MULTI-DOCUMENT TEXT SUMMARIZATION USING TEXT CLUSTERING FOR ARABIC LANGUAGE

SAMER ABDULATEEF WAHEEB

SCHOOL OF COMPUTING
UUM COLLEGE OF ARTS AND SCIENCES
UNIVERSITI UTARA MALAYSIA
2014

Permission to Use

In presenting this thesis in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:

Dean of Awang Had Salleh Graduate School of Arts and Sciences
UUM College of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok

Abstract (English)

The process of multi-document summarization is producing a single summary of a collection of related documents. In this work we focus on generic extractive Arabic multi-document summarizers. We also describe the cluster approach for multi-document summarization. The problem with multi-document text summarization is redundancy of sentences, and thus, redundancy must be eliminated to ensure coherence, and improve readability. Hence, we set out the main objective as to examine multi-document summarization salient information for text Arabic summarization task with noisy and redundancy information. In this research we used Essex Arabic Summaries Corpus (EASC) as data to test and achieve our main objective and of course its subsequent sub-objectives. We used the token process to split the original text into words, and then removed all the stop words, and then we extract the root of each word, and then represented the text as bag of words by TFIDF without the noisy information. In the second step we applied the K-means algorithm with cosine similarity in our experimental to select the best cluster based on cluster ordering by distance performance. We applied SVM to order the sentences after selected the best cluster, then we selected the highest weight sentences for the final summary to reduce redundancy information. Finally, the final summary results for the ten categories of related documents are evaluated using Recall and Precision with the best Recall achieved is 0.6 and Precision is 0.6.

Keywords: Multi-document text summarization, Arabic text summarization, Automatic text summarization, Text clustering.

Abstrak (Bahasa Malaysia)

Proses ringkasan multi-dokumen adalah menghasilkan ringkasan tunggal daripada beberapa dokumen yang berkaitan. Dalam disertasi ini kami memberi tumpuan kepada ringkasan multi-dokumen generik ekstraktif dalam Bahasa Arab. Kami juga menghuraikan pendekatan kluster bagi ringkasan berbilang dokumen. Permasalahan yang berkaitan dengan ringkasan multi-dokumen ialah lebihan ayat yang berulang, dan dengan itu, ianya mesti dikeluarkan daripada ringkasan bagi memastikan kepaduan dan meningkatkan kebolehbacaan ringkasan yang dihasilkan. Oleh itu, objektif utama disertasi ini ialah untuk memeriksa maklumat penting ringkasan pelbagai dokumen untuk teks Bahasa Arab dengan mengambilkira maklumat asing dan lebihan ayat. Dalam kajian ini kami menggunakan Essex Arabic Summaries Corpus (EASC) sebagai data bagi menguji dan mencapai matlamat utama kami dan seterusnya mencapai sub-objektif berikutnya. Kami menggunakan proses pertimbangan untuk mengasingkan teks asal ke dalam perkataan, dan kemudian mengeluarkan semua perkataan yang tidak signifikan, dan kemudian kata akar bagi setiap perkataan diekstrak, dan seterusnya mewakilkan teks dalam bentuk beg perkataan dengan TFIDF tanpa maklumat yang tidak diperlukan. Dalam langkah kedua kami menggunakan algoritma K-means dengan persamaan kosinus dalam percubaan untuk memilih kluster terbaik berdasarkan susunan kluster oleh prestasi jarak. Kami menggunakan SVM untuk menyusun ayat selepas memilih kluster yang terbaik, dan kemudian memilih ayat dengan pemberat paling tinggi bagi ringkasan akhir untuk mengurangkan maklumat lebihan. Akhirnya, keputusan ringkasan akhir bagi sepuluh kategori dokumen berkaitan dinilai menggunakan *Recall* dan *Precision* dengan *Recall* yang terbaik dicapai adalah 0.6 dan *Precision* ialah 0.6.

Kata kunci: Ringkasan teks multi-dokumen, ringkasan teks Bahasa Arab, ringkasan teks automatik, pengklusteran teks.

Acknowledgements

To the memory of my father …

To my mother …

To my family and brothers …

Thanks to Allah for giving me the help and power to accomplish this research. Without the grace of Allah, I was not able to accomplish this work.

Writing this master thesis would not have been possible without the help and continuous with people support me during this wonderful journey.

This thesis would not have seen the light without the endless support and enormous, guidance and patience of my supervisors, Dr. Husniza binti Husni, and A.P. Dr. Faudziah Ahmad for which my mere thanks expression likewise does not suffice.

My time as a master student at Utara University was great with lots of wonderful memories. The time when I started my master I was a bit worried of losing my social life but luckily that was not the case. I was able to split my time between my master, connecting with my friends, playing football, and travelling.

Last but not the least, I would like to offer thanks to all my friends at University Utara Malaysia. They each helped make my time during this master program more fun and interesting.

I would like to thank to all Universiti Utara Malaysia management especially College of Arts and Sciences staff and those who involved directly or indirectly in the dissertation. May Allah bless all of you.

Table of Contents

List of Figures

List of Tables

Glossary of Terms

Notation	**Description**
Natural Language Processing (NLP)	The science information branch that deals with natural language information.
Information extraction (IE)	A kind of information retrieval whose goal is to automatically extract structured information from unstructured documents.
Automatic Summarization	The creation of a shortened version of a text by a program of computer.
Extractive Summarization	Using IE for generating a system summary.
Generic-based Summary	A summary that presents an overall sense of a documents' contents.
Query-based Summary	A summary that presents the contents of a document that are related to a user's query.
Cluster	A similar group of objects growing closely together.
Clustering	The task of assigning a set of objects into groups (so called clusters) so that the objects in the same cluster are more similar to each other than to those in other clusters.
Hidden Markov Model	A statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved state.
Machine Learning	a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data.

Unsupervised learning	A machine learning task of inferring a function from unlabeled data.
Supervised learning	A machine learning task of inferring a function from supervised (labeled) training data.

CHAPTER ONE

INTRODUCTION

1.1 Introduction

This chapter presents an overview of the whole study. The first section describes the background of the study that leads to the implementation of the whole research. This is followed by the statement of problem, research question, research objectives, the scope, and the significance of the study.

1.2 Background of the Study

The availability of electronic documents in Arabic language on the internet is increasing exponentially. So everyone should take advantage of this information revolution. The excellent way to gain access to these documents and get the basic thought is to be able to extract the main idea and take advantage of them. For this reason, automatic text summarization has rapidly grown into main research area as demonstrated by the Document Understanding Conference (DUC), which started in 2001 and the Text Analysis Conference (TAC) (McKeown, 2011).

Automatic Text Summarization (ATS) is a procedure of examining the maximum salient information of related documents and transporting them in less space from the original text. On the other hand, Text Summarization (TS), which goal is to take out a reductive source text transformation to summary text through content condensation by selection and/or generality on what is significant in the original text (Lloret & Palomar, 2012). In principle, ATS is probable because of the naturally occurring redundancy in text and because significant salient information is spread unevenly in text documents. Examining the redundancy is a challenge which has not been resolved yet. On the other hand, the some quality of text summarization issues is still remaining for example redundancy, coherence, and grammaticality (Fukumoto, Sakai, & Suzuki, 2010; Lloret & Palomar, 2012).

There is no definition for redundancy and salience given that various user summaries may have various backgrounds, preferences, and tasks (Fukumoto, et al., 2010). Salience too relies on the structure of the source document or set. Meanwhile, information which the user knows should not be involved in final summary and at the same time information that is salient for a user may not be for others. It is very hard to achieve reliable judgments about final summary quality from human judges and thus this fact has made it problematic to evaluate automatic text summarization (Lloret & Palomar, 2012).

Text summarization is one of the natural language processing (NLP) applications which proposed to extract the most significant information from the document(s) and introduce it to the user. In this case of ATS tasks, most of the activities are focused on English and European language, as with DUC and TAC. However, in the Arabic language TAC and NLP lack resources for example Arabic lexicons and corpora (Ibrahim, Elghazaly, & Gheith, 2013).

In text summarization, the idea of classification is that one can distinguish between the following kinds of summaries: monolingual/multilingual, generic/query-based, extractive/abstractive, and single-document/multi document (see Figure 1.1). Most existing text summarizers work in an extractive approach, choosing parts of the original documents (e.g., words, sentences) that are believed to be more salient (Das & Martins, 2007; Larson, 2011). On the other hand, abstractive text summarization contains dynamic reformulation of the extracted content, including a deeper comprehension of the original text (Gholamrezazadeh, Salehi, & Gholamzadeh, 2009).

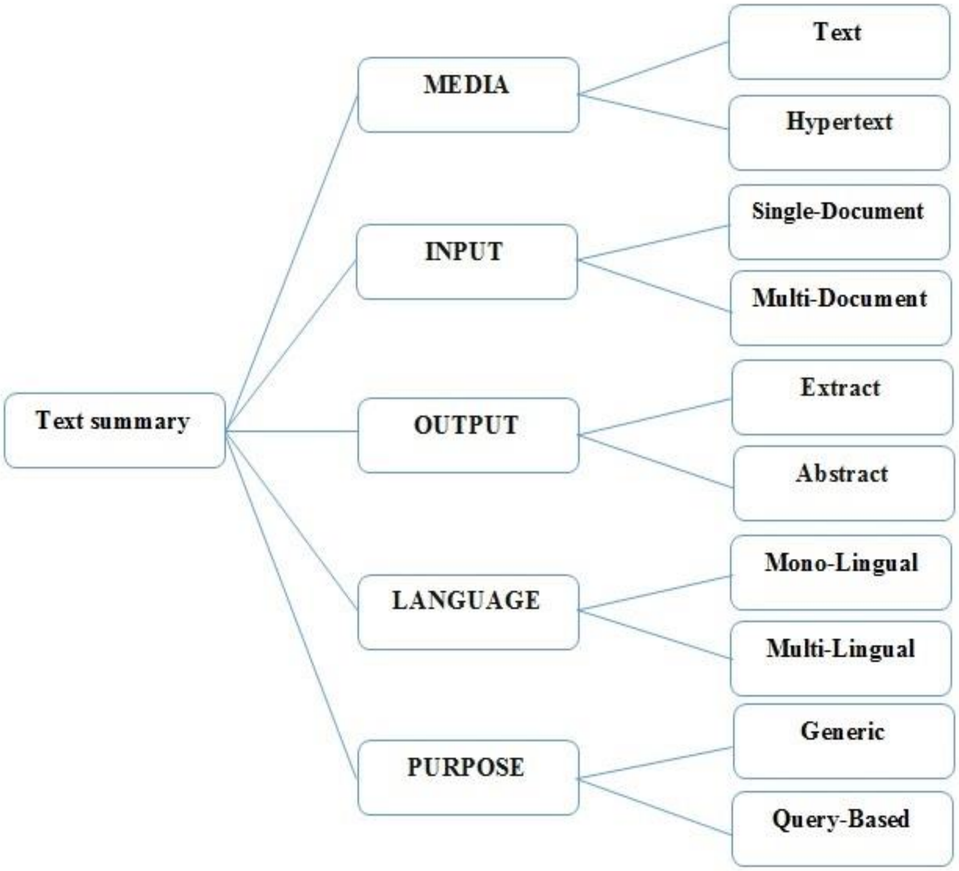


Figure 1.1: Summary types (Gholamrezazadeh, et al., 2009; Lloret & Palomar, 2012).

Query-based text summary is generated in reference to one user query (e.g., summary documents about an international conference focusing only on matters related to the environment) whereas generic text summaries attempt to exam salient information in textual without the context of a query. The variation among multi-document summarization (MDS) and single document summarization (SDS) is quite clear, though some of the kinds of problems that happen in MDS are qualitatively vary from the ones observed in single-document summarization (e.g., addressing redundancy across information sources and dealing with complementary and contradictory information) (Lloret & Palomar, 2012). This research will focuses on Arabic multi-document text summarization by extracting the information. In this work, the summarization approach proposed are generic as there is no query.

A number of evaluation methods for summarization have been developed and are typically categorized into two types (Larson, 2011). Intrinsic measures try to quantify the similarity of a text summarization with one or more summarization model produced by humans. Intrinsic measures include Recall, Precision, Sentence Overlap and F-measure. All of these metrics suppose that summaries have been produced in an extractive method (Das & Martins, 2007; Sobh, Darwish, & Fayek, 2009). Extrinsic measures contain using the summaries for a task (e.g., text classification, document retrieval, or question answering) (A. M. Azmi & Al-Thanyyan, 2012).

Usually, text summarization has been primarily applied to two kinds of text for example news stories and scientific papers. In both these domains, text summarization simply selects the first few sentences of document(s). The main aim of automatic text summarization ATS is to make the original text into a shorter form while preserving its information content and overall meaning. On the other hand, the system of multi document summarization in the news field. It extracts sentence that information signifies key gather from related documents and put into a document (Gholamrezazadeh, et al., 2009). There are attempts to summarize in other texts for example hypertext, email, fiction, video, image, and audio but they have been somewhat less successful (A. M. Azmi & Al-Thanyyan, 2012; Fan, Gao, Luo, Keim, & Li, 2008; He, Sanocki, Gupta, & Grudin, 1999; Sun et al., 2005; Zechner & Waibel, 2000).

Nowadays, text summarization researchers have also examined methods of text compression or simplification (Vishal &Gupta 2010). Actually, these approaches apply to a sentence at a time. Simple approaches include dropping insignificant words (adverbs). Complex approaches include rebuild the syntactic parse tree of the sentence to delete sections or rephrase units in shorter form (Lloret & Palomar, 2010).

1.3 Problem Statement

Multi Text Summarization (MTS) in overall is the procedure of summarizing a set of the related articles by summing up the most significant documents, making sure the documents arrangement is coherent by organizing them chronologically. On the other hand, ATS is the procedure of making a shorter, compact version of a text by using computers. For instance reducing a group of related documents into a shorter version of documents (words, sentences or paragraphs) employing automated techniques and tools.

The summarization should take the key contributions of the documents. In this case only the key sentences should be shown in the summarization and the procedure of determining those sentences is highly relying on the summarization technique used. The idea to choose sentences that carry the main idea of the document(s) is general between most of the summarization methods, various techniques and tools are applied to attempt to enhance the selection method (Lin & Hovy, 2002).

For extracting a meaningful structure of the data, a number of the next consecutive iterations can continually improve the clustering quality. This is achieved due to the generation less noisy data representations. The meaning of removing noisy information is keeping only the really and important information (Douzidia & Lapalme, 2004).

For multi-document summarization, the redundant sentences can be a challenge, and thus, redundancy must be eliminated to ensure coherence, and improve readability (Fukumoto, et al., 2010). The most significant part of a redundancy removal procedure is the measure of similarity. Fukumoto, et al. (2010) is focusing on an approach for redundancy elimination namely: cluster-based multi-document summarization approach (Jayashree, Murthy, & Anami, 2012). The advantage of cluster approach, is that efforts have been put into making the whole summarizing multi document process effective, which it is worth to determine the best clustering number, and clustering is better for reducing the number of redundant features (Kumar & Salim, 2011).

Text clustering can be potentially used to eliminate redundancy, where the extracted sentences are classified into sets of semantically related sentences. Fukumoto summarization research focused on discovering key sentences, which contain main information, from related documents (Fukumoto, et al., 2010). When making a comparison between two sentences, one of them is considered redundant if the similarity value among these two sentences is high (depending on a chosen similarity threshold) (El-Haj, Kruschwitz, & Fox, 2011). Thus, only one of the compared sentences should be chosen. The determination on which sentence should be selected and which one should not is based on the redundancy elimination tool or technique used (Fukumoto, et al., 2010). Working in multi text summarization raises questions on how to find solution for the noisy and redundancy problem without eliminating important sentences and which order should the extracted sentences be. Various approaches to arrange the extracted sentences contain sentence-position in the documents and the sentence order according to a support vector machine (SVM) (ranging from the highest weight to lowest weight), chronological order of the actions in the extracted sentences (El-Haj, et al., 2011).

1.4 Research Questions

Based on previous discussion, the research question for this study is concerning Arabic text summarization techniques as the following:

How to find solution for the noisy and redundancy problem in Arabic text without eliminating important sentences and which order should the extracted sentences be?

1.5 Research Objectives

The aim of this study is to examine multi-document summarization salient information for Arabic text summarization task with noisy and redundancy information. Therefore, to answer the research question, the following objectives have been identified:

- To analyze Arabic text in order to remove the noisy information.
- To implement the cluster approach for cluster order and redundancy elimination.
- To select sentences based on the order of clusters generated in the second objective.
- To evaluate the final result summary by using Recall and Precision.

1.6 Scope of Study

This study focuses to examine an approach to multi-document summarization namely: cluster-based multi-document summarization approaches to improve the result in real–world documents (95 online newspaper articles in related field) as a corpus in Arabic language called Essex Arabic Summaries Corpus (EASC 1.0) for noisy information and reducing redundant sentences. Additionally, this study will produce a useful output for empirically analyzing the approaches for implementation of multi text summarization.

1.7 Significance of Study

The main goal is to test clustering approach for multi-document Arabic text summarization that can summarize a group of related text documents written in Arabic language. Successful summarization approach needs a good guide to find the most significant sentences that are applicable to a particular criterion. Therefore, the cluster algorithm should work on extracting the most significant sentences from a set of related documents.

Additionally, the impact of this study is to guide researchers to better comprehend the tested text summarization approach in order to further improve results that can contribute to well rendering in the future.

1.8 Thesis Organization

This report of this research is organized into five chapters which include introduction, literature review, methodology, results & discussion, and conclusion. The following are the summarized contents for chapter one.

Chapter one presents the study background, problem statement, research questions, objectives, scope of the study, significance of the study, and research organization.

The rest of the organization of this thesis is as follows: Chapter two presents the detailed background of the various summarization techniques and methods and shows the key area of related work, and also this chapter also gives a detailed background on processing tools and Arabic language.

Chapter three presents a framework for automatic text summarization, and this chapter illustration how to solve for noisy and redundancy problem in Arabic language multi-document summarization, in addition to the summaries methodologies.

Chapter four presents implementations of Arabic extractive multi-document summary, moreover, this chapter also illustrations the evaluation results of the work done on multi-document by using Recall and Precision measure.

Finally, Chapter five concludes the entire work and discusses some limitation and give direction for future work in the field.

1.9 Summary

This chapter highlights the important research ideas based on the problem discussed. The objective of this study is to examine the text summarization cluster approach to reduce noisy and redundant sentences information. The scope of this study is limited to textual information in Arabic language environment.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter presents a general overview of the multi document summarization as well as some general approaches for summarization, Natural Language Processing for Arabic language, summarization system for Arabic and multi-document summarization evaluation measure.

2.2 Multi-Document Summarization

Multi-document summarization is an automated procedure to extract information from multiple text documents written on the same topic. Summary results can be used by individual users, for example professional information to consumers, to make this information familiar in a wide range of documents the automatic documents summarization was started before 55 years ago (Luhn, 1958).

The multi-document summarization function is much more difficult than summarizing a document and it's a very large function as this complexity arises from thematic diversity within a large set of documents (Christensen, Mausam, & Etzioni, 2013). The present execution contains development of a summarization technique that combines documents clustering and sentences clustering in these documents. The results achieved are important in reducing redundancy and efficiency to a large extent by using

Correctness and Performance measures for example F-measure, Precision and Recall. the results based on their new method which content from three phases neural network training, feature fusion, and sentence selection outperforms it reduces redundancy through the clustering (Deshpande & Lobo, 2013).

2.3 Multi-document text summarization approaches

There are many approaches have been examined, with considerable overlap between documents summarization approaches and methods (El-Haj, et al., 2011). The method of extraction for the text summarization is processing by many approaches. The rest of the following sections is focusing on the related works in different approaches, techniques, tools, and models used for automatic document summarization (Suanmali & Salim, 2009).

2.3.1 Machine -based approach

Machine learning Summarization approach have been used for text mining summarization, classification, and sentence ranking (Amini & Usunier, 2007; Fisher & Roark, 2007; Wang, Raghavan, Castelli, Florian, & Cardie, 2013).

A popular supervised learning method is Support Vector Machine (SVM) that recognizes and analyses data patterns. The method was applied on automatic summarization to rank sentences. On the other hand, The machine learning approach for text summary has the main advantage is that allows testing the features of high number of performance, for example statistical, lexical, and syntactic. Machine learning paradigms are used in various ways to learn which the most appropriate ones are. However, this approach also needs a fairly large training corpus in order to be able to get crucial results. Typically, the corpus contains of annotated source documents containing which sentences are important for the summary and which not or a set of human-written summaries. (Yang et al., 2011).

Schilder, Kondadadi, Leidner, and Conrad (2008) introduced FastSum a query-based summarizer based on word-frequency clusters features, documents and topics. As a query-based summarizer FastSum ranked the extracted sentences using regression Support Vectors Machine. On the other hand, the FastSum system steps are pre-processing and filtering, feature set, and training. They used Recall, Precision and F-measure to evaluate their work.

Using machine approach in Arabic language automatic text summarization is now starting to attract more attention. Boudabous, Maaloul, and Belguith, (2010) presented an automatic summarization method for Arabic documents. The method was based on applying a numerical approach that used a semi–supervised learning technique. They used SVM for the learning process. They performed a comparative study, using human experts, to evaluate their summarizer.

Ouyang, and Lu (2011) proposed a multi-document summarization systems based on an SVM model. The model was used to automatically combine features and sentences scores prior to summarization.

However, based on previous discussed this part is analyzed the machine -based approach based on the main idea proposed, the language(s), the system used, the evaluation measure, and the experimental results. The first research used Mono-Lingual English, the evaluation measure used F-measure43, Recall 33 and Precision 46, the experiment results came with category 1 the full FastSum system with aggressive filtering using all features, category 2 the FastSum system after feature engineering via LARS plus aggressive filtering, and category 3 A simple first sentence baseline with redundancy removal based on cosine similarity. And the main idea is to update summarization of English multi document Summarization (Schilder, Kondadadi, Leidner, & Conrad, 2008). The second paper used Mono-Lingual Arabic; the evaluation measure used ROUGE, the experiment results came with applied learning phase which rely on support vector machine algorithm. And they used their system called AIS (Arabic Intelligent Summarizer). And the main idea is based on applying a numerical approach that used a semi–supervised learning technique (Boudabous, Maaloul, and Belguith, 2010). The third research used Mono-Lingual English, the evaluation measure used ROUGE, the experiment results came with applied a varies type of learning models, called regression models, for query-focused multi-document summarization. And the main idea is to estimate the sentence significance in a document set to be summarized through a set of pre-defined features (Ouyang, and Lu,2011). Table 2.1 shows Machine -based approach summary.

Table 2.1: Machine -based approach summary.

Author (s) & Year	Language (s)	Technique (s)	Evaluation
(Schilder, Kondadadi, Leidner, & Conrad, 2008)	English	FastSum a query-based summarizer	Recall, Precision and F-measure
(Boudabous, Maaloul, & Belguith, 2010)	Arabic	semi–supervised learning	ROUGE
(Ouyang, & Lu, 2011)	English	Support vector machine	ROUGE

2.3.1.1 Clustering based approach

Clustering of data is the assignment of a set of observations into subsets, named clusters. Clustering has been applied to documents, sentences and words. As shown in Figure 2.1, clustering can in general be grouped into partition clustering and connectivity-based clustering (Lloret & Palomar, 2012).

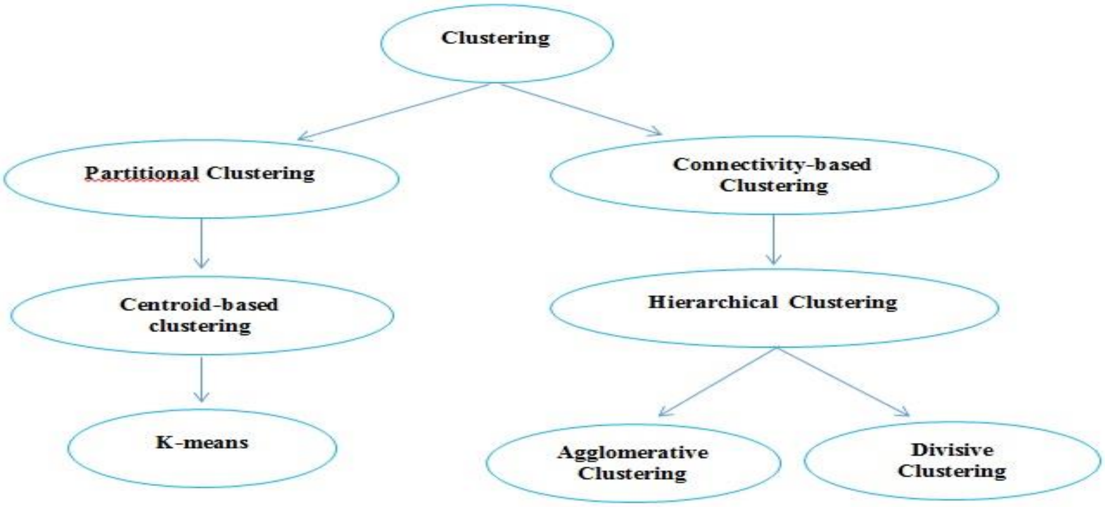


Figure 2.1: Techniques of Clustering (Kaur & Bhathal, 2013; Rai & Singh, 2010).

Liu, He, Ji, and Yang (2006) presented a cluster-based approach for Chinese multi-document summarization. It basically contains two stages: sentence clustering and selection of sentence. To cluster sentence, they propose two approaches for determining the number of automatic clustering: the first approach produce whole usage of the summary length constant via the user whereas the second approach is stabilization based, it can conclude the optimal number of cluster automatically. To select sentence, they show a global search approach, they choose a sentence based on its contribution to the rendering of all summary, and compared this approach with another local approach Term Frequency (TF), centroid sentence and Term Frequency Inverse Document Frequency (TF-IDF). To evaluate the summarization process, they suggest an extrinsic evaluation approach that relies on a classification job. Moreover, the approach of global sentence search and the number of automatic clusters discovery approach is useful to enhance the quality of summary as shown Figure 2.2, which illustrates the main idea of cluster approach.

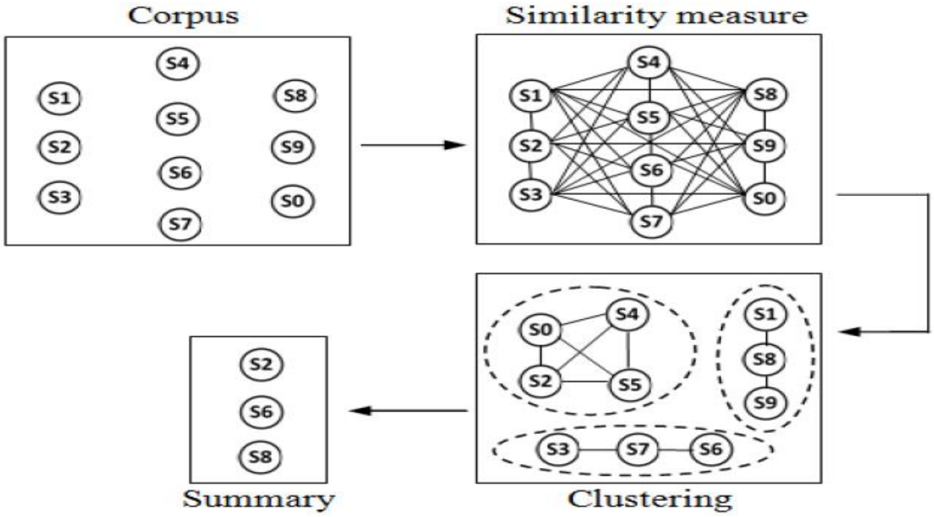


Figure 2.2: Cluster based summarization (Kumar & Salim, 2011).

Wan and Yang (2008) presented two models the first model is the Cluster-based HITS Model (Cluster HITS) (Hyperlink-Induced Topic Search), which regarded as the sentences as authorities and clusters as hubs in the HITS algorithm. The second model is Cluster-based Conditional Markov Random Walk Model (Cluster-based CMRW), which regarded as combines the cluster-level information into the link graph. These models rely on link analysis techniques. The whole multi document summarization outline contains from three stages: 1) Theme cluster detection 2) Sentence score computation 3) Summary extraction. To evaluate the summarization process, they suggest ROUGE (Recall-Oriented Understudy for Gisting Evaluation). The Cluster-based HITS Model is Validation to be more weakness from the cluster-based Conditional Markov Random Walk Model.

Agarwal, Reddy, Gvr, and Rosé (2011) presented a system called SciSumm to process scientific articles in multi-document summarization. The summarization is shown in the topic labeled clusters form, which offer article search based on the interest user topic. The way to get on final summary based on query to generate a summary called query-oriented fashion. In this paper suggest system SciSumm has four main modules: 1) TextTilling module: based on TextTilling algorithm; 2) Clustering module: Term Frequent based on text clustering algorithm; 3) Ranking module: the clusters are ordering based on the important for generating query by ranking module; and 4) Summarization presentation module: This module is used to show the ranked clusters to find the ranking module. The evaluation displays the SciSumm system for content selecting preferable another multi-document summarization system for multi document summarization.

Deshpande and Lobo (2013) presented a novel approach that outperforms the other approaches and it decreases redundancy by using clustering. The first clustering is based approach which groups, the similar a group of document into clusters and then sentences from each document cluster are clustered into clusters of sentence. The best scoring sentences from sentence clusters are chosen into the final summarization by find similarity between each sentence "cosine similarity measure" is used. Their results are evaluated using various extractive techniques by using correctness and Performance measures for example F-measure, Recall and Precision.

The typical partition algorithm that works well only on datasets that are isotropic clusters is K-means algorithm. This algorithm is popular because it is not complex to implement. Moreover, k-means would work well with large datasets (Rai & Singh, 2010). The k-means algorithm steps are given below (Rai & Singh, 2010):

- The k point will select randomly the k cluster will use to determine the centroids point.
- Determination each objects to the centroid closest to the other object in this way k exclusive cluster of object.
- New centroids of the clusters are calculated. For that reason average all attribute values of the objects belonging to the same centroid.
- Then the algorithm checks if the cluster centroids have changed. If yes start again. If not, cluster detection is finished.

The most popular cluster algorithm is K-means algorithm which is used to generate the summary, and also used in industrial and scientific applications. The cluster-based K-means algorithm advantage when using is that allows clustering the texts quickly. Moreover, clustering in automatic text summarization can be important for both selecting and extracting relevant sentences and eliminating redundancies. This algorithm works on initial the centroid points randomly based on the numbers of K groups, each object must contain to exactly one group and each group must belong at least one object. Moreover, the most important thing in K-means algorithm it’s how to decide the best number of K. Based on El-Haj and Hammo (2008), the approved the small number of cluster (one or two) when clustering a chosen sentence set to final summary is better than using five clusters or above. These objects are presented by TF and IDF (present the most important words in the original text). Then the cosine similarity measure is used to compare between them to select the significant object in the final summary (Deshpande & Lobo, 2013; Ghwanmeh, 2005; Gupta & Lehal, 2010; Kaur & Bhathal, 2013).

Schlesinger, and Conroy (2008) presented a multi-document summarizer system that used K-means clustering algorithm, in addition to other statistical models, to generate multi-document summaries for both Arabic and English languages.

Wan and Yan (2008) implemented a multi-document summarization technique using cluster-based link analysis. They used three different clustering detection algorithms including divisive clustering, k-means, and agglomerative, to generate multi-document summaries for English languages.

However, based on previous discussed this part is analyzed the clustering based approach based on the main idea proposed, the language(s), the system used, the evaluation measure, and the experimental results. The first research used Mono-Lingual Chinese, the evaluation measure used F-measure, Recall and Precision, the experiment results came with their summarization steps are efficient and effective for text summary. Furthermore, the automatic clusters number discovery method and the method of global sentence search are usefully to enhance the quality of summary, and the main idea is to enhance the performance of Chinese multi document Summarization (Liu, et al., 2006). The second research used Mono-Lingual English, the evaluation measure used ROUGE (Recall-Oriented Understudy for Gisting Evaluation), the experiment results came with the DUC2002 and DUC2001 datasets exhibit the better models efficiency and the cluster based Conditional Markov Random Walk Model (CMRWM) is approved to be more powerful from the Cluster-based HITS Model, and the main idea is two models were proposed in the procedure of sentence ranking. First is to join the cluster-level information into the link graph. Second is to consider the clusters and sentences such as authorities and hubs in the HITS algorithm to score sentences by (Wan & Yang, 2008). The third paper used Mono-Lingual English, the system used is SciSumm Agarwal, et al. (2011), the evaluation measure by F-measure, Recall and Precision, the experiment results came with the results showed that their systems perform importantly better three of the metrics ($p < .05$). On two metrics of SciSumm system executes marginally well ($p < .1$), and the main idea is the technique produces a summary in a query-oriented fashion with an unsupervised method so-called (SciSumm). The proposed method has four principal modules: clustering, ranking, text tilling and summery presentation by

(Agarwal, et al., 2011). The fourth article used Mono-Lingual English, the evaluation measure by F-measure, Recall and Precision, the experiment results came with the results illustration that their new method “Document and Sentence Clustering based Text Summarization” outperforms than other two methods document clustering and based on statistical features, and the core idea is compared the results developed by different extractive summarization techniques by using Performance and correctness measures by (Deshpande & Lobo, 2013). The Table 2.2 shows cluster approach summary.

Table 2.2: Cluster based approach summary.

Author (s) & Year	**Language (s)**	**Technique (s)**	**Evaluation**
(Liu, He, Ji, & Yang, 2006)	Chinese	Divisive clustering, k-means, & agglomerative algorithm	Recall, Precision& F-measure
(Wan & Yang, 2008)	English	(Cluster HITS) & (Cluster-based CMRW)	ROUGE
(Schlesinger, &Conroy, 2008)	Arabic and English	K-means clustering algorithm	ROUGE
(Agarwal, Reddy, Gvr, & Rosé, 2011)	English	SciSumm	Recall, Precision& F-measure
(Deshpande & Lobo, 2013)	English	Cluster	Recall, Precision& F-measure

2.3.2 Sentence co-relation based approach

Hariharan, Ramkumar, and Srinivasan (2012) examined and focused on two graphical approaches for multi document summarization namely: SentenceRank (Continuous) and SentenceRank (threshold) by Erkan and Radev 2004. This paper suggests improvements the above work examined two approaches by combining two more features to the current one. The first discounting approach was presented to form a summary which confirms less redundancy in sentences. The second location weight technique has been adopted to preserve significance based on the location they take. Two dataset have been evaluation by intrinsic method. Data set 1 has been produced manually from the 50 news paper documents together by them. Data set 2 is commercially available from DUC 2002 data. They used recall and precision parameters to evolution.

Tiedan Zhu (2012) shown the logical closeness criterion, this can be used for measuring the similarity among two sentences. Rely on the logical closeness, they propose an enhanced agglomerative algorithm to display the sentences order. Evaluation their augmented algorithm displays a development of the ordering over another baseline strategy. This paper highlighted on logical-closeness Instead of topical-closeness which is rely on synonymy and not powerful enough to measure the sentences coherence. They get results in their study based on the DUC 2006. The DUC 2006 datasets contain 50 document sets of various subjects and each subject contains 25 news documents. They select 20 such as testing datasets and 30 such as training datasets.

In other hand, based on previous discussed this part is analyzed the sentence co-relation based approach based on the main idea proposed, the language(s), the system used, the evaluation measure, and the experimental results. the first paper Hariharan, et al. (2012) used Mono-Lingual English, the evaluation measure by Recall and Precision, the experiment results came with Sentence rank method produces good results for both the datasets to evaluate measures, and the main idea is a link among two sentences is considered such as a vote cast from one sentence to another sentence. Sentences will be extracted rely on position, scores, casted votes, etc. to get the summary by (Hariharan, et al., 2012). The second research used Mono-Lingual English, the experiment results came with illustration that their method is impact for automatic summarization methods, and the main idea is highlighted on logical-closeness rather than topical-closeness which is rely on synonymy and not strong sufficient to measure the sentences coherence by (Tiedan Zhu, 2012). The Table 2.3 shows this approach summary.

Table 2.3: Sentence co-relation based approach summary.

Author (s) & Year	**Language (s)**	**Technique (s)**	**Evaluation**
(Hariharan, Ramkumar, &Srinivasan, 2012)	English	SentenceRank (Continuous) and SentenceRank (threshold)	Recall, Precision& F-measure
(Tiedan Zhu, 2012)	English	logical closeness criterion	

2.3.3 Time based approach

Wan (2007) presented the TimedTextRank algorithm for making the interim documents information relies on graph-ranking based algorithm. It was a preliminary study to prove the effectiveness of the suggested TimedTextRank algorithm to dynamic multi-document summarization. The TextRank algorithm makes of the connection among sentences and chooses sentences based on the recommendations or votes from nearby sentences, which is similar to HITS and PageRank. Then builds a similar graph to reproduce the relationships between wholly sentences in the set of document, based on the similar graph can compute the informativeness for each sentence score. The informativeness of sentence refers to what is the amount of information about the main topic in sentence. The sentences selected into the summary are with the highest informativeness scores in order to keep less redundancy in the final summary as possible. To evaluate the suggested TimedTextRank algorithm in real topic discovery system using user study. Then they recorded the publication time for each real topic. Then divided into the 5-point scale for each summary of each real topic from 9:00 am to 5:00 pm.

However, based on previous discussed this part is analyzed the time based approach based on the main idea proposed, the language(s), the system used, the evaluation measure, and the experimental results. paper by (Wan, 2007) used Mono-Lingual Chinese, and the main idea is The improvement of TextRank is unveiled called TimedTextRank with joining time dimension. This is relying on the show that for an evolving topic, new documents are usually more significant than previous documents. Table 2.4 shows Time based approach summary. On the other hand, the concepts of the

Time based approach based on previous documents have little bit important then new documents.

Table 2.4: Time based approach summary.

Author (s) & Year	Language (s)	Technique (s)	Evaluation
(Wan, 2007)	Chinese	TimedTextRank algorithm	Time

2.3.4 Graph based approach

Erkan and Radev (2004) presented graph-based approach in natural language processing. This research of multi-document summarization was focus, where the goal of this paper to generate multiple documents summary for related documents. The graph-based approach was used for determining sentence salience rely on centrality scoring of sentences. The centrality work is selecting the central sentences from the original text. The documents cluster show the sentences network which are related. The sentence which is more similar with other sentence this consider as salient or salient. The centrality definition is examining similarity among two sentences. To find the similarity they used bag of words model to present the sentence as a vector. Then they used MEAD system to implement their approach. They used 30 clusters in DUC 2003 then used ROUGE to evaluate their approach. See the Figure 2.3 which illustrates an example for this approach.

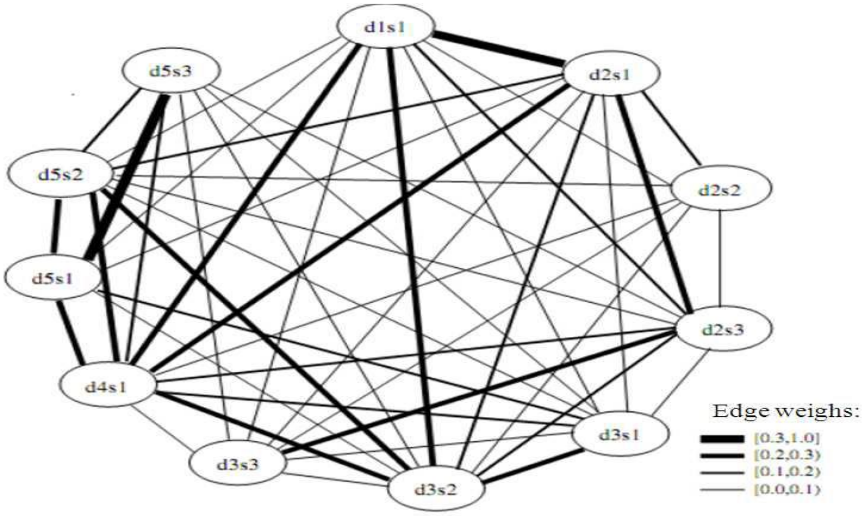


Figure 2.3: multi document node represents a sentence an example graph (Kumar & Salim, 2011).

Wan (2008) identifies the influence of document on the graph-based model for multi document summarization. The sentence-to-document and information of document-level the relationship is united into the algorithm of graph-based ranking. The graph-based model is basically a way of determining the significance of a peak within a graph based on overall information recursively drawn from a one-layer link sentences graph. The based graph document model is combined here to identify the impact of document by discovering sentence-to-document correlation and document significance into the ranking of sentence process. The experimental results on DUC2002 with 59 document sets and DUC2001 with 30 document sets prove the good efficiency of the suggested model.

However, based on previous discussed this part is analyzed the graph based approach based on the main idea proposed, the language(s), the system used, the evaluation measure, and the experimental results. The first article Erkan & Radev. (2004), used Multi-Lingual Arabic and English, the system used is MEAD, the evaluation measure by ROUGE, the results came with the centroid-based summarization comes with a good results. Furthermore, ROUGE result for various policies of MEAD system on 17% noisy DUC 2003 and 2004 data, and the main idea is to generate multiple documents summary for related documents. The graph-based approach was used for determining sentence salience rely on centrality scoring of sentences by (Erkan & Radev, 2004). The second paper used Mono-Lingual English, the evaluation measure by ROUGE, The experimental results on DUC2001 and 2002 exhibit the proposed model came with the better effectiveness, and the main idea is Two-link graph are both sentences and documents. It is assumed that the sentences which belong to a significant document, extremely correlated with the document, will be more probable to be selected into the final summary by (Wan, 2008). Table 2.5 shows graph based approach summary.

Table 2.5: Graph based approach summary.

Author (s) & Year	**Language (s)**	**Technique (s)**	**Evaluation**
(Erkan & Radev, 2004)	English	MEAD	ROUGE
(Wan, 2008)	English	graph-based ranking algorithm	ROUGE

2.3.5 Statistical based approach

Many summarization systems rely on statistical approach to extract relevant sentences (Berger & Mittal, 2000; Galanis & Malakasiotis, 2008).

Schlesinger, et al. (2008) presented CLASSY (Clustering, Linguistics, And Statistics for Summarization Yield), which is an automatic summarization to generate system which uses statistical and linguistic trimming methods to produce general summaries. CLASSY is an automatic summarization system, established for summarizing English documents. It uses trimming rules to shorten sentences in the document, examines sentences as being more or less likely to be involved in a summary, produces a summary for each document, selects sentences for a multi document summarization for a clustering of related documents, and then orders the chosen sentences for the final summarization. They used Document Understanding Conference DUC evaluations their final summary.

M. El-Haj, et al. (2011) show their general extractive Arabic and English multi document summarization. The produced Arabic multi document summarization using English extractive gold standards is evaluated by using machine translation. The summarization quality does not appear to affect by the translation procedure. In this paper the presenter three different models for delete the noisy information and redundancy elimination namely: Dice's Coefficient, Vector Space Model (VSM) and Latent Semantic Analysis (LSA). They collected 59 newspapers and newswire are related. They used ROUGE for evaluation their work.

Gupta, Chauhan, Garg, Borude, and Krishnan (2012) presented new tools for statistical approach to extract important text from a document set this tool called Kernel score text called KernelSum (KERNEL SUMMarizer). Kernel is examined as the most important paragraph of the source text. This approach is used to find the words frequency, ranking the sentences, scoring the sentences, to select the most important sentence. The Kernel-based composed the following efficient components namely:

1) Text pre-processor works on converting the Word Documents or HTML to plain text.
2) Separator of sentence split the sentences rely on some rules as ending point such as space and dot etc.
3) Separator of word isolates the words through some principles such as space.
4) Stop-words list delete in the regular English language such as the, an, a, from and of.
5) Frequency of word calculator computes the number of times a word appears in the document.
6) Using the TF-ISF (Term Frequency - Inverse Sentence Frequency) for scouring each sentence.
7) Algorithm ranking sums rank of each sentence according to the scores, heading sentence, length and location.
8) Summarization portion is chosen the sentences from the ranked list to get on the expected short of the original document.

Finally, they evaluate the Kernel-based system by KernelSum.

Azmi and Al-Thanyyan (2012) presented a new Hybrid approach when Integrates a sentence scoring scheme and Rhetorical Structure Theory (RST). They used RST for generating a primary summarization. In the primary summary assign a score to each sentence to the final summary. Then select sentences out of the incipient summarization with the maximizing the total score of the summarization. They used algorithm in eight steps namely: sentence segmentation, word segmentation, stop-words removal, root extraction, frequency computation, generation of the primary summary, sentence scoring, and finally they produce the summarization within the user specific size. Their data samples consist of 32 different documents of news articles. They used ROUGE (Recall and Precision) a commonly metric, for evaluating their automatically produced summarization.

Haboush and Al-Zoubi (2012) presented a new automatic Arabic text summarization model. The technique in this model is based on root word clustering. Their model implements root word cluster weight instead of the word weight itself. This model has seven steps namely: fed the document into the model; the model splits the original text into paragraphs number; then splits paragraphs to sentences; and sentences to words; the next step is to execute stemmer that discovers the root of each word in each sentence in the original text. Then find the Weight of each word, and then the model computes the score of each sentence, in Arabic the model select the important word for example (the most important thing: اهم الامور) then the sentence score is increase. Finally, the model takes the sentences with the maximum scores. With the resulting structure, two measures of Recall and Precision are used to evaluate.

Ibrahim and Elghazaly (2012) presented a Rhetorical Structure Theory RST such as a major natural text structure. It is applied for many languages such as English, Japanese and French but the RST still lack in the Arabic language. The researches were interested to analyze the Arabic grammars. The Rhetorical Structure Theory used to generate the relationship as a tree between the paragraphs. The used three statistical phases are implemented till get on rhetorical relations hypothesis in Arabic namely: Matching Relation Cue Phase, Statistical Verification Phase and Manual Review Phase. They used the statistical results to distinction between satellite and nucleus paragraphs. They process 30 articles to evaluate these articles used three measures namely: F-measure, recall, and precision.

Ibrahim, et al. (2013) presented a new hybrid summarization model for Arabic language, merging between Vector Space Model (VSM) and Rhetorical Structure Theory (RST). They used VSM to rank the important paragraphs relying on the feature of cosine similarity. Meanwhile they used also RST for discovering the important paragraphs rely on semantic criteria and functional. The framework of this research content from two parts the first one is The RST Sub-Model.

The second part is the support vector machine VSM Sub-Model content from many steps namely:

1) Vector Representation Process:

   Used the tf-idf weights to transfer the original text to vectors.

2) Length-Normalize:

   For completing the process of length normalization they divide the process of normalization into vector of title V (t) and vector of paragraph V (Ps).

3) Cosine Similarity:

The angle θ cosine among vectors (title V (t) and paragraphs V(Pn),

4) Ranking Paragraphs:

The final step the model will select the paragraph with high score.

Finally, they used precision to evaluate their work then they found when used this hybrid model can take the both advantages.

However, based on previous discussed this part is analyzed the statistical based approach based on the main idea proposed, the language(s), the system used, the evaluation measure, and the experimental results. The first research used Multi-Lingual Arabic and English, the CLASSY system is used, the evaluation measure by ROUGE, The experiment results show when presented documents group in both Arabic and English or any other language, using signature terms computed from by CLASSY system, for both the machine translate (MT) for Arabic documents and English, to generate quality summarization, and the main idea is (Clustering, Linguistics, And Statistics for Summarization Yield) CLASSY is an automatic, extract-generating, summarization system that uses statistical methods and linguistic trimming to produce generic summaries by (Schlesinger, et al., 2008). The second paper used Multi-Lingual Arabic and English, the evaluation measure by Recall and Precision, The implication testing exhibited the experiment results, by ROUGE measure, there are no implications among the Arabic and English summarizers, the main idea is the generic extractive Arabic and English multi-document summarizers. And the main found of this work is that automatic machine translation of English datasets into Arabic is a viable and economic alternative to the manual creation of Arabic datasets by (M. El-Haj, et al., 2011). The third paper used Mono-Lingual English, the Kernel-based system is used, the evaluation measure by KernelSum, the results show used this system to get on more coherent summarization, and the main idea is by simple statistical measures, Kernel is examined such as the most important passage of the source text that contains most frequent terms. It helps for example the guideline to select the other sentences for summary (M. V. Gupta, et al., 2012). The fourth research used Mono-Lingual Arabic, the RST-based system is used, the evaluation measure by ROUGE, and the main idea is offered a system for automatic extractive Arabic text summarization whereas The final summary can specify the size by the user (A. M. Azmi & Al-Thanyyan, 2012). The fifth used Mono-Lingual Arabic, the evaluation measures are Recall and precision, the results show, a suitable summarization levels have been recorded with the Recall 0.787 to Precision of 0.757 averages. Similar results of study used Arabic documents gave scores of 0.62 to 0.70, and the main idea is a new automatic Arabic text summarization model is discussed and offered (Haboush & Al-Zoubi, 2012). The sixth research, used Mono-Lingual Arabic, the evaluation measure by F-measure, Recall and Precision, The experimental results illustration the satellite paragraphs represent 80% and core paragraphs representing 20% from whole paragraphs inside corpus, and the main idea is Provides a framework to apply RST in Arabic, in order to enhance the ability of extracting the semantic behind the text (Ibrahim & Elghazaly, 2012). The final paper, used Mono-Lingual Arabic, the evaluation measure by precision, the hybrid model experimental results the average precision for output summarization is 71.6% which was 56.3% by RST, and the main idea is Shows a new hybrid model for Arabic text summarization, combining Vector Space Model (VSM) and Rhetorical Structure Theory (RST). The proposed model uses VSM for ranking the important paragraphs. The proposed model uses RST to discover the most important paragraphs (Ibrahim, et al., 2013). Table 2.6 shows this approach summary.

Table 2.6: Statistical based approach summary.

Author (s) & Year	**Language (s)**	**Technique (s)**	**Evaluation**
(Schlesinger, et al. 2008)	Arabic & English	CLASSY	ROUGE
(M. El-Haj, et al. 2011)	Arabic & English	Dice's Coefficient, Vector Space Model (VSM) and Latent Semantic Analysis (LSA)	ROUGE
(Gupta, Chauhan, Garg, Borude, &Krishnan, 2012)	English	Kernel	KernelSum
(Azmi and Al-Thanyyan, 2012)	Arabic	sentence scoring scheme & rhetorical structure theory	ROUGE
(Haboush and Al-Zoubi, 2012)	Arabic	Arabic text summarization model	Recall & Precision
(Ibrahim & Elghazaly, 2012)	Arabic	Rhetorical Structure Theory	F-measure, recall, & precision
(Ibrahim, et al. 2013)	Arabic	Vector Space Model & Rhetorical Structure Theory	precision

Table 2.7 shows the advantages and dis advantages for Machine -based approach, Clustering based approach, Sentence co-relation based approach, Time based approach, Graph based approach, and Statistical based approach.

Table 2.7 advantages and dis advantages for each approach.

Approaches	Advantages	Disadvantages
Machine -based approach	Allows testing the features of high number of performance	No
Clustering based approach	Summarizing multi document process effective. Clustering is better for reducing the number of redundant features. Clustering the texts quickly. Clustering in automatic text summarization can be important for both selecting and extracting relevant sentences and eliminating redundancies.	No
Sentence co-relation based approach	No	Link among two sentences is considered such as a vote cast from one sentence to another sentence. To depend on highlighted on logical-closeness rather than topical-closeness which is rely on synonymy and not strong sufficient to measure the sentences coherence.
Time based approach	No	The concepts of this approach based on previous documents have little bit important then new documents.

Graph based approach	No	It is assumed that the sentences which belong to a significant document, extremely correlated with the document.
Statistical based approach	No	Used machine translate (MT).

Hence, clustering based approach is chosen due the advantages of these techniques, and work on this approach obtains the better outcome for Arabic language.

2.4 The Arabic Natural Language Processing (ANLP)

The united nation considers Arabic as one of the six formal languages. The Arabic language alphabet contains 28 letters (أ ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي) and the direction of writing is from right to left (Albared, Omar, & Ab Aziz, 2009; Saad & Ashour, 2010). The Arabic native is nearly 300 million people in twenty-three countries, but so far, the number of research in natural language process (NLP) of Arabic language is less than other language such as English. Characteristics of the Arabic language are for providing the reader with the main knowledge necessary to understand the problems caused by the complex structure for this language. This is important when creating these problems obstacle to applications such as information retrieval and text clustering (Habash, 2010). The characters (letters) in Arabic that is associated with each other to form words. These characters contribute to four important types of words that contribute to the formation of sentences in Arabic language noun, verb, Adverbs and adjectives (Habash, 2010).

The name is a word name format (language) or more are used for identifying a person or an animal or other so as to distinguish it from the other. Names may be used for the definition of a group or a particular class is also used to identify individuals. Profile name is the name used to identify a particular person. Not limited to the use of the names on the rights, but extends to a wide range of animals, plants or objects. The verb is word indicate on an event that is associated with time. The verb is divided into past, present and future (Habash, 2010).

The most challenging of Arabic language in natural language processing, and attributed to four main reasons (Azmi & Al-Thanyyan, 2012):

1- Can write a set of characters in various ways and this based on the character (letter) position in the words.

2- The Arabic language is derivative and diacritical. Making Morphology difficult task.

3- Broken plurals are common: In linguistics, a broken plural is an irregular plural form of an adjective or noun found in the Semitic languages.

4- Words of Arabic are often unclear because to the system of tri-literal root in depend on features in Arabic, some complex processing in language takes long time as in natural languages when compared with the accomplishment in other Asian languages and English. The nature of these languages, although there is highlighted by Arabic language writing from right to left, capitalization for identifying proper names, abbreviations and acronyms. In addition, it has worth with machine readable dictionaries, lexicon, and corpora, which are trend to forward research in the various areas (Hammo, Abu-Salem, & Lytinen, 2002). To identify the words in Arabic language it is needful to identify the root of these words. In most cases the roots of words in the Arabic language is made up of three or four letters. Although it may be for some root words consists of more than four letters. Therefore, the Arabic word roots prefix, infix and suffix can be gathered for building derivations set (Azmi & Al-thanyyan, 2009). It is value mentioning that it is difficult to get on the root for any word in the Arabic language and it requires morphological analysis and detailed text Grammar. Usually, the Arabic words may not be derived from the root of the word, may have its own structure. Meanwhile the root can be a base of various words with cultured related meaning. For instance the root "laaba" لعب is used for many words relating for "playing", including "player" "malaab" ملعب and "laaeb" لاعب.

For finding the root of the word by removing prefixes, infixes and suffixes, of some letters annexed to this word. These characters (letters) may be at the beginning, middle or end of the word. To delete these sub parts of the first word is correspond to the current basic structures such as rhythms. When the main structure is found, then can delete the sub parts and summaries the word to its root. Table 2.8 shows an instance for this deletion process. Therefore, in this instance the root of all the noted words (علمية، علمتنا، علماء، علوم، تعليم، علمه، استعلامية، مدرسات، دارسون، المدارس) after deleting subparts is the unique root of (علم، دارس) "dares, aalm".

Table 2.8: Various words have different sub part and same root.

Word	Prefixes	Infixes	Suffixes	Meaning
المدارس	م + ل + ا	أ	-	Schools
علمية	-	-	ي+ ة	Scientific
دارسون	-	أ	و+ ن	Scholars
علمتنا	-	-	ت + ن + أ	Learned us
مدرسات	م	-	أ + ت	Teachers
علماء	-	-	ا + ء	Scientists
علوم	-	و	-	Sciences
تعليم	ت	ي	-	Teaching
علمه	-	-	ه	His science
استعلامية	أ + س + ت	أ	ي + ة	Informative

2.4.1 Arabic text Summarization Systems

A system for Arabic multi-document summarization called Lakhas by Douzidia & Lapalme, (2004). Is using techniques of extraction in order to get on ten word summarization of news articles. They used similarity coefficient technique for weighting each word. Arabic language involved for this work. They used “Lakhas” for generating a short summarization (≤ 75 bytes). They used four sentence-reduction methods to make them shorter and compress sentences namely: Name substitution, Removal of some type of words, Removal of part of sentence following some boundaries and Removal of indirect discourse. They used DUC-2004 dataset to translate the English into Arabic by machine translation tool. The main concern over Lakhas is that the process of reduction leds to loss of worthy information.

The second system called Clustering, Linguistics, and Statistics for Summarization Yield (CLASSY) (Conroy, O’Leary, & Schlesinger, 2006) is an automatic, extract producing, system of summarization which uses statistical methods and linguistic trimming to produce query-driven/ topic or generic summarization for multi document or documents clusters. Multilingual Arabic and English document were tested by CLASSY. The performance of this system is not so well when applied to Arabic and English original documents (Azmi & Al-Thanyyan, 2012; Schlesinger, et al., 2008).

The third common available toolkit for extractive multi-document summarization is called The MEAD system. While, the default of this system comes with a centroid-based summarization, its feature set can be expanded for implementation to any other method. The MEAD system contains of three components 1) the feature extraction is used to change any sentence in the original text into feature vector, 2) the feature vector is used to combine the outputs based on the weights of feature, and 3) the reranker is used for determining the relation type among the sentences in the pair. There are three main features for the MEAD system are position, centroid and length. The policies of the MEAD summarization system are 1) the command lines for all features, 2) the formula for converting the feature vector to a scalar, and 3) the command line for the reranker. The main problems in the MEAD system is poor user interface and many aspects or unfamiliar concepts for example the menu entries and wording of button labels. (Radev, Blair-Goldensohn, & Zhang, 2001).

2.5 Multi-document Summarization Evaluation

Evaluating the consistency and quality of a produced summary has confirmed to be a challenge (Fiszman, Demner-Fushman, Kilicoglu, & Rindflesch, 2009). One reason is that, in general, there is no clear ideal summary. The system evaluation may help in solving this problem. Two metrics types have been developed: content metrics and form metrics. Content metrics are more difficult for measuring. However, the system output is compared unit by unit or sentence by sentence to one or more human generated ideal summaries. Form metrics focus on grammaticality, overall organization and text coherence. They are usually measured on a point (Fiszman, et al., 2009).

Evaluation of summary approaches effort falls for either determining how reliable and adequate or how suitable a summarization is close to the source. Actually, approaches evaluation is divided into two kinds: The first kind of evaluation approach is extrinsic. The summarization quality are determined by users based on how complete some task is in summary for example should answer the questions related to the original text (Das & Martins, 2007). The second kind of evaluation is intrinsic by directly investigating the summarization can the users determine the quality. The summarization should cover the main ideas, or when the author written a summary of an idea can compares with the source text. However, all these procedures can be satisfactory and in most cases there is no summary is ideal for a specific document (Sobh, et al., 2009).

ROUGE is used for evaluation summary for automatically generated summarization by calculating n-gram. A high overlap score should show a high score of shared ideas among the two summaries. Summarization extractive approach is classification to allow us to use F-measure, precession and recall to evaluate the final summaries (Larson, 2011).

Giannakopoulos, Karkaletsis, Vouros, and Stamatopoulos (2008) suggest AutoSummENG approach for automatic evaluation, recently developed which has been confirmed to have a high relationship with human judgments. This approach varies from the others in three main features: (1) the approach used to calculate the similarity among summaries, (2) the representation selected for this extracted information, and (3) the kind of statistical information extracted. Here, the comparison of representations for establishes a degree of similarity among the graphs, and then the comparison among summaries is carried out by building first n-gram character graphs.

DEPEVAL(summ) the idea is similar to basic elements, and it contains of comparing dependency triples extracted from automatic summaries against the ones from model summaries, which is a dependency-based metric The main variance with basic elements is the parser used. Whereas Basic Elements uses Minipar, DEPEVAL(summ) is tested with different parsers, for example the Charniak parser (Owczarzak, 2009).

Katragadda (2010) suggests GEMS (GenerativeModelling for Evaluation of Summaries) to use for signature terms in order to analyze how they are captured in automatic summaries. The signature terms are calculated on the basis of part-of-speech tags, for example verbs and nouns, reference summaries terms and query terms. The distribution of the signature terms is calculated in the source document and then the probability of a summary being biased towards for example signature terms is found.

2.6 Corpora

The working on automatic multi-document summarization needs resources, for example data sets in order to perform an experiment. These include documents collection together with gold-standard summaries (human summary). This human summary is generated by human expert. These data sets allow judging the summaries performance and quality (El-Haj, et al., 2011). Therefore, resources for example corpus are significant for researchers working on the Arabic language (Al-Sulaiti & Atwell, 2006). There are many corpus in Arabic, this research focuses on some of them and selected one for the purpose of experimentation.

2.6.1 CCA corpus

The corpus of Contemporary Arabic (*CCA* Corpus) was released from the Eric Atwell and University of Leeds by Latifa Al-Sulaiti (Al-Sulaiti & Atwell, 2006). Their survey confirms that the existing corpora are too narrowly limited in source genre and type, and that there is a need for a freely-accessible modern Arabic corpus covering a wide range of text kinds. The corpus contains 293 text documents belonging to 1 of 5 categories (health & medicine 32, stories 58, science 70, tourist & travel 60, and autobiography 73) .The corpus includes 95,530 district keywords after stop words removal (Al-Sulaiti & Atwell, 2004; Al-Sulaiti & Atwell, 2006).

2.6.2 BBC Arabic corpus

The Arabic corpus from BBC Arabic website (www.bbcarabic.com) the corpus includes 4,763 text documents. Each text document belongs to 1 of 7 categories (art &

culture 122, international press 49, world news 1489, middle east news 2356, business & economy 296, science & technology 232, and sports 219). this corpus contains 1,860,786 (1.8M) words and 106,733 district keywords after stop words removal (Saad, 2010).

2.6.3 Aljazeera corpus

The corpus includes 1,500 text documents. Each text document belongs to 1 of 5 categories (art, economy, politics, science, and sport), each category includes 300 documents. The corpus includes 55,376 district keywords after stop words removal (Said, Wanas, Darwish, & Hegazy, 2009).

2.6.4 Khaleej-2004 corpus

Khaleej-2004 corpus was collected from Khaleej newspaper of the year 2004. The corpus includes 5,690 text documents. Each text document belongs to 1 of 4 categories (Economy 909, Local News 2398, International News 953, and Sport 1430). The corpus includes 122,062 district keywords after stop words removal (Abbas, Smaili, & Berkani, 2009a, 2009b).

2.6.5 CNN Arabic corpus

The CNN Arabic corpus from CNN Arabic website (www.cnnarabic.com) the corpus includes 5,070 text documents. Each text document belongs to 1 of 6 categories (science & technology 526, middle east news 1462, world news 1010, sports 762, entertainments 474, and business 836).The corpus contains 2,241,348 (2.2M) words and 144,460 district keywords after stop words removal (Saad, 2010).

2.6.6 Open Source Arabic Corpus (OSAC)

The OSAC is from multiple websites. The corpus includes 22,429 documents. Each document belongs to 1 of 10 categories (economics 3102, history 3233, education & family 3608, religious & fatwas 3171, sports 2419, heath 2296, astronomy 557, low 944, stories 726, and cooking recipes 2373). The corpus contains about 449,600 district keywords after stop words removal and 18,183,511 (18M) words (Saad, 2010).

2.6.7 Essex Arabic Summaries Corpus (EASC)

The document collection used in the creation of the multi-document summaries corpus was extracted from the Arabic language version of Wikipedia and two Arabic newspapers Alwatan from Saudi Arabia and Alrai from Jordan. These sources were chosen for the following reasons:

- They cover a range of topics from different subject areas (such as politics, economics, and sports), each with a credible amount of data.
- They are written by many authors from different backgrounds.
- They contain real text as would be written and used by native speakers of Arabic.

The documents of Wikipedia were chosen by asking students group to search on the Wikipedia website for arbitrary topics of their choice within given subject areas. The ten theme areas were: health, environment, tourism, art & music, education, religion, finance, science & technology, sports, and politics. The corpus includes 153 documents with a total number of words 18,264. Each document contains on average 380 words, with a minimum word-count of 116 words and a maximum of 971 words. The corpus contains about 2,360 sentences, 41,493 words, and 18,264 district keywords after stop words removal. Moreover, this corpus also include on the Gold-standard summaries (human summary) (M. El-Haj, et al., 2011; M. O. El-Haj & Hammo, 2008).

However, the human summary task for the documents in EASC was published as "Human Intelligence Tasks" (HITs). The assessor (worker) in Arabic language was asked to read and summarize a given articles by selecting what they considered to be the most important sentences that should make up the extractive summary. The sentences were showed to the users as an enumerated list, the sentences were numbered so the users could select the sentence numbers they believe should be in the summary. They were required to choice no more than half of the sentences in each article (El-Haj, et al., 2011; El-Haj & Hammo, 2008). This research will select the EASC as a data set to use in our experimental task because this corpus contained on the human summary, and this summary will help us to evaluate the final automatic summary.

2.7 Summary

This chapter presented a general overview of the multi document summarization general approaches namely: Clustering based approach, Sentence co-relation based approach, Time based approach, Graph based approach, statistical based approach, and machine based approach. Then it discusses natural language processing for Arabic language, summarization systems for Arabic language for example Lakhas , CLASSY and MEAD, presented many corpus in Arabic language, and finally, review two types of multi document summarization evaluation measure namely intrinsic and extrinsic.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

This chapter describes the methodology used in this research focusing on the multi text summarization approaches and tries to answer the research question and describes the way to achieve each objective.

3.2 Research Methodology

In this study, the methodology from (Radev, Hovy, & McKeown, 2002) as shown in Figure.3.1 is adopted. The methodology is conducted in five phases and is based on the objectives.

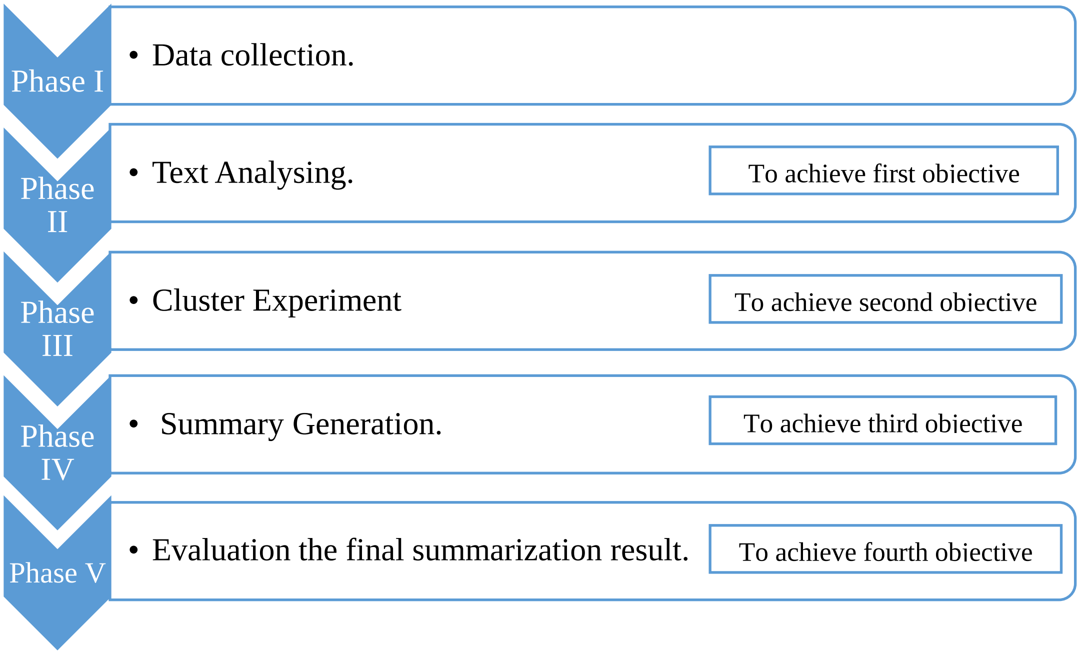

Phase I
• Data collection.
Phase II
• Text Analysing.
To achieve first objective
Phase III
• Cluster Experiment
To achieve second objective
Phase IV
• Summary Generation.
To achieve third objective
Phase V
• Evaluation the final summarization result.
To achieve fourth objective

Figure 3.1: Methodology phases (Radev, et al., 2002).

Phase I: Data collection

The data collection phase needs first determining what kind of data used in this research, and then gathering the data which content from ten categories. Each category contains ten documents except education, which contains seven documents and religion contains eight documents. Each document contains on average 380 words, with a minimum word-count of 116 words and a maximum of 971 words.. This study is looked to examine the corpus in Arabic language which generated by Mechanical Turk, called Essex Arabic Summaries Corpus (EASC) which is available in Essex University website. This corpus content from news related articles in Arabic language namely: health, environment, tourism, art & music, education, religion, politics, sports, finance, and science & technology.

Phase II: Text Analyzing

The second phase is to simplify text and this process comes with four main steps the first step is tokenization, the process of splitting the text into tokens (Algorithm 1) such as words, and involves the Arabic tokenization using additional Arabic punctuation characters for example the question mark (؟) which is not in the English language. The tokenize process used the following symbols: spaces and “.”, “,”, “؟”, “!”,“<”, “>”, “:”, to separate the original text into words (Azmi & Al-Thanyyan, 2012).

```
Input: Text Collection TC
Output: List of all extracted words $W_1$, $W_2$, ..., $W_h$ from TC
foreach D in TC do
    Begin;
    Split D into sentences;
    foreach S in D do
        Begin;
        Split S into words;
    End
End
$D_i$ represents the $i^{th}$ document in a Text Collection of n documents;
$S_j$ represents the $j^{th}$ sentence in $D_i$, m is the number of sentences in $D_i$ and j: 1 to m;
```

Algorithm 1: Algorithm of Tokenization (Azmi & Al-Thanyyan, 2012; El-Haj, et al., 2011).

The second deleted wholly stop words (Algorithm 2) from the text so that any text have only the nouns and the verbs. The stop words do not add any novel information to the textual (do not impact the meaning of the sentences if deleted). And it does not have a root, for example for these words are: (....في,من,هذا, هو, هي, الذي) (A. M. Azmi & Al-Thanyyan, 2012; Haboush & Al-Zoubi, 2012).

The third step implemented stemmer (Algorithm 2) which discoveries the root of each word in text in the original Arabic text. This means each word(verb or noun) in Arabic language has sub parts (infixes, suffixes, and prefixes) requirement to be deleted (A. M. Azmi & Al-Thanyyan, 2012; Haboush & Al-Zoubi, 2012).

The fourth step in this phase applied the TF–IDF (term frequency–inverse document frequency), to represent the significant words as a bag of word. The TF–IDF is a simple numerical statistic which reflects how significant a word is to a document in a corpus (Saad, 2010). For instance, consider a text containing 100 words wherein the word (كتاب) appears 3 times. Following the previously defined formulas, the term frequency (TF) for book is then 0.03 (3 / 100). Now, assume we have 10 million documents and (كتاب)seems in one thousands of these. Then, the inverse document frequency is calculated as log (10 000 000 / 1 000) = 4. The TF–IDF score is the product of these quantities: 0.03 × 4 = 0.12 (Davenport, 2012). In this phase will be represent the important vector words as (bag of word). Furthermore, in this phase we will achieve the first objective, and solve the noisy problem. See Figure 3.2 which includes on this phase steps.

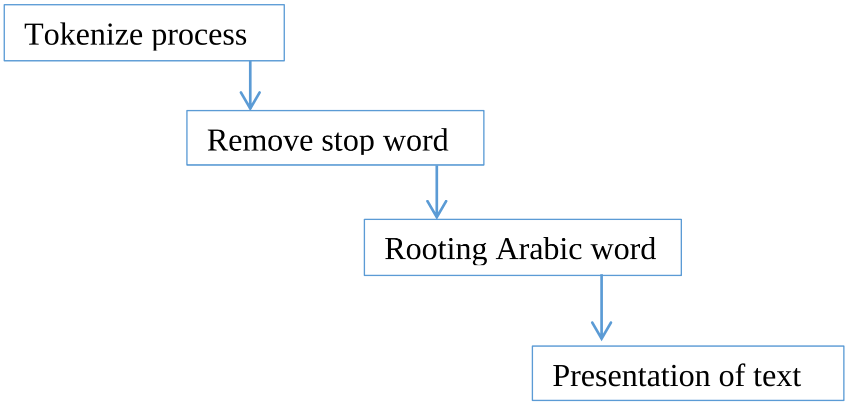


Figure 3.2: Phase II steps

```
Let G be the minimum-length of a word $W_a$ in a language V;

Let STOPS be the stop-word list for language V;

Input: Words extracted from the Tokeniser $W_1, W_2, ..., W_h$

Output: Stem of the input word

foreach $W \in S$ do
    Begin;
    If W Not In STOPS then
        Index W;
        If Length of $W<G$ then
            Word can not be stemmed;
        Else
            Remove prefixes, suffixes and infixes from W;
        End
    Else
        Do not index W;
    End
End
```

D_i represents the i^{th} document in a Text Collection of n documents;

S_j represents the j^{th} sentence in D_i, m is the number of sentences in D_i and j: 1 to m;

W_a represents the a^{th} word in S_j , b is the number of words in S_j and a: 1 to b. Words can be replaced with any other information (e.g., noun-phrases, named-entity definitions, etc).

Algorithm 2: Stop-word Removal and Stemming Algorithm (Azmi & Al-Thanyyan, 2012; El-Haj, et al., 2011)

The multi-document summarization usually ends up having redundant sentences in the generated summaries (El-Haj, et al., 2011). To address this problem this research chose to apply the cluster based approach.

This research focuses Arabic language cluster-based multi-document summarization techniques. The corpus EASC will use for Arabic language. In this research experiments will used a generic multi-document summarization for Arabic language, thus, this research is used clustering for redundancy elimination.

Phase III: Cluster Experiment

The third phase is to implement cluster algorithm (K-means) and this process comes with two main steps. The first step to implement K-means clustering is partitioned centroid-based clustering algorithm. This algorithm randomly chooses words as the initial centroid for each cluster. Then iteratively gives all words to the nearby clustering, and recalculates the centroid of each clustering, till the centroids no longer change. The main steps of K-means are 1) select k number of clusters to be determined, 2) select k objects randomly as the initial center of cluster, 3) assign each object to their neighboring cluster, Compute new clusters, i.e. Calculate mean points, and 4) until no alteration on centers of cluster (i.e. Centroids do not alteration location anymore) or no object changes its cluster sees Figure 3.3. The similarity among a word and a cluster centroid is calculated using the standard cosine measure that will apply to tokens within the word (Minaei-Bidgoli, Parvin, Alinejad-Rokny, Alizadeh, & Punch, 2014).

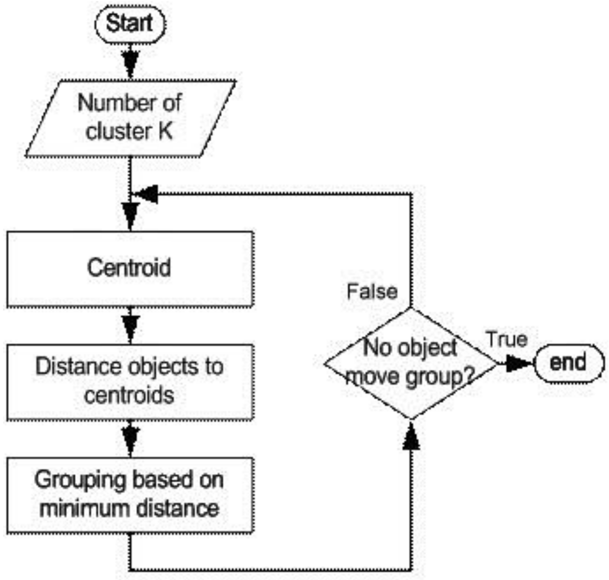


Figure 3.3: K-means steps.

Cosine similarity is a common word-to-word similarity metric used in several summarization tasks and clustering (Wan & Yang, 2008). Cosine similarity is a measure of similarity between two vectors by measuring the cosine of the angle between them sees Figure 3.4. Words are represented by a weights vector while cosine similarity of computing (Erkan & Radev, 2004; Wan & Yang, 2008).

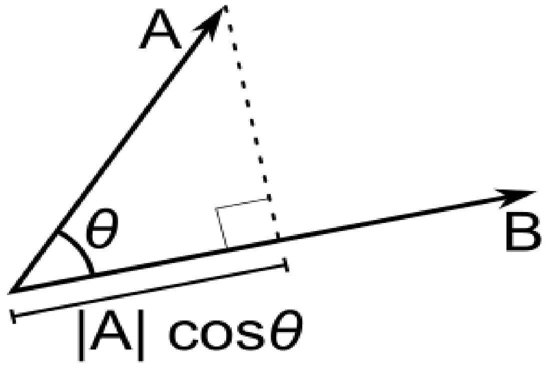


Figure 3.4: The angle of cosine measure.

The second step is ordering clusters, where algorithm of clustering does not assume any prior knowledge about the number of clusters to be formed and completely unsupervised, it is crucial to decide which cluster would contribute as the representative to the final summary. One way to use for clusters ordering which based on sentences counts, this mean which cluster contains large number of sentences is more important. But this method does not come with well performance because when cluster contain on short sentences, which increase only the size, but not the contents, and many clusters get on same size. Thus, to solve this problem we have suggest a method for cluster-ordering, which orders clusters based on the cluster distance performance. The output of this phase will be generating the primary summary by select the best cluster, and achieve the second objective. See Figure 3.5 which includes on this phase steps.

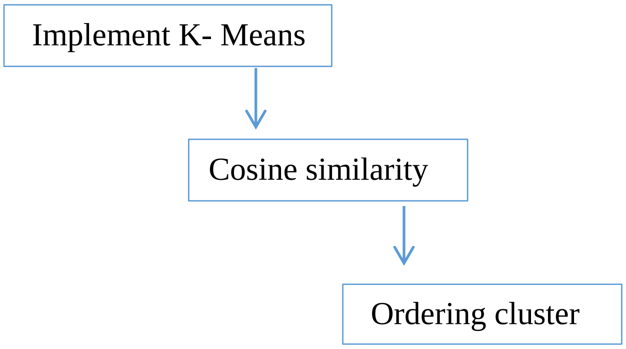


Figure 3.5: Phase III steps.

Phase IV: Summary Generation

The fourth phase is to generate the final summary by fusing and merging the information previously identified. After selected the documents in the best cluster, we will token this documents into sentences. The first step of this phase is transfer the sentences from nominal to numerical this process are used for changing the kind of non-numeric attributes to a numeric type. This process not only changes the type of selected attributes but it also maps all values of these attributes to numeric values. Binary attribute values are mapped to 0 and 1.

The second step is clustering the sentences in the each best cluster, K-means clustering. This is a partitioned centroid-based clustering algorithm. The algorithm randomly selects sentences as the initial centroid for each cluster. The K-means algorithm then iteratively assigns all sentences to the closest cluster, and recalculates the centroid of each cluster, until the centroids no longer change. For our experiments, the similarity between a sentence and a cluster centroid is calculated using the standard cosine measure applied to tokens within the words, and will use this cluster for label sentences.

The third step is sentences are ordered by using support vector machine. The weight by SVM process uses the coefficients of the normal vector of a linear SVM as attribute weights. Note that the attribute values still have to be numerical. This operator can be applied only on example sets with numerical label.

The four and last step in this phase, Based on sentences weights, then representative the most important sentence from the best cluster is select by the representative selection model which in depend on the best high weight. Then continue selecting the sentences from the best cluster in ordered list until a given summary length is reached. In this phase will be solve the redundancy problem by select the best high weight sentences, and we will achieve the third objective. Figure 3.6 shows this phase steps.

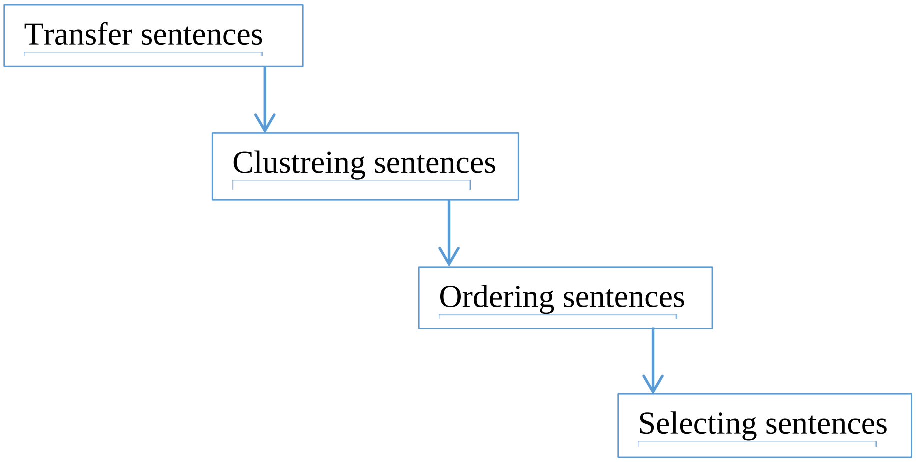


Figure 3.6: Phase IV steps.

Phase V: Evaluation of the final summarization result

The evaluation approaches are suitable in evaluating the trustfulness and usefulness of the summarization. Text summarization is evaluating the qualities, for example readability, eliminate redundancy, comprehensibility, and noisy information. There are two main measures to evaluate the quality of any approach that are recall and precision (Liu, et al., 2006; Hariharan, et al., 2012 ; Deshpande & Lobo, 2013) and they are used for specifying the similarity among the summary which is generated via human against the one generated via a system. With the structure of result, two measures of Recall and Precision are evaluated as (Wadhvani, Pateriya, & Roy, 2013):

Recall = correct/ (correct + missed)

Precision = correct/ (correct + wrong)

Wrong is given by the number of sentences presented in summarization and generated by system but is not involved in human produced summarization. Missed is given by the number of sentences which are not appeared in system produced summarization but presented in the summary generated via human. Correct is given by the number of sentences which are the same in both summary which are generated via system and human. Therefore, the Recall measures the number of appropriate sentences that the summarization system missed. While the Precision measures the number of appropriate sentences which are extracted by system (Haboush & Al-Zoubi, 2012). In this phase we will achieve the last objective.

3.4 Summary

This chapter discusses the fundamental steps that used to achieve the objectives set out for this work: data collection, text analyzing, experimental step, summary generation, and summary evaluation. The objectives are achieved through the application of this methodology based on the weight of each word in each sentence in the documents after removal of the stop word list and noise. Finally, recall and precision are used as metric to evaluate the final results thus achieving the final objective.

CHAPTER FOUR

RESULT AND DISCUSSION

4.1 Introduction

This chapter presents and analyzes the results of the study that have been obtained from the conducted experiments. The results are presented according to the phases described in Chapter 3. The experimentation is conducted using RapidMiner as a tool.

4.2 Experimental Results

4.2.1 Phase I – Data collection

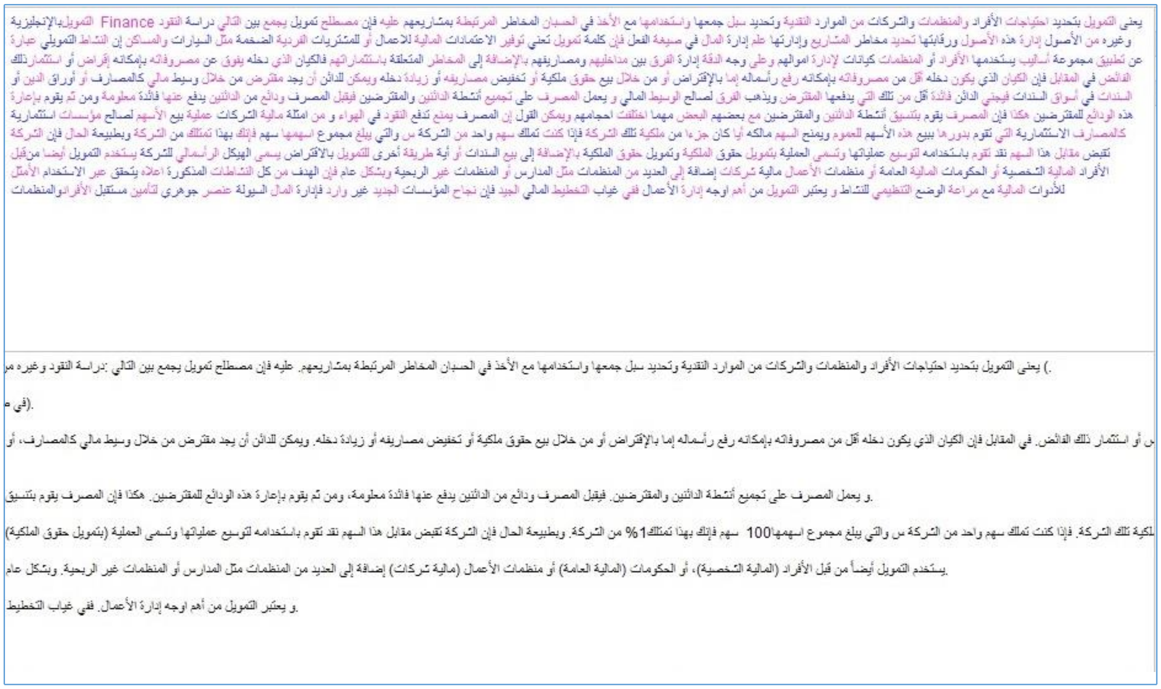
The first phase is completed as we have chosen the secondary data of Essex Arabic Summaries Corpus (EASC), which contain 95 documents after decrease the number of documents. The EASC comprises ten different categories namely, education, art and music, health, sport, finance, politics, religion, science and technology, environment, and tourisms. Each category contains ten documents except education, which contains seven documents and religion contains eight documents. Each document contains on average 380 words, with a minimum word-count of 116 words and a maximum of 971 words. We will use this corpus in the experiment phase. Table 4.1 shows the EASC corpus statistics.

Table 4.1: Statistics of EASC Corpus.

Corpus Name	Essex Arabic Summaries Corpus (EASC)
Documents number	95
Sentences number	1,652
Words number	29,045
Distinct words number	12,785
Gold-standard summaries number	10 (one for each category)

4.2.2 Phase II – Text Analyzing

We divided this section into four steps based on phase two in chapter three. The first step comes with tokenization, the process of splitting the document into tokens. The results of the tokenized text correspond to units whose character structures are recognizable, for example: regular expression to separate the original text into words. Figure 4.1 shows how the performed the token process.



يعنى التمويل بتحديد احتياجات الأفراد والمنظمات والشركات من الموارد النقدية وتحديد سبل جمعها واستخدامها مع الأخذ في الحسبان المخاطر المرتبطة بمشاريعهم عليه فإن مصطلح تمويل يجمع بين التالي دراسة النقود Finance التمويلبالإنجليزية
وغيره من الأصول إدارة هذه الأصول ورقابتها تحديد مخاطر المشاريع وإدارتها علم إدارة المال في صيغة الفعل فإن كلمة تمويل تعني توفير الاعتمادات المالية للاعمال أو للمشتريات الفردية الضخمة مثل السيارات والمساكن إن النشاط التمويلي عبارة
عن تطبيق مجموعة أساليب يستخدمها الأفراد أو المنظمات كيانات لإدارة اموالهم وعلى وجه الدقة إدارة الفرق بين مداخليهم ومصاريفهم بالإضافة إلى المخاطر المتعلقة باستثماراتهم فالكيان الذي دخله يفوق عن مصروفاته بإمكانه إقراض أو استثمارذلك
الفائض في المقابل فإن الكيان الذي يكون دخله أقل من مصروفاته بإمكانه رفع رأسماله إما بالإقتراض أو من خلال بيع حقوق ملكية أو تخفيض مصاريفه أو زيادة دخله ويمكن للدائن أن يجد مقترض من خلال وسيط مالي كالمصارف أو أوراق الدين أو
السندات في أسواق السندات فيجني الدائن فائدة أقل من تلك التي يدفعها المقترض ويذهب الفرق لصالح الوسيط المالي و يعمل المصرف على تجميع أنشطة الدائنين والمقترضين فيقبل المصرف ودائع من الدائنين يدفع عنها فائدة معلومة ومن ثم يقوم بإعارة
هذه الودائع للمقترضين هكذا فإن المصرف يقوم بتنسيق أنشطة الدائنين والمقترضين مع بعضهم البعض مهما اختلفت احجامهم ويمكن القول إن المصرف يمنع تدفع النقود في الهواء و من امثلة مالية الشركات عملية بيع الأسهم لصالح مؤسسات استثمارية
كالمصارف الاستثمارية التي تقوم بدورها ببيع هذه الأسهم للعموم ويمنح السهم مالكه أيا كان جزءا من ملكية تلك الشركة فإذا كنت تملك سهم واحد من الشركة س والتي يبلغ مجموع اسهمها سهم فإنك بهذا تمتلك من الشركة وبطبيعة الحال فإن الشركة
تقبض مقابل هذا السهم نقد تقوم باستخدامه لتوسيع عملياتها وتسمى العملية بتمويل حقوق الملكية وتمويل حقوق الملكية بالإضافة إلى بيع السندات أو أية طريقة أخرى للتمويل بالاقتراض يسمى الهيكل الرأسمالي للشركة يستخدم التمويل أيضا منقبل
الأفراد المالية الشخصية أو الحكومات المالية العامة أو منظمات الأعمال مالية شركات إضافة إلى العديد من المنظمات مثل المدارس أو المنظمات غير الربحية وبشكل عام فإن الهدف من كل النشاطات المذكورة اعلاه يتحقق عبر الاستخدام الأمثل
للأدوات المالية مع مراعاة الوضع التنظيمي للنشاط و يعتبر التمويل من أهم اوجه إدارة الأعمال ففي غياب التخطيط المالي الجيد فإن نجاح المؤسسات الجديد غير وارد فإدارة المال السيولة عنصر جوهري لتأمين مستقبل الأفرادوالمنظمات

.) يعنى التمويل بتحديد احتياجات الأفراد والمنظمات والشركات من الموارد النقدية وتحديد سبل جمعها واستخدامها مع الأخذ في الحسبان المخاطر المرتبطة بمشاريعهم. عليه فإن مصطلح تمويل يجمع بين التالي :دراسة النقود وغيره من

(في م.

س أو استثمار ذلك الفائض. في المقابل فإن الكيان الذي يكون دخله أقل من مصروفاته بإمكانه رفع رأسماله إما بالإقتراض أو من خلال بيع حقوق ملكية أو تخفيض مصاريفه أو زيادة دخله. ويمكن للدائن أن يجد مقترض من خلال وسيط مالي كالمصارف، أو

.و يعمل المصرف على تجميع أنشطة الدائنين والمقترضين. فيقبل المصرف ودائع من الدائنين يدفع عنها فائدة معلومة، ومن ثم يقوم بإعارة هذه الودائع للمقترضين. هكذا فإن المصرف يقوم بتنسيق

لكية تلك الشركة. فإذا كنت تملك سهم واحد من الشركة س والتي يبلغ مجموع اسهمها100 سهم فإنك بهذا تمتلك1% من الشركة. وبطبيعة الحال فإن الشركة تقبض مقابل هذا السهم نقد تقوم باستخدامه لتوسيع عملياتها وتسمى العملية (بتمويل حقوق الملكية)

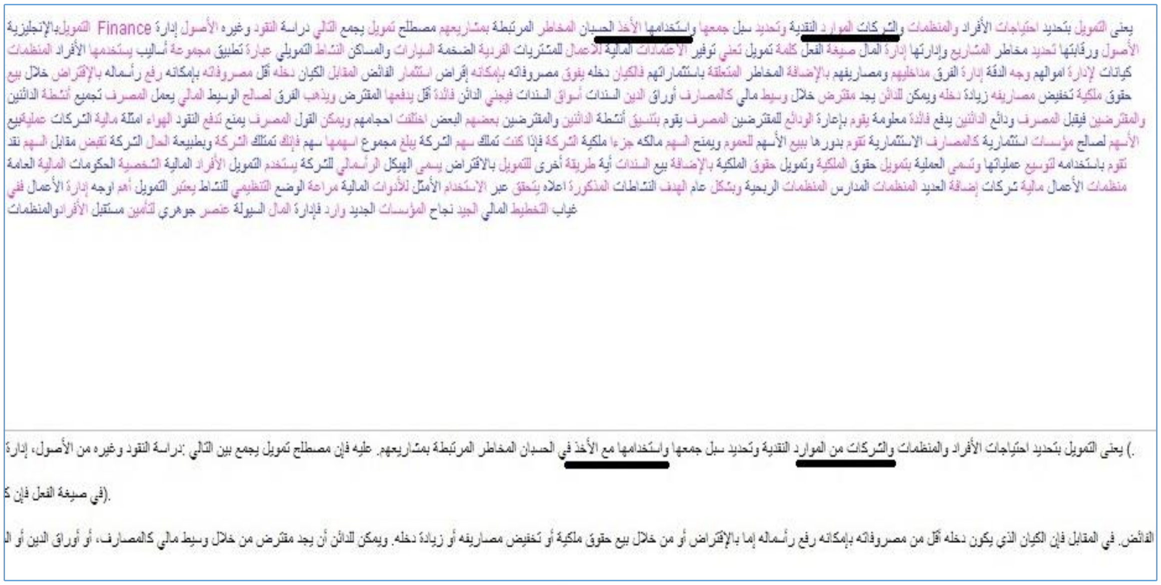
.يستخدم التمويل أيضاً من قبل الأفراد (المالية الشخصية)، أو الحكومات (المالية العامة) أو منظمات الأعمال (مالية شركات) إضافة إلى العديد من المنظمات مثل المدارس أو المنظمات غير الربحية. وبشكل عام

.و يعتبر التمويل من أهم اوجه إدارة الأعمال. ففي غياب التخطيط

Figure 4.1: Text tokenization.

The Figure 4.1 shows two main parts the bottom part which represented the original text in black color. And the top part which appeared the text after words token operation in blue and red color.

The second step of processing starts with removing wholly stop words from the text so that any text has only the verbs and nouns. The stop words do not add any new information to the textual. Stop-word lists differ based on the language, for example in Arabic language (مع، من). Figure 4.2 shows how we did the stop words remove.



يعنى التمويل بتحديد احتياجات الأفراد والمنظمات والشركات الموارد النقدية وتحديد سبل جمعها واستخدامها الأخذ الحسبان المخاطر المرتبطة بمشاريعهم مصطلح تمويل يجمع التالي دراسة النقود وغيره الأصول إدارة Finance التمويلبالإنجليزية
الأصول ورقابتها تحديد مخاطر المشاريع وإدارتها إدارة المال صيغة الفعل كلمة تمويل تعني توفير الاعتمادات المالية للأعمال للمشتريات الفردية الضخمة السيارات والمساكن النشاط التمويلي عبارة تطبيق مجموعة أساليب يستخدمها الأفراد المنظمات
كيانات لإدارة اموالهم وجه الدقة إدارة الفرق مداخليهم ومصاريفهم بالإضافة المخاطر المتعلقة باستثماراتهم فالكيان دخله يفوق مصروفاته بإمكانه إقراض استثمار الفائض المقابل الكيان دخله أقل مصروفاته بإمكانه رفع رأسماله بالإقراض خلال بيع
حقوق ملكية تخفيض مصاريفه زيادة دخله ويمكن للدائن يجد مقترض خلال وسيط مالي كالمصارف أوراق الدين السندات أسواق السندات فيجني الدائن فائدة أقل يدفعها المقترض ويذهب الفرق لصالح الوسيط المالي يعمل المصرف تجميع أنشطة الدائنين
والمقترضين فيقبل المصرف ودائع الدائنين يدفع فائدة معلومة يقوم بإعارة الودائع للمقترضين المصرف يقوم بتنسيق أنشطة الدائنين والمقترضين بعضهم البعض اختلفت احجامهم ويمكن القول المصرف يمنع تدفع النقود الهواء امثلة مالية الشركات عمليةبيع
الأسهم لصالح مؤسسات استثمارية كالمصارف الاستثمارية تقوم بدورها ببيع الأسهم للعموم ويمنح السهم مالكه جزءا ملكية الشركة فإذا كنت تملك سهم الشركة يبلغ مجموع اسهمها سهم فإنك تمتلك الشركة وبطبيعة الحال الشركة تقبض مقابل السهم نقد
تقوم باستخدامه لتوسيع عملياتها وتسمى العملية بتمويل حقوق الملكية وتمويل حقوق الملكية بالإضافة بيع السندات أية طريقة أخرى للتمويل بالاقتراض يسمى الهيكل الرأسمالي للشركة يستخدم التمويل الأفراد المالية الشخصية الحكومات المالية العامة
منظمات الأعمال مالية شركات إضافة العديد المنظمات المدارس المنظمات الربحية وبشكل عام الهدف النشاطات المذكورة اعلاه يتحقق عبر الاستخدام الأمثل للأدوات المالية مراعاة الوضع التنظيمي للنشاط يعتبر التمويل أهم اوجه إدارة الأعمال ففي
غياب التخطيط المالي الجيد نجاح المؤسسات الجديد وارد فإدارة المال السيولة عنصر جوهري لتأمين مستقبل الأفرادوالمنظمات

.) يعنى التمويل بتحديد احتياجات الأفراد والمنظمات والشركات من الموارد النقدية وتحديد سبل جمعها واستخدامها مع الأخذ في الحسبان المخاطر المرتبطة بمشاريعهم. عليه فإن مصطلح تمويل يجمع بين التالي :دراسة النقود وغيره من الأصول، إدارة

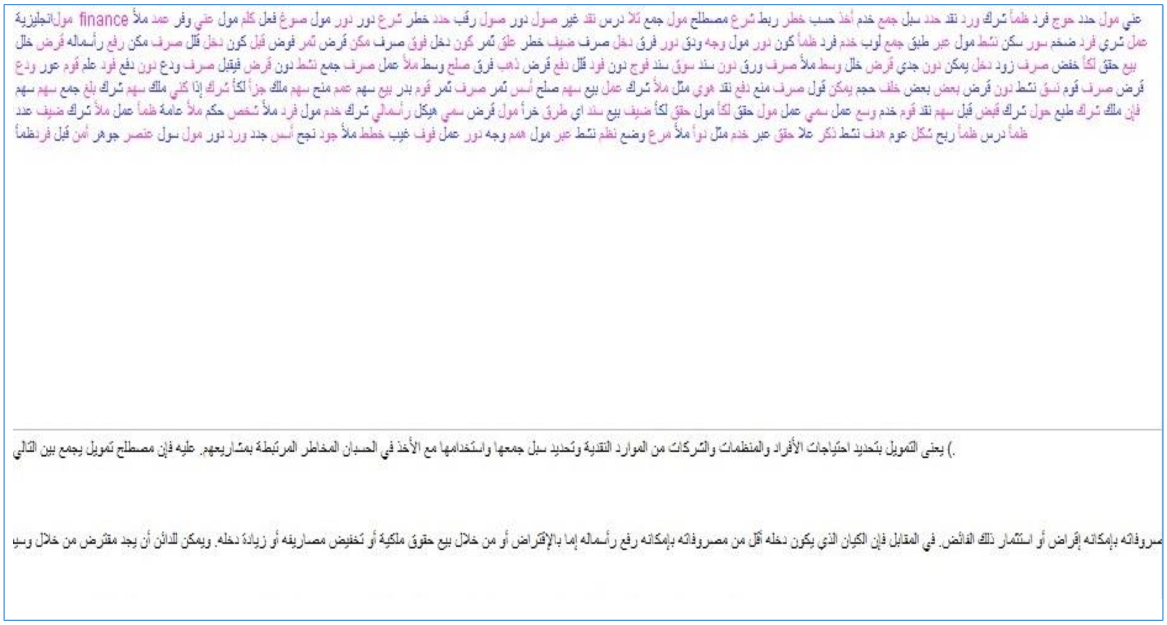
.(في صيغة الفعل فإن ك

الفائض. في المقابل فإن الكيان الذي يكون دخله أقل من مصروفاته بإمكانه رفع رأسماله إما بالإقراض أو من خلال بيع حقوق ملكية أو تخفيض مصاريفه أو زيادة دخله. ويمكن للدائن أن يجد مقترض من خلال وسيط مالي كالمصارف، أو أوراق الدين أو ال

Figure 4.2: Stop words removal.

Figure 4.2 illustrates the original text in the bottom part with black color and in the upper part the text after process with blue and red color. This figure shows also two samples from the stop word list which we are put under line (من، مع) and how we worked to remove these words.

The third processing step starts with stemmer which involved discoveries of the root of each word in text in the original Arabic text. This means each word in Arabic language has sub parts (infixes, suffixes, and prefixes) requirement to be removed. This is done by automatically stripping infixes, suffixes, and prefixes from words to obtain stems. For example, several words that are used to express a particular concept (e.g., المكتبة الكاتب الكتاب) can be grouped together and stemmed to work, since they all have the same conceptual meaning. Figure 4.3 illustrations how we did the words stemmer.



عني مول حدد حوج فرد ظمأ شرك ورد نقد حدد سبل جمع خدم أخذ حسب خطر ربط شرع مصطلح مول جمع ثلا درس نقد غير صول دور صول رقب حدد خطر شرع دور دور مول صوغ فعل كلم مول عني وفر عمد ملأ finance مولانجليزية
عمل شري فرد ضخم سور سكن نشط مول عبر طبق جمع لوب خدم فرد ظمأ كون دور مول وجه ودق دور فرق دخل صرف ضيف خطر طلق ثمر كون دخل فوق صرف مكن قرض ثمر فوض قبل كون دخل قلل صرف مكن رفع رأسماله قرض خلل
بيع حقق لكأ خفض صرف زود دخل يمكن دون جدي قرض خلل وسط ملأ صرف ورق دون سند سوق سند فوج دون فود قلل دفع قرض ذهب فرق صلح وسط ملأ عمل صرف جمع نشط دون قرض فيقبل صرف ودع دون دفع فود علم قوم عور ودع
قرض صرف قوم نسق نشط دون قرض بعض بعض خلف حجم يمكن قول صرف منع دفع نقد هوي مثل ملأ شرك عمل بيع سهم صلح أسس ثمر صرف ثمر قوم بدر بيع سهم عمم منح سهم ملك جزأ لكأ شرك إذا كثي ملك سهم شرك بلغ جمع سهم سهم
فإن ملك شرك طبع حول شرك قبض قبل سهم نقد قوم خدم وسع عمل سمي عمل مول حقق لكأ مول حقق لكأ ضيف بيع سند اي طرق خرأ مول قرض سمي هيكل رأسمالي شرك خدم مول فرد ملأ شخص حكم ملأ عامة ظمأ عمل ملأ شرك ضيف عدد
ظمأ درس ظمأ ربح شكل عوم هدف نشط ذكر علا حقق عبر خدم مثل دوأ ملأ سرع وضع نظم نشط عبر مول همم وجه دور عمل فوف غيب خطط ملأ جود نجح أسس جدد ورد دور مول سول عنصر جوهر أمن قبل فردظمأ

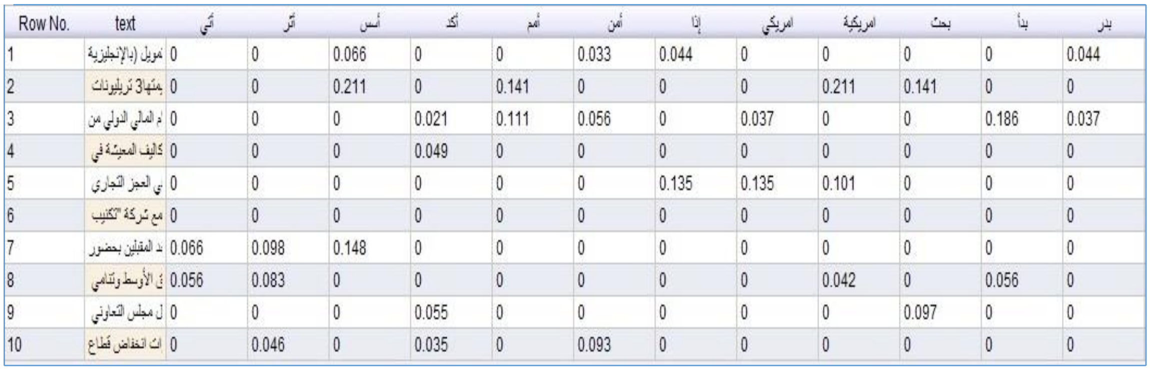
.) يعنى التمويل بتحديد احتياجات الأفراد والمنظمات والشركات من الموارد النقدية وتحديد سبل جمعها واستخدامها مع الأخذ في الحسبان المخاطر المرتبطة بمشاريعهم. عليه فإن مصطلح تمويل يجمع بين التالي

صروفاته بإمكانه إقراض أو استثمار ذلك الفائض. في المقابل فإن الكيان الذي يكون دخله أقل من مصروفاته بإمكانه رفع رأسماله إما بالإقتراض أو من خلال بيع حقوق ملكية أو تخفيض مصاريفه أو زيادة دخله. ويمكن للدائن أن يجد مقترض من خلال وسي

Figure 4.3: Stemming words.

Figure 4.2 illustrates the text with blue and red color in the above part, the text after process. And this figure shows also a sample from the stem. The second word from the right in the first line (التمويل) which represented by black color which means the funding, then when compare with the second word in first line which represented by red color, we found (مول) which means fund and how we worked to remove infixes, suffixes, and prefixes.

The fourth step of processing comes with applied the (term frequency–inverse document frequency) TF–IDF, to represent the most significant words as a bag of word. The TF–IDF is a simple numerical statistic which reflects how significant a word is to a document in a corpus. Figure 4.4 illustrates how we presented the most important words in the original text.



Row No.	text	أتي	أثر	أسس	أكد	أمم	أمن	إذا	امريكي	امريكية	بحث	بدأ	بدر
1	مويل (بالإنجليزية	0	0	0.066	0	0	0.033	0.044	0	0	0	0	0.044
2	يمتها3 تريليونات	0	0	0.211	0	0.141	0	0	0	0.211	0.141	0	0
3	م المالي الدولي من	0	0	0	0.021	0.111	0.056	0	0.037	0	0	0.186	0.037
4	كاليف المعيشة في	0	0	0	0.049	0	0	0	0	0	0	0	0
5	ي العجز التجاري	0	0	0	0	0	0	0.135	0.135	0.101	0	0	0
6	مع شركة "تكتيب	0	0	0	0	0	0	0	0	0	0	0	0
7	د المقبلين بحضور	0.066	0.098	0.148	0	0	0	0	0	0	0	0	0
8	ق الأوسط وتنامي	0.056	0.083	0	0	0	0	0	0	0.042	0	0.056	0
9	ل مجلس التعاوني	0	0	0	0.055	0	0	0	0	0	0.097	0	0
10	ات انخفاض قطاع	0	0.046	0	0.035	0	0.093	0	0	0	0	0	0

Figure 4.4: TFIDF vector of words.

When, we are finished all of these practical steps. Now we were able to solve the noisy information problem. Thus, we have the data ready to apply in the next phase.

4.2.3 Phase III: Cluster Experiment

We split this section into two steps based on phase three in the chapter three. The first step starts with, the algorithm used in the experiments, which is K-means clustering. This is a partitioned centroid-based clustering algorithm. The algorithm randomly selects words as the initial centroid for each cluster. The K-means algorithm then iteratively assigns all words to the closest cluster, and recalculates the centroid of each cluster, until the centroids no longer change. For our experiments, the similarity between a word and a cluster centroid is calculated using the standard cosine measure applied to tokens within the words. We run K-means clustering using number two of clusters. Clustering using a single cluster essentially results in words list which can be ranked according to similarity to the centroid of all words. Figure 4.5 shows the results for each category (from the ten aforementioned) inside the cluster algorithm (K-means).

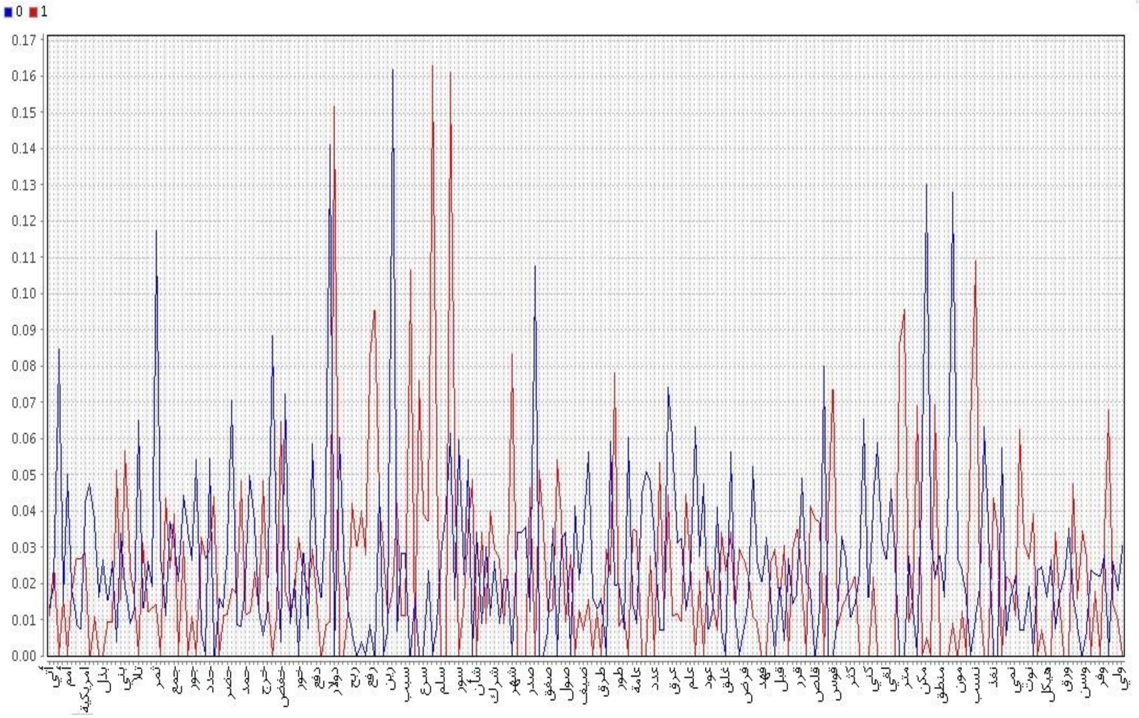


Figure 4.5: Cluster Results.

Figure 4.5 illustrates the number of clusters which are denoted by two different colors the blue represents the first cluster, and red represents the second cluster. The y axis illustrates the percentage of each word accrue in the clusters. Meanwhile, the x axis illustrates the most important words represented by term frequency and invers document frequency (TF IDF). Based on the result, the word سلم has the highest tf-idf value approximately with 0.16, while the word ربح has the lowest tf-idf value approximately with 0.01.

The second processing step comes with ranking the clusters. The ranking of clusters is done based on the cluster distance performance. The best cluster is defined as the one with low distance between the objects in each cluster. Table 4.2 shows the results of cluster ordering relying on distance performance, and which cluster will be contributed for final summary, and illustrations the number of document in the best clusters based on EASC categories.

Table 4.2: Order of clusters

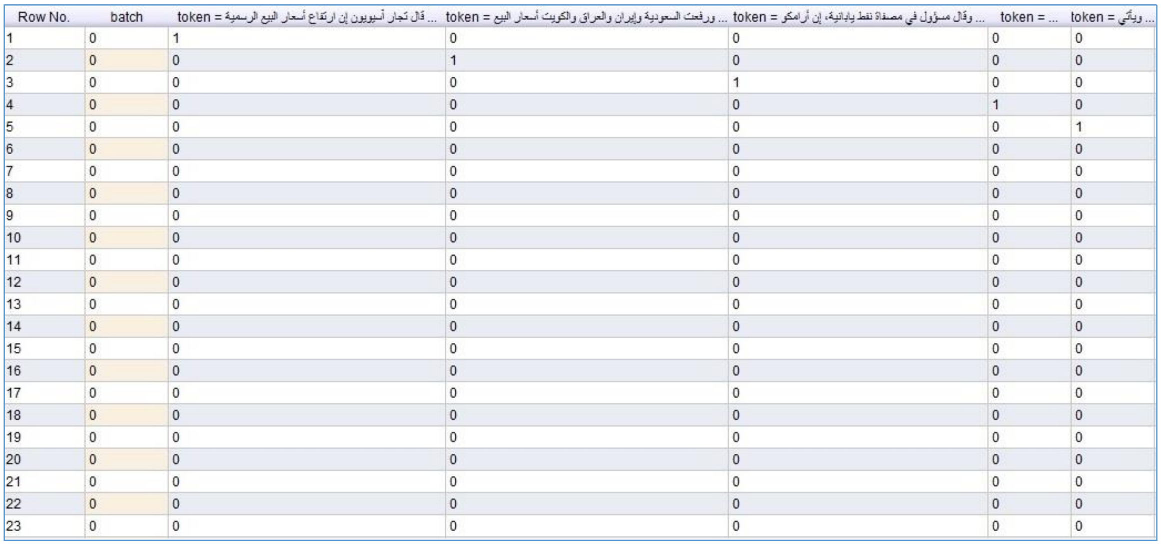
Name of Category	**Cluster Distance Performance**	**The Best Cluster**	**Number of Documents**
Art and Music	Cluster 0 = 0.695 Cluster 1 = 0.679	Cluster 1	1,2,3,4, and 10
Education	Cluster 0 = 0.640 Cluster 1 = 0.535	Cluster 1	5,6, and 7
Environment	Cluster 0 = 0.771 Cluster 1 = 0.485	Cluster 1	1,5, and 8
Finance	Cluster 0 = 0.676 Cluster 1 = 0.695	Cluster 0	1,2,3,7, and 9

Health	Cluster 0 = 0.584 Cluster 1 = 0.723	Cluster 0	2,3,4, and 10
Politics	Cluster 0 = 0.725 Cluster 1 = 0.613	Cluster 1	3,6,7, and 8
Religion	Cluster 0 = 0.534 Cluster 1 = 0.688	Cluster 0	1,3, and 5
Science and Technology	Cluster 0 = 0.710 Cluster 1 = 0.707	Cluster 1	5,7,8,9, and 10
Sport	Cluster 0 = 0.538 Cluster 1 = 0.717	Cluster 0	1,2, and 10
Tourisms	Cluster 0 = 0.661 Cluster 1 = 0.700	Cluster 0	2,3,4,5,7,and 9

4.2.4 Phase IV: Summary Generation

Four steps are involved based on phase four in the methodology chapter. After the selection of the document in the best clusters from the section 4.2.3, based on EASC categories, and the documents in the best cluster for each category are tokenized to sentence level. The first step in this phase starts with transferring the text from nominal to numerical coding to run the process.

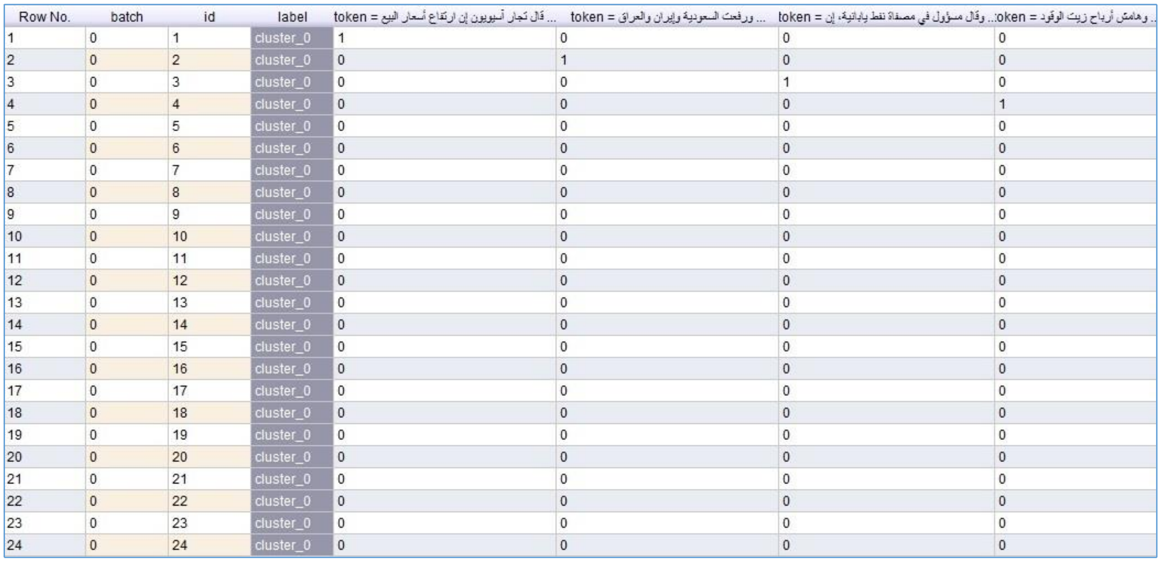
This parameter is selected, for all values of the nominal attribute, excluding the comparison group, a new attribute is created. The comparison group can be defined using the comparison group parameter. In every example, the new attribute which corresponds to the actual nominal value of that example gets value 1 and all other new attributes get value 0. If the value of the nominal attribute of this example corresponds to the comparison group, all new attributes are set to 0. Note that the comparison group is an optional parameter with 'dummy coding'. If no comparison group is defined, in every example the new attribute which corresponds to the actual nominal value of that example gets value 1 and all other new attributes get value 0. In this case, there will be no example where all new attributes get value 0. Figure 4.6 shows how binary attribute values are mapped to 0 and 1 for sentences.



Row No.	batch	token = قال تجار آسيويون إن ارتفاع أسعار البيع الرسمية ...	token = ورفعت السعودية وإيران والعراق والكويت أسعار البيع ...	token = وقال مسؤول في مصفاة نفط يابانية، إن أرامكو ...	token = ...	token = ويأتي ...
1	0	1	0	0	0	0
2	0	0	1	0	0	0
3	0	0	0	1	0	0
4	0	0	0	0	1	0
5	0	0	0	0	0	1
6	0	0	0	0	0	0
7	0	0	0	0	0	0
8	0	0	0	0	0	0
9	0	0	0	0	0	0
10	0	0	0	0	0	0
11	0	0	0	0	0	0
12	0	0	0	0	0	0
13	0	0	0	0	0	0
14	0	0	0	0	0	0
15	0	0	0	0	0	0
16	0	0	0	0	0	0
17	0	0	0	0	0	0
18	0	0	0	0	0	0
19	0	0	0	0	0	0
20	0	0	0	0	0	0
21	0	0	0	0	0	0
22	0	0	0	0	0	0
23	0	0	0	0	0	0

Figure 4.6: Sentences represented by binary map.

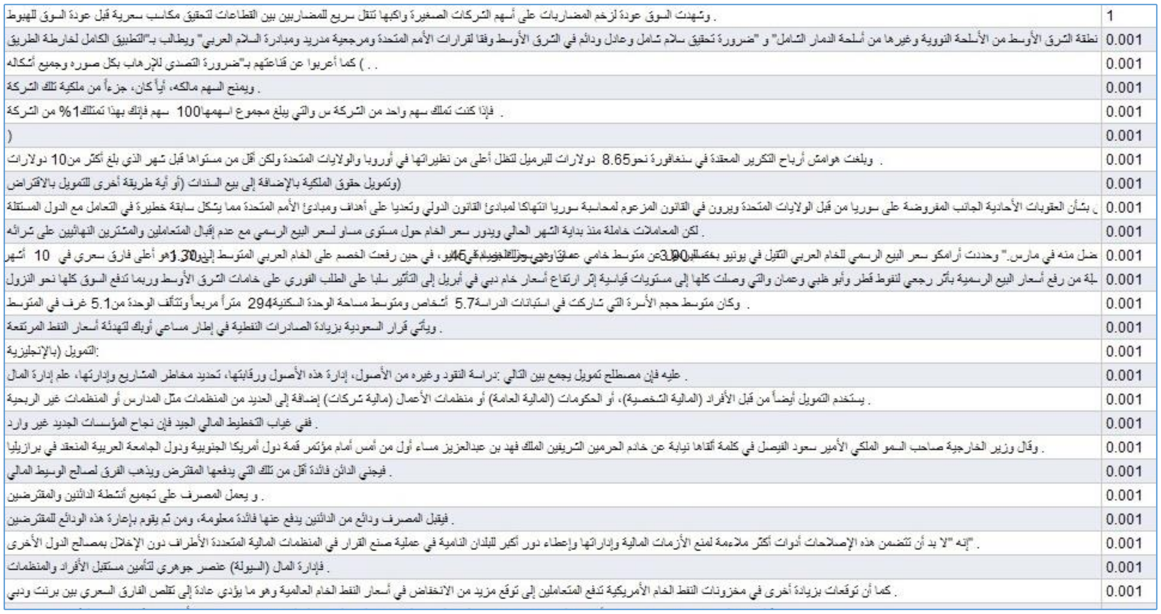
The second step starts with sentences cluster in the each best cluster, we ran K-means clustering using number two of clusters, and we used this cluster for label sentences. Figure 4.7 illustrations how the second cluster labeled all the sentences.



Row No.	batch	id	label	token = قال تجار آسيويون إن ارتفاع أسعار البيع ...	token = ورفعت السعودية وإيران والعراق ...	token = وقال مسؤول في مصفاة نفط يابانية، إن ...	oken = وهامش أرباح زيت الوقود ..
1	0	1	cluster_0	1	0	0	0
2	0	2	cluster_0	0	1	0	0
3	0	3	cluster_0	0	0	1	0
4	0	4	cluster_0	0	0	0	1
5	0	5	cluster_0	0	0	0	0
6	0	6	cluster_0	0	0	0	0
7	0	7	cluster_0	0	0	0	0
8	0	8	cluster_0	0	0	0	0
9	0	9	cluster_0	0	0	0	0
10	0	10	cluster_0	0	0	0	0
11	0	11	cluster_0	0	0	0	0
12	0	12	cluster_0	0	0	0	0
13	0	13	cluster_0	0	0	0	0
14	0	14	cluster_0	0	0	0	0
15	0	15	cluster_0	0	0	0	0
16	0	16	cluster_0	0	0	0	0
17	0	17	cluster_0	0	0	0	0
18	0	18	cluster_0	0	0	0	0
19	0	19	cluster_0	0	0	0	0
20	0	20	cluster_0	0	0	0	0
21	0	21	cluster_0	0	0	0	0
22	0	22	cluster_0	0	0	0	0
23	0	23	cluster_0	0	0	0	0
24	0	24	cluster_0	0	0	0	0

Figure 4.7: Sentences labeled.

The third step comes with ordered Sentences by using support vector machine SVM. We used this process to rank the sentences by selected descending parameter. This parameter is only available when the sort weights parameter is set to true. This parameter specifies the sorting order of the sentences according to their weights. Figure 4.8 shows the result of support vector machine.

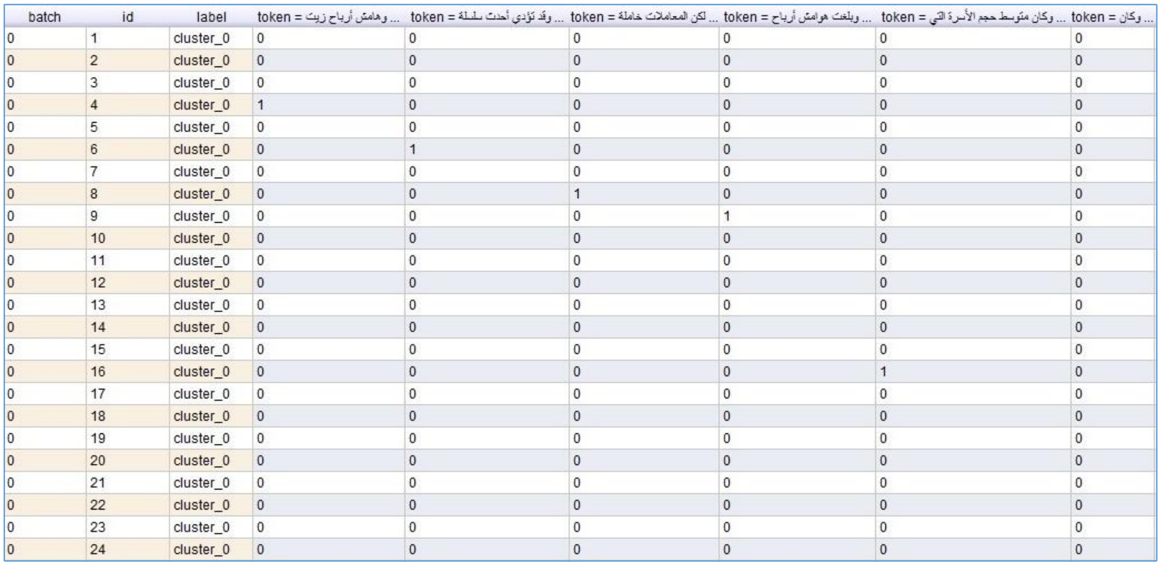


وشهدت السوق عودة لزخم المضاربات على أسهم الشركات الصغيرة واكبها تنقل سريع للمضاربين بين القطاعات لتحقيق مكاسب سعرية قبل عودة السوق للهبوط .	1
نطقة الشرق الأوسط من الأسلحة النووية وغيرها من أسلحة الدمار الشامل" و "ضرورة تحقيق سلام شامل وعادل ودائم في الشرق الأوسط وفقا لقرارات الأمم المتحدة ومرجعية مدريد ومبادرة السلام العربي" ويطالب بـ"التطبيق الكامل لخارطة الطريق	0.001
كما أعربوا عن قناعتهم بـ"ضرورة التصدي للإرهاب بكل صوره وجميع أشكاله) . .	0.001
ويمنح السهم مالكه، أياً كان، جزءاً من ملكية تلك الشركة .	0.001
فإذا كنت تملك سهم واحد من الشركة س والتي يبلغ مجموع اسهمها100 سهم فإنك بهذا تمتلك1% من الشركة .	0.001
)	0.001
وبلغت هوامش أرباح التكرير المعقدة في سنغافورة نحو8.65 دولارات للبرميل لتظل أعلى من نظيراتها في أوروبا والولايات المتحدة ولكن أقل من مستواها قبل شهر الذي بلغ أكثر من10 دولارات .	0.001
(وتمويل حقوق الملكية بالإضافة إلى بيع السندات (أو أية طريقة أخرى للتمويل بالاقتراض	0.001
ن بشأن العقوبات الأحادية الجانب المفروضة على سوريا من قبل الولايات المتحدة ويرون في القانون المزعوم لمحاسبة سوريا انتهاكا لمبادئ القانون الدولي وتعديا على أهداف ومبادئ الأمم المتحدة مما يشكل سابقة خطيرة في التعامل مع الدول المستقلة	0.001
لكن المعاملات خاملة منذ بداية الشهر الحالي ويدور سعر الخام حول مستوى مساو لسعر البيع الرسمي مع عدم إقبال المتعاملين والمشترين النهائيين على شرائه .	0.001
ضل منه في مارس." وحددت أرامكو سعر البيع الرسمي للخام العربي الثقيل في يونيو بخصم 3.90 عن متوسط خامي عمان ودبي 1.45 في حين رفعت الخصم على الخام العربي المتوسط إلى 1.30 وهو أعلى فارق سعري في 10 أشهر	0.001
لة من رفع أسعار البيع الرسمية بأثر رجعي لنفوط قطر وأبو ظبي وعمان والتي وصلت كلها إلى مستويات قياسية إثر ارتفاع أسعار خام دبي في أبريل إلى التأثير سلبا على الطلب الفوري على خامات الشرق الأوسط وربما تدفع السوق كلها نحو النزول	0.001
وكان متوسط حجم الأسرة التي شاركت في استبانات الدراسة5.7 أشخاص ومتوسط مساحة الوحدة السكنية294 متراً مربعاً وتتألف الوحدة من5.1 غرف في المتوسط .	0.001
ويأتي قرار السعودية بزيادة الصادرات النفطية في إطار مساعي أوبك لتهدئة أسعار النفط المرتفعة .	0.001
التمويل (بالإنجليزية:	0.001
عليه فإن مصطلح تمويل يجمع بين التالي :دراسة النقود وغيره من الأصول، إدارة هذه الأصول ورقابتها، تحديد مخاطر المشاريع وإدارتها، علم إدارة المال .	0.001
يستخدم التمويل أيضاً من قبل الأفراد (المالية الشخصية)، أو الحكومات (المالية العامة) أو منظمات الأعمال (مالية شركات) إضافة إلى العديد من المنظمات مثل المدارس أو المنظمات غير الربحية .	0.001
ففي غياب التخطيط المالي الجيد فإن نجاح المؤسسات الجديد غير وارد .	0.001
وقال وزير الخارجية صاحب السمو الملكي الأمير سعود الفيصل في كلمة ألقاها نيابة عن خادم الحرمين الشريفين الملك فهد بن عبدالعزيز مساء أول من أمس أمام مؤتمر قمة دول أمريكا الجنوبية ودول الجامعة العربية المنعقد في برازيليا .	0.001
فيجني الدائن فائدة أقل من تلك التي يدفعها المقترض ويذهب الفرق لصالح الوسيط المالي .	0.001
و يعمل المصرف على تجميع أنشطة الدائنين والمقترضين .	0.001
فيقبل المصرف ودائع من الدائنين يدفع عنها فائدة معلومة، ومن ثم يقوم بإعارة هذه الودائع للمقترضين .	0.001
إنه "لا بد أن تتضمن هذه الإصلاحات أدوات أكثر ملاءمة لمنع الأزمات المالية وإدارتها وإعطاء دور أكبر للبلدان النامية في عملية صنع القرار في المنظمات المالية المتعددة الأطراف دون الإخلال بمصالح الدول الأخرى" .	0.001
فإدارة المال (السيولة) عنصر جوهري لتأمين مستقبل الأفراد والمنظمات .	0.001
كما أن توقعات بزيادة أخرى في مخزونات النفط الخام الأمريكية تدفع المتعاملين إلى توقع مزيد من الانخفاض في أسعار النفط الخام العالمية وهو ما يؤدي عادة إلى تقلص الفارق السعري بين برنت ودبي .	0.001

Figure 4.8: Sentences Order.

Figure 4.8 shown the regularized weights parameter set to true, thus all the weights are normalized in the range 0 to 1.

The fourth step comes with representative the most important sentences from the best cluster for final summary were selected by the representative selection process which in depend on the best high weight with the top P percent attributes with highest weights are selected. P is specified by the p parameter, percentage of sentences to be selected. For example top P % with 15, then returned any sentence containing only sentences which were part of 15 % of the highest weight sentences, for example we used P with 0.2 to select the sentences for finance category the selection process selected the best 13 sentences from 64 sentences, this mean the system extracted approximately 20 %. Figure 4.9 illustrations the result of selection process.



batch	id	label	token = وهامش أرباح زيت ...	token = وقد تؤدي أحدث سلسلة ...	token = لكن المعاملات خاملة ...	token = وبلغت هوامش أرباح ...	token = وكان متوسط حجم الأسرة التي ...	token = وكان ...
0	1	cluster_0	0	0	0	0	0	0
0	2	cluster_0	0	0	0	0	0	0
0	3	cluster_0	0	0	0	0	0	0
0	4	cluster_0	1	0	0	0	0	0
0	5	cluster_0	0	0	0	0	0	0
0	6	cluster_0	0	1	0	0	0	0
0	7	cluster_0	0	0	0	0	0	0
0	8	cluster_0	0	0	1	0	0	0
0	9	cluster_0	0	0	0	1	0	0
0	10	cluster_0	0	0	0	0	0	0
0	11	cluster_0	0	0	0	0	0	0
0	12	cluster_0	0	0	0	0	0	0
0	13	cluster_0	0	0	0	0	0	0
0	14	cluster_0	0	0	0	0	0	0
0	15	cluster_0	0	0	0	0	0	0
0	16	cluster_0	0	0	0	0	1	0
0	17	cluster_0	0	0	0	0	0	0
0	18	cluster_0	0	0	0	0	0	0
0	19	cluster_0	0	0	0	0	0	0
0	20	cluster_0	0	0	0	0	0	0
0	21	cluster_0	0	0	0	0	0	0
0	22	cluster_0	0	0	0	0	0	0
0	23	cluster_0	0	0	0	0	0	0
0	24	cluster_0	0	0	0	0	0	0

Figure 4.9: Sentences selection process.

4.2.5 Phase V: Evaluation

This section shows the evaluation results of the work done on multi-document summarization presented in section 4.2.4. Both automatically generated summary and human generated summary were used in the evaluation. The evaluation results of the multi-document summarization redundancy elimination and experiments techniques are presented and compared with the results of other summarization techniques and systems as reviewed in the literature.

The evaluation results of the cluster-based summarization experiments presented in section 4.2.4 is illustrated in Table 4.3. This table shows the Arabic summarizers results when applied K-means clustering for redundancy elimination based on EASC categories.

Table 4.3: Recall and Precision results.

EASC categories	**Recall**	**Precision**
Art and Music	0.57	0.4
Education	0.5	0.5
Environment	0.4	0.14
Finance	0.42	0.25
Health	0.6	0.5
Politics	0.5	0.55
Religion	0.5	0.37
Science and Technology	0.57	0.5
Sport	0.5	0.6
Tourisms	0.24	0.13

The main outcome of our experiments appears to be the fact that a simple centroid based similarity clustering with a cluster when performing summarization could be considered an alternative way to the use of different cluster numbers. The work by Radev et al. (2000, 2004) considered a variable number of clusters, whereas our experiments demonstrate that for the given test fixed number of the closeness to the centroid (to identify the significant words) can produce summaries with similar quality.

4.3 Discussion

We tested 95 documents (text) representation for EASC corpus. We used our objectives to discusses our results, the first objective is to analyze Arabic text in order to remove the noisy information. We implemented text preprocessing with words token, stop words list removal, stemming, and text presentation after we remove the noisy problem. The second objective is to implement the cluster approach for cluster order and redundancy elimination. We applied the K-means clustering algorithm with cosine similarity to get on the clusters documents based on K equal two, then we ordered the clusters based on cluster distance performance to obtain on the best cluster. The third objective is to select sentences based on the order of clusters generated in the second objective. We used SVM to order the sentences based on their weight from 1 to 0, then we selected the best sentences weight to final summary to achieve the order sentences. The fourth objective is to evaluate the final result summary by using Recall and Precision. For the final results based on ten categories, we found the Health and Sport category obtains the highest score for the Recall and Precision.

We have focused on techniques of extractive summarization. We have compared the results developed by various extractive techniques by using correctness and performance measures for example Recall and Precision. Based on our results the tool which is used outperforms the other technique and tools, and reduced redundancy because clustering. Table 4.4 shows the comparison.

Table 4.4: Comparison results based on Recall and Precision results.

Authors name	Data set	Technique	Recall & Precision results
(Deshpande & Lobo, 2013)	English	Clustering	0.47, 0.66
(El-Haj, et al., 2011).	EASC	Vector Space Model , Latent Semantic Analysis & Dice's Coefficient	0.36, 0.37
(Deshpande & Lobo, 2013)	English (second data collection)	Clustering	0.45, 0.51
Our experimental	EASC	Clustering	0.6, 0.6

4.4 Summary

This chapter discusses the results obtained. The results were presented according to each phase as described in methodology chapter. We did our experiment with word-clustering based multi-document summarization. For this experiment, Essex Arabic Summaries Corpus EASC as input. We did some steps under process documents operation in RapidMiner, namely: token operation to split the text into words, deleted all stop words, we did the stem operation to extract the root for each of these words, and we are used TFIDF to present the most important words. However, we have data ready to feed in the K-means algorithm after we decided K equals two. We used cluster distance performance to determine which cluster contributed for the final summary. We applied support vector machine to order the sentences by their weight, the range of weight 0 to 1.

Then we selected the sentences based on highest weight. To give the fully understanding for our experimental we divided into three models based on RapidMiner operations, the first one is called text cluster, the second one is sentences token, and the third model called sentences section. On other hand, when we compared our work with (Deshpande & Lobo, 2013; El-Haj, et al., 2011) works. Thus, we found slight improvement based on recall and precision measures.

CHAPTER FIVE

CONCLUSION AND FUTURE WORK

5.1 Conclusion

In this work we focused on Arabic multi-document summarization and addressed the text summarization issue in particular, redundancy elimination, noisy information, and sentence ordering. We investigated clustering for multi-document Arabic summarization. We explored how clustering can be applied for multi-document summarization in addition to for redundancy elimination within this process. We used various parameter settings including the cluster order and the sentences selection method applied in the summarization of extractive process.

This study is based on four objectives to solve the noisy problem, redundancy elimination, sentences ordering, and evaluates the final result. The first objective is to analyze Arabic text in order to remove the noisy information and we achieved this objective in phase two in chapter three by tokening words operation, stop words list operation, extract the root of each word, and represented the most important words by TF-IDF.

The second objective is to implement the cluster approach for cluster order and redundancy elimination and we achieved this objective in phase three in methodology chapter by applied the K-means algorithm with cosine similarity and we used the cluster distance performance to order the clusters to select the best cluster.

The third objective is to select sentences based on the order of clusters generated in the second objective and we achieved this objective in phase four the methodology by transferring the text from nominal to numerical coding to run the process, we used second cluster step for label sentences, ordered Sentences by using support vector machine SVM, and representative the most significant sentences from the best cluster for final summary based on the highest weight.

The forth objective is to evaluate the final result summary by using Recall and Precision and we achieved this objective in phase five of the methodology using Recall and Precision to measure.

By using recall and precision metrics we were able to measure the effect of applying different tools and methods. One of our main findings is that selecting sentences with highest weight of all sentences in the collection of related documents gives the better recall and precision scores. The work in this research is eclectic as a result of trying for demonstrating that summarization in Arabic is potential, and can imitate what is done in English, European, and Asian languages. Researchers on Arabic multi-document summarization now have resources, tools and results that can be used for Arabic multi-document summarization to progress this area of research.

5.2 Future Work

Improvement and future works can be directed to the implementation of more enhanced clustering for improving results. Experimenting with many language-specific features, for example anaphoric resolution, textual entailment, and morphological parsers is an open research for more enhancements in the future.

Arabic Natural Language Processing researchers were not successful yet in attempting the Arabic abstractive summarization field. Abstractive summarization requires an understanding of the original text and regenerating it in a shorter version. This varies from extractive summarization as it involves the use of Natural Language Generation (NLG) tools to paraphrase the corpus using novel sentences or words. The lack of Arabic resources and NLG tools made it hard for researchers to successfully tackle this field. This can be solved by building Arabic NLG tools and resources including Arabic language models, lexicons and Word Net for developing Arabic abstractive summarizers that can produce cohesive sentences.

Through more Arabic news available on news websites for example CNN and BBC Arabic, a real-time summarizer can be built to create abstractive summaries for continuing events, the summarizer can work incrementally, where the created summary will be updated as long as the event is still going and more news are being generated.

Arabic automatic summarization researches have a plenty of room in this area. Improving the current Arabic summarization techniques and methods is highly dependent on availability of more Arabic resources and tools and the advancing the work on Arabic NLP.

REFERENCES

Abbas, M., Smaili, K., & Berkani, D. (2009a). Comparing TR-Classifier and KNN by using Reduced Sizes of Vocabularies. *Culture, 1*, 210.

Abbas, M., Smaili, K., & Berkani, D. (2009b). *A trigger-based classifier*. Paper presented at the The 2nd Int. Conf. on Arabic Language Resources and Tools (MEDAR 2009).

Agarwal, N., Reddy, R. S., Gvr, K., & Rosé, C. P. (2011). Towards multi-document summarization of scientific articles: making interesting comparisons with SciSumm. *ACL HLT 2011*, 8.

Al-Sulaiti, L., & Atwell, E. (2004). *Designing and developing a corpus of contemporary Arabic*. University of Leeds (School of Computing).

Al-Sulaiti, L., & Atwell, E. (2006). The design of a corpus of Contemporary Arabic. *International Journal of Corpus Linguistics, 11*(2).

Albared, M., Omar, N., & Ab Aziz, M. J. (2009). *Classifiers combination to arabic morphosyntactic disambiguation*. Paper presented at the Electrical Engineering and Informatics, 2009. ICEEI'09. International Conference on.

Amini, M., & Usunier, N. (2007). A contextual query expansion approach by term clustering for robust text summarization.

Azmi, A., & Al-thanyyan, S. (2009). *Ikhtasir—A user selected compression ratio Arabic text summarization system*. Paper presented at the Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009. International Conference on.

Azmi, A. M., & Al-Thanyyan, S. (2012). A text summarizer for Arabic. *Computer Speech & Language, 26*(4), 260-273.

Berger, A., & Mittal, V. O. (2000). *Query-relevant summarization using FAQs.* Paper presented at the Proceedings of the 38th Annual Meeting on Association for Computational Linguistics.

Boudabous, M. M., Maaloul, M. H., & Belguith, L. H. (2010). Digital learning for summarizing Arabic documents *Advances in Natural Language Processing* (pp. 79-84): Springer.

Christensen, J., Mausam, S. S., & Etzioni, O. (2013). *Towards Coherent Multi-Document Summarization.* Paper presented at the Proceedings of Association for Computational Linguistics pages 1163–1173, Atlanta, Georgia.

Conroy, J. M., O'Leary, D. P., & Schlesinger, J. D. (2006). CLASSY Arabic and English multi-document summarization. *Multi-Lingual Summarization Evaluation, 2006*.

Das, D., & Martins, A. F. (2007). A survey on automatic text summarization. *Literature Survey for the Language and Statistics, 4*, 192-195.

Davenport, M. (2012). Introduction to Modern Information Retrieval. *Journal of the Medical Library Association: JMLA, 100*(1), 75.

Deshpande, A. R., & Lobo, L. (2013). Text Summarization using Clustering Technique. *International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue8*

Douzidia, F. S., & Lapalme, G. (2004). Lakhas, an Arabic summarization system. *Proceedings of Document Understanding Conferences 2004*.

El-Haj, M., Kruschwitz, U., & Fox, C. (2011). *Multi-document Arabic text summarisation.* Paper presented at the Computer Science and Electronic Engineering Conference (CEEC), 2011 3rd.

El-Haj, M. O., & Hammo, B. H. (2008). *Evaluation of query-based Arabic text summarization system.* Paper presented at the Natural Language Processing and Knowledge Engineering, 2008. NLP-KE'08.

Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *(JAIR), 22*(1), 457-479.

Fan, J., Gao, Y., Luo, H., Keim, D. A., & Li, Z. (2008). *A novel approach to enable semantic and visual image summarization for exploratory image search.* Paper presented at the Proceedings of the 1st ACM international conference on Multimedia information retrieval.

Fisher, S., & Roark, B. (2007). *Feature expansion for query-focused supervised sentence ranking.* Paper presented at the Document Understanding (DUC 2007) Workshop Papers and Agenda.

Fiszman, M., Demner-Fushman, D., Kilicoglu, H., & Rindflesch, T. C. (2009). Automatic summarization of MEDLINE citations for evidence-based medical treatment: A topic-oriented evaluation. *Journal of biomedical informatics, 42*(5), 801-813.

Fukumoto, F., Sakai, A., & Suzuki, Y. (2010). *Eliminating redundancy by spectral relaxation for multi-document summarization.* Paper presented at the Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing.

Galanis, D., & Malakasiotis, P. (2008). *Aueb at tac 2008.* Paper presented at the Proceedings of the TAC 2008 Workshop.

Gholamrezazadeh, S., Salehi, M. A., & Gholamzadeh, B. (2009). A comprehensive survey on text summarization systems. *9*, 1-6.

Ghwanmeh, S. H. (2005). Applying Clustering of hierarchical K-means-like Algorithm on Arabic Language. *International Journal of Information Technology, 3*(3).

Giannakopoulos, G., Karkaletsis, V., Vouros, G., & Stamatopoulos, P. (2008). Summarization system evaluation revisited: N-gram graphs. *ACM Transactions on Speech and Language Processing (TSLP), 5*(3), 5.

Gupta, M. V., Chauhan, M. P., Garg, S., Borude, M. A., & Krishnan, S. (2012). An Statistical Tool for Multi-Document Summarization. *International Journal of Scientific and Research Publications 2*(5).

Gupta, V., & Lehal, G. S. (2010). A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence, 2*(3), 258-268.

Habash, N. Y. (2010). Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies, 3*(1), 1-187.

Haboush, A., & Al-Zoubi, M. (2012). Arabic Text Summerization Model Using Clustering Techniques *World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 2, No. 3, 62 – 67, 2012.*

Hammo, B., Abu-Salem, H., & Lytinen, S. (2002). *QARAB: A question answering system to support the Arabic language.* Paper presented at the Proceedings of the ACL-02 workshop on Computational approaches to semitic languages.

Hariharan, S., Ramkumar, T., & Srinivasan, R. (2012). Enhanced Graph Based Approach for Multi Document Summarization. *The International Arab Journal of Information Technology*, 4460-4411.

He, L., Sanocki, E., Gupta, A., & Grudin, J. (1999). *Auto-summarization of audio-video presentations.* Paper presented at the Proceedings of the seventh ACM international conference on Multimedia (Part 1).

Ibrahim, A., & Elghazaly, T. (2012). *Arabic text summarization using Rhetorical Structure Theory.* Paper presented at the Informatics and Systems (INFOS), 2012 8th International Conference on.

Ibrahim, A., Elghazaly, T., & Gheith, M. (2013). A Novel Arabic Text Summarization Model Based on Rhetorical Structure Theory and Vector Space Model.

Jayashree, R., Murthy, S., & Anami, B. (2012). *Categorized Text Document Summarization in the Kannada Language by sentence ranking.* Paper presented at the Intelligent Systems Design and Applications (ISDA), 2012 12th International Conference on.

Katragadda, R. (2010). GEMS: generative modeling for evaluation of summaries *Computational Linguistics and Intelligent Text Processing* (pp. 724-735): Springer.

Kaur, R., & Bhathal, G. S. (2013). A Survey of Clustering Techniques. *International Journal of Advanced Research in Computer Science and Software Engineering, 3*(5).

Kumar, Y. J., & Salim, N. (2011). Automatic multi document summarization approaches. *Journal of Computer Science, 8*(1), 133.

Larson, M. (2011). Automatic Summarization. *Foundations and Trends® in Information Retrieval, 5*(3), 235-422.

Lin, C.-Y., & Hovy, E. (2002). *From single to multi-document summarization: A prototype system and its evaluation.* Paper presented at the Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.

Liu, D.-X., He, Y.-X., Ji, D.-H., & Yang, H. (2006). *A Novel Chinese Multi-Document Summarization Using Clustering Based Sentence Extraction.* Paper presented at the Machine Learning and Cybernetics, 2006 International Conference on.

Lloret, E., & Palomar, M. (2010). Challenging Issues of Automatic Summarization: Relevance Detection and Quality-based Evaluation. *Informatica (Slovenia), 34*(1), 29-35.

Lloret, E., & Palomar, M. (2012). Text summarisation in progress: a literature review. *Artificial Intelligence Review, 37*(1), 1-41.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, *2*(2), 159-165.

McKeown, A. N. a. K. (2011). Automatic Summarization. *The Essence of Knowledge, 5*.

Minaei-Bidgoli, B., Parvin, H., Alinejad-Rokny, H., Alizadeh, H., & Punch, W. F. (2014). Effects of resampling method and adaptation on clustering ensemble efficacy. *Artificial Intelligence Review, 41*(1), 27-48.

Ouyang, Y., Li, W., Li, S., & Lu, Q. (2011). Applying regression models to query-focused multi-document summarization. *Information processing & management, 47*(2), 227-237.

Owczarzak, K. (2009). *Depeval (summ): dependency-based evaluation for automatic summaries.* Paper presented at the Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1.

Radev, D. R., Blair-Goldensohn, S., & Zhang, Z. (2001). Experiments in single and multi-document summarization using MEAD. *Ann Arbor, 1001*, 48109.

Radev, D. R., Hovy, E., & McKeown, K. (2002). Introduction to the special issue on summarization. *Computational linguistics, 28*(4), 399-408.

Rai, P., & Singh, S. (2010). A survey of clustering techniques. *International Journal of Computer Applications, 7*(12), 156-162.

Saad, M. K. (2010). Open Source Arabic Language and Text Mining Tools. *International Conference on Electrical and Computer Systems (EECS'10).*

Saad, M. K., & Ashour, W. (2010). Arabic Morphological Tools for Text Mining. *Corpora, 18*, 19.

Said, D., Wanas, N. M., Darwish, N. M., & Hegazy, N. (2009). *A study of text preprocessing tools for Arabic text categorization.* Paper presented at the The Second International Conference on Arabic Language.

Schilder, F., Kondadadi, R., Leidner, J. L., & Conrad, J. G. (2008). *Thomson reuters at tac 2008: Aggressive filtering with fastsum for update and opinion summarization.* Paper presented at the Proceedings of the first Text Analysis Conference, TAC-2008.

Schlesinger, J. D., O'leary, D. P., & Conroy, J. M. (2008). Arabic/English multi-document summarization with CLASSY—the past and the future *Computational Linguistics and Intelligent Text Processing* (pp. 568-581): Springer.

Sobh, I., Darwish, N., & Fayek, M. (2009). *Evaluation Approaches for an Arabic Extractive Generic Text Summarization System.* Paper presented at the proceeding of 2nd International Conference on Arabic Language Resource and Tools.

Suanmali, L., & Salim, N. (2009). Literature Reviews for Multi-Document Summarization.

Sun, J.-T., Shen, D., Zeng, H.-J., Yang, Q., Lu, Y., & Chen, Z. (2005). *Web-page summarization using clickthrough data.* Paper presented at the Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval.

Tiedan Zhu, X. Z. (2012). An Improved Approach to Sentence Ordering For Multi-document Summarization. *IACSIT Press, Singapore, 25*.

Vishal Gupta , G. S. L. (2010). A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence, 2*(3), 258-268.

Wadhvani, R., Pateriya, R., & Roy, D. (2013). A Topic-driven Summarization using K-mean Clustering and Tf-Isf Sentence Ranking. *International Journal of Computer Applications, 79*.

Wan, X. (2007). *TimedTextRank: adding the temporal dimension to multi-document summarization.* Paper presented at the Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.

Wan, X. (2008). *An exploration of document impact on graph-based multi-document summarization.* Paper presented at the Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Wan, X., & Yang, J. (2008). *Multi-document summarization using cluster-based link analysis.* Paper presented at the Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.

Wang, L., Raghavan, H., Castelli, V., Florian, R., & Cardie, C. (2013). *A Sentence Compression Based Framework to Query-Focused Multi-Document Summarization.* Paper presented at the Proceedings of ACL.

Yang, Z., Lin, Y., Wu, J., Tang, N., Lin, H., & Li, Y. (2011). Ranking support vector machine for multiple kernels output combination in protein–protein interaction extraction from biomedical literature. *Proteomics, 11*(19), 3811-3817.

Zechner, K., & Waibel, A. (2000). *DIASUMM: Flexible summarization of spontaneous dialogues in unrestricted domains.* Paper presented at the Proceedings of the 18th conference on Computational linguistics-Volume 2.