Text Summarization

An Extractive Method

A report submitted for the final project for the course IST 664 - Natural Language Processing, Spring 2019 under the guidance of Ms. Lu Xiao

Syracuse University

Submitted by:

Yesaswi Avula – SUID: 516460317
Dinesh Kumar Dhamotharan – SUID: 586563818
Harsh Rajesh Darji – SUID: 810131016
Namitha Chandrashekaraiah Reddy – SUID: 249983216
Sakthi Priya Rajendran – SUID: 954370010

Contents

1. Introduction

Have you at any point condensed an extensive record into a short passage? To what extent did you take? Physically producing an outline can be tedious and repetitive. Pro-grammed content synopsis guarantees to defeat such challenges and enable you to create the key thoughts in a bit of composing effectively. Or have you ever tried the portable application inshorts? It's an imaginative news application that changes over news articles into a 60-word rundown.What's more, that is actually what we will realize in this project — Text Summarization.

Text summarization is the technique for generating a concise and precise summary of voluminous texts while focusing on the sections that convey useful information, and without losing the overall meaning.

Automatic text summarization aims to transform lengthy documents into shortened versions, something which could be difficult and costly to undertake if done manually. Machine learning algorithms can be trained to comprehend documents and identify the sections that convey important facts and information before producing the required sum-marized texts.

2. What is the need for text summarization?

With the present blast of information coursing the computerized space, which is for the most part non-organized literary information, there is a need to create programmed content synopsis apparatuses that enable individuals to get bits of knowledge from them effectively. At present, we appreciate fast access to huge measures of data. Be that as it may, the greater part of this data is repetitive, irrelevant, and may not pass on the planned significance. For instance, on the off chance that you are searching for explicit data from an online news article, you may need to burrow through its substance and invest a great deal of energy removing the pointless stuff before getting the data you need. Hence, utilizing programmed content summarizers fit for extricating valuable data that forgets inessential and immaterial information is getting to be crucial. Executing synopsis can upgrade the lucidness of archives, lessen the time spent in looking into for data, and consider more data to be fitted in a specific territory.

3. Advantages of Text Summarization

1. Summarizing reduces perusing time
2. While investigating reports, outlines make the determination procedure simpler
3. Summarization improves the adequacy of ordering

4. Summarization calculations are less one-sided than human summarizers
5. Personalized summaries are useful in question-answering systems as they provide personalized information
6. Utilizing programmed or Summarization frameworks empowers business theoretical administrations to build the quantity of content archives they can process

Furthermore, applying text summarization reduces reading time, accelerates the pro-cess of researching for information, and increases the amount of information that can fit in an area. Content Summarization strategies are openly confined into abstractive and extractive synopsis.

In our project we are going to understand all the methods of Text Summarization and study Extractive summarization method in depth. An extractive method comprises of choosing essential sentences, passages, and so forth, from the first original copy and linking them into a shorter structure. The noteworthiness of sentences is emphatically founded on factual and phonetic highlights of sentences.

4. Various approaches for Text Summarization

There are various approaches to text summarization. The Text summarization approaches can be categorized in different ways according to the various measures and features. The measures and features are related to input, output, purpose and result in different types of summary. In the following various methods have been considered in general classification.

4.1 Summarization based on details

Further classified into classes namely:

Indicative summarization

Informative summarization

In indicative summary the main idea of the text is presented, and it is usually about 5 to 10 percent of the original text. This kind of abstraction is used to encourage the reader to read the original text. For example, a brief summary of a movie or a story in the context of its advertising which only leads to further questions and encourage the readers to watch the film and read the story. Informative summary consists of the main abstraction and the important issues of the text. This kind of summary is between 20 and 30 percent of the original text and contains all the main points of the text.

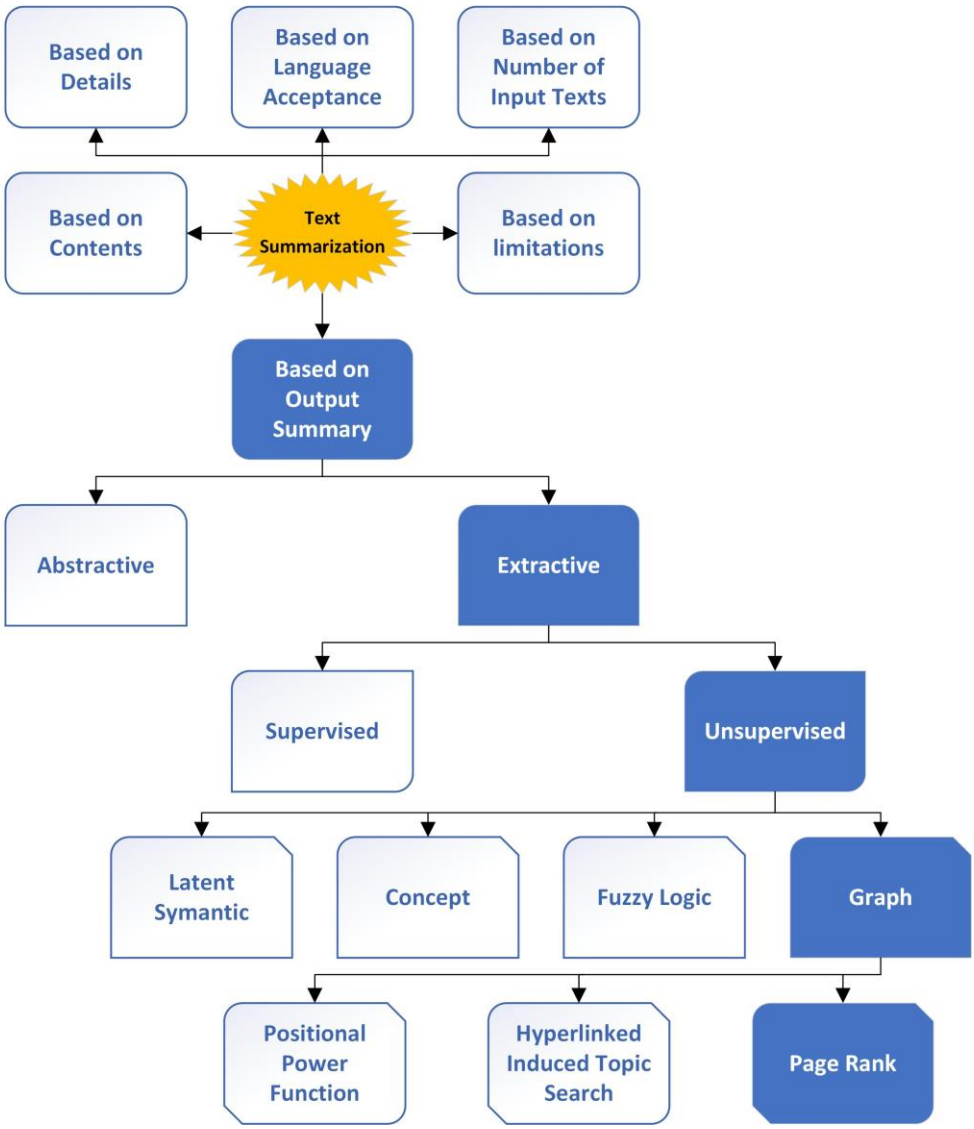


Figure 4.1: Various Approaches in Text Summarization

4.2 Summarization based on contents

Another classification in text summarization is content base classes. This type of classi-fication can be further divided into two categories:

Generic

Query-Based

Generic summarization is not depended on topic of text and it is assumed that the reader does not have any basic knowledge about the text. Generic summary contains all aspects and important issues of main text and readers are able to achieve a thorough understanding of the subject without prior knowledge of the text. Query-Based summary,

assumes that the reader has a general knowledge about the topic and just looking for specific information in the text. In this case based on the user's question, a related summary is created.

4.3 Summarization based on limitation

This category is based on the limitations of input text. The mentioned group has three categories:

Independent

Domain dependent

Genre specific

Independent summarization is something like generic summary, accepted every text of each field and generate a general summary regardless of the text scope or type.

Domain dependent summarization accepts texts with specific field of literature and type. There are many specific text patterns such as News, science text, fiction, sports. Generated summary of domain dependent systems are according to the input text type.

Genre specific summarization is trying to summary much more specialized field. This group can be abstract specific literature such as sports news, texts related to the field of geography, political News and etc.

4.4 Summarization based on the number of input texts

Summarization systems can be divided other two main categories based on their input text. The mentioned categories are

Single document summarization

Multi document summarization

The input of single-document summarization systems is only one text but multi- document summarization, which is very popular these days, is the improvement of single document summarization to collections of related documents. The main propose of multi- document summarization is summarizing texts and removing redundancy and considering the similarities and differences in the information content of different documents.

Multi document text summarization accepts multiple documents with common scope in a different perspective and closely tied to answering systems and search-based sum-marization.

There are two major approaches to summarize multiple documents:

The first approach uses the usual methods of single document summarization and summarizes each document separately. Then combining all of the summaries together and tried to remove redundancy by overlap similar sentences to produce the final summary. There are some proposed methods too which behave created single summarize as inputs and then merge them to create main summary.

The second approach is specifically designed for multiple documents. In the mentioned approaches, all documents assume as one document and all the important sentences are extracted from all of the documents together using methods like graphs or clustering. This approach is more challenging, intelligent and complicated. An example of the second approach is the SUMMONS system which extracts and com-bines information from multiple sources and passes them to a language generation component and produces the final summary. In general, the complexity of a single document model is far less than multi document models.

4.5 Summarization based on language acceptance

A text summarization system also can be divided to

Mono-language

Multi-language

Mono language summarization only accepts documents in a specific language such as English, Persian. A multi-language summarization is adopted by different languages and can be able to summarize them.

4.6 Summarization based on output summary

Text Summarization methods can be classified into

Extractive summarization

Abstractive summarization

An extractive summarization method consists of selecting important sentences or paragraphs from the original text and gathering them into shorter text. The importance of sentences is decided based on some statistical and linguistic features of sentences, extract and placed in the output text. An abstractive summarization attempts to extract the main concept of the text in clear natural language without necessity to use text phrases. Each abstractive summarization consists of comprehension part to interpret the text and find the new concepts and production part to generate new shorter text with most important information from the original document. In this method, sentences could be omitted or changed, or even new sentences could be generated. It should be noted that this method is very complicated and even more complicated than machine translation. In this research the more emphasis is on extractive text summarization as it takes into consideration the

entire document to produce the summary unlike the other approaches that consider only certain amount of the document.

5. Comparison of Models

Categories	Methodology	Concept	Advantages	Limitations
Unsupervised Learning Approaches	Graph based approach	Construction of graph to capture relationship between sentences	1)Captures redundant information. 2)Improves coherency	Doesn't focus on issues such as dangling anaphora problem
Unsupervised Learning Approaches	Concept oriented approach	Importance of sentences calculated based on the concepts retrieved from external knowledge based	incorporation of similarity measures to reduce redundancy	Dangling anaphora and verb referents not considered
Unsupervised Learning Approaches	Fuzzy Logic based approach	Summarization based on fuzzy rule using various sets of features	improved quality in summary by maintaining coherency	membership functions and work of the fuzzy logic system
Supervised Learning Approaches	Machine Learning approach Bayes rule	Large set of training data improves the sentence selection for summary	Large set of training data improves the sentence selection for summary	Human interruption required for generating manual summaries
Supervised Learning Approaches	Artificial Neural Network	Trainable summarization - neural network is trained, pruned and generalized to filter sentences and classify them as "summary" or "non-summary sentence"	The network can be trained according to the style of human reader. The set of features can be altered to reflect user's need and requirements	I)Neural Network is slow in training phase and also in application phase. 2) I t is difficult to determine how the net makes decision. 3) Requires human interruption for training data
Supervised Learning Approaches	Conditional Random Fields (CRF)	Statistical modelling approach which uses CRF as a sequence labelling problem	Identifies correct features and provides better representation of sentences and groups terms appropriately into its segments	1) focuses on domain specific which requires an external domain specific corpus for training step. 2) Limitation is that linguistic features are not considered

Figure 5.1: Comparison Table

6. Features for Extractive Text Summarization

6.1 Word Level Features

1. Title Word Features:
   The sentences in the first report which comprises of words referenced in the title have more prominent opportunities to add to the last outline since they fill in as pointers of the topic of the record. Eg: If the title of the document is “Automation in Healthcare Industry” then words like automation, healthcare appearing in the content are given greater significance

2. Content Word Feature:
   Keywords are basic in distinguishing the significance of the sentence. The sentence that comprises of primary keywords is in all probability incorporated into the last synopsis. The substance (keywords) words will be words that are noun, action words, adverbs and adjectives

3. Upper Case Word Features:
   The words which are in capitalized, for example, "UNESCO" are viewed as signifi-cant words and those sentences that comprise of these words are named significant with regards to sentence determination for the last synopsis

4. Cue Phrase Features:
   Cue expressions are words and expressions that show the structure of the archive stream and it is utilized as an element in sentence choice. The sentence that contains sign expressions (for example "end result", "since", "this data", "outline", "create", "because" and so forth.) are for the most part to be incorporated into the last rundown

5. Biased Word Features:
   The sentences that comprise of biased words are almost certain significant. The biased words are a rundown of the predefined set of words that might be area(theme) explicit. They are generally significant words that portray the topic of the archive

6.2 Sentence Level Features

1. Sentence length feature:
   The sentence length plays a significant job in recognizing key sentences. Shorter writings don’t pass on fundamental data and exceptionally long sentences likewise need not be incorporated into the outline. The standardized length of the sentence is determined as the proportion between various words in the sentence to the quantity of words in the longest sentence in the report

2. Sentence Location Feature:
   The sentences that happen first and foremost and the end some portion of the record

are no doubt significant since most reports are progressively (hierarchy) organized with significant data in the first place and the bottom of the passages

3. Sentence-to-sentence cohesion:
   The cohesion between sentences for every sentence(s), the similarity between s and alternative sentences are calculated which are summed up and coarse value of the aspect is obtained for s. The feature values are normalized between [0, 1] where value closer to 1.0 indicates a higher degree of cohesion between sentences
4. Paragraph Location Feature:
   Similar to Sentence Location, passage location additionally assumes a critical job in choosing key sentences. A Higher score is given out to the passage in the fringe segments (starting and end sections of the record)

7. Extractive Text Summarization Methods

Extractive Text Summarization methods can be broadly classified as

1. Unsupervised Learning methods
2. Supervised learning methods

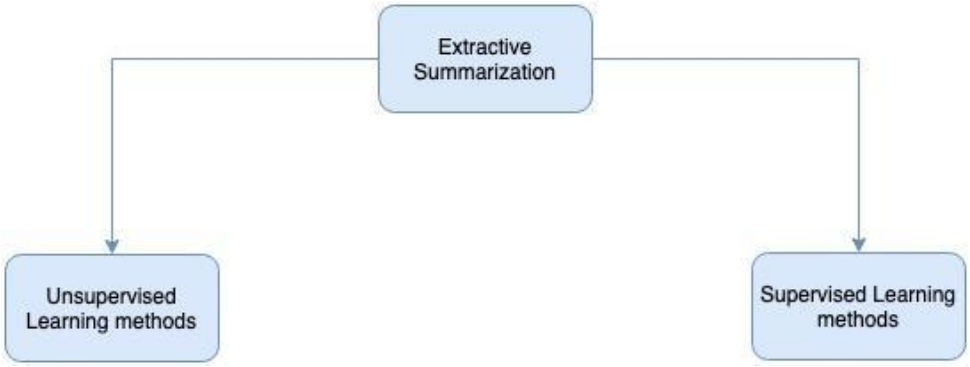


Figure 7.1: Overview of extractive text summarization

7.1 Unsupervised Learning Methods

The unsupervised approaches requires the most sophisticated algorithm to provide com-pensation for the lack of human knowledge and does not need human summaries (user input) in deciding the important features of the document. Unsupervised summaries pro-vide a higher level of automation compared to supervised model and are more suitable for processing Big Data. Unsupervised learning models have proved successful in text summarization task.

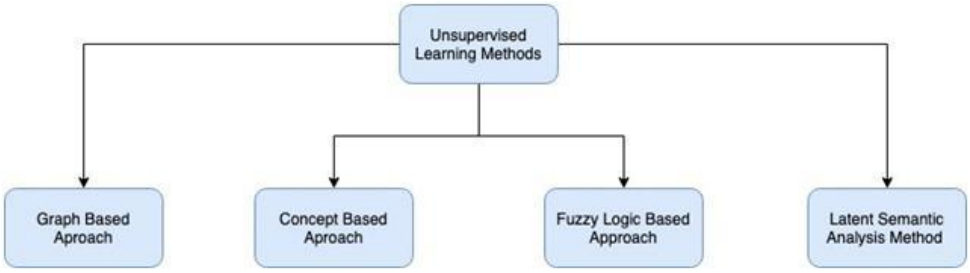


Figure 7.2: Overview of Unsupervised Learning Methods

7.1.1 Graph Based Approach

Graph-based models are extensively used in document summarization since graphs can ef-ficiently represent the document structure. Extractive text summarization using external knowledge from Wikipedia incorporating bipartite graph framework has been used. An iterative ranking algorithm is proposed which is efficient in selecting important sentences and also ensures coherency in the final summary. PageRank is a Graph-based approach where the salience of the sentence is determined by the concept of cosine similarity. The sentences in the document are represented as a graph and the edges between the sentences represents weighted cosine similarity values. The sentences are clustered into groups based on their similarity measures and then the sentences are ranked based on their PageRank scores similar to PageRank algorithm except that the similarity graph is undirected in PageRank method.

7.1.2 Concept Based Approach

In concept-based approach, the concepts are extracted from a piece of text from external knowledge base such HowNet and Wikipedia. In the methodology proposed, the impor-tance of sentences is calculated based on the concepts retrieved from HowNet instead of words. A conceptual vector model is built to obtain a rough summarization and similarity measures are calculated between the sentences to reduce redundancy in the final summary. A good summarizer focuses on higher coverage and lower redundancy. Ramanathan et al proposed Wikipedia-based summarization which utilizes graph structure to produce summaries. The method uses Wikipedia to obtain concept for each sentence and builds a sentence-concept bipartite graph as already mentioned in Graph-based summarization. The basic steps in concept-based summarization are:

Retrieve concepts of a text from external knowledge base (HowNet, WordNet, Wikipedia)

Build a conceptual vector or graph model to depict relationship between concept and sentences

Apply ranking algorithm to score sentences

Generate summaries based on the ranking scores of sentences

7.1.2.1 Advantages

Incorporation of similarity measures to reduce redundancy

7.1.2.2 Limitations

Dangling anaphora and verb referents not considered

7.1.3 Fuzzy Logic Based Approach

The fuzzy logic approach mainly contains four components: defuzzifier, fuzzifier, fuzzy knowledge base and inference engine. The textual characteristics input of Fuzzy logic approach are sentenced length, sentence similarity etc which is later given to the fuzzy system. Fuzzy logic approach is used for automatic text summarization which is the initial step, the text document is pre-processed followed by feature extraction (Title features, Sentence length, Sentence position, Sentence-sentence similarity, term weight, Proper noun and Numerical data. The summary is generated by ordering the ranked sentences in the order they occur in the original document to maintain coherency. The proposed method shows improvement in the quality of summarization but issues such as dangling anaphora are not handled.

7.1.3.1 Advantages

Improved quality in summary by maintaining coherency

7.1.3.2 Limitations

Membership functions and work of the fuzzy logic system

7.1.4 Latent Symantec Analysis Methods

Latent Semantic Analysis (LSA) is a method which excerpt hidden semantic structures of sentences and words that are popularly used in text summarization task. It is an unsupervised learning approach that does not demand any sort of external or training knowledge. LSA captures the text of the input document and excerpt information such as words that frequently occur together and words that are commonly seen in different sentences. A high number of common words amongst the sentences illustrate that the sentences are semantically related. Singular Value Decomposition (SVD), is a method used to find out the interrelations between words and sentences which also has the competence of noise reduction that helps to improve accuracy. SVD, when enforced to document word matrices, can group documents that are semantically associated to one other despite them sharing no common words. The set of words that ensue in connected text is also connected within the same peculiar dimensional space. LSA technique is applied to excerpt the subject-related words and important content conveying sentences from report. The advantage of adopting LSA vectors for summarization over word vectors is that conceptual rela-tions as represented in the human brain are naturally captured in the LSA. Choice of the

representative sentence from every scale of the capacity ensures relevancy of sentence to the document and ensures non-redundancy. LSA works with text data and the principal ambit due to the collection of topics can be located.

7.2 Supervised Learning Methods

Supervised extractive summarization related techniques are based on a classification approach at sentence level where the system learns by examples to classify between summary and non-summary sentences. The major drawback with the supervised approach is that it requires manually created summaries by a human to label the sentences in the original training document enclosed with "summary sentence" or "non-summary sentence" and it also requires more labeled training data for classification.

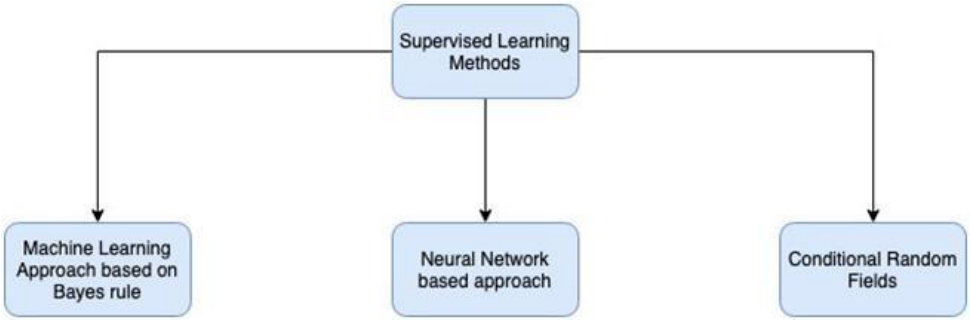


Figure 7.3: Overview of Supervised Learning Methods

7.2.1 Machine Learning Approach based on Bayes Rule

A set of training documents along with its extractive summaries is fed as input to the training stage. The machine learning approach views classification problem in text sum-marization. The probability of classification are learned from the training data by the following Bayes rule : where s rep- resents the set of sentences in the document and fi represents the features used in classification stage and S represents the set of sentences in the summary. P (s E< SII1,h,h,....In) represents the probability of the sentences to be included in the summary based on the given features possessed by the sentence.

7.2.1.1 Advantages

Large set of training data improves the sentence selection for summary

7.2.1.2 Limitations

Human interruption required for generating manual summaries

7.2.2 Neural Network based approach

In the approach proposed in, RankNet algorithm using neural nets to identify the important sentences in the document. It uses a two-layer neural network with back propagation trained using RankNet algorithm. The first step involves labeling the training data using a machine- learning approach and then extract features of the sentences in both test set and train sets which is then inputted to the neural network system to rank the sentences in the document. Another approach [17] uses a three-layered feed-forward neural network which learns in the training stage the characteristics of summary and non-summary sen-tences. The major phase is the feature fusion phase where the relationship between the features are identified through two stages:

Eliminating infrequent features

Collapsing frequent features after which sentence ranking is done to identify the important summary sentences

This system gets the Wikipedia articles as input followed by tokenization and stem-ming. The pre-processed passage is sent to the feature extraction steps,which is based on multiple features of sentences and words. The scores obtained after the feature extraction are fed to the neural network, which produces a single value as output score, signifying the importance of the sentences.

7.2.2.1 Advantages

The network can be trained according to the style of human reader. The set of features can be altered to reflect user’s need and requirements

7.2.2.2 Limitations

1. Neural Network is slow in training phase and also in application phase.
2. It is difficult to determine how the net makes decision.
3. Requires human interruption for training data

7.2.3 Conditional Random Fields

Conditional Random Fields are a statistical modeling approach that focuses on machine learning to provide a structured prediction. The proposed system overcomes the issues faced by non-negative matrix Factorization (NMF) methods by incorporating conditional random fields (CRF) to identify and extract correct features to determine the important sentence of the given text. The approach specified in uses CRF as a sequence labelling problem and also captures interaction between sentences through the features extracted for each sentence and it also incorporates complex features such as LSA scores and lilT score, but the limitation is that linguistic features are not considered.

7.2.3.1 Advantages

Identifies correct features and provides better representation of sentences and groups terms appropriately into its segments

7.2.3.2 Limitations

1. Focuses on domain specific which requires external domain specific corpus for train-ing step
2. Limitation is that linguistic features are not considered

8. Our Approach - Graph based TextRank Algorithm

1. Initially, we take a text input and tokenize it into sentences using nltk sentence tokenizer
2. Preprocessing
   (a) Usual texts in the internet has apostrophes which are used to mark possessions as in that's, doesn't, that's etc. And to mark letters omitted in contractions such as you're for you are. Such apostrophes need to be eliminated and the full word needs to be substituted in the text which we do using regular expressions. Where we substitute most common contractions such as I'm, he's, that's, she's etc. into a full word which will facilitate removal of stop words
   (b) Removing stop words from the tokenized sentences. We derived stop words from nltk stop word corpus
   (c) The overlap of two sentences can be determined simply as the number of com-mon tokens between the lexical representations of the two sentences, or it can be run through syntactic filters, which only count words of a certain syntactic category. Moreover, to avoid promoting long sentences, we are using a normal-ization factor, and divide the content overlap of two sentences with the length of each sentence
   (d) To establish connections (edges) between sentences, we are defining a "similar-ity" relation, where "similarity" is measured as a function of content overlap. Such a relation between two sentences can be seen as a process of "recommen-dation": a sentence that addresses certain concepts in a text, gives the reader a "recommendation" to refer to other sentences in the text that address the same concepts, and therefore a link can be drawn between any two such sentences that share common content

The overlap of two sentences can be determined simply as the number of common tokens between the lexical representations of the two sentences, or it can be run through syntactic filters, which only count words of a certain syntactic category. Moreover, to avoid

promoting long sentences, we are using a normalization factor, and divide the content overlap of two sentences with the length of each sentence. Formally, given two sentences Si and Sj, with a sentence being represented by the set of Ni words that appear in the sentence: Si = W1i, W2i, ..., Wi , the similarity of Si and Sj is defined as:

$$Similarity(S_i, S_j) = \frac{|W_k|W_k \in S_i \& W_k \in S_j|}{log(|S_i|)+log(|S_j|)}$$

The resulting graph is highly connected, with a weight associated with each edge, indicating the strength of the connections between various sentence pairs in the text2. The text is therefore represented as a weighted graph, and consequently we are using the weighted graph-based ranking formulae. The graph can be represented as: (a) simple undirected graph; (b) directed weighted graph with the orientation of edges set from a sentence to sentences that follow in the text (directed forward); or (c) directed weighted graph with the orientation of edges set from a sentence to previous sentences in the text (directed backward). After the ranking algorithms is run on the graph, sentences are sorted in reverse order of their score, and the top ranked sentences are selected for inclusion in the summary.

9. Related Work

Utilizing the pros and cons in product reviews as training data, we apply a machine-learning approach, using neural networks. Neural models have become popular for text classification tasks (Socher et al., 2013; Zhang et al., 2015; Conneau et al., 2016) in recent years. Text summarization can also be done using other methods like Fuzzy Logic and Latent Semantic. But the best method differs for each application of the summarization. So, text summarization cannot be generalized, and each application should be carried out with different methods. Automated text summarization in interpreting and organiz-ing scholarly work will help build applications for integration with digital libraries and reference management tools.

10. Conclusions

The most important sentences give more information about the text as they have a higher score. The algorithm is highly portable to other domains, genres or languages. It checks the connections between the entities in a text and applies recommendation to them. It is simpler and ecient than supervised learning techniques as the data do not need to be trained. This is because, in supervised method, the model should be trained with some specific data and then tested on other data after proper validation. This is a generalized approach and the validation is not same for all types of data. Hence it is riskier unlike unsupervised approach.

11. Result

11.1 Example 1:

11.1.1 Input Text

"Austin , Texas– Committee approval of Gov. Price Daniel's " abandoned property " act seemed certain Thursday despite the adamant protests of Texas bankers .Daniel per-sonally led the fight for the measure , which he had watered down considerably since its rejection by two previous Legislatures , in a public hearing before the House Committee on Revenue and Taxation .Under committee rules , it went automatically to a subcommittee for one week .But questions with which committee members taunted bankers appearing as witnesses left little doubt that they will recommend passage of it .Daniel termed " extremely conservative " his estimate that it would produce 17 million dollars to help erase an anticipated deficit of 63 million dollars at the end of the current fiscal year next Aug. 31 .He told the committee the measure would merely provide means of enforcing the escheat law which has been on the books " since Texas was a republic " .It permits the state to take over bank accounts , stocks and other personal property of persons missing for seven years or more .The bill , which Daniel said he drafted personally , would force banks , insurance firms , pipeline companies and other corporations to report such prop-erty to the state treasurer .The escheat law cannot be enforced now because it is almost impossible to locate such property , Daniel declared .It permits the state to take over bank accounts , stocks and other personal property of persons missing for seven years or more .The bill , which Daniel said he drafted personally , would force banks , insurance firms
, pipeline companies and other corporations to report such property to the state trea-surer .The escheat law cannot be enforced now because it is almost impossible to locate such property , Daniel declared .Dewey Lawrence , a Tyler lawyer representing the Texas Bankers Association , sounded the opposition keynote when he said it would force banks to violate their contractual obligations with depositors and undermine the confidence of bank customers ." If you destroy confidence in banks , you do something to the economy "
, he said ." You take out of circulation many millions of dollars " .Rep. Charles E. Hughes of Sherman , sponsor of the bill , said a failure to enact it would amount " to making a gift out of the taxpayers' pockets to banks , insurance and pipeline companies " .His contention was denied by several bankers , including Scott Hudson of Sherman , Gaynor B. Jones of Houston , J. B. Brady of Harlingen and Howard Cox of Austin .Cox argued that the bill is " probably unconstitutional " since , he said , it would impair contracts
.He also complained that not enough notice was given on the hearing , since the bill was introduced only last Monday .Austin , Texas– Senators unanimously approved Thursday the bill of Sen. George Parkhouse of Dallas authorizing establishment of day schools for the deaf in Dallas and the four other largest counties .The bill is designed to provide spe-cial schooling for more deaf students in the scholastic age at a reduced cost to the state
.There was no debate as the Senate passed the bill on to the House .It would authorize the Texas Education Agency to establish county-wide day schools for the deaf in counties of 300,000 or more population , require deaf children between 6 and 13 years of age to

attend the day schools , permitting older ones to attend the residential Texas School for the Deaf here .Operating budget for the day schools in the five counties of Dallas , Harris , Bexar , Tarrant and El Paso would be $451,500 , which would be a savings of $157,460 yearly after the first year's capital outlay of $88,000 was absorbed , Parkhouse told the Senate .The TEA estimated there would be 182 scholastics to attend the day school in Dallas County , saving them from coming to Austin to live in the state deaf school .Dallas may get to hear a debate on horse race parimutuels soon between Reps. V. E. (Red) Berry and Joe Ratcliff .While details are still to be worked out , Ratcliff said he expects to tell home folks in Dallas why he thinks Berry's proposed constitutional amendment should be rejected .“ We're getting more ' pro ' letters than ' con ' on horse race betting ” , said Ratcliff .“ But I believe if people were better informed on this question , most of them would oppose it also .I'm willing to stake my political career on it ” .Rep. Berry , an ex-gambler from San Antonio , got elected on his advocacy of betting on the ponies .A House committee which heard his local option proposal is expected to give it a favorable report , although the resolution faces hard sledding later .The house passed finally , and sent to the Senate , a bill extending the State Health Department's authority to give planning assistance to cities .The senate quickly whipped through its meager fare of House bills approved by committees , passing the three on the calendar .One validated acts of school districts .Another enlarged authority of the Beaumont Navigation District .The third amended the enabling act for creation of the Lamar county Hospital District , for which a special constitutional amendment previously was adopted .Without dissent , senators passed a bill by Sen. A. R. Schwartz of Galveston authorizing establishment in the future of a school for the mentally retarded in the Gulf Coast district .Money for its construction will be sought later on but in the meantime the State Hospital board can accept gifts and donations of a site .Two tax revision bills were passed .One , by Sen. Louis Crump of San Saba , would aid more than 17,000 retailers who pay a group of miscellaneous excise taxes by eliminating the requirement that each return be notarized .Instead , retailers would sign a certificate of correctness , violation of which would carry a penalty of one to five years in prison , plus a $1,000 fine .It was one of a series of recommendations by the Texas Research League .The other bill , by Sen. A. M. Aikin Jr. of Paris , would relieve real estate brokers , who pay their own annual licensing fee , from the $12 annual occupation license on brokers in such as stocks and bonds .Natural gas public utility companies would be given the right of eminent domain , under a bill by Sen. Frank Owen 3 , of El Paso , to acquire sites for underground storage reservoirs for gas .Marshall Formby of Plainview , former chairman of the Texas Highway Commission , suggested a plan to fill by appointment future vacancies in the Legislature and Congress , eliminating the need for costly special elections .Under Formby's plan , an appointee would be selected by a board composed of the governor , lieutenant governor , speaker of the House , attorney general and chief justice of the Texas Supreme Court .Austin , Texas– State representatives decided Thursday against taking a poll on what kind of taxes Texans would prefer to pay .An adverse vote of 81 to 65 kept in the State Affairs Committee a bill which would order the referendum on the April 4 ballot , when Texas votes on a U.S. senator .Rep. Wesley Roberts of Seminole , sponsor of the poll idea , said that further delay in the committee can kill the bill .The West Texan reported that he had finally gotten Chairman Bill Hollowell of the committee to set it for public hearing

on Feb. 22 .The proposal would have to receive final legislative approval , by two-thirds majorities , before March 1 to be printed on the April 4 ballot , Roberts said .Opponents generally argued that the ballot couldn't give enough information about tax proposals for the voters to make an intelligent choice .All Dallas members voted with Roberts , except Rep. Bill Jones , who was absent .Austin , Texas– Paradise lost to the alleged water needs of Texas' big cities Thursday .Rep. James Cotten of Weatherford insisted that a water development bill passed by the Texas House of Representatives was an effort by big cities like Dallas and Fort Worth to cover up places like Paradise , a Wise County hamlet of 250 people .When the shouting ended , the bill passed , 114 to 4 , sending it to the Senate , where a similar proposal is being sponsored by Sen. George Parkhouse of Dallas .Most of the fire was directed by Cotten against Dallas and Sen. Parkhouse .The bill would increase from 5; 000; 000to15,000,000 the maximum loan the state could make to a local water project .Cotten construed this as a veiled effort by Parkhouse to help Dallas and other large cities get money which Cotten felt could better be spent providing water for rural Texas .Statements by other legislators that Dallas is paying for all its water program by local bonds , and that less populous places would benefit most by the pending bill , did not sway Cotten's attack .The bill's defenders were mostly small-town legislators like J. W. Buchanan of Dumas , Eligio (Kika) De La Garza of Mission , Sam F. Collins of Newton and Joe Chapman of Sulphur Springs .“ This is a poor boy's bill ” , said Chapman .“ Dallas and Fort Worth can vote bonds .This would help the little peanut districts ” .Austin , Texas– A Houston teacher , now serving in the Legislature , proposed Thursday a law reducing the time spent learning “ educational methods ” .Rep. Henry C. Grover , who teaches history in the Houston public schools , would reduce from 24 to 12 semester hours the so-called “ teaching methods ” courses required to obtain a junior or senior high school teaching certificate .The bill would increase from 5; 000; 000to15,000,000 the maximum loan the state could make to a local water project .Cotten construed this as a veiled effort by Parkhouse to help Dallas and other large cities get money which Cotten felt could better be spent providing water for rural Texas .Statements by other legislators that Dallas is paying for all its water program by local bonds , and that less populous places would benefit most by the pending bill , did not sway Cotten's attack

.The bill's defenders were mostly small-town legislators like J. W. Buchanan of Dumas , Eligio (Kika) De La Garza of Mission , Sam F. Collins of Newton and Joe Chapman of Sulphur Springs .“ This is a poor boy's bill ” , said Chapman .“ Dallas and Fort Worth can vote bonds .This would help the little peanut districts ” .Austin , Texas– A Houston teacher , now serving in the Legislature , proposed Thursday a law reducing the time spent learning “ educational methods ” .Rep. Henry C. Grover , who teaches history in the Houston public schools , would reduce from 24 to 12 semester hours the so-called

“ teaching methods ” courses required to obtain a junior or senior high school teaching certificate .The bill would increase from 5; 000; 000to15,000,000 the maximum loan the state could make to a local water project .Cotten construed this as a veiled effort by Parkhouse to help Dallas and other large cities get money which Cotten felt could better be spent providing water for rural Texas .Statements by other legislators that Dallas is paying for all its water program by local bonds , and that less populous places would benefit most by the pending bill , did not sway Cotten's attack .The bill's defenders were mostly small-town legislators like J. W. Buchanan of Dumas , Eligio (Kika) De La Garza

of Mission , Sam F. Collins of Newton and Joe Chapman of Sulphur Springs .“ This is a poor boy’s bill ” , said Chapman .“ Dallas and Fort Worth can vote bonds .This would help the little peanut districts ” .Austin , Texas– A Houston teacher , now serving in the Legislature , proposed Thursday a law reducing the time spent learning “ educational methods ” .Rep. Henry C. Grover , who teaches history in the Houston public schools , would reduce from 24 to 12 semester hours the so-called “ teaching methods ” courses required to obtain a junior or senior high school teaching certificate .The bill would in-crease from $5,000,000 to $15,000,000 the maximum loan the state could make to a local water project .Cotten construed this as a veiled effort by Parkhouse to help Dallas and other large cities get money which Cotten felt could better be spent providing water for rural Texas .Statements by other legislators that Dallas is paying for all its water program by local bonds , and that less populous places would benefit most by the pending bill , did not sway Cotten’s attack .The bill’s defenders were mostly small-town legislators like J. W. Buchanan of Dumas , Eligio (Kika) De La Garza of Mission , Sam F. Collins of Newton and Joe Chapman of Sulphur Springs .“ This is a poor boy’s bill ” , said Chapman .“ Dallas and Fort Worth can vote bonds .This would help the little peanut districts ” .Austin , Texas– A Houston teacher , now serving in the Legislature , proposed Thursday a law reducing the time spent learning “ educational methods ” .Rep. Henry C. Grover , who teaches history in the Houston public schools , would reduce from 24 to 12 semester hours the so-called “ teaching methods ” courses required to obtain a junior or senior high school teaching certificate .A normal year’s work in college is 30 semester hours .Grover also would require junior-senior high teachers to have at least 24 semester hours credit in the subject they are teaching .The remainder of the 4-year college requirement would be in general subjects .“ A person with a master’s degree in physics , chemistry , math or English , yet who has not taken Education courses , is not permitted to teach in the public schools ” , said Grover .College teachers in Texas are not required to have the Education courses .Fifty-three of the 150 representatives immediately joined Grover as co-signers of the proposal .Paris , Texas (sp.)– The board of regents of Paris Junior Col-lege has named Dr. Clarence Charles Clark of Hays , Kan. as the school’s new president .Dr. Clark will succeed Dr. J. R. McLemore , who will retire at the close of the present school term .Dr. Clark holds an earned Doctor of Education degree from the University of Oklahoma .He also received a Master of Science degree from Texas A I College and a Bachelor of Science degree from Southwestern State College , Weatherford , Okla. .In ad-dition , Dr. Clark has studied at Rhode Island State College and Massachusetts Institute of Technology .During his college career , Dr. Clark was captain of his basketball team and was a football letterman .Dr. Clark has served as teacher and principal in Oklahoma high schools , as teacher and athletic director at Raymondville , Texas , High School , as an instructor at the University of Oklahoma , and as an associate professor of education at Fort Hays , Kan. , State College .He has served as a border patrolman and was in the Signal Corps of the U.S. Army .Denton , Texas (sp.)– Principals of the 13 schools in the Denton Independent School District have been re-elected for the 1961-62 session upon the recommendation of Supt. Chester O. Strickland .State and federal legislation against racial discrimination in employment was called for yesterday in a report of a “ blue ribbon ” citizens committee on the aid to dependent children program .The report , culminating a year long study of the ADC program in Cook county by a New York City

welfare consulting firm , listed 10 long range recommendations designed to reduce the soaring ADC case load .The report called racial discrimination in employment “ one of the most serious causes of family breakdown , desertion , and ADC dependency ” In ad-dition , Dr. Clark has studied at Rhode Island State College and Massachusetts Institute of Technology .During his college career , Dr. Clark was captain of his basketball team and was a football letterman .Dr. Clark has served as teacher and principal in Oklahoma high schools , as teacher and athletic director at Raymondville , Texas , High School , as an instructor at the University of Oklahoma , and as an associate professor of education at Fort Hays , Kan. , State College .He has served as a border patrolman and was in the Signal Corps of the U.S. Army .Denton , Texas (sp.)– Principals of the 13 schools in the Denton Independent School District have been re-elected for the 1961-62 session upon the recommendation of Supt. Chester O. Strickland .State and federal legislation against racial discrimination in employment was called for yesterday in a report of a “ blue ribbon ” citizens committee on the aid to dependent children program .The report , culminating a year long study of the ADC program in Cook county by a New York City welfare consulting firm , listed 10 long range recommendations designed to reduce the soaring ADC case load .The report called racial discrimination in employment “ one of the most serious causes of family breakdown , desertion , and ADC dependency ” .“Calls for extensionOther recommendations made by the committee are :Extension of the ADC program to all children in need living with any relatives , including both parents , as a means of preserving family unity .Research projects as soon as possible on the causes and prevention calls for extensionOther recommendations made by the committee are :Exten-sion of the ADC program to all children in need living with any relatives , including both parents , as a means of preserving family unity .Research projects as soon as possible on the causes and prevention must solve problem ”The monthly cost of ADC to more than 100,000 recipients in the county is 4.4 million dollars , said C. Virgil Martin , president of Carson Pirie Scott Co. , committee chairman .“ We must solve the problems which have forced these people to depend upon ADC for subsistence ” , Martin said .The volume of ADC cases will decrease , Martin reported , when the community is able to deal effectively with two problems : Relatively limited skills and discrimination in employment because of color .These , he said , are “ two of the principal underlying causes for family breakups leading to ADC ” .Calls for extensionOther recommendations made by the committee are :Extension of the ADC program to all children in need living with any relatives , including both parents , as a means of preserving family unity .Research projects as soon as possible on the causes and prevention of dependency and illegitimacy ."

11.1.2 Output Summarized text

Austin, Texas – Paradise lost to the alleged water needs of Texas’ big cities Thursday. Rep. James Cotten of Weatherford insisted that a water development bill passed by the Texas House of Representatives was an effort by big cities like Dallas and Fort Worth to cover up places like Paradise, a Wise County hamlet of 250 people. Paris, Texas (sp.) The monthly cost of ADC to more than 100,000 recipients in the county is 4.4 million dollars, said C. Virgil Martin, president of Carson Pirie Scott Co., committee chairman.

11.1.3 Graph - Sentence Number vs Importance

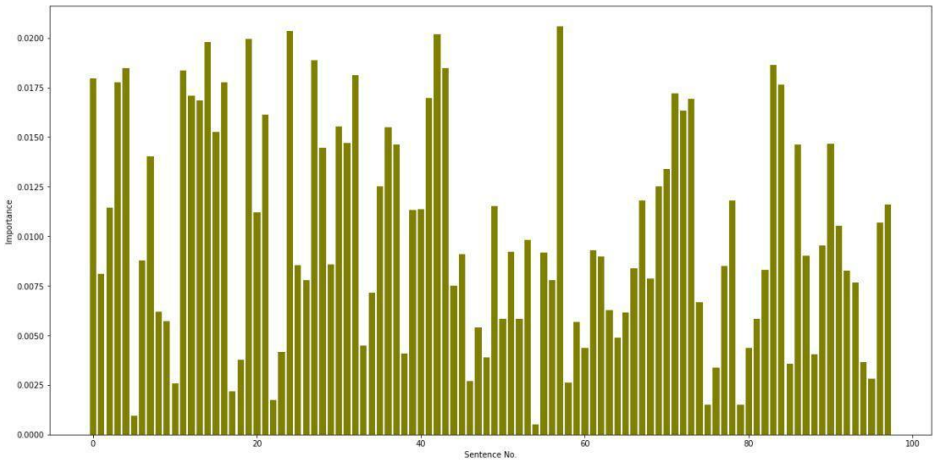


Figure 11.1: Sentence Number vs Importance

11.1.4 Word Clouds

In the following word clouds, the input is the brown corpus and the output word cloud has the most summarized words based on the ranking of the page rank. The output has the words that have been identified to have high importance from the page rank algorithm. The importance of the words is directly proportional to the size of those words in the output word cloud

Figure 11.2: Brown Corpus - Input Word Cloud

Figure 11.3: Brown Corpus - Output Word Cloud

11.2 Example 2:

11.2.1 Input Text

The most positive element to emerge from the Oslo meeting of North Atlantic Treaty Organization Foreign Ministers has been the freer , franker , and wider discussions , animated by much better mutual understanding than in past meetings .This has been a working session of an organization that , by its very nature , can only proceed along its route step by step and without dramatic changes .In Oslo , the ministers have met in a climate of candor , and made a genuine attempt to get information and understanding one another's problems .This atmosphere of understanding has been particularly noticeable where relations are concerned between the " colonialist " powers and those who have never , or not for a long time , had such problems .The nightmare of a clash between those in trouble in Africa , exacerbated by the difficulties , changes , and tragedies facing them , and other allies who intellectually and emotionally disapprove of the circumstances that have brought these troubles about , has been conspicuous by its absence .Explosion avoidedIn the case of Portugal , which a few weeks ago was rumored ready to walk out of the NATO Council should critics of its Angola policy prove harsh , there has been a noticeable relaxation of tension .The general , remarkably courteous , explanation has left basic positions unchanged , but there has been no explosion in the council .There should even be no more bitter surprises in the UN General Assembly as to NATO members' votes , since a new ad hoc NATO committee has been set up so that in the future such topics as Angola will be discussed in advance .Canada alone has been somewhat out of step with the Oslo attempt to get all the allied cars back on the track behind the NATO locomotive .Even Norway , despite daily but limited manifestations against atomic arms in the heart of this northernmost capital of the alliance , is today closer to the NATO line .On the negative side of the balance sheet must be set some disappointment that the United States leadership has not been as much in evidence as hoped for .One diplomat described the tenor of Secretary of State Dean Rusk's speeches as " inconclusive " .But he hastened to add that , if United States policies were not always clear , despite Mr. Rusk's

analysis of the various global danger points and setbacks for the West , this may merely mean the new administration has not yet firmly fixed its policy .Exploratory moodA certain vagueness may also be caused by tactical appreciation of the fact that the present council meeting is a semipublic affair , with no fewer than six Soviet correspondents accredited .The impression has nevertheless been given during these three days , despite Mr. Rusk's personal popularity , that the United States delegation came to Oslo in a somewhat tentative and exploratory frame of mind , more ready to listen and learn than to enunciate firm policy on a global scale with detailed application to individual danger spots
.The Secretary of State himself , in his first speech , gave some idea of the tremendous . . .

11.2.2 Output Summarized text

The most positive element to emerge from the Oslo meeting of North Atlantic Treaty Organization Foreign Ministers has been the freer , franker , and wider discussions , animated by much better mutual understanding than in past meetings .How effective have Kennedy administration first foreign policy decisions been in dealing with Communist aggression ? ?He was critical of what he feels is President Kennedy's tendency to be too conciliatory .Mr. Nixon , for his part , would oppose intervention in Cuba without specific provocation

11.2.3 Graph - Sentence Number vs Importance

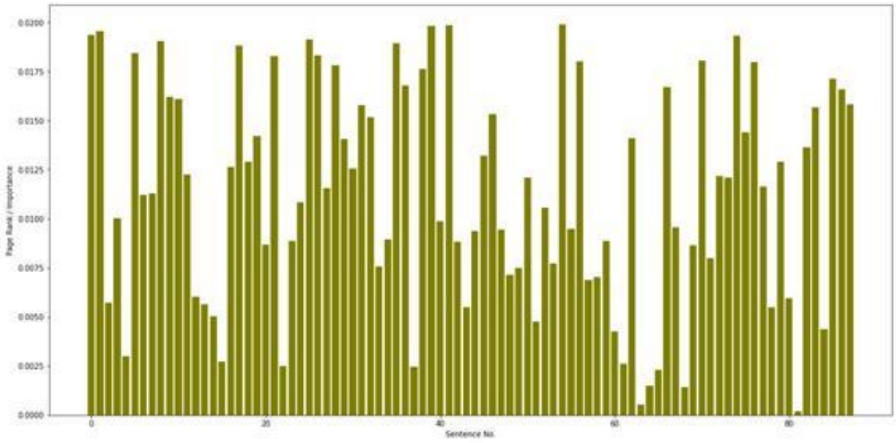


Figure 11.4: Sentence Number vs Importance

11.2.4 Word Clouds

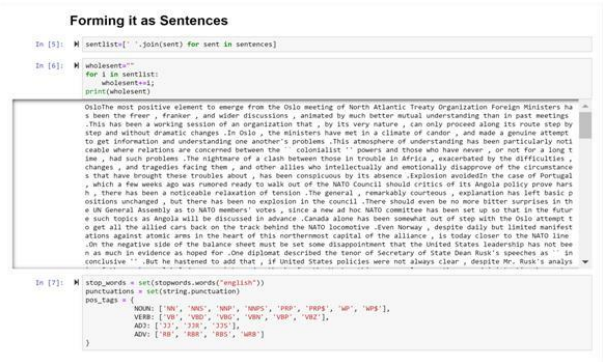
Figure 11.5: Brown Corpus - Input Word Cloud

Figure 11.6: Brown Corpus - Output Word Cloud

12. Appendix

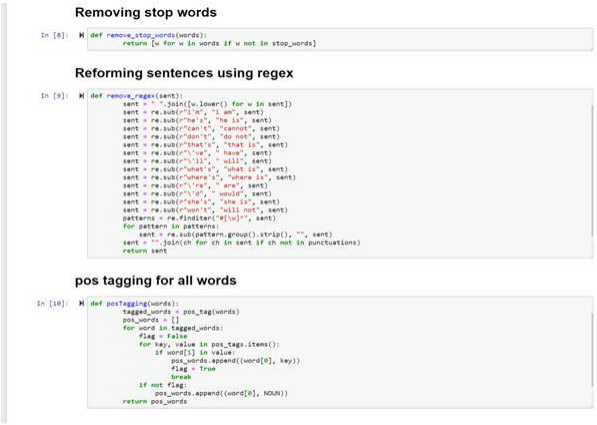
```
In [1]: import re
        import string
        import numpy as np
        import matplotlib.pyplot as plt
        from nltk import pos_tag
        from nltk.tokenize import sent_tokenize, word_tokenize
        from nltk.stem.wordnet import WordNetLemmatizer
        from nltk.corpus.reader.wordnet import NOUN, VERB, ADJ, ADV
        from nltk.corpus import brown, stopwords
        from nltk.cluster.util import cosine_distance
        from operator import itemgetter
        %matplotlib

        Using matplotlib backend: Qt5Agg

In [2]: np.seterr(divide='ignore', invalid='ignore')
Out[2]: {'divide': 'warn', 'over': 'warn', 'under': 'ignore', 'invalid': 'warn'}

Loading Corpus

In [3]: sentences = brown.sents('ca04')

In [4]: len(sentences)
Out[4]: 88
```

Figure 12.1: Code Snap-Shot 1



```
Forming it as Sentences

In [5]: sentlist=[' '.join(sent) for sent in sentences]

In [6]: wholesent=""
        for i in sentlist:
            wholesent+=i;
        print(wholesent)

        OsloThe most positive element to emerge from the Oslo meeting of North Atlantic Treaty Organization Foreign Ministers ha
        s been the freer , franker , and wider discussions , animated by much better mutual understanding than in past meetings
        .This has been a working session of an organization that , by its very nature , can only proceed along its route step by
        step and without dramatic changes .In Oslo , the ministers have met in a climate of candor , and made a genuine attempt
        to get information and understanding one another's problems .This atmosphere of understanding has been particularly noti
        ceable where relations are concerned between the `` colonialist '' powers and those who have never , or not for a long t
        ime , had such problems .The nightmare of a clash between those in trouble in Africa , exacerbated by the difficulties ,
        changes , and tragedies facing them , and other allies who intellectually and emotionally disapprove of the circumstance
        s that have brought these troubles about , has been conspicuous by its absence .Explosion avoidedIn the case of Portugal
        , which a few weeks ago was rumored ready to walk out of the NATO Council should critics of its Angola policy prove hars
        h , there has been a noticeable relaxation of tension .The general , remarkably courteous , explanation has left basic p
        ositions unchanged , but there has been no explosion in the council .There should even be no more bitter surprises in th
        e UN General Assembly as to NATO members' votes , since a new ad hoc NATO committee has been set up so that in the futur
        e such topics as Angola will be discussed in advance .Canada alone has been somewhat out of step with the Oslo attempt t
        o get all the allied cars back on the track behind the NATO locomotive .Even Norway , despite daily but limited manifest
        ations against atomic arms in the heart of this northernmost capital of the alliance , is today closer to the NATO line
        .On the negative side of the balance sheet must be set some disappointment that the United States leadership has not bee
        n as much in evidence as hoped for .One diplomat described the tenor of Secretary of State Dean Rusk's speeches as `` in
        conclusive '' .But he hastened to add that , if United States policies were not always clear , despite Mr. Rusk's analys

In [7]: stop_words = set(stopwords.words("english"))
        punctuations = set(string.punctuation)
        pos_tags = {
                    NOUN: ['NN', 'NNS', 'NNP', 'NNPS', 'PRP', 'PRP$', 'WP', 'WP$'],
                    VERB: ['VB', 'VBD', 'VBG', 'VBN', 'VBP', 'VBZ'],
                    ADJ: ['JJ', 'JJR', 'JJS'],
                    ADV: ['RB', 'RBR', 'RBS', 'WRB']
        }
```

Figure 12.2: Code Snap-Shot 2



```
Removing stop words

In [8]: def remove_stop_words(words):
            return [w for w in words if w not in stop_words]

Reforming sentences using regex

In [9]: def remove_regex(sent):
            sent = " ".join([w.lower() for w in sent])
            sent = re.sub(r"i'm", "i am", sent)
            sent = re.sub(r"he's", "he is", sent)
            sent = re.sub(r"can't", "cannot", sent)
            sent = re.sub(r"don't", "do not", sent)
            sent = re.sub(r"that's", "that is", sent)
            sent = re.sub(r"\'ve", " have", sent)
            sent = re.sub(r"\'ll", " will", sent)
            sent = re.sub(r"what's", "what is", sent)
            sent = re.sub(r"where's", "where is", sent)
            sent = re.sub(r"\'re", " are", sent)
            sent = re.sub(r"\'d", " would", sent)
            sent = re.sub(r"she's", "she is", sent)
            sent = re.sub(r"won't", "will not", sent)
            patterns = re.finditer('#(\w)*', sent)
            for pattern in patterns:
                sent = re.sub(pattern.group().strip(), "", sent)
            sent = "".join(ch for ch in sent if ch not in punctuations)
            return sent

pos tagging for all words

In [10]: def posTagging(words):
             tagged_words = pos_tag(words)
             pos_words = []
             for word in tagged_words:
                 flag = False
                 for key, value in pos_tags.items():
                     if word[1] in value:
                         pos_words.append((word[0], key))
                         flag = True
                         break
                 if not flag:
                     pos_words.append((word[0], NOUN))
             return pos_words
```

Figure 12.3: Code Snap-Shot 3

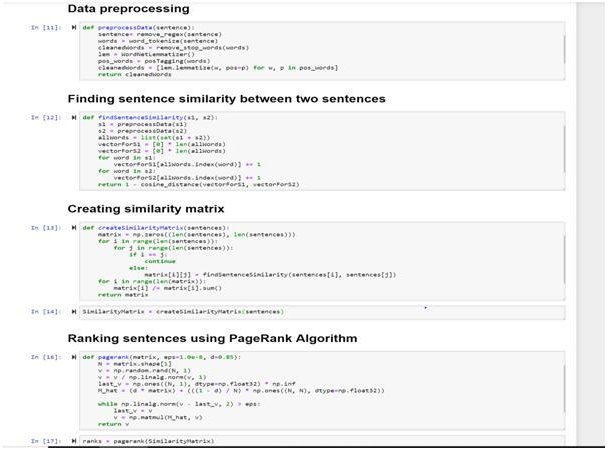


Figure 12.4: Code Snap-Shot 4

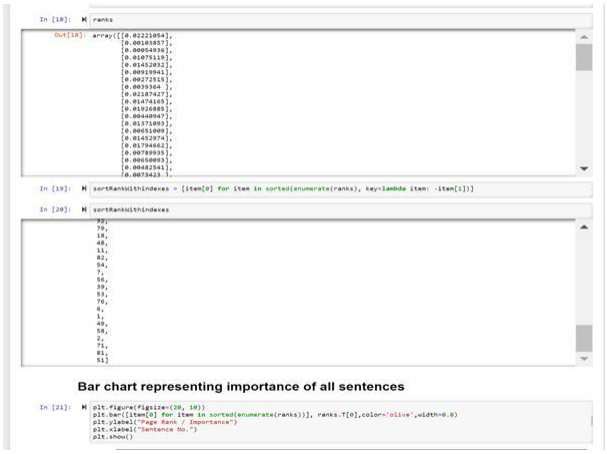


Figure 12.5: Code Snap-Shot 5

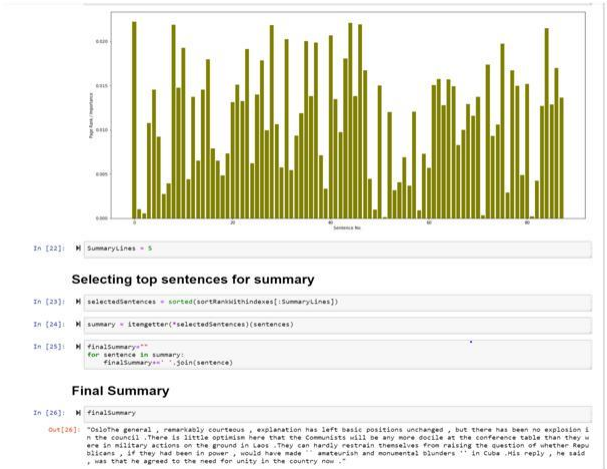


Figure 12.6: Code Snap-Shot 6

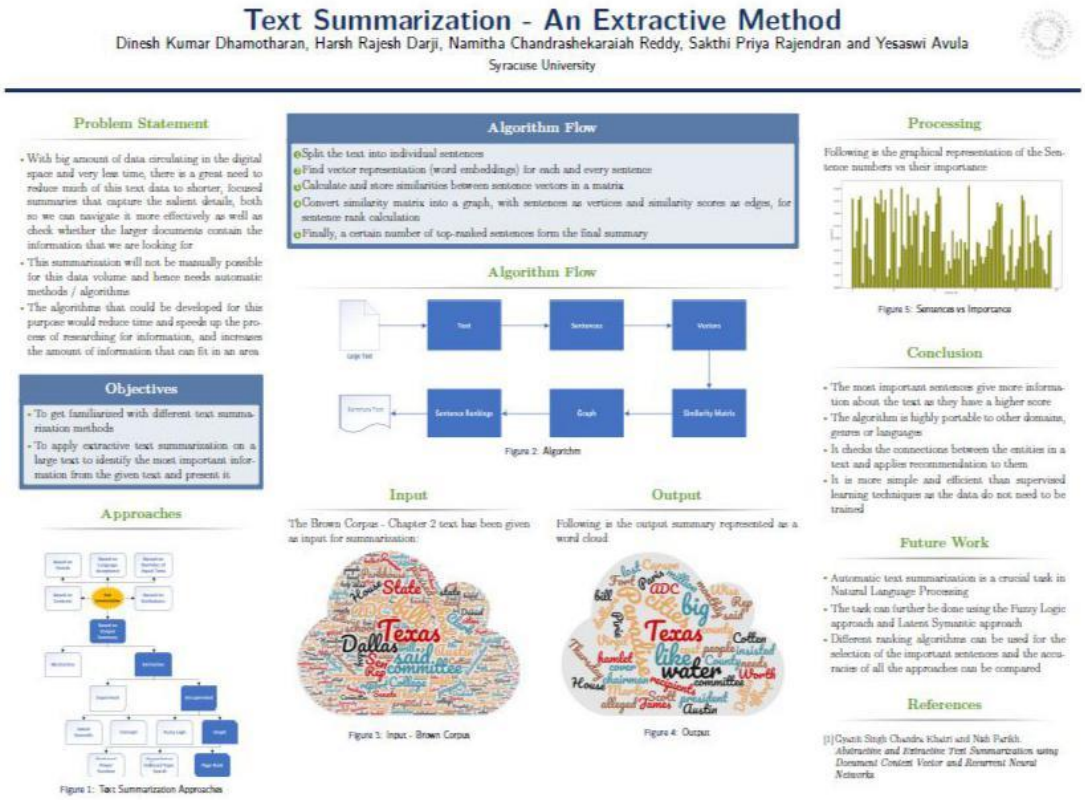


Figure 12.7: Poster

Bibliography

[1] Text Summarization in 5 steps, https://becominghuman.ai/text-summarization-in-5-steps-using-nltk-65b21e352b65

[2] Introduction to Text Summarization, https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/

[3] Unsupervised Text Summarization, https://medium.com/jatana/unsupervised-text-summarization-using-sentence-embeddings-adb15ce83db1

[4] Text Summarization Measures, http://www.cai.sk/ojs/index.php/cai/article/viewFile/37/24

[5] Graph-based Ranking Algorithms for Sentence Extraction, https://dl.acm.org/citation.cfm?id=1219064

[6] A survey on extractive text summarization, https://ieeexplore.ieee.org/document/7944061/