benjamins Publishing Co., Amsterdam/Philadelphia, 1-5.
- [3] **BRANDOW, R., MITZE, K., RAU, L. F.** (1995), Automatic Condensation of Electronic Publications by Sentence Selection., *Information Processing & Management*, 675–685, Vol 31(5).
- [4] **BAXENDALE, P.B.** (1958). Machine-Made Index for Technical Literature—An Experiment. *IBM Journal* (October) 354–361.
- [5] **DONLAN, D.** (1980). Locating Main Ideas in History Textbooks. *Journal of Reading*, 24, 135–140.
- [6] **LIN, C. Y., HOVY, E. H.** (1997), Identifying Topics by Position, *Applied Natural Language Processing Conference*, 283–290.
- [7] **TEUFEL, S., MOENS, M.** (1997), Sentence Extraction as a Classification Task, *ACL/EACL97-WS*, Madrid, 58-65.
- [8] **EDMUNDSON, H. P.** (1969), New Methods in Automatic Extracting, *Journal of the Association for Computing Machinery*, 264–285, Vol 16(2).
- [9] **LUHN, H. P.** (1958), The Automatic Creation of Literature Abstracts, *IBM Journal of Research and Development*, 159-165, Vol 2(2).

- [10] **SALTON, G.** et. al. (1997), Automatic Text Structuring and Summarization, *Information Processing and Management*, Vol 33(2).
- [11] **BOGURAEV, B., KENEDY, C.** (1997), Saliency-Based Content Characterisation of Text Documents, *Advances in Automatic Text Summarization*, MIT Press, 2-9.
- [12] **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.
- [13] **BERRY, M. W.** (1992), Large Scale Singular Value Computations, *International Journal of Supercomputer Applications*, 13-49, Vol 6.
- [14] **BERRY, M. W., DUMAIS, S. T., O'BRIEN, G.W.** (1995), Using Linear Algebra for Intelligent Information Retrieval, *SIAM: Review*, 573-595, Vol 37.
- [15] **LANDAUER, T. K., FOLTZ, P. W., & LAHAM, D.** (1998), Introduction to Latent Semantic Analysis, *Discourse Processes*, 25, 259-284.
- [16] **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.
- [17] **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.
- [18] **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.
- [19] **HEYER, L.J., KRUGLYAK, S. AND YOOSEPH, S.**, Exploring Expression Data: Identification and Analysis of Coexpressed Genes, *Genome Research* 9:1106-1115.

- [20] **MANI, I.** (2001), Morphological-Level Approaches, *Automatic Summarization*, John Benjamins Publishing Co., Amsterdam/Philadelphia, 181-182.
- [21] **SALTON, G., MCGILL, M. J.** (1983), *Introduction to Modern Information Retrieval*, McGraw Hill Book Co., New York.
- [22] **KARAKAYNAK, S.** (2009), *Development of Tool for Managing Semantic Text Content*, M.S. Thesis, The Graduate School of Natural and Applied Sciences, Cankaya University, Ankara.
- [23] **ERCAN, G.** (2006), *Automated Text Summarization and Keyphrase Extraction*, M. S. Thesis, Institute of Engineering and Science, Bilkent University, Ankara.
- [24] **LOVINS, J. B.** (1968), Development of a Stemming Algorithm, *Mechanical Translation and Computational Linguistics* 11, 22-31.
- [25] **PORTER, M. F.** (1980), An Algorithm for Suffix Stripping, *Program*, 14(3), 130–137.
- [26] <http://tartarus.org/~martin/PorterStemmer/>
- [27] <https://zemberek.dev.java.net/>
- [28] **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.
- [29] **STEINBERGER, J., JEZEK, K.** (2004), Using Latent Semantic Analysis in Text Summarization and Summary Evaluation, *In Proc. ISIM '04*, 93-100.
- [30] <http://duc.nist.gov/duc2004/>
- [31] **LIKOWSKY, K. C.** (2004), Summarization Experiments in DUC 2004, *In Proceedings of the HLT-NAACL Workshop on Automatic Summarization*, Boston.

[32] <http://berouge.com>

[33] **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

English 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

English Stop Words (cont.)

am	everybody	mightn't	shalt	vice
american	everyone	mine	shan't	vis-avis
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

English Stop Words (cont.)

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

Turkish Stop Words

acaba	bu	iki	nereye	trilyon
altmış	buna	ile	niye	tüm
altı	bunda	ise	niçin	ve
ama	bundan	için	on	veya
bana	bunu	katrilyon	ona	ya
bazı	bunun	kez	ondan	yani
belki	da	ki	onlar	yedi
ben	daha	kim	onlardan	yetmiş
benden	dahi	kimden	onları	yirmi
beni	de	kime	onların	yüz
benim	defa	kimi	onu	çok
beş	diye	kırk	otuz	çünkü
bin	doksan	milyar	sanki	üç
bir	dokuz	milyon	sekiz	şey
biri	dört	mu	seksen	şeyden
birkaç	elli	mü	sen	şeyi
birkez	en	mı	senden	şeyler
birşey	gibi	mi	seni	şu
birşeyi	hem	nasıl	senin	şuna
biz	hep	ne	siz	şunda
bizden	hepsi	neden	sizden	şundan
bizi	her	nerde	sizi	şunu
bizim	hiç	nerede	sizin	

APPENDIX B

ROUGE SCORES

Top ROUGE Results with K-Means Clustering

Term %	Rank-k %	Cluster No	R1_AF	R2_AF	R3_AF	R4_AF	RL_AF	RW_12_AF
10	10	1	24 0.33489	18 0.07082	15 0.02392	15 0.01008	19 0.29728	15 0.13465
10	10	2	23 0.33741	21 0.06941	18 0.02286	18 0.00951	16 0.30004	15 0.1344
10	20	1	24 0.33489	18 0.07082	15 0.02392	15 0.01008	19 0.29728	15 0.13465
10	30	1	24 0.33489	18 0.07082	15 0.02392	15 0.01008	19 0.29728	15 0.13465
10	30	2	19 0.34117	22 0.06807	20 0.02154	22 0.00815	16 0.30065	15 0.13497
10	30	3	17 0.34433	18 0.0706	18 0.02279	19 0.0091	14 0.30535	13 0.13599
10	40	1	24 0.33489	18 0.07082	15 0.02392	15 0.01008	19 0.29728	15 0.13465
10	40	2	26 0.33203	23 0.06489	22 0.02045	22 0.00838	21 0.29481	17 0.13217
10	40	3	26 0.33196	23 0.067	20 0.02179	19 0.00868	21 0.2949	19 0.13189
10	50	1	24 0.33489	18 0.07082	15 0.02392	15 0.01008	19 0.29728	15 0.13465
10	50	2	27 0.33103	23 0.06548	21 0.02099	23 0.0079	21 0.29445	17 0.13221

Top ROUGE Results with K-Means Clustering (cont.)

Term %	Rank-k %	Cluster No	R1_AF	R2_AF	R3_AF	R4_AF	RL_AF	RW_12_AF
10	60	1	24 0.33489	18 0.07082	15 0.02392	15 0.01008	19 0.29728	15 0.13465
10	60	2	25 0.33328	23 0.06561	21 0.02076	21 0.00843	20 0.2962	17 0.13284
10	60	3	25 0.33294	23 0.06609	21 0.02114	19 0.00917	21 0.29494	19 0.13171
10	70	1	24 0.33489	18 0.07082	15 0.02392	15 0.01008	19 0.29728	15 0.13465
10	70	2	23 0.33724	21 0.06927	19 0.0224	19 0.00914	16 0.29989	14 0.13535
10	70	3	18 0.3424	21 0.06929	20 0.02205	18 0.00921	14 0.30399	13 0.13625
10	80	1	24 0.33489	18 0.07082	15 0.02392	15 0.01008	19 0.29728	15 0.13465
10	80	2	21 0.33871	20 0.07016	18 0.02283	19 0.00911	16 0.30081	14 0.13533
10	80	3	19 0.34021	20 0.0704	16 0.0234	17 0.0098	15 0.30206	14 0.13514
10	90	1	24 0.33489	18 0.07082	15 0.02392	15 0.01008	19 0.29728	15 0.13465
10	90	2	25 0.33331	22 0.06765	21 0.02075	22 0.00831	20 0.2963	16 0.13327
10	100	1	24 0.33489	18 0.07082	15 0.02392	15 0.01008	19 0.29728	15 0.13465
10	100	3	23 0.33711	22 0.06714	20 0.02201	18 0.00952	19 0.2973	17 0.13281
20	10	1	27 0.32858	22 0.06727	20 0.02173	19 0.00873	23 0.29161	19 0.1318
20	10	2	25 0.33309	23 0.06689	21 0.0211	21 0.00849	19 0.29822	15 0.1337
20	30	2	24 0.33472	23 0.06579	21 0.02066	23 0.00792	19 0.29809	15 0.13384
20	30	3	25 0.33351	26 0.06156	24 0.01941	22 0.0082	19 0.29791	17 0.13285
20	40	2	21 0.33826	21 0.06909	19 0.02239	19 0.00894	15 0.30239	13 0.13601
20	40	3	21 0.33974	22 0.0683	18 0.02278	15 0.01003	15 0.30176	15 0.13456

Top ROUGE Results with K-Means Clustering (cont.)

Term %	Rank-k %	Cluster No	R1_AF	R2_AF	R3_AF	R4_AF	RL_AF	RW_12_AF
20	50	3	23 0.3362	25 0.06358	24 0.01925	23 0.00782	17 0.29863	16 0.13333
20	60	2	23 0.33527	23 0.06701	19 0.02256	18 0.00962	16 0.29886	15 0.13417
20	70	2	21 0.33853	20 0.06981	19 0.02253	18 0.00954	16 0.30096	15 0.13499
20	70	3	21 0.33919	22 0.06769	20 0.02142	19 0.0088	16 0.30131	14 0.13534
20	80	2	23 0.3371	23 0.06609	20 0.02183	19 0.00912	16 0.3009	15 0.13484
20	90	1	27 0.32858	22 0.06727	20 0.02173	19 0.00873	23 0.29161	19 0.1318
20	90	2	25 0.33262	25 0.0633	24 0.01981	21 0.00843	20 0.29605	17 0.13246
20	90	3	18 0.3423	22 0.06728	20 0.02175	21 0.00843	14 0.30452	15 0.13478
20	100	1	27 0.32858	22 0.06727	20 0.02173	19 0.00873	23 0.29161	19 0.1318
20	100	2	27 0.33102	23 0.06707	20 0.02222	18 0.00923	21 0.29503	17 0.13301
20	100	3	23 0.33566	25 0.06296	21 0.02078	19 0.00867	16 0.29878	15 0.13369
30	10	1	27 0.32833	23 0.06675	20 0.02178	19 0.00881	23 0.29137	19 0.1315
30	10	2	21 0.33819	22 0.06882	19 0.02256	18 0.00942	16 0.30051	15 0.13483
30	10	3	21 0.33841	22 0.06857	20 0.02212	19 0.00877	16 0.30074	15 0.1348
30	20	2	21 0.33892	22 0.06845	20 0.02222	19 0.00888	15 0.30164	13 0.13592
30	20	3	27 0.33061	23 0.0642	24 0.0192	22 0.00804	19 0.29659	17 0.13269
30	30	1	27 0.32833	23 0.06675	20 0.02178	19 0.00881	23 0.29137	19 0.1315
30	30	2	25 0.33332	25 0.06189	24 0.01923	24 0.00772	20 0.29594	17 0.13267
30	40	1	27 0.32833	23 0.06675	20 0.02178	19 0.00881	23 0.29137	19 0.1315

Top ROUGE Results with K-Means Clustering (cont.)

Term %	Rank-k %	Cluster No	R1_AF	R2_AF	R3_AF	R4_AF	RL_AF	RW_12_AF
30	80	3	27 0.3312	23 0.06433	20 0.02174	19 0.00913	21 0.29471	20 0.13132
30	100	1	27 0.32833	23 0.06675	20 0.02178	19 0.00881	23 0.29137	19 0.1315
40	10	1	27 0.32793	22 0.06721	20 0.022	19 0.00893	23 0.29094	19 0.13156
40	20	1	27 0.32793	22 0.06721	20 0.022	19 0.00893	23 0.29094	19 0.13156
40	20	2	23 0.33684	23 0.06634	20 0.02218	18 0.00971	16 0.29908	15 0.13445
40	20	3	25 0.33333	26 0.06116	26 0.01807	25 0.00717	19 0.29749	17 0.13251
40	40	2	27 0.32804	23 0.06508	20 0.02204	18 0.00926	22 0.29266	19 0.13155
40	70	2	23 0.33656	23 0.06459	21 0.0208	22 0.00834	18 0.29831	15 0.13398
40	80	2	27 0.33121	23 0.06554	21 0.02114	21 0.00843	21 0.29395	18 0.13207
50	10	3	26 0.33239	25 0.06247	25 0.01869	24 0.00739	20 0.29601	17 0.13249
50	20	2	25 0.33288	23 0.06488	21 0.021	20 0.00857	20 0.29648	17 0.13301
60	10	2	23 0.33677	23 0.06479	22 0.02026	22 0.00818	19 0.29764	17 0.13279
60	20	2	26 0.33143	25 0.06324	22 0.02019	19 0.00863	21 0.29526	17 0.13296
60	40	2	25 0.33278	25 0.06366	21 0.02062	19 0.00884	21 0.29509	17 0.13277
60	70	2	23 0.33732	23 0.06649	21 0.02103	19 0.00871	17 0.29869	15 0.13466
60	90	2	26 0.3315	23 0.06513	21 0.02117	19 0.00867	21 0.29499	17 0.13261
70	30	2	25 0.33305	25 0.06291	24 0.01952	24 0.00778	21 0.29508	17 0.13223
70	40	3	23 0.33613	23 0.06453	22 0.02007	22 0.00826	16 0.29918	17 0.13296
70	60	2	26 0.33172	26 0.06168	24 0.01923	22 0.00813	21 0.29489	18 0.13204

Top ROUGE Results with K-Means Clustering (cont.)

Term %	Rank-k %	Cluster No	R1_AF	R2_AF	R3_AF	R4_AF	RL_AF	RW_12_AF
70	70	2	23 0.33578	22 0.06771	20 0.02199	19 0.00888	18 0.29828	15 0.13447
70	80	2	21 0.33849	23 0.06544	21 0.02065	22 0.00822	16 0.2989	15 0.13357
70	90	2	25 0.33297	25 0.06327	23 0.01996	23 0.00798	21 0.2947	17 0.13256
70	100	3	22 0.33799	23 0.06411	22 0.02031	21 0.00854	16 0.29999	15 0.13397
80	10	2	27 0.32964	23 0.06414	21 0.02109	21 0.00847	22 0.2935	17 0.13251
80	20	2	23 0.33539	23 0.06516	21 0.02086	22 0.00823	19 0.29726	17 0.13312
80	30	2	23 0.33745	23 0.06529	21 0.0208	22 0.00838	20 0.29646	17 0.13281
80	40	2	25 0.33253	25 0.0638	22 0.02051	22 0.00813	21 0.29509	17 0.13311
80	40	3	23 0.33613	23 0.06536	20 0.0215	18 0.00978	19 0.29803	16 0.13337
80	60	2	23 0.33737	23 0.06678	20 0.02159	18 0.00924	16 0.29986	15 0.13429
80	70	2	23 0.33577	23 0.06446	21 0.02083	22 0.00828	20 0.29573	17 0.13261
80	90	3	25 0.33332	25 0.06215	25 0.01874	25 0.00703	20 0.29584	17 0.13303
90	60	3	23 0.33603	25 0.06343	25 0.01886	24 0.0074	19 0.29785	17 0.13304
90	90	3	24 0.33448	25 0.06269	22 0.0204	19 0.00863	19 0.2976	17 0.13305
100	40	2	26 0.33236	25 0.06317	22 0.01999	23 0.00799	21 0.29466	17 0.13235
100	60	2	27 0.3299	23 0.06523	21 0.02057	23 0.00784	22 0.29272	18 0.13208
100	80	2	27 0.32685	23 0.06492	21 0.02126	18 0.00936	23 0.28957	21 0.13048
100	90	2	27 0.33064	23 0.06556	21 0.02079	19 0.0089	21 0.29455	17 0.13287

Top ROUGE Results with QT Clustering

Term %	Rank-k %	Threshold Value	R1_AF	R2_AF	R3_AF	R4_AF	RL_AF	RW_12_AF
10	20	0.3	25 0.33337	26 0.0611	26 0.01667	25 0.00713	21 0.29447	20 0.13087
10	20	0.4	26 0.33225	26 0.05962	26 0.01648	28 0.00551	20 0.2959	19 0.13165
10	20	0.5	19 0.34138	25 0.06261	24 0.01908	25 0.00719	16 0.30119	15 0.1341
10	20	0.6	23 0.3375	23 0.06507	24 0.01992	24 0.00758	16 0.30055	17 0.13288
10	20	0.7	23 0.33688	24 0.06386	24 0.01946	24 0.00756	16 0.29971	17 0.13264
10	20	0.8	24 0.33452	25 0.06296	25 0.01874	25 0.00707	19 0.29763	19 0.13164
10	20	0.9	23 0.33535	25 0.06335	24 0.01887	25 0.00707	16 0.29907	17 0.13229
10	30	0.3	23 0.33715	25 0.0624	25 0.01874	24 0.0073	16 0.30033	15 0.13424
10	30	0.4	19 0.34071	25 0.06368	24 0.01961	22 0.00806	16 0.29952	16 0.13321
10	30	0.8	23 0.33664	23 0.06408	24 0.01891	24 0.00727	20 0.29615	19 0.13158
10	30	0.9	23 0.33618	23 0.06464	24 0.01896	25 0.00721	20 0.29573	19 0.13152
10	40	0.4	26 0.3324	25 0.06243	22 0.01999	22 0.00809	20 0.29552	20 0.13098
10	50	0.4	23 0.33543	25 0.06362	22 0.02025	24 0.00778	20 0.29594	18 0.13207
10	50	0.5	22 0.33755	25 0.06225	25 0.01864	26 0.00662	20 0.29637	19 0.13191
10	60	0.4	23 0.33553	25 0.0629	24 0.01971	24 0.00778	19 0.29779	18 0.13203
10	60	0.5	21 0.33929	23 0.065	24 0.01992	24 0.00757	19 0.29825	17 0.13279
10	70	0.5	23 0.33575	26 0.0607	26 0.01781	26 0.00677	19 0.29785	19 0.13173
10	80	0.5	23 0.33568	26 0.0608	26 0.01772	26 0.00677	20 0.29643	19 0.13166

Top ROUGE Results with QT Clustering (cont.)

Term %	Rank-k %	Threshold Value	R1_AF	R2_AF	R3_AF	R4_AF	RL_AF	RW_12_AF
10	90	0.9	26 0.33171	26 0.05999	26 0.01773	23 0.0078	22 0.29295	20 0.13106
10	100	0.5	23 0.33643	25 0.06314	24 0.01894	26 0.00677	16 0.29928	17 0.1328
10	100	0.6	25 0.33285	23 0.06655	20 0.02192	17 0.00979	20 0.29537	17 0.13254
10	100	0.7	26 0.33193	23 0.06683	20 0.0219	18 0.00977	21 0.29426	17 0.13222
10	100	0.8	26 0.33216	23 0.06677	20 0.02206	15 0.00993	21 0.2946	17 0.13251
10	100	0.9	24 0.33374	22 0.06711	20 0.02213	15 0.00993	20 0.29593	17 0.133
20	10	0.5	22 0.3376	26 0.06138	26 0.0181	24 0.00722	19 0.29824	17 0.13264
20	10	0.6	25 0.33297	26 0.06003	26 0.01703	24 0.0074	19 0.29758	18 0.13205
20	20	0.5	21 0.33842	26 0.06119	24 0.01901	22 0.00824	16 0.29886	17 0.13304
20	30	0.5	24 0.33466	28 0.05756	26 0.0169	26 0.00642	20 0.29611	17 0.13218
20	40	0.5	19 0.34183	23 0.0666	24 0.0192	24 0.00731	15 0.30224	14 0.13509
20	50	0.5	19 0.34049	25 0.06369	24 0.01993	20 0.00862	14 0.30343	14 0.13502
20	60	0.5	19 0.34002	25 0.06381	22 0.02	21 0.00851	19 0.29784	17 0.1326
20	70	0.5	23 0.33539	25 0.06337	24 0.0192	23 0.0079	21 0.29501	20 0.13132
20	80	0.6	25 0.3329	26 0.06053	25 0.01859	23 0.0078	20 0.29609	17 0.13223
20	80	0.7	26 0.33179	26 0.05996	25 0.01851	23 0.00785	21 0.29497	19 0.13177
20	80	0.8	26 0.33179	26 0.05996	25 0.01851	23 0.00785	21 0.29497	19 0.13177
20	80	0.9	26 0.33179	26 0.05996	25 0.01851	23 0.00785	21 0.29497	19 0.13177

Top ROUGE Results with QT Clustering (cont.)

Term %	Rank-k %	Threshold Value	R1_AF	R2_AF	R3_AF	R4_AF	RL_AF	RW_12_AF
30	10	0.6	26 0.33163	26 0.05984	26 0.01752	24 0.0075	21 0.29434	20 0.13117
30	20	0.4	23 0.33731	23 0.06402	26 0.01767	26 0.00582	20 0.29624	17 0.13243
30	50	0.4	25 0.33292	25 0.06269	24 0.0193	25 0.0072	21 0.29425	20 0.13126
30	60	0.5	23 0.33625	26 0.06131	24 0.01945	22 0.00842	21 0.29496	19 0.13155
40	20	0.4	19 0.34062	25 0.06327	25 0.01886	24 0.0073	19 0.29713	19 0.13194
40	50	0.4	21 0.33946	22 0.06885	20 0.0218	19 0.00912	16 0.29962	17 0.13283
50	20	0.7	27 0.3305	26 0.05956	26 0.01809	25 0.0070	22 0.29371	20 0.13095
50	30	0.6	26 0.33219	26 0.06171	25 0.01851	24 0.0076	21 0.29404	20 0.13106
50	40	0.4	28 0.32121	26 0.05975	22 0.02012	19 0.0088	26 0.2841	24 0.12662
50	90	0.5	25 0.33294	26 0.06174	24 0.0197	19 0.00898	20 0.29626	17 0.1322
60	10	0.4	23 0.33614	26 0.06002	26 0.01669	26 0.00643	20 0.29653	17 0.1323
60	40	0.5	21 0.33835	23 0.06676	21 0.02139	19 0.00912	14 0.30359	13 0.13579
60	50	0.5	26 0.33172	25 0.06271	24 0.01957	21 0.00845	22 0.29333	20 0.13124
80	10	0.4	24 0.33478	25 0.06207	26 0.01694	26 0.00597	19 0.2978	17 0.13255
80	30	0.5	25 0.3326	26 0.06084	26 0.0183	24 0.00734	22 0.29369	20 0.13135
80	40	0.5	23 0.33668	26 0.06156	26 0.01786	24 0.00725	20 0.29588	19 0.13184
90	10	0.4	23 0.33682	26 0.06172	26 0.01695	26 0.00616	19 0.29687	19 0.13187
90	30	0.5	26 0.33242	23 0.06399	22 0.01998	22 0.00821	20 0.29537	19 0.13168
100	40	0.5	24 0.33451	23 0.06572	17 0.02291	15 0.01007	17 0.29869	15 0.13374

Top ROUGE Results with Agglomerative Hierarchical Clustering

Term %	Rank-k %	Cluster No	R1_AF	R2_AF	R3_AF	R4_AF	RL_AF	RW_12_AF
10	10	1	23 0.33507	23 0.06628	20 0.02164	19 0.00883	18 0.2983	15 0.13373
10	30	2	27 0.3313	23 0.06558	21 0.02098	22 0.00837	21 0.29487	19 0.13155
10	40	1	26 0.33206	25 0.06248	24 0.01986	21 0.00847	21 0.29426	20 0.13129
10	40	2	27 0.3311	23 0.06468	24 0.01957	24 0.00755	22 0.29322	20 0.1309
10	60	1	27 0.33073	24 0.06389	22 0.02018	19 0.00875	22 0.29286	21 0.13045
10	70	1	26 0.33202	25 0.06351	24 0.01966	20 0.00859	22 0.29331	21 0.13078
10	80	1	27 0.32887	25 0.06271	24 0.01966	19 0.00869	23 0.28983	23 0.12975
10	90	1	27 0.32899	26 0.06167	24 0.01953	20 0.00857	23 0.29137	23 0.12989
10	100	1	27 0.32896	25 0.06338	22 0.02006	19 0.00864	23 0.29163	22 0.13006
20	20	1	27 0.32993	25 0.06188	25 0.0187	25 0.00706	23 0.29231	20 0.1312
20	30	1	24 0.33408	26 0.06099	25 0.01834	25 0.00688	20 0.29581	18 0.13206
20	60	1	26 0.33232	26 0.06105	25 0.01834	25 0.00715	22 0.29359	20 0.13131
20	90	1	26 0.33208	26 0.06141	25 0.01874	24 0.00736	23 0.29251	20 0.13106
30	10	1	25 0.33361	26 0.05993	26 0.01788	24 0.00737	21 0.29496	19 0.13153
30	20	1	24 0.33459	26 0.05965	26 0.01749	26 0.00676	19 0.29666	19 0.13167
30	40	1	24 0.33383	26 0.06119	26 0.01765	25 0.00685	21 0.29456	20 0.13134
30	50	1	23 0.33555	25 0.06236	25 0.01842	25 0.00705	20 0.29601	18 0.13205
30	70	1	25 0.3336	26 0.06036	26 0.01759	25 0.00685	21 0.29442	19 0.1315

Top ROUGE Results with Agglomerative Hierarchical Clustering (cont.)

Term %	Rank-k %	Cluster No	R1_AF	R2_AF	R3_AF	R4_AF	RL_AF	RW_12_AF
30	80	1	24 0.33434	26 0.06099	26 0.01721	26 0.00681	21 0.29526	19 0.13171
30	90	1	26 0.33222	26 0.06134	26 0.01756	25 0.00704	21 0.29435	20 0.13136
30	100	1	23 0.3364	25 0.06361	25 0.01838	24 0.00727	19 0.29666	17 0.13268
40	60	1	24 0.33433	25 0.06235	25 0.01854	25 0.00705	20 0.29545	19 0.13174
40	80	1	25 0.333	26 0.06114	26 0.01813	25 0.00711	21 0.29443	20 0.13092
60	80	1	27 0.32736	23 0.06454	21 0.02094	20 0.00858	23 0.28847	23 0.1294
60	90	2	29 0.32034	26 0.05925	24 0.01961	19 0.00892	24 0.28556	23 0.12837
70	10	2	29 0.32022	26 0.0612	21 0.02084	19 0.00889	24 0.28543	23 0.12747
70	90	2	30 0.31239	28 0.05722	24 0.0189	19 0.00887	26 0.27696	26 0.12385
80	60	1	29 0.32009	25 0.06359	24 0.01982	20 0.00855	26 0.28216	23 0.12716
80	80	1	29 0.31795	25 0.06236	25 0.01854	24 0.00765	26 0.28054	24 0.12642
80	100	1	28 0.32156	25 0.06302	25 0.01878	24 0.0076	26 0.28383	23 0.12775

(Meanings of titles are shown in Figure 5.22)

APPENDIX C

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Alim, Suat
Nationality: Turkish (TC)
Date and Place of Birth: 15 January 1983, Sivas
Marital Status: Single
Phone: +90 536 285 49 08
email: suat@infopark.com.tr

EDUCATION

Degree	Institution	Graduation Year
MS	Çankaya University Computer Engineering	2009
BS	Çankaya University Computer Engineering	2005
High School	Kongre High School, Sivas	1999

WORK EXPERIENCE

Year	Place	Enrollment
2007- Present	İnfopark Software and Consultancy	Project Manager
2005- 2007	İnfopark Information Technologies	Software Specialist

FOREIGN LANGUAGES

English

HOBBIES

Basketball, Movies, Motor Sports.