Text Mining Techniques and an Application on Natural Language Processing by Using R

Daniel Onyeka Onwochei

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Applied Mathematics and Computer Science

Eastern Mediterranean University
February 2019
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Assoc. Prof. Dr. Ali Hakan Ulusoy
Acting Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science in Applied Mathematics and Computer Science.

Prof. Dr. Nazim Mahmudov
Chair, Department of Mathematics

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Master of Science in Applied Mathematics and Computer Science.

Asst. Prof. Dr. Mehmet Ali Tut
Supervisor

Examining Committee

1. Prof. Dr. Rashad Aliyev ____________________

2. Asst. Prof. Dr. Hüseyin Lort ____________________

3. Asst. Prof. Dr. Mehmet Ali Tut ____________________

ABSTRACT

In our current society, technology is advancing at a very high pace, and these new inventions also generates large amount of data. Data is now increasing at an exponential rate, and this alarming growth rate has led to difficulty in getting and retrieving specific information from the web. Automatic summarization systems can help to resolve this information overload problem in an effective way. It easily identifies the important points from a document to produce a concise summary. Thus, the thesis investigates the extractive-based approach in generation of a summary from single documents/texts.

In the study, an extractive-based summarization framework (EBSF) was designed, also, an extractive-based text summarization system has been developed, evaluated and its workflow described. The framework implements several techniques and the summarization system generates extractive summaries from news articles using an extractive-based summarization technique which is based on the TextRank algorithm. Results from the various program testing shows that the summaries generated using our extractive-based summarization system offers an excellent tradeoff between time/length and accuracy. In this study, the summaries from the designed summarizing system, tends to be concise and contain less extraneous material.

Keywords: Big Data, Data Mining, Information Extraction, Natural Language Processing, Summarization, Text Mining.

ÖZ

Mevcut toplumumuzda, teknoloji çok hızlı ilerliyor ve bu yeni buluşlar da büyük miktarda veri üretiyor. Veriler artık üstel bir oranda artmakta ve bu endişe verici büyüme hızı, web'den belirli bilgilerin elde edilmesinde ve alınmasında zorluklara neden olmuştur. Otomatik özetleme sistemleri, bu bilgi aşırı yük sorununu etkili bir şekilde çözmeye yardımcı olabilir. Kısa bir özeti oluşturmak için bir belgedeki önemli noktaları kolayca tanımlar. Bu nedenle, tez, tek bir belgeden / metinlerden bir özet özeti çıkarmaya dayalı yaklaşımı incelemektedir.

Çalışmada, çıkartma temelli bir özetleme çerçevesi (EBSF) tasarlandı, ayrıca çıkartma temelli bir metin özetleme sistemi geliştirildi, değerlendirildi ve iş akışı tanımlandı. Çerçeve, çeşitli teknikleri uygular ve özetleme sistemi, TextRank algoritmasına dayanan bir çekişme tabanlı özetleme tekniğini kullanarak haber makalelerinden çekişme özetleri oluşturur. Çeşitli program testlerinden elde edilen sonuçlar, ekstraktif tabanlı özetleme sistemimizi kullanarak oluşturulan özetlerin zaman / uzunluk ve doğruluk arasında mükemmel bir değişim sunduğunu göstermektedir. Bu çalışmada, tasarlanan özetleme sistemindeki özetler özlü olma eğilimindedir ve daha az yabancı materyal içerir.

Anahtar Kelimeler: Büyük Veri, Veri Madenciliği, Bilgi Çıkarma, Doğal Dil İşleme, Özetleme, Metin Madenciliği.

DEDICATION

To My Amazing Family and Friends.

ACKNOWLEDGMENT

I would like to express my deepest gratitude to God for his Grace and Mercies shown throughout my graduate studies.

I would sincerely like to thank Asst. Prof. Dr. Mehmet Ali Tut, for his commitment, supervision and consistent support before and throughout this work.

My appreciation goes to every member of my thesis committee and the department of Applied Mathematics and Computer Science.

I remain indebted and grateful to my family for their consistent support and prayers. I love you all.

Finally, I would like to thank my friends for their support and encouragement during the period of this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
EBSF	Extraction Based Summarization Framework
IDF	Inverse Document Frequency
LSA	Latent Sematic Analysis
MMR	Maximal Marginal Relevance
NLP	Natural Language Processing
TC	Term Count
TF-IDF	Term Frequency Inverse Document Frequency
TF-ISF	Term Frequency- Inverse Sentence Frequency
VS	Vector Space

Chapter 1

INTRODUCTION

1.1 Background and Definitions

In our current society, technology is advancing at a really high pace, and these new inventions additionally generates great amount of knowledge. Based on this increased rate of information, the creation of big data came to be. Big Data simply means sophisticated knowledge. Many organizations have collected and held on to immense quantity of knowledge. However, they are unable to find valuable information hidden within the data by reworking these data into valuable and useful information (Han and Kamber, 2012). One very important facet of organization resource that provides organizations competitive advantage is information and this has given rise to knowledge management (KM) initiatives. Managing these sophisticated information is not an easy task and as such, several techniques are available to process and analyse these complex information. Storing bulk of unprocessed information in a storage space will cause wastage of storage space.

Furthermore, since the past 3 decades, there are many effort to perfect the idea of data discovery in databases and data processing. Several organizations now use information technology tools to manage data. Specifically, some of these organizations have started incorporating this approach to promote their marketing. According to (Han and Kamber, 2012), data mining plays an integral part in knowledge management. Data mining is basically the process of deriving from enormous datasets knowledge that is relevant with the aid of some tools.

1.1.1 Data Mining

Data mining is the process of removing hidden information considered useful from available data sets, and this can be considered infeasible and tasking if done manually. Data mining discloses unrevealed relationships from large amount of data by application of various advanced statistical and machine learning methods. The Results and models produced by data mining methods could serve as an effective support for knowledge management (KM) (Lukasz and Petr, 2016). Furthermore, it is now of necessity to develop better algorithmic processes and methods to extract relevant information from these big data which are also available in digital form and are in unstructured textual form. Therefore, text mining and information extraction as an area of research has become popular.

1.1.2 Text Mining

Text mining is defined according to (Srivastava and Sahami, 2009) as the non-trivial extraction of relevant piece of information that was previously not known but considered useful from large amount of unstructured data (text). Text Mining is the process of extracting interesting and meaningful information from natural language text. Text mining is also a process of evaluating text to extract useful and purposeful information. It includes a procedure of sorting out the info content, inferring designs inside the organized information, and lastly assessment and interpretation of the output (Fan and Li, 2013).

Basic text mining technologies includes (Zanasi, 2007):

- **Information Retrieval:** This manages the total field of information processing, it finds information resources from a given group of data sets;

- **Information Extraction:** Its task is majored on the removal of semantic information from the text;

- **Categorization:** Simply is the act of allocating textual documents to predefined set of topics based on their content. It is a collection of text documents, the process of ascertaining the accurate topic for each document;

- **Clustering:** It tends to find intrinsic structures in information, and arrange them into significant subgroups for further study and analysis;

- **Summarizations**: This entails an automated based process of generating a summary i.e. simply a shortened version of a document without neglecting its relevant information.

Information retrieval is considered as the simplest form of text mining. Major areas of text mining include: text clustering, automatic text classification, text summarization, topic extraction for a given set of texts and topic trends evaluation in text streams.

1.1.3 Text Mining and Data Mining

Text mining is sometimes likened to data mining but they are both different. Data mining deals with structured data from databases, on the other hand, text mining deals with unstructured or semi-structured data sets.

1.1.4 Text Mining Applications

Some of the popular text mining application areas include: academics, bioinformatics, copyright and customer profile analysis, internet security, risk management, management of knowledge, contextual advertising, spam filtering and analysis of data from social media.

1.1.5 Text Summarization

It is the process of reducing the size (irrelevant content) of a text document with software, as aforementioned earlier, in order to create a compressed version with the relevant points of the original document. Text summarization technique involves the process of reducing large document into smaller or shorter document. I.e. summarization handles the process of highlighting the significant parts of a text document and generates reasonable summaries which conveys the actual purpose (intent) of the document. (Suneetha, Parvez and Sameen, 2012).

Automatic data summarization is considered as a piece of machine learning and information mining methods. The strategy in content rundown is to recognize a segment of the record that contains the important data of the whole set. At the end of the day, the summarizer makes a whole report, by recognizing the most educational or essential sentences. Summarization systems are now applicable in such a large number of businesses today (Yatsko, 2010).

Automatic summarization has two methods to it. These are extraction and abstraction based methods. Extraction based text summarization is an approach used in recognizing sentences that are viewed as of high importance (rank) from the document based on word and sentence highlights and in this way blends them to produce a rundown. At the end of the day, extractive techniques work by choosing a subset of existing words, expressions, or sentences from the original content to shape the rundown. Text features that are important are extracted from the given document and based on this, a decision model is used to decide the degree of significance of each sentence based on its rated features. On the other hand, the method of abstraction develops an internal semantic representation, and with the aid of natural language generation techniques, it produces a summary that would be considered similar to that of a human. Summaries generated by abstractive method would possibly include verbal innovations. Over the years, most of the conducted researches have been geared towards the direction of the extractive methods, which are appropriate for image collection summarization and video summarization (Duy, Guilherme, Hurdle and religious mystic, 2016).

Furthermore, with the large increase in the textual information that we receive on a daily basis, text summarization system may well be useful in ascertaining the foremost vital contents of the text in a short time. It is now being introduced in fields were there's need to decrease the amount of transferred information. Text summarizations systems are now being deployed in diverse fields and different sectors of our everyday life, and it is however showing its importance by drastically reducing the time spent in finding the most important contents of a given text in a quick period of time (Yatsko and Vishnyakov, 2007).Furthermore, while information retrieval and different forms of text mining oftentimes makes use of word stemming, a lot of refined and sophisticated techniques from Natural language process (NLP) have been rarely used.

1.1.6 Natural Language Processing

Natural language processing (NLP) is basically a route for PCs to examine, comprehend, and get significance from human language in a brilliant and helpful way, it is considered as a subfield of software engineering, data designing, and artificial intelligence that deals with the interactions between computers and human (natural) languages (Steven Bird and Ewan Klein and Edward Loper, 2009).

With proper deployment of NLP, developers can compose and structure information to perform assignments, for example automatic summarization, speech recognition, translation, topic segmentation, named entity recognition, sentiment analysis and relationship extraction.

In NLP, R programming language plays an interestingly vital role in exploring big data, and also for computationally intense learning analytics. It has quite a large number of algorithms used for natural language processing.

Thus, in this study, a text summarization system will be designed using R programming language. The summary of the document will be created based on the degree of importance of the sentences in the document.

1.2 Design of the Study

Currently, finding information over the web is quite tasking especially when the information has to be specific, because the number of documents retrieved, wouldn't make deriving the specific information easy. Thus, this study seeks to design and develop an automatic text summarization system that can easily identify the important points from a document to produce a concise summary using R programming language. Also, the text summarization process is categorized into abstraction and extraction, with respect to their formation. In the case of abstraction, summaries are created in a way that they act as substitute for a given text. I.e. it has the ability to introduce into the summary words and phrases which were not originally present in the original document, but replaces the words with same meaning. However in the case of Extraction, the sentences are lifted from the original document, and then used to generates the summary. That is, sentences are chosen from the original document based on their level of relevance (importance), and these selected sentences are then used to generate the summary.

In this thesis, we would be employing the extraction method for summarization. Purely extractive summaries often tends to yield much accurate outputs in comparison to automatic abstractive summaries. This is considered so, based on the fact that lexical analysis and a data dictionary are used to find the related meaning of words in a document in abstractive summarization methods and those data dictionaries are required in other to cope with problems such as natural language generation.

1.2.1 Benefits

Due to the explosive growth of the web at large, prompting the route through several large amount of documents in order to discover intriguing information would be a herculean task and a waste of time and effort to carry out this search manually, and so, automatic text summarization has become a necessary tool for internet users. Thus, a study of this sort using R programming language, can facilitate in extracting the sentences considered vital from a source document when compared to other commercially available text summarizers. In fact, the most vital advantages of text summarization system is that it brings down the time spent in reading a full document and provide fast direction to the actual information in a document. The text summarization system finds the most vital text units and yields them as the summary of the source document. Furthermore, the automatic text summarization system which is developed dependent on the advantages of different assets in type of consolidated model will deliver a decent outline (summary) for the original document.

1.2.2 Scope

In this thesis, the following will be covered:

- An overview of the concept of text summarization and a review of existing text summarization systems, will be provided.

- Proposal of an extraction-based automatic summarization system and the development of a text summarization system using R programming language; description of the software design and software outputs evaluation.

1.2.3 Tool

R-Programming Language: The choice of programming language for the implementation of the text summarization system is R programming language. R is a language quite suitable for statistical computing, handling graphics, developing Artificial Intelligence (AI) applications and excellent choice for natural language processing.

Chapter 2

LITERATURE REVIEW

2.1 Overview of Text Summarization

A summary is a text according to (Hovy, 2005) extracted from a given set of texts, that possesses an essential part of information gotten from the source text(s), and this summary tends to be half but not more than half of the original text(s).

According to Kan, McKeown and Klavans (2011), text summarization can be described as the process of highlighting and the selection of the most vital set of knowledge (information) from an original document to give an abridged version for a particular source.

Automatic text summarization is the use of the computer system is achieving the aforementioned task of text summarization. According to Edmundson (1999), the output from the summarization system can either be an extractive or abstractive output. It is extractive if the summary can play the role of being considered as an alternative for the original document (Edmundson, 1999).

Below are some few existing related works on text summarization.

Bhargava, and Sharma (2016) proposed a graph based method that produces summaries of a document. Abstractive based summaries is generated to convey the idea in the source document. According to the authors, text summarization is the process of removing redundant data from a document and vital pieces of information are selected as the summary in the shortest possible way. The authors further posited that with the number of data available on social networks, it is now needful to carry out analysis on these texts for seeking information that will be more useful to people.

Jishma, Sunitha and Ganesha (2016), discussed about various works carried out using ontology for abstractive text summarization. According to the authors, with wide use of Internet and the occurrence of huge information, a robust text summarization system is needed to ease the reading of these information. Automatic summarization systems shorten these files by taking out the parts that are of necessity most important. Summarization is generally grouped into: extractive and abstractive. Abstraction based summarization requires an actual comprehension of the source text and a semantically related text is generated. Abstractive summarization requires a good knowledge of NLP tasks. There are various techniques in use for abstractive summarization.

Mehdi, Trippe and Gutierrez (2017) reviewed and described the main approaches to automatic text summarization. The investigation explicitly checked on the different forms for rundown and depict the effectiveness and deficiencies of the different strategies. The investigation was required because of the ongoing blast in the measure of content information from an assortment of sources. As indicated by the creators, the volume of content is a priceless wellspring of data and learning which should be effectively condensed to be valuable.

Marcu (1999) in his study, utilized comprehensive inquiry to create the best Extract from guaranteed (Abstract, Text) tuple, where the best Extract contains a lot of provisos from Text that have the most noteworthy comparability to the given Abstract. In addition, Donaway (2000) utilized thorough inquiry to make all the sentence concentrates of length three beginning with 15 TREC Documents, so as to pass judgment on the execution of a few rundown assessment measures proposed in their paper.

Kanitha and Mubarak (2016) reviewed existing extractive content summarization models. Additionally, various calculations were examined and their assessments were clarified. The embodiment of the investigation is to watch the idiosyncrasies of existing extractive outline models and to locate a decent methodology that manufactures another content synopsis framework. As indicated by the creators, the accessibility of online data demonstrates a need of proficient content rundown framework. The content outline framework can utilize either the extractive and abstractive techniques. The noteworthy sentences are chosen in the extractive outline technique and this choice depends on sentence positioning strategies, while with the abstractive rundown framework, the real thought on the report is comprehended and create a general thought of the subject.

Saeedeh, Mohsen and Bahareh (2010) displayed a scientific categorization of synopsis frameworks and characterizes the most critical criteria for a rundown which can be produced by a framework. Moreover, unique strategies for content rundown just as principle ventures for synopsis process was talked about. The investigation additionally experience primary criteria for assessing a content synopsis. As per the creators, content synopsis frameworks are among the most alluring exploration territories these days. Outline frameworks gives offices to distinguishing the significant purposes of data in an archive and all things considered, the client will invest less energy perusing the whole report.

Hakan and Rada (2010), broke down the theme identification phase of single-archive programmed content rundown in four different zones, comprising of scholarly, logical, newswire and legal documents. The examination explicitly investigated the outline space of every area by means of a thorough inquiry methodology, and find the probability density function (pdf) of the ROUGE score disseminations for every area. The pdf was then used to ascertain the percentile rank of extractive summarization systems.

Ani and Kathleen (2012) gave a wide diagram of existing methodologies dependent on these refinements, with specific consideration on how portrayal, sentence scoring or outline determination techniques adjust the general execution of the summarizer. The examination additionally call attention to a portion of the quirks of the assignment of outline which have presented difficulties to machine learning approaches for the issue, and a portion of the recommended arrangements. As per the creators, content is spoken to by a different arrangement of conceivable markers of significance which don't go for finding topicality. These markers are assembled, utilizing machine learning calculations, to score the significance of each sentence. At long last, an outline is produced dependent on the sentences that has the most elevated positioning.

Chieze, Atefeh, Farzindar and Guy (2010) displayed a data framework that orders and condenses content. The investigation clearly portrayed the utilization of a blend of etymology mindful transductor and XML advancements for bilingual data extraction from decisions in both English and French dialects inside a lawful data and condensing framework. Additionally given in the investigation were the principle difficulties and how they were handled by plainly isolating dialect and space subordinate terms and vocabularies.

Gholamrezazadeh, Salehi and Gholamzadeh (2009) introduced a scientific classification of outline frameworks and characterizes the most essential criteria for a synopsis which can be produced by a framework. Furthermore, extraordinary techniques for content rundown just as fundamental strides for synopsis process was examined. The examination likewise experienced fundamental criteria for assessing a content rundown. As indicated by the creators, content summarization frameworks are among the most appealing exploration zones these days. Outline frameworks gives the choice of determining the real purposes of writings and all things considered the client or user will invest less energy in perusing the entire record.

Verma, WeiLu and Chen (2016) proposed a user query based text summarization system. According to the authors, as tremendous measures of learning are made quickly, compelling data get to turns into an imperative issue. Particularly some basic zones, for example, prescription and fund, efficient recovery of brief and applicable data is very required. Accordingly, in the investigation, the rundown framework planned is uniquely tweaked to abridge restorative records.

Thanh and Philipp (2016) displayed a methodology for translating watchword inquiries utilizing foundation learning accessible in ontologies. In light of a couple of suppositions about how individuals depict their data needs, a methodology was exhibited which makes an interpretation of a catchphrase question into a Description Logics (DL) conjunctive inquiry which can be surveyed as for a basic knowledge base

(KB). One key issue the strategy experiences is the way that it doesn't mull over that watchwords can be hazy as to names in the philosophy however it just considers the primary coordinating cosmology component to begin the procedure.

In Peroni, Motta and d'Aquin (2016), the authors address the issue of recognizing the ideas in a metaphysics, which best condense what the philosophy is about. A few measures were considered, and furthermore, a few calculations were created, to distinguish key ideas of a cosmology. The criteria territories include: name effortlessness which favours ideas that are named with basic names while punishing mixes; essential dimension which estimates how "focal" an idea is in the scientific categorization of the philosophy; thickness considers ideas which are luxuriously described with properties and ordered connections; inclusion intends to guarantee that no critical piece of the metaphysics is ignored; and fame distinguishes ideas that are normally utilized. The outline results, for example real ideas, were evaluated against human assessors' synopses, alluded to as ground truth.

Nesrine (2015) presented an exhaustive system for building a space explicit cosmology by applying two methodologies for metaphysics accomplishment so as to make the area philosophy. The principal approach was to make a little area explicit centre cosmology from starting and so as to recover space related website pages, a web crawler is connected to this metaphysics. The second methodology takes a settled thesaurus as a fundamental vocabulary reference set and changes over it to a philosophy portrayal.

Zhang (2016) proposed a novel way to deal with programmed metaphysics synopsis dependent on RDF Sentence Graph. The creators made a correlation of five diverse centrality estimations in estimating the remarkable quality of RDF sentence and characterized a reward-penalty re-ranking algorithm to make the synopses comprehensive. The near investigation uncovered that weighted in-degree centrality measures and a few eigenvector centralities all have great execution in producing qualified synopses after re-ranking. Besides, the outcomes from the examination demonstrates that cosmology outline approach was doable and promising.

Likewise, numerous specialists have tried to utilize ontology to enhance the procedure of outline. Greater part of the records on the web are space related on the grounds that they bargains on comparative subject or occasion. Every space has its very own insight structure and that can be all around signified by cosmology.

Lee (2014) fuzzy ontology with fuzzy concepts is acquainted for Chinese news rundown with model dubious data and henceforth can more readily portray the area information. In this strategy, the space specialists initially characterize the area philosophy for news occasions. Next, the report pre-processing stage creates the important terms from the news corpus and the Chinese news lexicon. At that point, the term classifier orders or characterize the critical terms dependent on occasions of news. Moreover, the fluffy surmising stage produces the participation.

Mithun and Munirathnam (2012) displayed a semi-automatic development of an ontology library for the topics characterized in the National Intelligence Priorities Framework (NIPF). They utilize a best in class apparatus named Jaguar-KAT for information procurement and space understanding, with diminished manual mediation to produce NIPF ontologies stacked with rich semantic substance. The apparatus by configuration manufactures area explicit ontologies from content. Wellspring of contribution to Jaguar incorporate content from MS Word, PDF and HTML website pages, and so forth likewise, the philosophy/information base produced by Jaguar incorporates ontological pecking order, ideas and logical learning base.

2.2 Review of Existing Text Summarization Systems

Luhn (1958) created the first automatic text summarizer for summarize technical articles. Each sentence in the document was ranked with respect to their word frequency and phrase frequency. The word frequencies were calculated after processing the stemming and stop word removal. According to the author, the word frequency explains the level of importance of a sentence. All sentences are ranked on their level of importance and each sentence get a rank score. The sentences that are considered top ranked are then highlighted and selected as the summary sentences.

Boxendale (1958) proposed a position method for sentence extraction. The author argued that some important sentences are placed in some fixed positions. The author checked two hundred (200) paragraphs in newspaper articles and 85% of the paragraphs, the topic sentence come first and 7% come last. Based on this, the author posited that in newspaper articles the first sentence in each paragraph got high chance to include in summary.

Edmundson (1969) developed a new method in automatic summarization. This approach calculates the candidate sentence by adding some sentence scoring parameters such as cue phrases, title plus heading, sub heading words, keywords and sentence location. This sentence scoring parameters are used to get the top ranked sentences. The stop words are taken away from the source document. This approach also gives high score to title word, heading and sub-heading words which are included in the sentences.

Kupiec, Pedersen and Chen (1995) developed a trainable document summarizer. The trainable document summarizer executes sentence extraction on the basis of some sentence weighting methods. The important methods used in this summarizer are:

- Sentence length cut-off feature, this feature is used to remove or exclude sentences which contains less than the pre-specified number of words.
- Cue words and phrases related sentences are added.
- The foremost sentence in each paragraph is included.
- Thematic words; the most frequent words are included.

Thus the rating of sentences are executed based on the aforementioned features and thereby selecting high scored sentences as the summary sentences.

Brandow, Mitze and Rau (1995) designed a text extraction system named ANES. The ANES text extraction system is a domain-independent summary system for summarizing news articles.

Barzilay and Elahadad (1997) develop a summarizer based on lexical chain method. The sentences are generated using a collection of the related words which form a lexical chain. The lexical chain connects the semantically similar terms with the different parts of source document. The lexical chain developed by the authors was done using a WordNet.

Boguraev, Branimir and Kennedy (1997) develop a single document and domain independent system using linguistic techniques. The system selects sentences based on title word, noun phrases and topic related sentences.

Kan-Yen and Kathleen (1999) designed and developed a summarization system which follows a question answering approach. The system is a two staged system which takes a question and then summarizes the source text and generates an answer to the question. The system uses a named entity extractor to find the significant or important term of the document.

Lin and Hovy (1999) developed a machine learning model for summarization making use of decision trees instead of a naive Bayes classifier. The text summarization system developed produces summaries of the web documents based on both extractive and abstractive based summaries. The text summarization system first identifies the main topics of the document using the chain of lexically connected sentences. The statistical approaches such as cue phrases, position, word frequency, numerical data, proper name, etc. are used for extractive-based summary.

Barzilay, McKeown and Elhadad (1999) developed a multi document summarization system named Multigen. The system recognized differences and similarities across the documents by applying the statistical approaches. It extracted high weight sentences that denote important aspect of information in the set of related documents. This is done by applying the machine learning algorithm to group paragraph sized chunks of text in related topics. Sentences from these groups are parsed and the resultant trees are merged together to produce the logical representations of the regularly occurring concepts. Matching concepts are selected based on the linguistic knowledge such as part-of-speech, stemming, synonymy and verb classes.

Jing, Hongyan and McKeown (2000) developed a cut and paste system designed to understand the key concepts of the sentences. These key ideas are then put together to create new sentences. The system first copies the surface form of these major ideas and place them into the summary sentences. The major and significant idea are identified by probabilities learnt from a training corpus and lexical links.

Swesum (2000) create summaries from Swedish or English texts either the newspaper or academic domains based on weighted word level features of sentences. The system uses statistical, linguistic and heuristic approaches to generate summary. The approaches include: Baseline, Title, First sentence, Word frequency, Sentence length, Position score, Numerical data and Proper names, etc. The first sentence in the processed article gets the highest score. The system uses the formula: 1/n, where n is the line number. The first sentence in the paragraph of the processed article gets the high score. The system uses a combination of function on the above formula to generate the required summary sentences.

Radev, Jing, Stys and Tam (2001) designed and developed a text summarization system named MEAD. The system computed the score of a sentence on the basis of a centroid score. The centroid score is created based on tf-idf values, closeness to the initial sentence of the file (document), position of the sentence in the file, the length of sentence etc. The highest ranked sentences forms the summary.

Radev (2001) web based summarizer system named Webinessence, it is an improved version of the MEAD summarizer. The design of the system is in two stages. The first stage, the system collects uniform resource locations (URLs) from the different web pages and extracts the news articles. The second stage groups the data from different documents. A centroid algorithm is used for identifying the representative sentences. The system, while generating the final summary avoid repetition.

Balabantary, Sahoo, and Swain (2012) developed a text summarization system using term weights. The authors designed and developed a statistical method to summarize the source text. The system first split the sentences into tokens and then remove the stop words. After the process of stop words removal, a value with regards to weight is given to each single term. This weight is thereby calculated based on the frequency of a term in the sentence divided by frequency of term in the document. Additional score are added to the weight of terms which are shown in bold, italic, underlined or any combination of these. The individual sentences are then ranked based on their weight value that is calculated as weight of individual term divided by total number of terms in that sentence. Conclusively top ranked sentences includes the initial sentence of the initial paragraph of the input text are extracted to form the summary generated.

Landauer and Dumais (1997) developed a text summarization system using Latent Semantic Analysis (LSA) framework. LSA is a method for extracting the hidden semantic representation of terms, sentences, or documents. LSA is an unsupervised method for extracting the semantics of terms by examining the co-occurrence of words. The method starts with the representation of input documents as a word by sentence matrix A. Each row denotes a word from the document and each column denotes sentence in the document. So in a given matrix, A=mXn with ‘m’ words and ‘n’ sentences. An application of the Singular Value Decomposition also known as SVD from linear algebra is implemented on matrix A. SVD of the given matrix mXn has a formal definition of A=U∑VT. Matrix U serves as the mXn matrix for numbers that are considered real (real numbers). Matrix ∑ is diagonal nXn matrix. The VT matrix is nXn matrix each row denoted as sentences.

Pourvali and Abadeh (2012) designed a text summarization system using an approach based on lexical chains method and the exact meaning of each word in the text is determined by using WordNet and Wikipedia. The score of a sentence is determined by the number and type of relation in the chains. The highest chains sentences are extracted and generated as final summary sentences.

Khushboo, Dharaskar and Chandak (2010) proposed a method based on graph based algorithms for text summarization. This approach creates a graph from the source text. The nodes are denoted as sentences and the edge denotes the semantic relation between sentences. The weight of each node is computed and the highest ranking sentences are extracted and generated for final summary.

2.3 Summary of Literature Review

Vast majority of the early work on text summarization reviewed were based on single document summarization in technical articles. Due to lack of powerful computers and technological developments, such early summarization systems consider some simple surface level features of sentences like word frequency, position, length of the sentence, etc. However, with the development of Artificial Intelligence (AI) technology, recent summarization systems depend on AI technology. Thus, this study proposes a way to deal with the issue of enhancing content selection in automatic text summarization by using R programming language. This method is a trainable summarizer, which considers several features, for each sentence to create summaries using R programming language. The system also analyses the document providing graphical representation of the summarized sentences.

Chapter 3

OVERVIEW OF SUMMARIZATION WORKFLOW AND THE EXTRACTION-BASED SUMMARIZATION FRAMEWORK

3.1 Introduction

In this section, we would elaborate on the summarization workflow, develop and describe an extraction-based summarization framework (EBSF), which would make use of diverse methods in creating a summary.

3.2 Summarization Workflow

The processes involve in automatic text summarization is multistage in nature. Figure 3.1 below illustrate an overall extraction based summarization workflow.

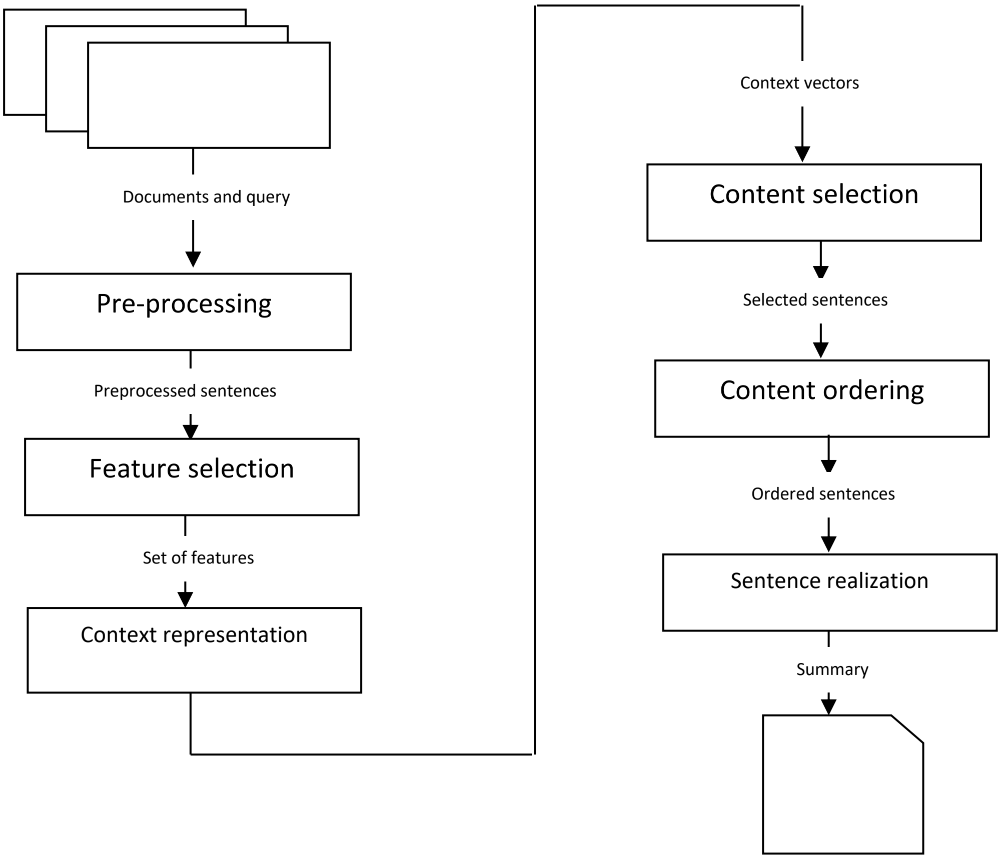


Figure 3.1: General Extraction-Based Summarization Workflow

Based on this, the file inserted to the system (summarizer) could either be a one or set of multiple documents. Irrespective of the number of documents, there are basically six sub-tasks in the summarization processes. These six sub-tasks are discussed below:

Pre-processing Task: This task collects a non-analyzed set of unstructured data which is then loaded to the system and then introduces required procedures to remove non useful text components (such as non-content words and punctuations) for the summarization.

Feature Selection Task: This task recognizes words, phrases, features, topic terms or topic signatures in a pre-processed text and such information forms valuable characteristics of textual contexts.

Context Representation Task: The features obtained from the feature selection task is used in this stage. The features are further represented in a context form suitable for further processing.

Content Selection Task: This task recognizes contexts that should be contained in a summary by computing the similarities between textual contexts.

Context Ordering Task: This task positions contexts highlighted in the aforementioned subtask of content selection to make a comprehensible and readable text.

Sentence Realization Task: This task further processes the sub-sentence level with the sole purpose of improving readability and clarity of a text.

The aforementioned tasks are further discussed elaborately in the subsequent sections.

3.2.1 Pre-processing

The pre-processing stage involves the following steps: stemming of words, removal of stop words, and segmentation of texts and expansion of query.

Word Stemming: This process involves the decreasing of a word from its inflected forms to its root form in order to make the data less sparse. Word stemming are usually applied in jobs that deals with information retrieval.

Stop Words Removal: At this stage words that have high frequency and do not carry any particular information are removed from the document to bring down the number of features. Some of the words included in a stop words list include pronouns, prepositions, conjunctions, and auxiliary verbs. In other words, all words in the stop words list are removed from the source document during this stage.

Query Expansion: The content of a query helps in knowing which aspect of a document is important. Usually when these queries are expanded, it provides more clues for establishing the important aspects of a document. Several approaches using different external sources have been suggested for this task. This include WordNet which can recognize related words in a query.

Text Segmentation: Sentences are referred to as a context in extraction-based summarization. Text segmentation involves the act of dividing an input document into an independent piece of information. In extraction-based summarization, identifying sentence boundaries is usually a difficult task, especially when there are periods used to denote abbreviations. Extractive-based text summarization demands high accuracy sentence boundary detection methods in order to enhance coherence and readability of a summary. Text segmentation are generally done using either rule-based approach or machine learning approach.

Feature Selection: Lexical features like words and phrases are mostly used in automatic text summarization and such series of features is regarded as N-gram. N denotes amount of terms in the series. Features or topic signatures in text categorization and automatic text summarization makes use of these N-grams. Also, single words known as unigrams and two-word phrases known as bigrams are mostly used in extractive-based text summarization. Every word and phrases have different amount of information and apart from stop words removal approach, there are other different approaches that can be applied to remove features that are not important from a piece of information and also to weight or score the important features according to their importance. More sophisticated weighting techniques such as the term frequency and inverse document frequency (TF-IDF) and log-likelihood ratio are used. Statistical information in distributing the words in the texts is the technique used by both. The TF-IDF is a weighting function that is simply calculated by the multiplication of a pair of components namely: the term frequency (TF) and the inverse document frequency (IDF). The Term Frequency defines each word significance in a file, on the other hand, IDF specifies significance of a given word throughout the total collection of files. Frequent occurrences of a word in a particular file but not in other files, is regarded or marked to be insignificant and such word is given a high weight score.

Context Representation: Unlike with normal data or word processing where data reserved in a PCs memory are rendered as a series of computer bits, in tasks concerned with NLP, the emphasis is placed on the text interpretation instead of the syntactical form. Presently, the Vector Space (VS) model is considered as the highest textual representation model being utilized.

Similarity Measures: Textual similarity is related with the semantic closeness of two textual contexts. A set of documents are considered similar when they possess similar concepts. In automatic text summarization, similarity metrics are utilized when it comes to the aspect of centrality-based context selection and in the area of unnecessary text recognition.

Content Selection: This task is basically aimed at the recognition of a set of sentences that are considered to have significant information based on their relevance, redundancy and length, with the sole goal of identifying which of the sentences in the input files are important to take into the synopsis. Content selection in extractive-based summarization is addressed by making use of a supervised or unsupervised approach.

a.) Supervised Methods: These techniques utilizes a classifier that is prepared on a given accumulation of records (documents) with related extracts. These set of related extracts generates the possibility of labelling the sequence of text in the given file with a binary value: 1 (representing its inclusion in the summary), or 0 (stating its exclusion from the summary).

b.) Unsupervised Methods: The unsupervised contents election methods collects important strings of words without training on a given collection of labelled files. The algorithms used by this summarization method, mostly are either centroid-based or centrality-based. In centroid-based, the idea utilized is essentially to highlight Informative sentences (Sentences which possesses words that contain information considered relevant) these are also regarded as topic signatures. The value of information contained in the remaining set of words is calculated using popular weighting schemes.

Redundancy Removal: This is the process of removing documents on the same topic which contains huge amount of similar information. It is considered a key problem. The maximal marginal relevance (MMR) model proposed by Carbonell and Goldstei (1998), is a model quite effective for redundancy removal.

Context Ordering: Content ordering is a method aimed at the organization of selected contexts into a reasonable summary.

Sentence Realization: This techniques changes the sentences that is as of now incorporated into the summary so as to influence the summary by making it brief just as to enhance the accuracy and meaningfulness of the summary.

3.3 Extractive-Based Summarization System Overview

Extractive-based summarization framework (EBSF) designed and implemented in this study is designed to perform single-document summarization tasks. Also, the summarizer implemented is purely extractive, which implies that the summarizer does not implement sentence realization techniques due to the complexity of such techniques and limited time for the study. In addition to the above, ordering of sentences is also not performed. Thus, the basic function of the applied extraction-based summarization system is basically content selection mechanism (i.e., the framework chooses which sentences ought to be included in the summary. In other words, every single sentence is given a score, thereafter, sentences with scores that are considered high are chosen. The general workflow of the applied system is given in figure 3.2 below:

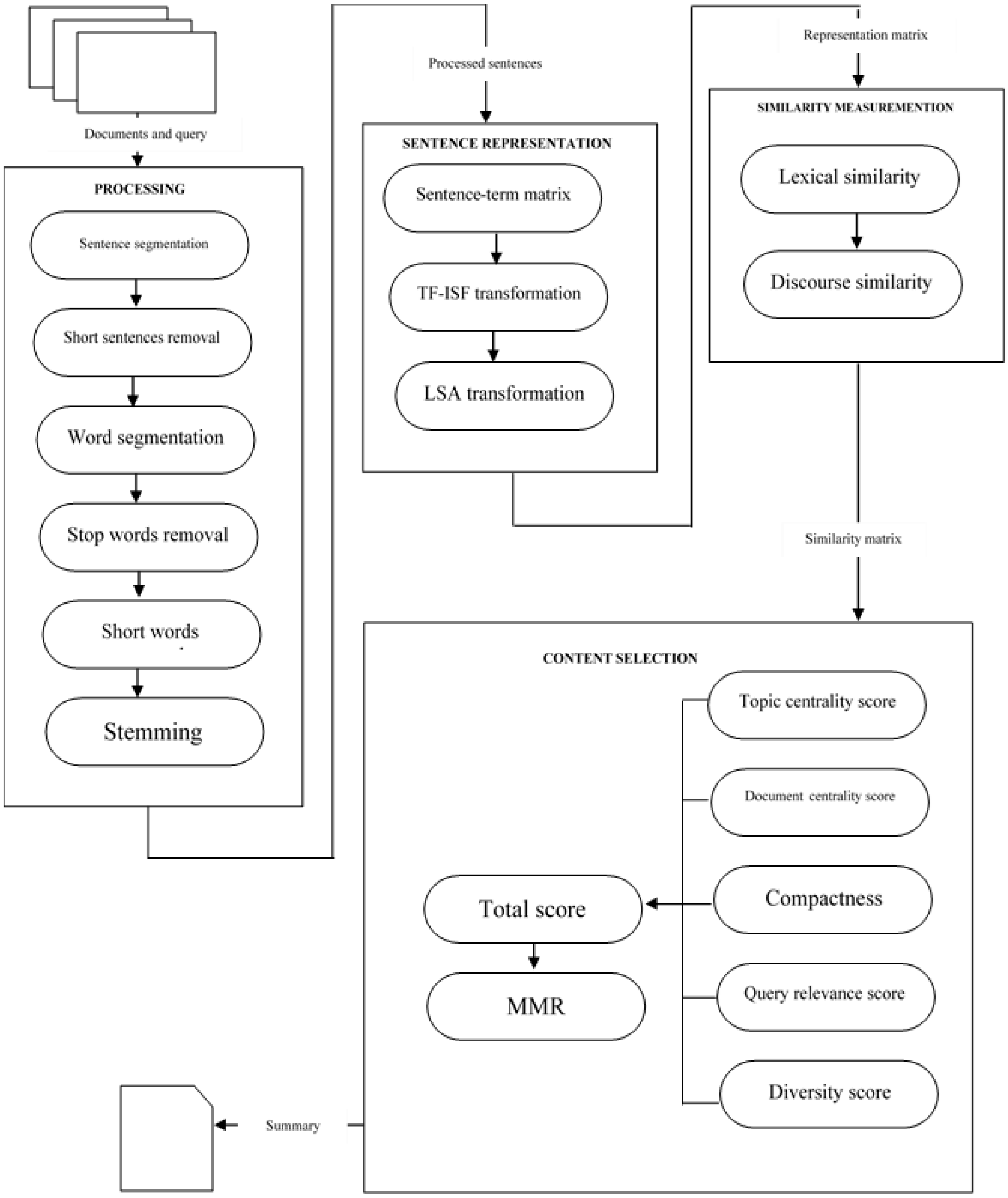


Figure 3.2: The workflow of the Extraction-Based Summarization Framework.

3.3.1 Pre-processing

The pre-processing stage involves the following routines:

1. Sentence segmentation
2. Short sentence removal
3. Word segmentation
4. Stop word removal
5. Short word removal
6. Stemming

3.3.2 Sentence Representation

Sentence representation and selection of features are closely coupled and they are discussed in the following subsection. Choosing distinctive words from pre-processed sentences is considered as obtaining features. A context vector constitutes every sentence. Then, a matrix for representation is accumulated by the vectors. The framework implements the following representation models:

Term count (TC): This represents sentences as vectors in which components are the supreme frequency of words in the sentence.

Term frequency-inverse sentence frequency (TF-ISF): here, sentences are used instead of documents.

3.3.3 Similarity Measurement

Similarity is registered between each pair of context vectors from a representation matrix. In getting centralities and locating queries in appropriate sentences, Similarities are utilized in the synopsis procedure:

Jaccard similarity coefficient: This is a set-based similarity metric which is utilized for deciding likenesses between sentences denoted with TC representation.

3.3.4 Sentence Selection

This is considered as the concluding venture during the development of a synopsis in EBSF. The determination procedure primarily depends on likenesses connecting sentences. Every given sentence is allotted five diverse scores:

Query relevance score is considered a likeness connecting a sentence and a query; strings of words like the inquiry gets scores that are higher. It speaks to how applicable a specific sentence is to the data asked for by the query.

Document centrality score is used in the measurement of the centrality of a sentence within the given document in which there is occurrence of this sentence.

Topic centrality score this is used to measure the centrality of a sentence in a given set of documents considered as input.

Compactness score is based on the length of a sentence; longer sentences get lower scores. This score may be utilized properly when asked to make a choice between sentences of diverse sizes that communicates related information.

Diversity score is done in agreement to a closeness connecting a candidate sentence and sentences effectively incorporated into the outline; higher scores are assigned to less comparative sentences. This derived score is then utilized in the maximal marginal relevance (MMR) calculation, so as to avoid redundancy.

Chapter 4

IMPLEMENTATION AND EVALUATION

4.1 Implementation

The summarizer has been implemented in R programming language and it is designed primarily for web articles/ news summarization. The program can run as a batch file or at interactive modes under R studio. Furthermore, the software was designed using a modular approach. The diverse sets of modules were conclusively linked up to form this coherent system.

For simplicity and ease, the processing functions of the extractive-based text summarization system has been decomposed into sub-tasks.

A module by definition is considered a modest-sized sub-program, which possesses the ability to function independently, however if it be absent from a system should only disable its unique task to that program. The source program listing is given in Appendix A of this report.

It is also important to note that this study implements TextRank, which is an extractive summarization based method using R programming language. The TextRank does the following:

TextRank reads the text, parse it, remove the unnecessary stop words, and tokenize it and counting the occurrence of the words and phrases and then it starts weighting based on the words and phrases level of occurrences. After weights are done, the weights are normalize between 0 and 1 by TextRank. Then the system finds the most important words, phrases or sentences. Furthermore, the algorithm for the entire extractive-based text summarization system is given below.

Text Pre-Processing and Information Extraction are the phases used in executing the algorithm. Selection of document having being carried out by the user is sent to the file classification module, which discerns the document type before reading its contents.

Phase 1: Text Pre-Processing

Text pre-processing is carried out with the aid of the following sub-tasks:

1. **Syntactic analysis:** in view of the unverifiable arrangement of unstructured content, choosing the beginning and ending of a giving information is required. By and by we accept that full stop symbol denotes the finish of sentence. Any string of characters up to full stop symbol is treated as one sentence.

2. **Tokenization:** This is a very vital stage for processing text, in this step, the sentences are broken into tokens, which may consist of words, numbers, symbols etc. these forms the pieces of a sentence.

3. **Semantic analysis:** Semantic analysis module simply translates the job each word plays in a sentence. It then allots a tag to each for example noun, verb, adjective, adverb, etc. For this module to be executed, part-of-speech (POS) tagger is used. Part-of-speech tagging is a way of getting the understanding the part played by each word and after that, passing a proper tag to it.

4. **Stop word removal:** Stop words are fundamentally words that happen habitually, however, they have no criticalness when we assume control over all importance of the sentence in a natural language text. These words are expelled before the file is considered for next level of processing.

5. **Stemming:** This is associated with the task of finding the root form of a certain word in a given text document. After the stemming process the text is thereby forwarded into the Information extraction module.

Phase 2: Information extraction

Information extraction uses the pre-processed text to discover significant information using algorithms based on machine learning designed and implemented by us.

This part of the implementation includes the modules shown below:

1. The system performs the information retrieval by calculating the term frequency–inverse document frequency (tf–idf) value. The tf-idf is considered as one of the most key weighting schemes used to show importance of a word in a document.

2. The system calculates each sentence score.

3. Once the above two processes are finalized, the system generates summary.

4.1.1 Implementation Examples

Below, we present two examples summaries using two articles from http://www.time.com website. The R program codes for summarizing each of the considered article is given in Appendix A of this report. The program codes can be run at the command prompt of the R studio or as a batch file.

Furthermore, the summarizer is domain dependent, however, it can be easily adapted to or modified to suite other websites. The summarizer program was used to derive summaries of series of articles in the website used as case study (Times.com). Results from the various program testing shows that the summaries generated using our extractive-based summarization system offers an excellent trade-off between time/length and accuracy. The summaries from the summarizer system designed in this study tend to be concise and contain less extraneous material.

Example 1:

TITLE: Fitbit's Newest Fitness Tracker Is Just for Kids

Fitbit is launching a new fitness tracker designed for children called the Fitbit Ace, which will go on sale for $99.95 in the second quarter of this year.

The Fitbit Ace looks a lot like the company’s Alta tracker, but with a few child-friendly tweaks. The most important of which is Fitbit’s new family account option, which gives parents control over how their child uses their tracker and is compliant with the Children’s Online Privacy Protection Act, or COPPA. Parents must approve who their child can connect with via the Fitbit app and can view their kid’s activity progress and sleep trends, the latter of which can help them manage their children’s bedtimes.

Like many of Fitbit’s other products, the Fitbit Ace can automatically track steps, monitor active minutes, and remind kids to move when they’ve been still --for too long. But while Fitbit’s default move goal is 30 minutes for adult users, the Ace’s will be 60 minutes, in line with the World Health Organization’s recommendation that children between the ages of five and 17 get an hour of daily physical activity per day. Fitbit says the tracker is designed for children eight years old and up.

Fitbit will also be introducing a Family Faceoff feature that lets kids compete in a five-day step challenge against the other members of their family account. The app also will reward children with in-app badges for achieving their health goals. Fitbit’s new child-friendly fitness band will be available in blue and purple, is showerproof, and should last for five days on a single charge.

> The Ace launch is part of Fitbit's broader goal of branching out to new audiences. The company also announced a new smartwatch on Tuesday called the Versa, which is being positioned as an everyday smartwatch rather than a fitness-only device or sports watch, like some of the company's other products.
>
> Above all else, the Ace is an effort to get children up and moving. The Centers for Disease Control and Prevention report that the percentage of children and adolescents affected by obesity has more than tripled since the 1970's. But parents who want to encourage their children to move already have several less expensive options to choose from. Garmin's $79.99 Vivofit Jr. 2, for example, comes in themed skins like these Minnie Mouse and Star Wars versions, while the wristband entices kids to move by reflecting their fitness achievements in an accompanying smartphone game. The $39.99 Nabi Compete, meanwhile, is sold in pairs so that family members can work together to achieve movement milestones.

Evaluation of Example 1

With respect to the program operations, the program first identifies eighteen (18) sentences in example 1 article as follows:

Sentence

1 fitbit is launching a new fitness tracker designed for children called the fitbit ace, which will go on sale for $99.95 in the second quarter of this year.

2 the fitbit ace looks a lot like the company's alta tracker, but with a few child-friendly tweaks.

3 the most important of which is fitbit's new family account option, which gives parents control over how their child uses their tracker and is compliant with the children's online privacy protection act, or coppa.

4 parents must approve who their child can connect with via the fitbit app and can view their kid's activity progress and sleep trends, the latter of which can help them manage their children's bedtimes.

5 like many of fitbit's other products, the fitbit ace can automatically track steps, monitor active minutes, and remind kids to move when they've been still for too long.

6 but while fitbit's default move goal is 30 minutes for adult users, the ace's will be 60 minutes, in line with the world health organization's recommendation that children between the ages of five and 17 get an hour of daily physical activity per day.

7 fitbit says the tracker is designed for children eight years old and u.
8 fitbit will also be introducing a family faceoff feature that lets kids compete in a five-day step challenge against the other members of their family account.
9 the app also will reward children with in-app badges for achieving their health goals.
10 fitbit’s new child-friendly fitness band will be available in blue and purple, is showerproof, and should last for five days on a single charge.
11 the ace launch is part of fitbit’s broader goal of branching out to new audiences.
12 the company also announced a new smartwatch on tuesday called the versa, which is being positioned as an everyday smartwatch rather than a fitness-only device or sports watch, like some of the company’s other products.
13 above all else, the ace is an effort to get children up and moving.
14 the centers for disease control and prevention report that the percentage of children and adolescents affected by obesity has more than tripled since the 1970’s.
15 but parents who want to encourage their children to move already have several less expensive options to choose from.
16 garmin’s $79.99 vivofit jr. 2, for example, comes in themed skins like these minnie mouse and star wars versions, while the wristband entices kids to move by reflecting their fitness achievements in an accompanying smartphone game.
17 the $39.99 nabi compete, meanwhile, is sold in pairs so that family members can work together to achieve movement milestones.
18 contact us at editors@time.com.

The program also generate Textrank score for each sentence as follows:

Sentence Number	**Textrank Score**
1	0.111643324
2	0.085730442
3	0.066121154
4	0.050253498
5	0.075974049
6	0.054394061
7	0.102599180
8	0.056483280
9	0.047589916
10	0.047649885

11	0.047714158
12	0.027480163
13	0.072902023
14	0.035755763
15	0.058923419
16	0.028204609
17	0.021834719
18	0.008746356

Based on the text rank scores, the top five (5) most important sentences is shown below:

Top five (5) most important sentences:

1. fitbit is launching a new fitness tracker designed for children called the fitbit ace, which will go on sale for $99.95 in the second quarter of this year.
2. fitbit says the tracker is designed for children eight years old and up.
3. the fitbit ace looks a lot like the company's alta tracker, but with a few child-friendly tweaks.
4. like many of fitbit's other products, the fitbit ace can automatically track steps, monitor active minutes, and remind kids to move when they've been still for too long.
5. above all else, the ace is an effort to get children up and moving.

Next, the top most important three (3) and bottom least important three (3) sentences as shown below:

Top three (3) most important sentences:

1. fitbit is launching a new fitness tracker designed for children called the fitbit ace, which will go on sale for $99.95 in the second quarter of this year.
2. fitbit says the tracker is designed for children eight years old and up.
3. the fitbit ace looks a lot like the company's alta tracker, but with a few child-friendly tweaks.

Bottom three (3) least important sentences:

1. contact us at editors@time.com.
2. the $39.99 nabi compete, meanwhile, is sold in pairs so that family members can work together to achieve movement milestones.
3. the company also announced a new smartwatch on tuesday called the versa, which is being positioned as an everyday smartwatch rather than a fitness-only device or sports watch, like some of the company’s other products.

From the report on top five (5) most important sentences and that of the top three (3) most important sentences, it shows that the top three (3) most important sentences are exactly the sentences reported in the first three (3) of the first five (5) most important sentences. However, with the bottom sentences, the situation is different. The word “fitbit” was not included in any of the bottom three least important sentences, rather the sentences tend to focus on some other things such as referencing another product in the second most important sentence.

Furthermore, the summarizer can as well represent the level of importance of each sentence in the summarized article graphically. For our study case, the graph shown in figure 4.1 below plotted by the summarizer, tend to show where the most important sentences appear.

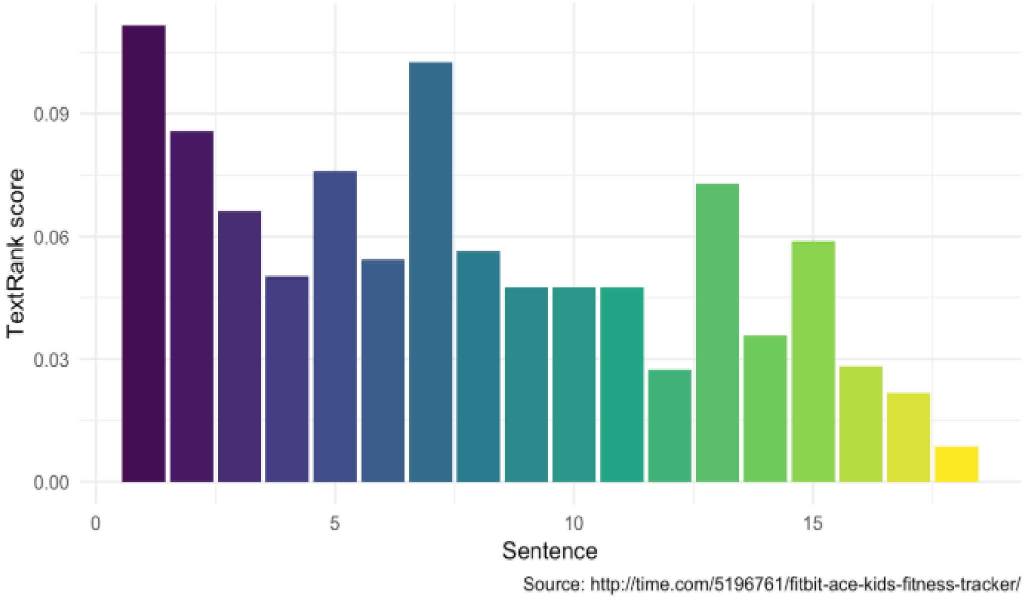


Figure 4.1: Graph Showing TextRank Score for Each Identified Sentence in the Article 1

From figure 4.1, it shows that the very first sentence tend to have the highest text rank score and this implies that it is the most important sentence in the processed article, followed by the seventh (7) sentence and followed by the second (2) sentence and next, the fifth (5) sentence. It also reveals that the four most informative sentences appear within first half of sentences in the processed article.

Example 2:

TITLE: How an Oil Boom in West Texas Is Reshaping the World

My view from the window seat of a small regional jet landing in Midland, Texas, is either a testament to the advances of human civilization or a sign of its impending demise, depending on your perspective. Countless oil wells, identified by their glowing red flames, dot the dark landscape.

We are descending into the Permian Basin, the heart of American oil country, where the massive oil and gas boom is changing not just Texas but also the nation and the world.

This year the region is expected to generate an average of 3.9 million barrels per day, roughly a third of total U.S. oil production, according to the U.S. Department of Energy. That's enough to make the U.S., as of late

2018, the world's largest producer of crude. The windfall has turned a nation long reliant on foreign oil into a net exporter in a few short years.

Not even the plunge in oil prices in recent months, which led some companies to scale back their plans for the Permian, has stopped the enthusiasm. Analysts predict the region's output will expand in coming years, thanks to cost-reducing advances in hydraulic fracturing, better known as fracking, to release oil from shale, plus changes in U.S. export policy. By 2025, U.S. oil production is expected to equal that of Saudi Arabia and Russia combined, according to the International Energy Agency (IEA).

The power of the Permian oil and gas boom is easy to spot in the basin itself, which stretches across more than 75,000 sq. mi. of scrubby ranchland in West Texas and New Mexico. So-called man camps–hastily constructed short-term housing for oil-field workers–have sprung up everywhere, amid new luxury construction projects and shiny billboards advertising Rolexes to laborers pulling in six-figure salaries. But the impact extends far beyond the region.

During the past three years, the boom in these parts has transformed the U.S. economy, upended the international energy industry, undermined global environmental efforts and tilted the balance of power among Beijing, Moscow and Washington. In places like Saudi Arabia, uncertainty over future oil profits driven by rising U.S. production contributed to a rethinking of the economy. In theory, less reliance on Saudi oil also gives the U.S. more leverage in other areas, like the war in Yemen, although the Trump Administration hasn't prioritized such efforts. The vast new U.S. oil reserves have provided cover for the imposition of tough sanctions against nations like Iran and Venezuela, moves that at other times might have crippled global supply. And around the world, the boom in the U.S. has inspired other countries to race to develop their own shale resources. "In a shale revolution world, no country is an island," says Fatih Birol, who leads the IEA. "Everyone will be affected."

The question is how. Presidents Donald Trump and Barack Obama have championed the nation's growing oil and gas markets. Abundant new shale reserves have driven economic growth and regional job creation while reducing costs for American consumers and manufacturers.

But analysts across the political spectrum caution that the energy windfall presents profound challenges as well. Neither energy markets nor national security are simple, and they overlap in complex ways here. In the long term, the boom actually threatens to undermine bipartisan efforts to establish U.S. energy independence. It could destabilize international partnerships, make the U.S. vulnerable to trade retaliation and raise formidable new hurdles in the ongoing effort to curb climate change. The nation's response to the opportunities and risks raised by the Permian Basin boom will shape our economic, environmental and geopolitical prospects for generations.

There's nothing quite like oil country in boom times. Across the Permian, gas stations, retail shops and fast-food restaurants advertise perks like $15-per-hour pay and 401(k) benefits as they compete to lure workers. Bare-bones motels charge hundreds of dollars a night. Local restaurants, patronized by women clutching designer handbags, charge $18 for a salad.

The surge in production in the Permian came at a propitious time. In the aftermath of the 2008 recession, oil demand spiked just as drilling technology unlocked layers of rich shale. Locals are eager to tout the spoils. In Odessa, Texas, entrepreneur Toby Eoff shows me the defunct theater the city is paying to renovate to house stage productions. Next door, Eoff and his wife are building a $79 million Marriott and conference center. Collin Sewell, who runs a group of car dealerships in the region, points out the window of his brand new office to the lot full of Ford trucks his employees serviced that day. When I visited him in September, his sales were up 50% from 2016."When it's good, it's awesome," Sewell says.

A couple hours away, in Hobbs, N.M., population 37,000, Mayor Sam Cobb gave me a tour of a brand new, $61 million recreation center, supported by the city's growing tax base. It's 158,000 sq. ft., with two four-story-tall water slides that loom over a giant pool, a soccer field and basketball and racquetball courts. Residents work out on Technogym equipment, the Rolls-Royce of exercise gear.

While previous oil booms have ended in busts that devastated the region, local officials say this time is different. In the past, high oil prices fueled short-lived enthusiasm that dwindled when the price of crude dropped. But recently, drillers have flocked to the Permian despite low oil prices, in part because fracking and other technological advances have made extraction so cheap. Drillers strike crude in areas inaccessible just years ago. "We're not looking for hydrocarbons, because the hydrocarbons are there," says Vicki

Hollub, CEO of Occidental Petroleum. “The Permian will continue for many years to come.” A report from the Federal Reserve Bank of Dallas estimates that new Permian oil wells break even around $50 a barrel–far less than the $80 that Saudi Arabia spends on average, according to the International Monetary Fund, to extract the same quantity of crude. “We don’t use the B word,” says Bobby Burns, president of the Midland Chamber of Commerce. “Boom doesn’t really describe it.” The Permian, he says, will be a force for a generation.

The main problem at this point, energy executives say, is there’s not enough infrastructure to handle all the oil and gas coming out of the ground. As a result, many drillers simply burn off valuable natural gas rather than capturing and selling it. Companies are also struggling to ship oil. In 2017, more than a quarter of U.S. oil exports–112 million barrels of crude–left from the port of Corpus Christi, Texas. The world would have taken much more, which is why a $327 million expansion is under way, the centerpiece of a slew of projects that could double the port’s export capacity in the coming years. When it’s completed later in 2019, a new crude-oil pipeline planned by a partnership of three companies will link the Permian oil fields and Corpus Christi, winding some 730 miles through Texas backcountry, picking up cargo along the way. It’s expected to transport 550,000 barrels of crude every day to ships that will carry it around the globe.

Environmental groups have opposed the new pipelines and expansions. The more oil and gas that’s pulled from the earth, transported, exported and burned, they argue, the faster the climate warms. But energy executives point to the region’s vast reserves and to demand. “It’s got to go somewhere,” says Brad Barron, the CEO NuStar Energy, a pipeline company operating in the Permian and Corpus Christi.

All this has come with costs. The man camps and other temporary housing facilities have been marred by crime and drug abuse. Home prices have soared. Roads and highways, many designed for ranchers, have become overrun by trucks and tankers, making them some of the most dangerous in the country. (During a violent storm in September, I pulled to the side of the highway in Andrews County, Texas, for half an hour, uncomfortable with careening big rigs in low visibility.)

But the most detrimental effects may be the hardest to see. Some locals, like Sharon Wilson, worry about the ramifications of nonstop fracking operations in their backyard. A Texas native and former oil company employee, Wilson is an organizer with Earthworks, a Washington-based environmental group. Using an infrared camera to capture images of gas leaks, she says she regularly detects dangers leaking from wells, including methane, a greenhouse gas that is responsible for about a quarter of global warming worldwide. Some locals have also suffered from exposure to other substances, she says, including benzene, a chemical in crude classified as a carcinogen. While it's difficult to measure the broad scale of the Permian's environmental impact, or its localized effect on health, Wilson says the leaks in the basin today are the worst she has seen in her years of tracking leaks. "Nothing can even come close," she says. "It's unimaginable what's happening out there."

The Permian boom transformed America's place in global energy markets almost overnight. Until recently, federal law forbade American producers from exporting crude oil at all, a policy holdover from the 1970s, when energy shortages racked the nation. But in the first decade of the new millennium, fracking and horizontal drilling opened vast new reserves of previously untapped oil. Public policy changed too. In December 2015, Obama signed a bill negotiated by congressional leaders that lifted the four-decade ban on exporting crude.

The resulting explosion in oil production has remade swaths of the U.S. economy and acted as something of a nationwide stimulus package. By helping keep the price of oil and gas low, domestic energy production aided other industries as well, tamping down the cost of air travel, trucking and even agricultural goods, because of the reduced cost of the diesel fuel many farmers rely on.

But determining the economic value of all this new oil and gas requires complex calculations. Using oil production as a pillar of the economy sounds good when prices are low, but it could hurt down the road. Consider energy security. Policy experts across the ideological spectrum have long insisted the best way to curb oil dependency is to develop diverse energy sources. That's why President George W. Bush, who lived in Midland as a child, signed a bill imposing more stringent fuel-economy standards, and it's one of many reasons Obama embraced funding for renewable energy. It also explains why GOP lawmakers who champion fossil fuels and express skepticism about climate change have also supported research into alternative energy.

These measures make sense no matter how much oil the U.S. produces, especially because the price of crude is sensitive to global events. Instability in Iraq or a burst pipeline in Canada can bump the price at the pump for U.S. consumers. The abundance generated by the Permian may obscure the risks posed by our reliance on oil, says Jason Bordoff, who advised the Obama Administration on energy and climate policy and now heads Columbia University's Center on Global Energy Policy. "It's not just boom and bust that hurts," he says. "Volatility itself hurts people."

A gas flare shown here in May 2018 burns off excess gas in Midland, Texas, part of the Permian Basin, where an oil boom has remade the landscape.

A gas flare shown here in May 2018 burns off excess gas in Midland, Texas, part of the Permian Basin, where an oil boom has remade the landscape.
Benjamin Lowy—Getty Images

Moreover, while the U.S. is now a net oil exporter, the boom may have given political and industry leaders a false sense of security. One reason is that the U.S., despite the treasure trove beneath the Permian, doesn't have the ability to process much of what it produces. Many of the refineries sprinkled across the Gulf Coast aren't built to work with U.S. crude oil, which is lighter than the product the U.S. imports from countries like Venezuela and Canada. As long as that's the case, the U.S. has to continue importing oil even if, in theory, it's producing enough of its own.

For decades, safeguarding access to imported oil has been a pillar of U.S. foreign policy. That's meant carefully maintaining relations with petrostates like Saudi Arabia and using the military to ensure stability in resource-rich regions. Under Trump and Obama, the U.S. has sought to strengthen ties with countries that import oil and gas as well. It's unclear how this new dynamic will shape U.S. relations abroad.

George David Banks, a former Trump Administration energy adviser, says the Permian windfall has expanded U.S. soft power. "The whole transformation has put us back into a role to help define the policy of global energy production and not just as a consumer," Banks says. But others analysts worry that if the economy becomes dependent on crude exports, the U.S. could become increasingly vulnerable to retaliatory tactics. Washington officials saw a glimpse of that in 2018, when China threatened to impose tariffs on U.S. oil and gas amid the trade war between the two superpowers.

All these questions set aside the primary challenge arising from the U.S.'s newfound oil reserves. Burning fossil fuels causes climate change, and the more oil and gas the U.S. produces and exports, the faster the world will warm. Many Americans see the issue through a moral lens: by drilling in the Permian Basin today, we contribute to a sicker world for future generations. Research has shown the world needs to halve greenhouse gas emissions by about 2030 to keep temperatures from rising to unsafe levels. That will be hard enough without unabated drilling in the Permian or anywhere else.

Trump has dismissed such concerns. Since taking office, his Administration has systematically slashed environmental regulations and sought to open vast new areas to drilling, including coastlines and federal lands. In theory, those moves help U.S. oil and gas companies. And while many industry leaders have praised the policies, others see reasons for long-term alarm. As Europe and the rest of the world impose increasingly stringent regulations on imported oil and gas, American energy companies–particularly large multinational corporations–could find themselves pariahs on the global market. French President Emmanuel Macron suggested last year that high-carbon products from countries that aren't committed to addressing climate change could face additional trade barriers, an idea that has gained attention among stakeholders working to deal with the issue.

Perhaps it's no surprise then that a handful of energy companies, including Shell and Chevron, have begun to change some of their climate policies, including calling for measures like a carbon price. ExxonMobil asked the Environmental Protection Agency (EPA) in December to uphold an Obama-era rule on methane emissions that the Trump Administration has sought to weaken. "Reasonable regulations help," the company said in a letter to the EPA.

Hollub, the CEO of Occidental, whose company is one of the biggest drillers in the Permian, has chosen to focus on capturing and storing CO[subscript 2], even as the Trump Administration has rejected calls for a carbon tax or other forms of carbon pricing. The company benefits from a tax incentive for doing so. But Hollub also says she sees a long-term strategic advantage. "Ultimately, there will be a carbon price," she told me in an interview at the company's Houston headquarters.

A carbon price is just one way to manage the boom that struck the Permian. And whether you see the boom as

a testament to human ingenuity, a threat to civilization
or maybe a little of both, it needs to be managed.

Evaluation of Example 2

Running example 2 article through the summarizer system, the program identified one hundred and thirty two (132) sentences. The identified sentences are shown below:

Sentence

1 my view from the window seat of a small regional jet landing in midland, texas, is either a testament to the advances of human civilization or a sign of its impending demise, depending on your perspective.
2 countless oil wells, identified by their glowing red flames, dot the dark landscape.
3 we are descending into the permian basin, the heart of american oil country, where the massive oil and gas boom is changing not just texas but also the nation and the world.
4 this year the region is expected to generate an average of 3.9 million barrels per day, roughly a third of total u.s. oil production, according to the u.s.
5 department of energy.
6 that's enough to make the u.s., as of late 2018, the world's largest producer of crude.
7 the windfall has turned a nation long reliant on foreign oil into a net exporter in a few short years.
8 not even the plunge in oil prices in recent months, which led some companies to scale back their plans for the permian, has stopped the enthusiasm.
9 analysts predict the region's output will expand in coming years, thanks to cost-reducing advances in hydraulic fracturing, better known as fracking, to release oil from shale, plus changes in u.s. export policy.
10 by 2025, u.s. oil production is expected to equal that of saudi arabia and russia combined, according to the international energy agency (iea).
11 the power of the permian oil and gas boom is easy to spot in the basin itself, which stretches across more than 75,000 sq. mi. of scrubby ranchland in west texas and new mexico.
12 so-called man camps–hastily constructed short-term housing for oil-field workers–have sprung up everywhere, amid new luxury construction projects and shiny billboards advertising rolexes to laborers pulling in six-figure salaries.

13 but the impact extends far beyond the region.

14 during the past three years, the boom in these parts has transformed the u.s. economy, upended the international energy industry, undermined global environmental efforts and tilted the balance of power among beijing, moscow and washington.

15 in places like saudi arabia, uncertainty over future oil profits driven by rising u.s. production contributed to a rethinking of the economy.

16 in theory, less reliance on saudi oil also gives the u.s. more leverage in other areas, like the war in yemen, although the trump administration hasn't prioritized such efforts.

17 the vast new u.s. oil reserves have provided cover for the imposition of tough sanctions against nations like iran and venezuela, moves that at other times might have crippled global supply.

18 and around the world, the boom in the u.s. has inspired other countries to race to develop their own shale resources.

19 "in a shale revolution world, no country is an island," says fatih birol, who leads the iea.

20 "everyone will be affected."

21 the question is how.

22 presidents donald trump and barack obama have championed the nation's growing oil and gas markets.

23 abundant new shale reserves have driven economic growth and regional job creation while reducing costs for american consumers and manufacturers.

24 but analysts across the political spectrum caution that the energy windfall presents profound challenges as well.

25 neither energy markets nor national security are simple, and they overlap in complex ways here.

26 in the long term, the boom actually threatens to undermine bipartisan efforts to establish u.s. energy independence.

27 could destabilize international partnerships, make the u.s. vulnerable to trade retaliation and raise formidable new hurdles in the ongoing effort to curb climate change.

28 the nation's response to the opportunities and risks raised by the permian basin boom will shape our economic, environmental and geopolitical prospects for generations.

29 oil workers, known as roughnecks, extracting oil at a midland rig in may 2018; many laborers earn six-figure salaries because of high demand benjamin lowy—getty images
there's nothing quite like oil country in boom times.

30 across the permian, gas stations, retail shops and fast-food restaurants advertise perks like $15-per-hour pay and 401(k) benefits as they compete to lure workers.

31 bare-bones motels charge hundreds of dollars a night.

32 local restaurants, patronized by women clutching designer handbags, charge $18 for a salad.

33 the surge in production in the permian came at a propitious time.

34 in the aftermath of the 2008 recession, oil demand spiked just as drilling technology unlocked layers of rich shale.

35 locals are eager to tout the spoils.

36 in odessa, texas, entrepreneur toby eoff shows me the defunct theater the city is paying to renovate to house stage productions.

37 next door, eoff and his wife are building a $79 million marriott and conference center.

38 collin sewell, who runs a group of car dealerships in the region, points out the window of his brand new office to the lot full of ford trucks his employees serviced that day.

39 when i visited him in september, his sales were up 50% from 2016."

40 when it's good, it's awesome," sewell says.

41 a couple hours away, in hobbs, n.m., population 37,000, mayor sam cobb gave me a tour of a brand new, $61 million recreation center, supported by the city's growing tax base.

42 it's 158,000 sq. ft., with two four-story-tall water slides that loom over a giant pool, a soccer field and basketball and racquetball courts.

43 residents work out on technogym equipment, the rolls-royce of exercise gear.

44 while previous oil booms have ended in busts that devastated the region, local officials say this time is different.

45 in the past, high oil prices fueled short-lived enthusiasm that dwindled when the price of crude dropped.

46 but recently, drillers have flocked to the permian despite low oil prices, in part because fracking and other technological advances have made extraction so cheap.

47 drillers strike crude in areas inaccessible just years ago.

48 "we're not looking for hydrocarbons, because the hydrocarbons are there," says vicki hollub, ceo of occidental petroleum.

49 "the permian will continue for many years to come."

50 a report from the federal reserve bank of dallas estimates that new permian oil wells break even around $50 a barrel–far less than the $80 that saudi arabia

spends on average, according to the international monetary fund, to extract the same quantity of crude.

51 “we don’t use the b word,” says bobby burns, president of the midland chamber of commerce.

52 “boom doesn’t really describe it.”

53 the permian, he says, will be a force for a generation.

54 the main problem at this point, energy executives say, is there’s not enough infrastructure to handle all the oil and gas coming out of the ground.

55 as a result, many drillers simply burn off valuable natural gas rather than capturing and selling it.

56 companies are also struggling to ship oil.

57 in 2017, more than a quarter of u.s. oil exports–112 million barrels of crude–left from the port of corpus christi, texas.

58 the world would have taken much more, which is why a $327 million expansion is under way, the centerpiece of a slew of projects that could double the port’s export capacity in the coming years.

59 when it’s completed later in 2019, a new crude-oil pipeline planned by a partnership of three companies will link the permian oil fields and corpus christi, winding some 730 miles through texas backcountry, picking up cargo along the way.

60 it’s expected to transport 550,000 barrels of crude every day to ships that will carry it around the globe.

61 environmental groups have opposed the new pipelines and expansions.

62 the more oil and gas that’s pulled from the earth, transported, exported and burned, they argue, the faster the climate warms.

63 but energy executives point to the region’s vast reserves and to demand.

64 “it’s got to go somewhere,” says brad barron, the ceo nustar energy, a pipeline company operating in the permian and corpus christi.

65 all this has come with costs.

66 the man camps and other temporary housing facilities have been marred by crime and drug abuse.

67 home prices have soared.

68 roads and highways, many designed for ranchers, have become overrun by trucks and tankers, making them some of the most dangerous in the country.

69 (during a violent storm in september, i pulled to the side of the highway in andrews county, texas, for half an hour, uncomfortable with careening big rigs in low visibility.)

70 but the most detrimental effects may be the hardest to see.

71 some locals, like sharon wilson, worry about the ramifications of nonstop fracking operations in their backyard.

72 a texas native and former oil company employee, wilson is an organizer with earthworks, a washington-based environmental group.

73 using an infrared camera to capture images of gas leaks, she says she regularly detects dangers leaking from wells, including methane, a greenhouse gas that is responsible for about a quarter of global warming worldwide.

74 some locals have also suffered from exposure to other substances, she says, including benzene, a chemical in crude classified as a carcinogen.

75 while it's difficult to measure the broad scale of the permian's environmental impact, or its localized effect on health, wilson says the leaks in the basin today are the worst she has seen in her years of tracking leaks.

76 "nothing can even come close," she says.

77 "it's unimaginable what's happening out there."

78 pipes at nustar energy's facility in corpus christi, texas, a key port linking oil from the permian basin to the world.

79 brandon thibodeaux—the new york times/redux the permian boom transformed america's place in global energy markets almost overnight.

80 until recently, federal law forbade american producers from exporting crude oil at all, a policy holdover from the 1970s, when energy shortages racked the nation.

81 but in the first decade of the new millennium, fracking and horizontal drilling opened vast new reserves of previously untapped oil.

82 public policy changed too.

83 in december 2015, obama signed a bill negotiated by congressional leaders that lifted the four-decade ban on exporting crude.

84 the resulting explosion in oil production has remade swaths of the u.s. economy and acted as something of a nationwide stimulus package.

85 by helping keep the price of oil and gas low, domestic energy production aided other industries as well, tamping down the cost of air travel, trucking and even agricultural goods, because of the reduced cost of the diesel fuel many farmers rely on.

86 but determining the economic value of all this new oil and gas requires complex calculations.

87 using oil production as a pillar of the economy sounds good when prices are low, but it could hurt down the road.

88 consider energy security.

89 policy experts across the ideological spectrum have long insisted the best way to curb oil dependency is to develop diverse energy sources.

90 that’s why president george w.

91 bush, who lived in midland as a child, signed a bill imposing more stringent fuel-economy standards, and it’s one of many reasons obama embraced funding for renewable energy.

92 it also explains why gop lawmakers who champion fossil fuels and express skepticism about climate change have also supported research into alternative energy.

93 these measures make sense no matter how much oil the u.s. produces, especially because the price of crude is sensitive to global events.

94 instability in iraq or a burst pipeline in canada can bump the price at the pump for u.s. consumers.

95 the abundance generated by the permian may obscure the risks posed by our reliance on oil, says jason bordoff, who advised the obama administration on energy and climate policy and now heads columbia university’s center on global energy policy.

96 “it’s not just boom and bust that hurts,” he says.

97 “volatility itself hurts people.”

98 a gas flare shown here in may 2018 burns off excess gas in midland, texas, part of the permian basin, where an oil boom has remade the landscape.

99 benjamin lowy—getty images moreover, while the u.s. is now a net oil exporter, the boom may have given political and industry leaders a false sense of security.

100 one reason is that the u.s., despite the treasure trove beneath the permian, doesn’t have the ability to process much of what it produces.

101 many of the refineries sprinkled across the gulf coast aren’t built to work with u.s. crude oil, which is lighter than the product the u.s. imports from countries like venezuela and canada.

102 as long as that’s the case, the u.s. has to continue importing oil even if, in theory, it’s producing enough of its own.

103 for decades, safeguarding access to imported oil has been a pillar of u.s. foreign policy.

104 that’s meant carefully maintaining relations with petrostates like saudi arabia and using the military to ensure stability in resource-rich regions.

105 under trump and obama, the u.s. has sought to strengthen ties with countries that import oil and gas as well.

106 it's unclear how this new dynamic will shape u.s. relations abroad.

107 george david banks, a former trump administration energy adviser, says the permian windfall has expanded u.s. soft power.

108 "the whole transformation has put us back into a role to help define the policy of global energy production and not just as a consumer," banks says.

109 but others analysts worry that if the economy becomes dependent on crude exports, the u.s. could become increasingly vulnerable to retaliatory tactics.

110 washington officials saw a glimpse of that in 2018, when china threatened to impose tariffs on u.s. oil and gas amid the trade war between the two superpowers.

111 all these questions set aside the primary challenge arising from the u.s.'s newfound oil reserves.

112 burning fossil fuels causes climate change, and the more oil and gas the u.s. produces and exports, the faster the world will warm.

113 many americans see the issue through a moral lens: by drilling in the permian basin today, we contribute to a sicker world for future generations.

114 research has shown the world needs to halve greenhouse gas emissions by about 2030 to keep temperatures from rising to unsafe levels.

115 that will be hard enough without unabated drilling in the permian or anywhere else.

116 trump has dismissed such concerns.

117 taking office, his administration has systematically slashed environmental regulations and sought to open vast new areas to drilling, including coastlines and federal lands.

118 in theory, those moves help u.s. oil and gas companies.

119 and while many industry leaders have praised the policies, others see reasons for long-term alarm.

120 as europe and the rest of the world impose increasingly stringent regulations on imported oil and gas, american energy companies–particularly large multinational corporations–could find themselves pariahs on the global market.

121 french president emmanuel macron suggested last year that high-carbon products from countries that aren't committed to addressing climate change could face additional trade barriers, an idea that has gained attention among stakeholders working to deal with the issue.

122 perhaps it's no surprise then that a handful of energy companies, including shell and chevron, have begun to change some of their climate policies, including calling for measures like a carbon price.

123 exxonmobil asked the environmental protection agency (epa) in december to uphold an obama-era rule on methane emissions that the trump administration has sought to weaken.
124 “reasonable regulations help,” the company said in a letter to the epa.
125 hollub, the ceo of occidental, whose company is one of the biggest drillers in the permian, has chosen to focus on capturing and storing co[subscript 2], even as the trump administration has rejected calls for a carbon tax or other forms of carbon pricing.
126 the company benefits from a tax incentive for doing so.
127 but hollub also says she sees a long-term strategic advantage.
128 “ultimately, there will be a carbon price,” she told me in an interview at the company’s houston headquarters.
129 a carbon price is just one way to manage the boom that struck the permian.
130 and whether you see the boom as a testament to human ingenuity, a threat to civilization or maybe a little of both, it needs to be managed.
131 write to justin worland at justin.worland@time.com.
132 this appears in the january 14, 2019 issue of time.

The program also generated Textrank score for each sentence in article 2 as follows:

Sentence Number	**Textrank Score**
1	0.003922449
2	0.007620912
3	0.017725103
4	0.013805983
5	0.006662167
6	0.008937662
7	0.009256823
8	0.011587431
9	0.011514318
10	0.015034181
11	0.013952074
12	0.007638033
13	0.002303783
14	0.010915678
15	0.012983470
16	0.012922092
17	0.010775555
18	0.010192597
19	0.004034653
20	0.001182033
21	0.001182033
22	0.011702276

23	0.004616526
24	0.005550859
25	0.006044263
26	0.011573531
27	0.006105858
28	0.008358370
29	0.009657790
30	0.006449441
31	0.002395806
32	0.003543477
33	0.007775313
34	0.008491111
35	0.002095167
36	0.002665963
37	0.002576202
38	0.003273802
39	0.001594195
40	0.005068512
41	0.002882282
42	0.002983002
43	0.001182033
44	0.010440610
45	0.011341862
46	0.012577697
47	0.004151042
48	0.002566042
49	0.007279994
50	0.010145333
51	0.002575697
52	0.004774838
53	0.006500971
54	0.013415147
55	0.004168611
56	0.010572840
57	0.014953462
58	0.003558451
59	0.012649176
60	0.006238199
61	0.002402446
62	0.010793181
63	0.007114025
64	0.011737218
65	0.001555757
66	0.001435826
67	0.002073942
68	0.001968701
69	0.003593500
70	0.001182033
71	0.003296036
72	0.011286738

73	0.004470462
74	0.004651626
75	0.005372467
76	0.001182033
77	0.004438915
78	0.013344277
79	0.010100225
80	0.011757239
81	0.009998017
82	0.002341503
83	0.004460758
84	0.012116576
85	0.010629674
86	0.011599878
87	0.011039181
88	0.007369755
89	0.010983860
90	0.003248054
91	0.007275969
92	0.005488251
93	0.015093600
94	0.006518292
95	0.012888721
96	0.009154000
97	0.001821060
98	0.016885998
99	0.013268045
100	0.008803999
101	0.011619001
102	0.016736693
103	0.013332407
104	0.002881743
105	0.016392124
106	0.008644580
107	0.012729765
108	0.007667691
109	0.008405617
110	0.012476299
111	0.012448586
112	0.015534941
113	0.007551791
114	0.005016972
115	0.006973300
116	0.002403389
117	0.005251730
118	0.018186142
119	0.003121355
120	0.012953266
121	0.003810213
122	0.009084584

123	0.005106725
124	0.003070265
125	0.007967933
126	0.003177823
127	0.003066964
128	0.002893519
129	0.010318861
130	0.004393113
131	0.001182033
132	0.002213957

Based on the text rank scores, the top five (5) most important sentences in article 2 is shown below:

<u>Top five (5) most important sentences:</u>

1. in theory, those moves help u.s. oil and gas companies.
2. we are descending into the permian basin, the heart of american oil country, where the massive oil and gas boom is changing not just texas but also the nation and the world.
3. a gas flare shown here in may 2018 burns off excess gas in midland, texas, part of the permian basin, where an oil boom has remade the landscape.
4. as long as that’s the case, the u.s. has to continue importing oil even if, in theory, it’s producing enough of its own.
5. under trump and obama, the u.s. has sought to strengthen ties with countries that import oil and gas as well.

Next, the top most important three (3) and bottom least important three (3) sentences in article 2 is shown below:

<u>Top three (3) most important sentences:</u>

1. "in theory, those moves help u.s. oil and gas companies."
2. "we are descending into the permian basin, the heart of american oil country, where the massive oil and gas boom is changing not just texas but also the nation and the world."
3. "a gas flare shown here in may 2018 burns off excess gas in midland, texas, part of the

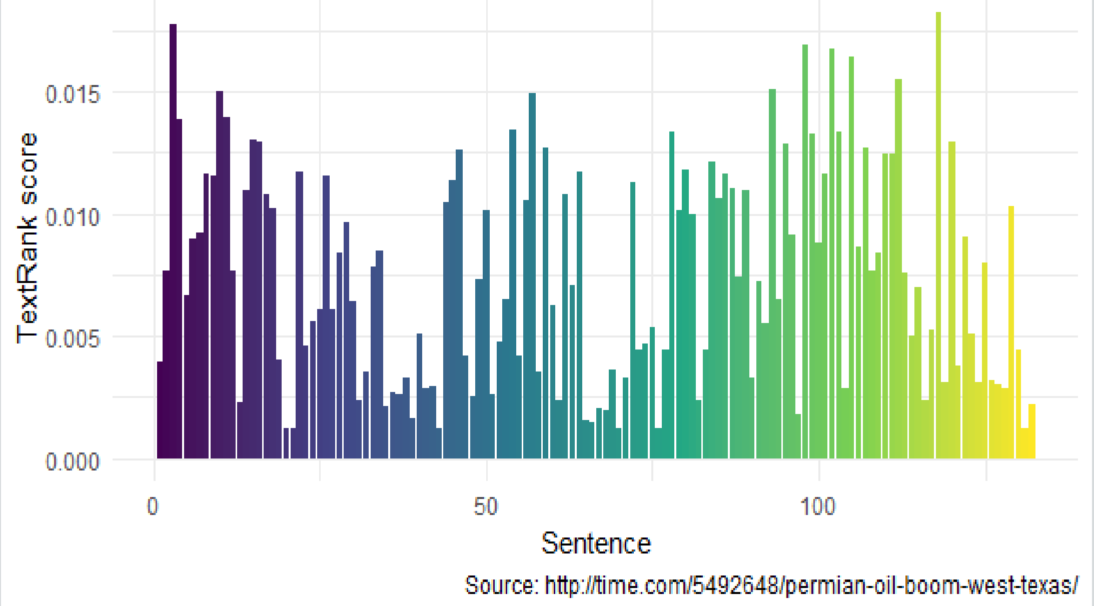
permian basin, where an oil boom has remade the landscape."

Bottom three (3) least important sentences:

1. “everyone will be affected."
2. "the question is how."
3. "residents work out on technogym equipment, the rolls-royce of exercise gear."

Also, like with article 1 example, from the report on top five (5) most important sentences and that of the top three (3) most important sentences, it shows that the top three (3) most important sentences are exactly the sentences reported in the first three (3) of the first five (5) most important sentences, while it is different with the bottom three least important sentences.

Furthermore, the graph shown in figure 4.2 below plotted by the summarizer, tend to show where the most important sentences appear.



Source: http://time.com/5492648/permian-oil-boom-west-texas/

Figure 4.2: Graph Showing TextRank Score for Each Sentence Identified in Article 2.

From figure 4.2, it shows that the 118th sentence tend to have the highest text rank score, followed by the 3rd sentence and followed by the 102nd sentence and this implies that the sentences are the most important sentences in the processed article.

Chapter 5

CONCLUSION

5.1 Conclusion

So far in this study, an extractive-based text summarization system have been studied, developed and evaluated. The study specifically described the work flow of extractive-based summarization and also provided a detailed study and discussion on some of the techniques used.

The extractive-based summarization system generates extractive summaries from news articles using the extractive-based summarization technique. That is simply highlighting the important sentences from the original text. The significance of a sentence is determined based on statistical and linguistic features of the sentence. Also, in the ESBF, the major aspect of the procedure is the pointing out of texts that are considered noteworthy.

The TextRank algorithm was used in achieving the summarization process. TextRank works by considering the entirety of information gotten from the text (graph). With this, it builds a relationship for the entities in that text, then based on this, a concept of recommendation is then implemented. Determining sentence importance in a text, is done by the process of recommendation, whereby a sentence suggests or recommends another sentence based on their similarity in making the overall concept of the text understandable. A high score is assigned to any sentence that is consistently recommended by other sentences because they tend to be more informative for that given text. TextRank algorithm has a peculiar significance of being highly portable in respect to other domains, languages and intense linguistic knowledge is not all that needful.

Our extraction-based summarization system designed and developed also represent summaries graphically. These graph-based representations tend to describe the level of importance of each sentence in the article being summarized.

Furthermore, design and implementation of an extractive based summarization framework (EBSF) was carried out in the thesis. The above framework applies various techniques used in the extraction-based summarization and thereby creating summaries for different articles from the website used as case study. Among the techniques used in the extractive-based summarization method include: word stemming, stop words removal, query expansion, text segmentation, feature selection, context representation, similarity measures, content selection and redundancy removal. Each of the aforementioned technique is key in actualizing the implementation of the extractive-based text summarization system.

5.2 Further Study

Automatic text summarization is a very wide area of research. Due to time and cost constraints, few areas were considered in this study. Thus, the following issues can be considered in future research.

i.) One major drawback of TextRank algorithm used in this thesis is the aspects of the algorithm ignoring the semantic similarity among texts in the document, which results in the poor quality of the summary being generated by the system. In view of this drawback, future research can consider combining both semantics and TextRank algorithms. The semantic algorithm will produce a semantic graph that represents the document in such a way that edges between sentences are based on semantic similarity between sentences and the sentences are ranked by applying PageRank to the resulting graph. A summary is formed by selecting the top ranked sentences, using a threshold based on required size of the summary.

ii.) In our system, only unigrams were deployed as features. Future research can make use of unigrams, bigrams and skip-bigrams in order to improve performance of the summarization system.

iii.) Furthermore, ordering sentences and sentence realization were not applied in this system. Though their implementations is difficult, but they are required for generation of coherent summaries. Also, sentence realization is required, because it will enhance the performance of the system with respect to the extractive summaries generated. With the implementation of sentence realization, sentences generated as summaries, will be more concise while preserving the important information.

REFERENCE

Aliguliyev, R. M. & Isazade, N. R. (2013*). Multiple documents summarization based on evolutionary optimization algorithm, Expert Systems with Applications, 40(5), 1675-1689.*

Aliguliyev, R. M., Hajirahimova, M. S. & Mehdiyev, C. A. (2011). *Maximum coverage and minimum redundant text summarization model, Expert Systems with Applications, 38(12).*

Ani, N. & Kathleen (2012). *A survey of text summarization techniques, Computational Linguistics, 4(3).*

Balabantary, R. C., Sahoo, D. K., Sahoo. B. & Swain, M. (2012). *Text summarization using term weights, International Journal of Computer Applications, 38(1).*

Barzilay, R. & Elhadad, M. (1997). *Using lexical chains for text summarization, Journal of machine learning research, 3(3).*

Baxendale, P. B. (1958). *Machine-made index for technical literature: an experiment, IBM Journal, 2(1), 354-361.*

Berg-Kirkpatrick, T., Gillick, D., & Klein, D. (2011). *Jointly learning to extract and compress, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, Association for Computational Linguistics, pp. 481-490.*

Bhargava, R., Sharma, Y., & Sharma, G. (2016). *ATSSI: Abstractive text summarization using sentiment infusion, Twelfth International Multi-Conference on Information Processing (IMCIP).*

Boguraev, S., Branimir, H., & Kennedy, C. (1997). *Sailence-based content characterisation of text documents. In proceedings of ACL'97 Workshop on Intelligent, Scalable Text Summarization, Madrid, Spain*

Brandow, R., Mitz. K., & Rau, L. F. (1995). *Automatic condensation of electronic publications by sentence selection, Information Processing and Management, 31(5), 657-688.*

Cantú, F. J., & Ceballos, H. G. (2010). *A framework for fostering multidisciplinary research collaboration and scientific networking within University environs. In J. Liebowitz (Ed.), Knowledge Management Handbook: Collaboration and Social Networking, CRC Press, 207-217.*

Chengcheng, L. (2010). *Automatic text summarization based on rhetorical structure theory, Expert Systems with Applications, 4(2).*

Chieze, E., Farzindar, A., & Lapalme, G. (2010). *An automatic system for summarization and information extraction of legal information, Expert Systems with Applications, 4(2).*

Donaway, R. L. (2000). *A comparison of rankings produced by summarization evaluation measures. In NAACLANLP 2000 Workshop on Automatic summarization, Morristown, NJ, USA, 69-78.*

Duy, D., Guilherme, D. F., Hurdle, J. F., & Siddhartha, J. (2016). *Extractive text summarization system to aid data extraction from full text in systematic review development, Journal of Biomedical Informatics, 64(1), 265-272.*

Edmundson, H. P. (1969). *New methods in automatic extracting, Journal of the ACM, 16(2), 264-285.*

Edmundson, H. P. (1999). *New methods in automatic extracting. In Mani, I. and Marbury, M. (eds.) Advances in Automatic Text Summarization. MIT Press.*

Elena Lloret (2015). *Text summarization: An overview, Expert Systems with Applications, 7(3).*

Erkan, G., & Radev, D. R. (2004*). Lexrank: Graph-based lexical centrality as salience in text summarization, Journal of Artificial Intelligence, 22(1), pp. 457-479.*

Fan, J., & Li, H. (2013). *Challenges of big data analysis, National Science Review, 1(2), 293-314.*

Gholamrezazadeh, S., Salehi, M. A., & Gholamzadeh, B. (2009). *A comprehensive survey on text summarization systems, Computational Linguistics, 3(2).*

Hakan, C., & Rada, M. (2010). *Quantifying the limits and success of extractive summarization systems across domains, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, Los Angeles, California, 903-911.*

Han, J., & Kamber, M (2012). *Data mining: Concepts and techniques, 3rd Edition, Kaufmann Publishers, Morgan, Boston.*

Harabagiu, S., & Lacatusu, F. (2005). *Topic themes for multi-document summarization, in Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval. ACM.*

Hovy, E. H. (2005). *Automated text summarization. In R. Mitkov (ed), The Oxford Handbook of Computational Linguistics, Oxford University Press.*

Jing, H., Hongyan, J., & McKeown, K. R. (2000). SweSum - *A text summarizer for Swedish hercules dalianis, Expert Systems with Applications, 2(1).*

Jishma, M., Sunitha, C., & Ganesha, A. (2016). *A study on ontology based abstractive summarization, Procedia Computer Science, 87(1), 32-37.*

Kan, M. Y., & McKeown, K. (2011). *Information extraction and summarization: Domain independence through focus types. Technical report, Computer Science Department, Columbia University.*

Kanitha, D. K., Muhammad, D., & Mubarak, N. (2016). *Comparison of text summarizer in Indian languages, International Journal of Advanced Trends in Engineering and Technology (IJATET), 3(1).*

Kan-Yen, M., & Kathleen, M. (1999). *Information extraction and summarization: Domain independence through focus types. Technical report, Computer Science Department, Columbia University, New York.*

Khushboo, S. T., Dharaskar, R. V., & Chandak, M. B. (2010*). Graph-based algorithms for text summarization. Emerging trends in engineering & technology, Expert Systems with Applications, 4(2).*

Kupiec, J., Pedersen, J., & Chen, F. (1995). *A trainable document summarizer. In Proceedings of the 18th annual international ACMSIGIR conference on research and development in information retrieval (SIGIR'95), Seattle, WA, USA, 68–73.*

Landauer, T., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge, Psychological Review, 104(1), 211-240.

Lee, C. S. (2014). A fuzzy ontology and its application to news summarization, Systems, Man, and Cybernetics, 35(2), 859-880

Lin, C. Y., & Hovy, E. (1999). *Automatic evaluation of summaries using n-gram concurrence statistics, Language Technology Conference, Edmonton, Canada.*

Lloret, E., & Palomar, M. (2012). *Text summarization 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.*

Lukasz, H., & Petr, K. (2016). *A knowledge management approach to data mining process for business intelligence, Industrial Management & Data Systems, 108(5), 622-634.*

Mani, I., & Maybury, M. T. (1999). *Advances in automatic text summarization, The MIT Press.*

Marcu, D. (1999). *The automatic construction of large-scale corpora for summarization research, in SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, 137-144.*

Mehdi, A., Trippe, E. D., & Gutierrez, J. B. (2017). *Text summarization techniques: A brief survey, Computational Linguistics, 7(1).*

Mithun, B., & Munirathnam, S. (2012*). Automatic ontology creation from text for national intelligence priorities framework (NIPF), Lymba Corporation Richardson, TX, 75080, USA.*

Nenkova, A., & McKeown, K. (2012*). A survey of text summarization techniques, Information Processing & Management, 47(2), 227-237.*

Nesrine, B. M. (2015). *Combining semantic search and ontology learning for incremental web ontology engineering, National School of Computer Sciences, University of Manouba, Manouba.*

Padr´o Cirera, L., Fuentes, M. J., & Alonso, L. (2004). *Approaches to text summarization: Questions and answers. Revista Iberoamericana de Inteligencia Artificial, (22), 79-102.*

Peroni, S., Motta, E., & D'Aquin, M. (2016*). Identifying key concepts in an ontology, through the integration of cognitive principles with statistical and topological* *measures, ASWC '08 Proceedings of the 3rd Asian Semantic Web Conference on The Semantic Web, 242-256.*

Pourvali, M., & Abadeh, M. S. (2012). *Automated text summarization base on lexical chain and graph using of word net and wikipedia knowledge base, IJCSI International Journal of Computer Science, 3(9).*

Radev, D. (2001). *A common theory of information fusion from multiple text sources, step one: Cross-document structure. In Proceedings, 1st ACL SIGDIAL Workshop on Discourse and Dialogue, Hong Kong, October.*

Radev, D. R., Jing, H., & Budzikowska, M. (2000). *Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies, In ANLP/NAACL Workshop on Summarization, Seattle, April.*

Radev, D., Jing, M., Stys, M., & Tam, D. (2001*). Centroid-based summarization of multiple documents. Information Processing and Management, 40(1), 919-938.*

Saggion, H., & Poibeau, T. (2013*). Automatic text summarization: Past, present and future, Information Processing & Management, 51(3), 65-73.*

Sparck, J. (2007). *Automatic summarizing: The state of the art, Information Processing & Management, 43(6), 1449-1481.*

Sparck, J. K. (1999*). Automatic summarizing: Factors and directions. In Mani, I. and Marbury, M. (eds.), Advances in Automatic Text Summarization, MIT Press.*

Srivastava, A., & Sahami, M. (2009). *Text mining. Classification, clustering, and applications, Boca Raton.*

Steven, B., Ewan, K., & Edward, L. (2009). *Analyzing text with the natural language toolkit, Butterworth Heinemann, Boston.*

Suneetha, M. S., Pervez, M. Z., & Fatima, S. S. (2012). *A novel automatic text summarization system with feature terms identification, India Conference (INDICON), Annual IEEE.*

Thanh, T., & Philipp, C. (2016). *Ontology-based interpretation of keywords for semantic search, Institute AIFB, Universität Karlsruhe, Germany.*

Vanderwende, L., Suzuki, H., Brockett, C., & Nenkova, A. (2007). *Beyond sumbasic: Task focused summarization with sentence simplification and lexical expansion, Information Processing & Management, 43(6), 1606-1618.*

Verma, R., WeiLu, N., & Chen, P. (2016). *A semantic free text summarization system using ontology knowledge, IEEE Transactions on Information Technology in Biomedicine, 5(4), 261-270.*

Wan, X., & Yang, J. (2008). *Multi-document summarization using cluster-based link analysis, Information Processing & Management, 44(4), 312-321.*

Yatsko, A. (2010). *On the discretization of continuous features for classification. School of Information Technology and Mathematical Sciences, University of Ballarat Conference.*

Yatsko, A., & Vishnyakov, K. (2007*). Data mining for exploring hidden patterns between KM and its performance, Knowledge-Based Systems, 23(1), 397-401.*

Yih, W., Goodman, J., Vanderwende, L., & Suzuki, H. (2007). *Multidocument summarization by maximizing informative content-words, IJCAI, 4(2).*

Zhang, X. (2016). *Ontology summarization based on rdf sentence graph, In: 16th Inter. World Wide Web Conference Banff, Alberta, Canada, May 8-12.*

APPENDIX

Appendix A: Sample Source Program Listing

```
#load needed packages

library(tidyverse)

library(tidytext)

library(textrank)

library(rvest)

library(jsonlite)

#url to scrape

url <- "http://time.com/5497131/paul-manafort-accused-sharing-polling-data/"

#read page html

article <- read_html(url) %>%

  html_nodes('div[class="padded"]') %>%

  html_text()

#loading article into a tibble and tokenize according to sentences which is done by setting token = "sentences" in unnest_tokens.

article_sentences <- tibble(text = article) %>%

  unnest_tokens(sentence, text, token = "sentences") %>%

  mutate(sentence_id = row_number()) %>%

  select(sentence_id, sentence)
```

```
#re-tokenize to get words.
article_words <- article_sentences %>%
  unnest_tokens(word, sentence)
#remove the stop words in article_words
article_words <- article_words %>%
  anti_join(stop_words, by = "word")
#store processed top number of specified sentences in article_summary
article_summary <- textrank_sentences(data = article_sentences,
 terminology = article_words)
#generate text summary
article_summary
#generating the text rank scores for each sentence
article_summary[["sentences"]]
#extracting the top 3 sentences
article_summary[["sentences"]] %>%
  arrange(desc(textrank)) %>%
  slice(1:3) %>%
  pull(sentence)
#extracting the bottom 3 sentences
```

```
article_summary[["sentences"]] %>%

  arrange(textrank) %>%

  slice(1:3) %>%

  pull(sentence)

#graphical representation of ranked sentences

article_summary[["sentences"]] %>%

  ggplot(aes(textrank_id, textrank, fill = textrank_id)) +

  geom_col() +

  theme_minimal() +

  scale_fill_viridis_c() +

  guides(fill = "none") +

  labs(x = "Sentence",

       y = "TextRank score",

       title = " “

       subtitle = '     ‘

      caption = "Source: http://time.com/5497131/paul-manafort-accused-sharing-polling-data/")
```