Project Report on

Extractive Text Summarization

at

U. V. Patel College of Engineering

Ganpat University | U.V. Patel College of Engineering

॥ विद्यया समाजोत्कर्षः ॥

Internal Guide:

Prof. Pravesh Patel

Prepared By:

Mr. Nishith Patel (19012011050)
Mr. Priyansh Patel (19012011051)

B.Tech. Semester VII
(Computer Engineering)
Nov-Dec **2022**

Submitted to,
Department of Computer Engineering
U.V. Patel College of Engineering
Ganpat University, Kherva - 384 012

U.V. PATEL COLLEGE
OF
ENGINEERING

Ganpat University
॥ विद्यया समाजोत्कर्षः ॥
U.V. Patel College of Engineering

24/11/2022

CERTIFICATE

TO WHOM SO EVER IT MAY CONCERN

This is to certify that Mr. Nishith J. Patel (19012011050) student of **B.Tech. Semester VII (Computer Engineering)** has completed his full semester on site project work titled “**Extractive Text Summarization**” is satisfactory. Department of Computer Engineering, Ganpat University, Kherva, Mehsana in the year 2022.

College Project Guide
Prof. Pravesh Patel

Dr. Paresh M. solanki,
Head,Computer Engineering

U.V. PATEL COLLEGE OF ENGINEERING

24/11/2022

CERTIFICATE

TO WHOM SO EVER IT MAY CONCERN

This is to certify that Mr. Priyansh S. Patel (19012011051) student of **B.Tech. Semester VII (Computer Engineering)** has completed his full semester on site project work titled “**Extractive Text Summarization**” is satisfactory. Department of Computer Engineering, Ganpat University, Kherva, Mehsana in the year 2022.

College Project Guide
Prof. Pravesh Patel

Dr. Paresh M. solanki,
Head,Computer Engineering

ACKNOWLEDGEMENT

This satisfaction that successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation it made it possible, whose constant guidance and encouragement crown all efforts with success. We are grateful to our guide **Prof. Pravesh Patel** for the guidance, inspiration and constructive suggestions that helpful us in the preparation of this project. We also thank our colleagues who have helped in successful completion of the project.

NISHITH PATEL [19012011050]
PRIYANSH PATEL [19012011051]

ABSTRACT

Since the amount of information on the internet is growing rapidly, it is not easy for a user to find relevant information for his/her query. To tackle this issue, much attention has been paid to Automatic Document Summarization. The key point in any successful document summarizer is a good document representation. Automatic text summarization is a major area of research in the domain of information systems. Most of the methods requires domain knowledge to produce a coherent and meaningful summary. In Extractive text summarization, sentences are scored on some features. Many feature-based scoring methods have been proposed for extractive automatic text summarization by researchers.

Text summarization is the process of automatically creating a shorter version of one or more text documents. It is an important way of finding relevant information in large text libraries or in the Internet. Essentially, text summarization techniques are classified as Extractive and Abstractive. Extractive techniques perform text summarization by selecting sentences of documents according to some criteria. Abstractive summaries attempt to improve the coherence among sentences by eliminating redundancies and clarifying the contest of sentences. In terms of extractive summarization, sentence scoring is the technique most used for extractive text summarization.

INDEX

1. INTRODUCTION 1

1.1 Classification of Text Summarization 2

1.2 Extractive Text Summarization Flow Chart 4

2. PROJECT SCOPE 5

2.1 Purpose 5

2.2 Problem Statement 5

2.3 Overview 5

2.4 Objective 5

2.5 Tools & Technology 6

3. FEASIBILITY STUDY: 7

3.1 Study of Current System: 7

3.2 Problem &Weakness of Current System: 7

3.3 Requirement of New System: 7

3.4 Technical Feasibility: 7

3.5 Economic Feasibility: 8

3.6 Operational Feasibility: 8

3.7 Literature Survey: 8

4. HARDWARE AND SOFTWARE REQUIREMENT: 9

5. PROJECT PLAN: 10

6. SYSTEM DESIGN: 11

6.1. Flow of the Extractive Text Summarization: 11

6.2 User Interface: 21

7. Result Analysis of Algorithms: 22

8. Conclusion and Future Work: 27

9. Reference: 28

List of Figures

Figure 1: Type of Text Summarization....2
Figure 2: Flow of Extractive Text Summarization 4
Figure 3: Flow of Extractive Summarization....12
Figure 4: Flow of Cuckoo Search....20
Figure 5: Home Page....21
Figure 6: Summary Page 21
Figure 7 : Graphs....23

List of Tables

Table 1: Hardware & Software Requirement 9
Table 2: Project Plan..10
Table 3: Result Using Sentence Scoring Method 22
Table 4:Result After Combination..23
Table 5: Result After Cuckoo Search Algorithm..23

1. INTRODUCTION

Text summarization is a method used to generate summaries of text documents by extracting important information from the summarized document. It is used extensively in summarizing the search engine results, providing brief version of large documents in which abstracts are not present. A summary of a text document is useful for humans to quickly extract the important information in the text.

The rapid growth of the Internet yielded a massive increase of the amount of information available, especially regarding text documents (e.g. news articles, electronic books, scientific papers, blogs, etc.). Text summarization involves two types, abstractive and extractive. In ABSTRACTIVE it considers human knowledge and understands the text in order to generate the summary whereas in EXTRACTIVE the important sentence is picked from the text as a summary. However multiple documents can also be summarized by choosing the proper integration algorithms.

The text summarization using sentence scoring involves four phases: - Pre-Processing, Sentence Scoring, Sentence Ranking, Summary Extraction. The text document is selected and data in it is segmented into sentences and tokens during the pre-processing stage. Each sentence is evaluated in sentence scoring by considering the linear combination of multiple parameters like frequency, sentence position, cue words, similarity with title, sentence length and proper noun. The sentences are ranked with respect to the scores. The summary is generated by selecting the required number of sentences with highest rank and it is made sure that no consecutive sentences are selected.

The Text Summarization is very helpful in daily tasks. Some of the day to day real life applications of text summarization are Media monitoring, Newsletters, Search marketing and SE, Financial research, Social media marketing, Question answering and bots, Books and literature, Email overload, Science and R&D, Automated content creation.

1.1 Classification of Text Summarization

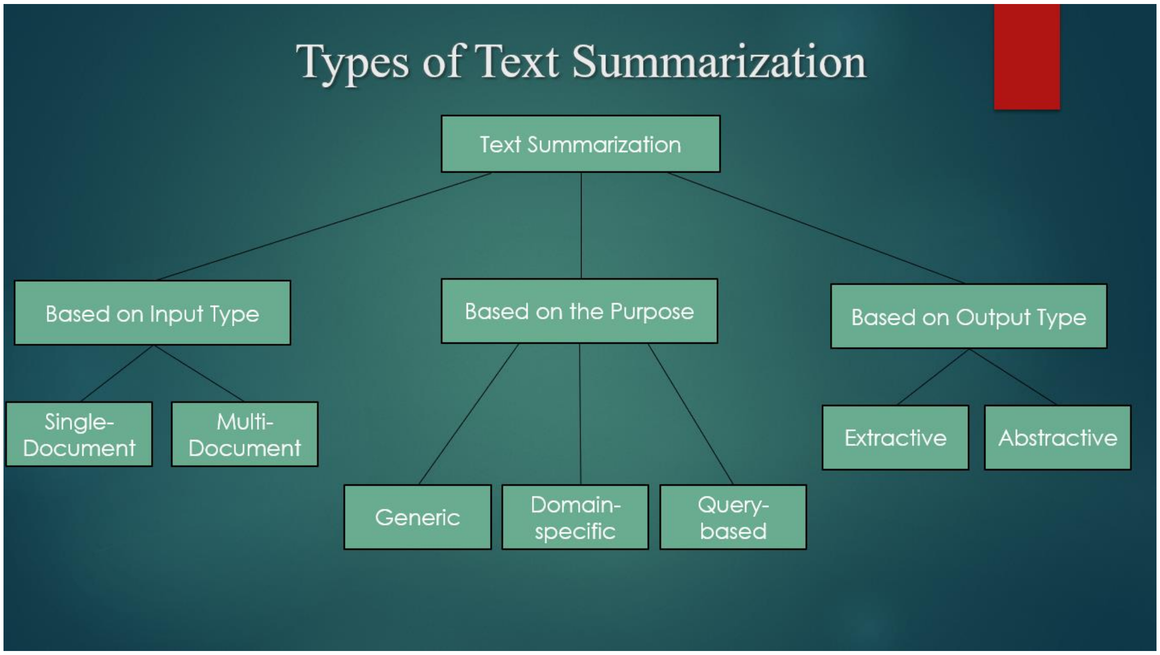


Figure 1: Type of Text Summarization

• **Single Document Summarization**

This is basically self-explanatory. Single document summarizers aim to summarize one single document. Given a single document to generate a summary.

• **Multiple Document Summarization**

Given a group of documents to generate a summary out of it. Multiple Document or multiple text summarization includes multiple documents and the final paper has to contain summarized information from all the documents.

• **Generic Summarization**

Summarize the content of a document. Generic summarizations are not programmed to make any assumptions like the domain-specific or query-based summarizers. It just condenses or summarizes the information from the source document.

• **Domain Specific**

Text summarization aims at extracting the essential information from a text to produce a shorter version, such as generating headlines for news and subject lines for emails. Domain knowledge is used in domain-specific summarization. Domain-specific summarizers can be integrated with specific context, knowledge, and words. For example, models can be integrated with words used in medical science so that it can better understand scientific articles on medical science and summarize them.

• **Query-focused Summarization**

Summarize a document with respect to an information need expressed in a user query. Query-based summaries mostly contain information about natural language questions. This is similar to Google's search results. Sometimes we type in questions on the search bar and Google shows us websites or articles that have answers to our questions. It shows us a snippet or summary of an article related to the question we searched.

• **Extractive Text Summarization**

Extraction-based summarization is a simple process. The important words and phrases are taken out of the original text and compiled together to make the summary. There is no rephrasing or using synonyms in this summarization process. The words are taken out as they are and slightly rearranged to give the sentence a structure. Because there is no use of synonyms and no rephrasing, it makes the summarization process easier. Create the summary from phrases or sentences in the given documents

• **Abstractive Text Summarization**

Abstraction-based summarization is more complex than extraction-based summarization. It takes out the original and important sentence from a text document and rephrases it with proper synonyms. That way, it will look like a completely different text but have the same meaning as the original text.

That is why it is difficult because figuring out the right synonyms and rephrasing by keeping the meaning the same is tough. Express the idea in the source document using different word

1.2 Extractive Text Summarization Flow Chart

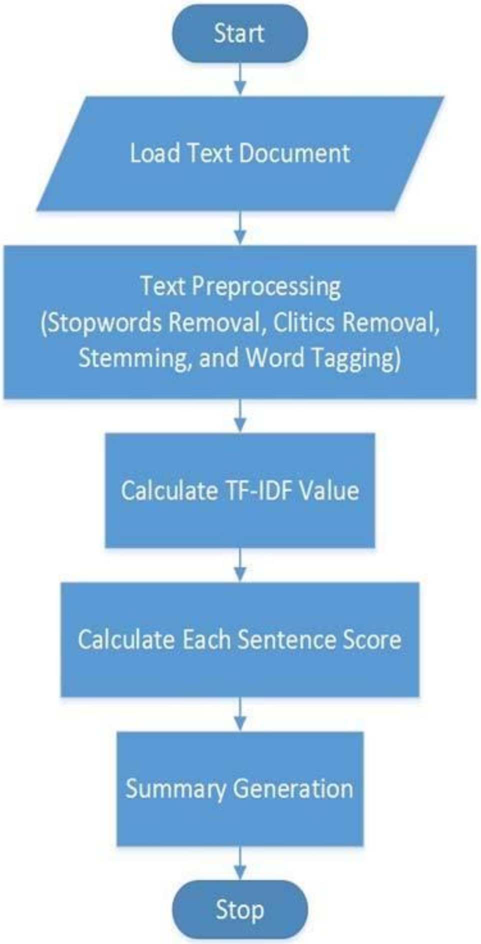

Start
Load Text Document
Text Preprocessing
(Stopwords Removal, Clitics Removal,
Stemming, and Word Tagging)
Calculate TF-IDF Value
Calculate Each Sentence Score
Summary Generation
Stop

Figure 2: Flow of Extractive Text Summarization

2. PROJECT SCOPE

The main contribution of this work is proposing an unsupervised approach for extractive single-document summarization. The rapid growth of the Internet yielded a massive increase of the amount of information available, especially regarding text documents (e.g. news articles, electronic books, scientific papers, blogs, etc.).Hence, Automatic text summarization is a process of condensing the text document automatically by computer system such that the produced summary is no longer than a specified threshold [typically one-third of the size of original text] and covers all the information contents of it which helps user to read the whole document very easily

2.1 Purpose

With the present explosion of data circulating the digital space, which is mostly non-structured textual data, there is a need to develop automatic text summarization tools that allow people to get insights from them easily. Currently, we enjoy quick access to enormous amounts of information. However, most of this information is redundant, insignificant, and may not convey the intended meaning. Therefore, using automatic text summarizers capable of extracting useful information that leaves out inessential and insignificant data is becoming vital. Implementing summarization can enhance the readability of documents, reduce the time spent in researching for information, and allow for more information to be fitted in a particular area.

2.2 Problem Statement

Exponential growth of textual documents on web in the last few decades forced researchers to find ways so that users can save their time and resources in finding relevant information. On the other side the information contents should also be preserved. The produced text summary should be a true condensed replica of the original text. Text summarization is a method used to generate summaries of text documents by extracting important information from the summarized document.

2.3 Overview

The amount of information available has significantly increased because of the Internet's explosive expansion, particularly in the area of text materials. Text summarizing is a technique for creating summaries of text documents by taking the key details out of the document to be summarized. It is frequently used to provide a concise rendition of lengthy texts without abstracts in search engine results. Humans can quickly obtain the key details from a text document by reading the summary.

2.4 Objective

This study's key contribution is the unsupervised method it suggests for extractive single-document summarization. The Internet's explosive expansion has resulted in a tremendous rise in the amount of information available, particularly for text documents (e.g. news articles, electronic books, scientific papers, blogs, etc.). As a result, automatic text summarization is the process of automatically compressing a text document by a computer system so that the produced summary is no longer than a specified size and covers all of its information contents, allowing the user to read the entire document very easily.

2.5 Tools & Technology

- Python – 3.0
- HTML5
- CSS3
- JavaScript
- BootStrap-5

3. FEASIBILITY STUDY:

The main aim of the feasibility study is to determine whether it would be financially and technically feasible to develop the product. This project will require the use of DUC 2001 and 2002 datasets. The people should become familiar with new technology and became a part of the text summarization.

3.1 Study of Current System:

Nowadays the people are using technology very much in every aspect like for studying, for fun, for entertainment etc. In market many sites are available for text summarization. They provide the features of text summarization but with some limitations. This problem has given the new solution that is associated to data mining and machine learning which returns query specific information from large set of offline documents and represents as a single document to the user. So, automated summarization is an important area in Natural Language Processing (NLP) research. Automated summarization provides single document summarization and multi-document summarization.

3.2 Problem &Weakness of Current System:

In today's time people are using technology very much in everything, so the product should be able to serve the users efficiently anytime anywhere without any delay or loading. The available platforms nowadays are sometime getting leggy or are taking very much time to load data or the content user want to see, which is very frustrating as a user. So, this type of situation is very difficult for user to access the content. The replacement of the product become very tedious job because this process takes time.

3.3 Requirement of New System:

- To remove problems of the current system we require new system which has to fulfill the user requirements and satisfy them.
- The new platform should be faster for better user experience.
- Then new platform must run smoothly in order to satisfy the user.
- The user data should be secured.
- Easy understanding of the application so that everyone can easily access it.

3.4 Technical Feasibility:

Technical Feasibility current resources both hardware software along with required technology are analyzed/assessed to develop project. This technical feasibility study gives report whether there exists correct required resources and technologies which will be used for project development. Along with this, feasibility study also analyzes technical skills and capabilities of technical team, existing technology can be used or not, maintenance and upgradation is easy or not for chosen technology. We are using visual studio toolkit in which we are developing our project in Python. We are pretty familiar about these technologies and our hardware also supports this technology.

3.5 Economic Feasibility:

In Economic Feasibility study cost and benefit of the project is analyzed. Means under this feasibility study a detail analysis is carried out about what will be cost of the project for development which includes all required cost for final developments like hardware and software resource required, design and development cost and operational cost and so on. After that it is analyzed whether project will be beneficial in terms of finance for organization or not. In our project we are using minimum resources and open languages for which we don’t have to invest much. Our main hardware needed for the development of this project is basic, so we can easily afford the development of this project. We have used Python language which is open source.

3.6 Operational Feasibility:

In Operational Feasibility degree of providing service to requirements is analyzed along with how much easy product will be to operate and maintenance after deployment. Operational feasibility also checks, whether the end user will adapt the new system or not. If user does not understand or is not able to work with this app then development of this system is waste of time and money. In this system I will also provide user guide that will help user to learn about it and access it easily. We are making our project as simple as possible so that the user find it easy to work with and can easily access it.

3.7 Literature Survey:

There are many prominent works in Text Summarization from the past few years. Earlier works dealt mainly with Single Document Text Summarization. Now that the technology has increased as well as computing power has increased which paved the path for a faster, more effective and more accurate way of processing documents when compared with the earlier methods.

An improved method of text summarization for web contents using lexical chain with semantic-related terms proposes an improved extractive text summarization method for documents by enhancing the conventional lexical chain method to produce better relevant information. Then, firstly investigated the approaches to extract sentences from the document(s) based on the distribution of lexical chains which learns the characteristics of the assigned keywords from the training data set. A new method of ensemble ranking algorithm is used to rank sentences TFIDF is used to word count and word level feature extraction.

4. HARDWARE AND SOFTWARE REQUIREMENT:

Developer Side Requirements	Client-Side Requirements
Hardware Requirement: • Computer/Laptop with minimum 4GB RAM and minimum 246GB Hard drive. • High speed internet connectivity • Google Colab, jupyter **Software Requirement:** • Browser • OS: Windows 10	• Any type of device which has browser with supports of html5, css3 and 4G internet connectivity.

Table 1: Hardware & Software Requirement

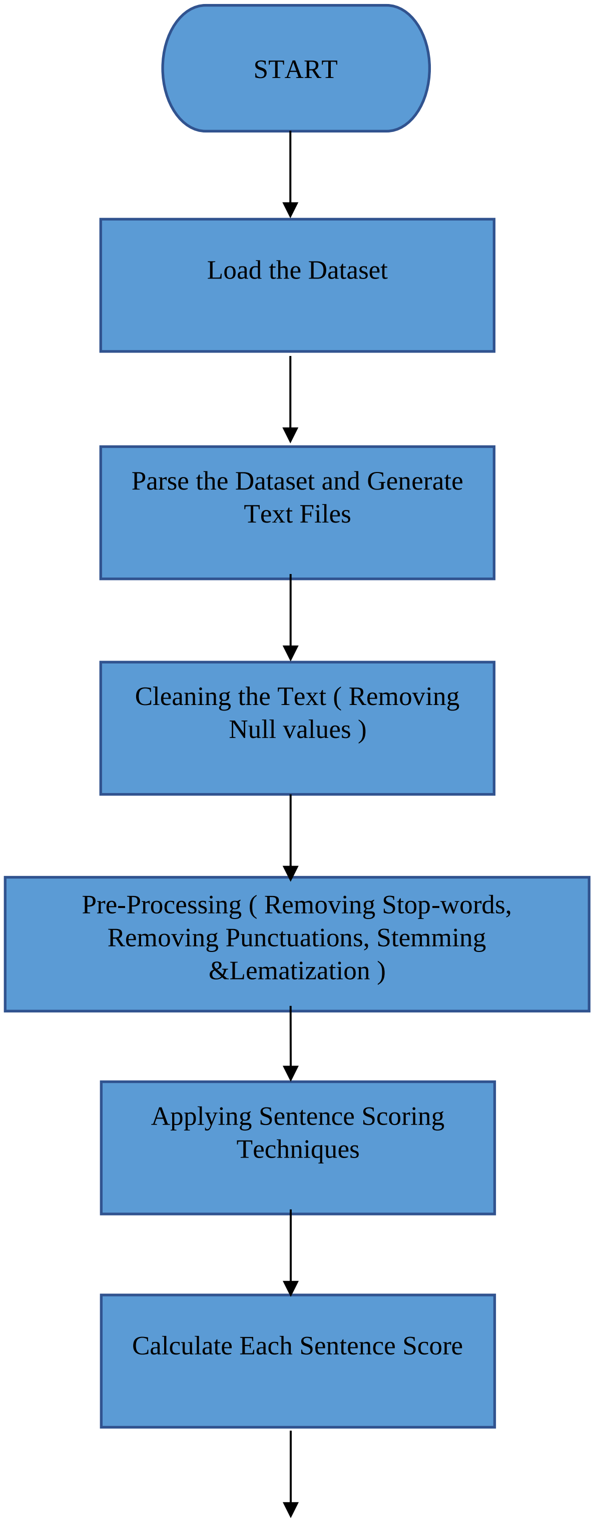
5. PROJECT PLAN:

Development Phase	Timeline	Description
Project Initialization	Jul-Aug	Researching and Exploring other similar products
Searching appropriate dataset	Jul-Aug	Searching appropriate dataset
Requirement Gathering	Jul-Aug	Researching and exploring other similar products
Validate & Update	Aug-Sep	Validating the previous work done and updating as per requirement.
Development	Sep-Oct	Project Phase 1 Development

Table 2: Project Plan

6. SYSTEM DESIGN:

6.1. Flow of the Extractive Text Summarization:


START
Load the Dataset
Parse the Dataset and Generate Text Files
Cleaning the Text (Removing Null values)
Pre-Processing (Removing Stop-words, Removing Punctuations, Stemming &Lematization)
Applying Sentence Scoring Techniques
Calculate Each Sentence Score

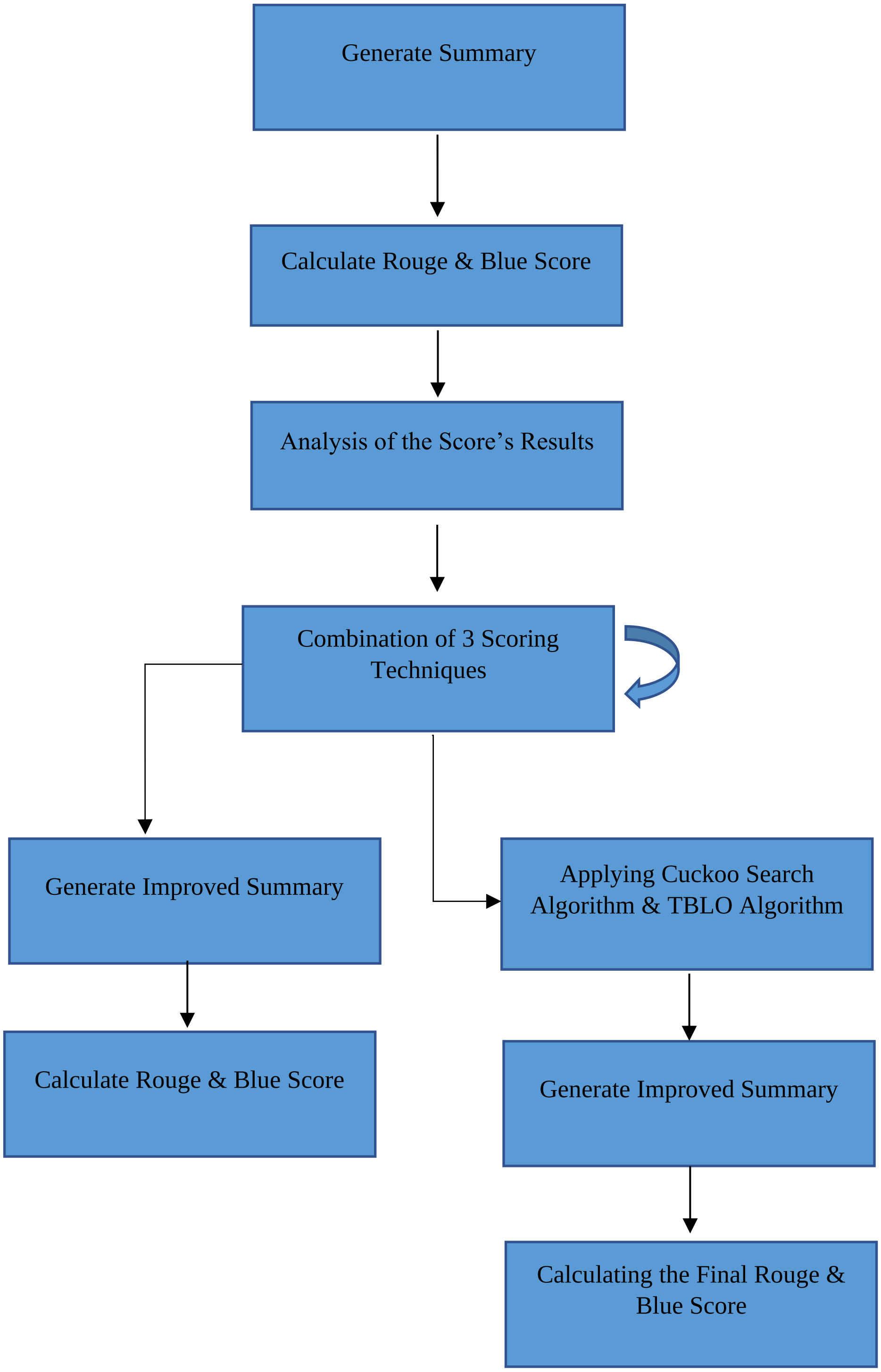

Generate Summary
Calculate Rouge & Blue Score
Analysis of the Score's Results
Combination of 3 Scoring Techniques
Generate Improved Summary
Applying Cuckoo Search Algorithm & TBLO Algorithm
Calculate Rouge & Blue Score
Generate Improved Summary
Calculating the Final Rouge & Blue Score

Figure 3: Flow of Extractive Summarization

Dataset Description:

To further progress in summarization and enable researchers to participate in large-scale experiments, the National Institute of Standards and Technology (NIST) continued an evaluation in the area of text summarization called the **Document Understanding Conference** (DUC). DUC is part of a Defense Advanced Research Projects Agency (DARPA) program, Translingual Information Detection, Extraction, and Summarization (TIDES), which specifically calls for major advances in summarization technology, both in English and from other languages to English (cross-language summarization). The basic design for the evaluation follows ideas in a summarization road map that was created by a committee of researchers in summarization, headed by Daniel Marcu. It also profited from the experiences of DUC 2001.

Plans called for the creation of reference data (documents and summaries) for testing. DUC 2001 data was available as training data for DUC 2002, once the short application had been submitted (see "Call for participation" below) and the required permission forms were signed. No additional training data was created due to the shortened schedule. The test data for DUC 2002 was distributed at the end of March and results were due for evaluation mid-April. The DUC 2002 workshop was held as part of the ACL-2002 Automatic Summarization Workshop, July 11-12, 2002 in Philadelphia to discuss these results and to make plans for further evaluations.

Parsing the dataset:

Parsing is the process of extracting the useful content from the given dataset. In this project we have used DUC 2001 dataset where files are there in xml format. We have used “Beautiful soup (Python Library for pulling out data from XML file)” for extracting the Text which is the actual news which we are going to use for summarization.

Generating Text Files:

The modified Parser also generates the text files from the extracted text data from the dataset. This is the actual data on which we are going to apply the text summarization.

Cleaning the data:

In this process the data is cleaned by removing the null values from the dataset and also dealt with the abnormal data.

Pre-Processing:

In this step we have removed stop-words, punctuation marks from the data. We have also applied stemming and lemmatization to make the data more clean for further processing.

Scoring Techniques:

Proper noun
Sentence length character
Sentence length words

Sentence position
Word frequency
Numerical value
Named entity
Iterative query score
Cue words

Word frequency:

As the name suggests the more the frequency of the word in the sentence the higher will be it's score. In other words the sentence containing the most frequent words of the document has high chances of getting selected for the Final summary. This is based on the assumption that the higher the frequency of the word in the text , it is more likely to be related to the subject of the document.

$S(L) = N(w) / N(d)$

Where,
$N(w)$ = Sum of the frequency of the words of the sentence
$N(d)$ = Sum of the frequency of the words of the document.

Proper noun:

Proper noun refers to an individual, place or organization. It is considered to be carrying greater information from the rest of the words.
The Sentence containing a higher number of proper nouns is more likely to be selected for the final summary. It is a specialization of the upper case method.

The score is calculated as the number of proper nouns of the sentence to the number of proper noun of the document.

$S(P) = N(P) / N(D)$

Where,
$N(P)$ = Number of the proper nouns of the sentence
$N(D)$ = Number of the proper nouns of the document.

Sentence length character:

This method is used for calculating the sentence score lengthwise. The average sentence length is calculated by following function given below:

$Avg(L)= max(L)+min(L)/2$

Where,
$max(L)$ = maximum length of the sentence (character-wise)
$min(L)$ = minimum length of the sentence (character-wise)

$Score= (|Sl -Avg(L) |) / Msl$
Where,
Sl= Sentence length
Avg(L) = The average sentence length of the document
Msl = The maximum sentence length

Sentence length words:

This method is used for calculating the sentence score lengthwise. The average sentence length is calculated by following function given below:

$Avg(L)= max(L)+min(L)/2$

Where,
max(L) = maximum length of the sentence (word-wise)
min(L) = minimum length of the sentence (word-wise)

$Score= (|Sl -Avg(L) |) / Msl$

Where,
Sl= Sentence length
Avg(L) = The average sentence length of the document
Msl = The maximum sentence length

Sentence position:

It is assumed to be the most important feature for sentence scoring. Normally a few sentences at the beginning and at the end are considered to be more important than the other sentences. They are more likely to be selected for the final summary.

Numerical value:

Numerical values represent important figures. According to this method the sentence containing numerical values are considered to be more important than other sentences.

$S(L) = Nnv / T$

Where,
Nnv = Number of numerical values in the sentence
T = Total number of numerical values in the document.

Named entity:

The sentence containing more number of named entities will be considered to be having more weight than other sentences.

S(L)= Net / T
Where,
Net = Number or named entity of the sentence
T = Number of named entities of the document.

Iterative query score:

In this technique first we calculate the frequency of the words contained in the document and then the words are sorted based on their frequency. Then the top frequent words are chosen from the sorted list which is known as the initial keyword set. Then the sentence containing the higher number of the initial keywords are considered to be more important than other sentences and they have high chances to get selected for the final summary.

Score= Nit / T

Where,
Nit = Number of the initial keywords of the sentence
T = Total number of initial keywords of the document.

Cue words :

In this technique the sentences are scored based on the number of cue words it contains. For using this technique a set of cue words like ‘In brief’ ,’Summing up’, ‘thus’, ‘to conclude’ has to prepared.

Score = Ncw / T
Where,
Ncw = Number of cue words in the sentence
T = Number of cue words of the document

Calculating Sentence Score:

After Applying different scoring techniques we calculate the individual scores of each sentences. This scores are now used for generation of the final summary.

Generate Summary:

In this process the sentences having highest score’s are selected and then used for generating the final summary.

Calculate Rouge & Blue Score:

In this process we compare our generated summary with the golden summary and find the rouge and blue score which tells us how much our system generated summary is accurate in comparison with the golden summary.

Analysis of Score's Result:

In the process we analyze the Rouge & Blue score and after analyzing the score we will select combinations of 3 techniques for further score improvement.

Combinations of 3 Scoring Techniques:

In this process we take combinations of scoring techniques having the highest score in individual methods which will improve the accuracy of the generated summary.

Generate Improved Summary:

In this process the sentences having highest score's are selected and then used for generating the improved summary.

Calculate Rouge & Blue Score:

In this process we compare our generated summary with the golden summary and find the rouge and blue score which tells us how much our system generated summary is accurate in comparison with the golden summary.

Applying Cuckoo & Tblo Score:

In this process we will apply the optimization algorithm which will improve the sentence score by using heuristic function of cuckoo and tblo algorithm.

Calculate Rouge & Blue Score:

In this process we compare our generated summary with the golden summary and find the rouge and blue score which tells us how much our system generated final summary is accurate in comparison with the golden summary.

Cuckoo Search Algorithm:

Cuckoo search is an optimization algorithm used in Natural Language Processing. It was inspired by the obligate brood parasitism of some cuckoo species by laying their eggs in the nests of host birds of other species. Some host birds can engage direct conflict with the intruding cuckoos.

For example, if a host bird discovers the eggs are not their own, it will either throw these alien eggs away or simply abandon its nest and build a new nest elsewhere. Cuckoo search idealized such breeding behavior, and thus can be applied for various optimization problems. It has been shown that cuckoo search is a special case of the well-known (μ + λ)-evolution strategy.

Each egg in a nest represents a solution, and a cuckoo egg represents a new solution. The aim is to use the new and potentially better solutions (cuckoos) to replace a not-so-good solution in the nests. In the simplest form, each nest has one egg. The algorithm can be extended to more complicated cases in which each nest has multiple eggs representing a set of solutions.

CS is based on three idealized rules:

- Each cuckoo lays one egg at a time, and dumps its egg in a randomly chosen nest.
- The best nests with high quality of eggs will carry over to the next generation.
- The number of available hosts nests is fixed, and the egg laid by a cuckoo is discovered by the host bird with a probability in range (0,1). In this case, the host bird can throw the egg away/abandon the nest, and build a completely new nest.

Calculation for Levy's Flight – Levy's Flight provide a random walk,

$$x_i^{t+1} = x_i^t + \propto \bigoplus Levy(\lambda)$$

Random Step can be drawn from a Levy's Distribution,

$$Levy \sim u \ = \ t^{-\lambda} \, (1 \ < \ \lambda \ \leq \ \ 3$$

Step Size can be Expressed as,

$$s \ = \ \frac{\sigma_u \ * \ u}{|v|^{1/\beta}}$$

Standard Deviation Calculated as,

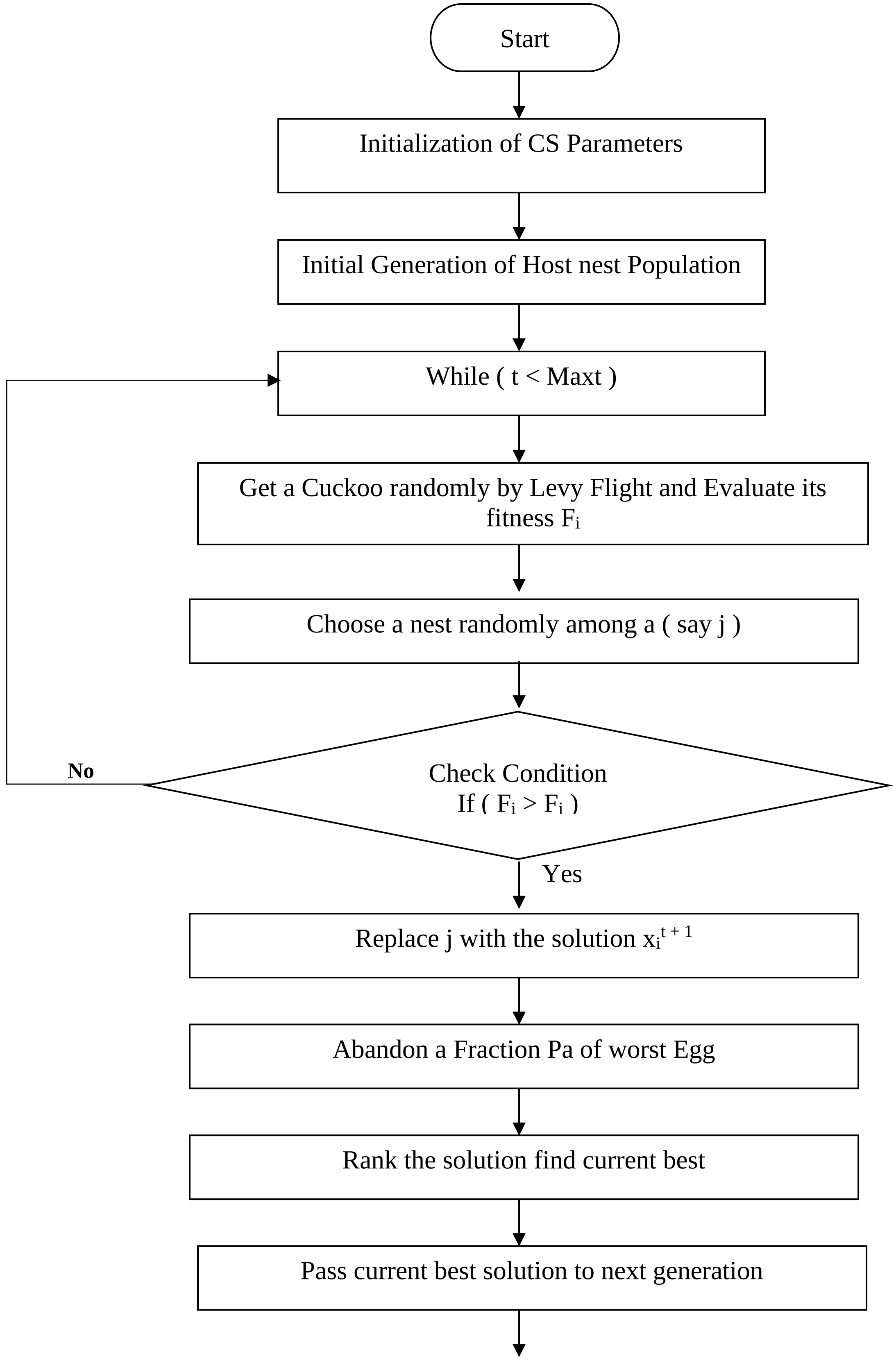
$$\sigma_u = \left(\frac{\Gamma(1 \ + \ \beta) \ * \ \mathrm{Sin}(\pi \ * \ \beta/2)}{\Gamma((1 \ + \ \beta)/\ 2) \ * \ \beta \ * 2^{(\beta \ - \ 1)/2}} \right)^{1/\beta}$$

$$\sigma_u \ = \ 0.6966$$

Fitness Function Calculated as,

F(x) = x1 * 2 – x1 * x2 + x2 ^ 2 + 2 * x1 + 4 * x2 + 3

Cuckoo Search algorithm Flow:


Start
Initialization of CS Parameters
Initial Generation of Host nest Population
While (t < Maxt)
Get a Cuckoo randomly by Levy Flight and Evaluate its fitness F_i
Choose a nest randomly among a (say j)
Check Condition
If ($F_i > F_i$)
No
Yes
Replace j with the solution x_i^{t+1}
Abandon a Fraction Pa of worst Egg
Rank the solution find current best
Pass current best solution to next generation

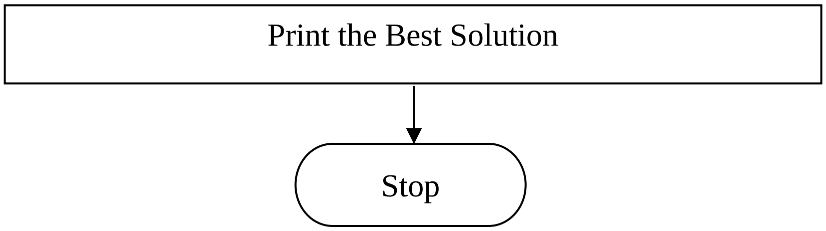


Figure 4: Flow of Cuckoo Search

Rouge Score:

The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scoring algorithm evaluates the similarity between a candidate document and a collection of reference documents. Use the ROUGE score to evaluate the quality of document translation and summarization models.

Recall:

The recall counts the number of overlapping n-grams found in both the model output and reference then divides this number by the total number of n-grams in the reference.

$$\text{Recall} = \frac{\text{Number of n-gram found in model and reference}}{\text{Number of n-gram in reference}}$$

Precision:

To avoid this, we use the precision metric – which is calculated in almost the exact same way, exact same way, but rather than dividing by the reference n-gram count, we divide by the model n-gram count.

$$Precision = \frac{\text{Number of n-gram found in model and reference}}{\text{Number of n-gram in reference}}$$

F1- Score:

Now that we both the recall and precision values, we can use them to calculate out ROUGE score like so:

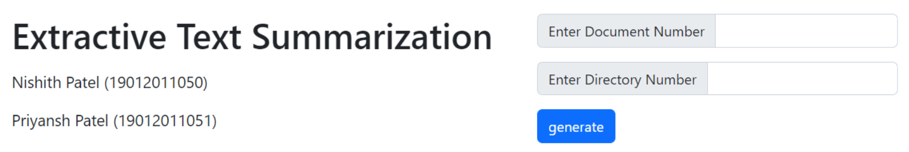
$$F1\,Score = 2 * \frac{precision * recall}{precision + recall}$$

Bleu:

Bleu is a measurement of the difference between an automatic translation and human-created reference translations of the same source sentence. The BLEU algorithm compares consecutive phrases of the automatic translation with the consecutive phrases it finds in the reference translation, and counts the number of matches, in a weighted fashion. These matches are position independent. A higher match degree indicates a higher degree of similarity with the reference translation, and higher score. Intelligibility and grammatical correctness aren't taken into account.

6.2 User Interface:

Home page:



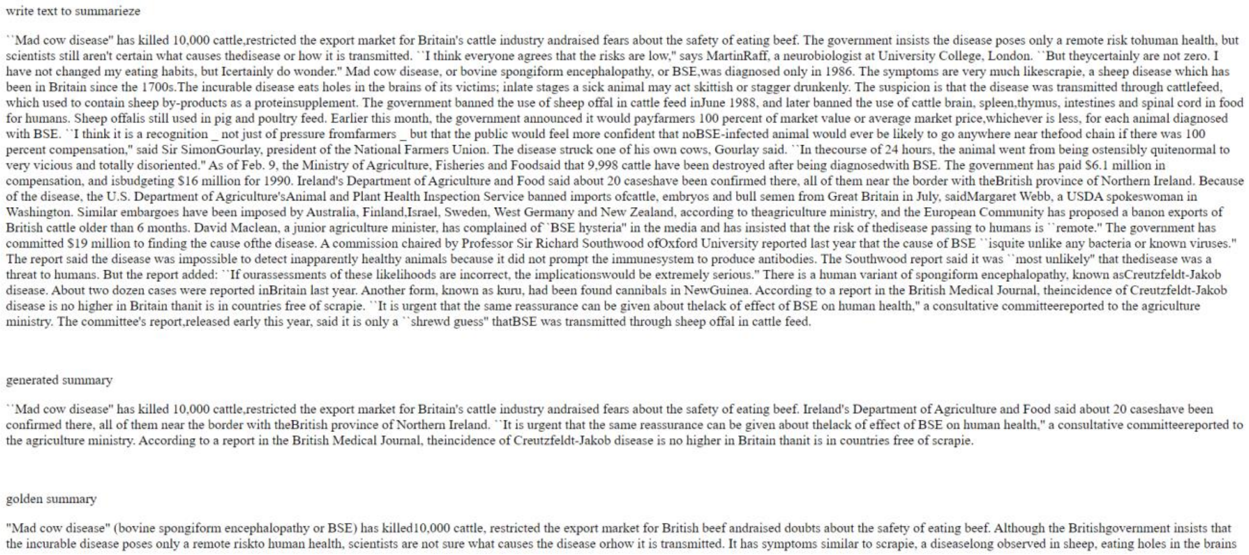
Extractive Text Summarization

Nishith Patel (19012011050)

Priyansh Patel (19012011051)

Figure 5: Home Page

Summary page:



write text to summarieze

``Mad cow disease'' has killed 10,000 cattle,restricted the export market for Britain's cattle industry andraised fears about the safety of eating beef. The government insists the disease poses only a remote risk tohuman health, but scientists still aren't certain what causes thedisease or how it is transmitted. ``I think everyone agrees that the risks are low,'' says MartinRaff, a neurobiologist at University College, London. ``But theycertainly are not zero. I have not changed my eating habits, but Icertainly do wonder.'' Mad cow disease, or bovine spongiform encephalopathy, or BSE,was diagnosed only in 1986. The symptoms are very much likescrapie, a sheep disease which has been in Britain since the 1700s.The incurable disease eats holes in the brains of its victims; inlate stages a sick animal may act skittish or stagger drunkenly. The suspicion is that the disease was transmitted through cattlefeed, which used to contain sheep by-products as a proteinsupplement. The government banned the use of sheep offal in cattle feed inJune 1988, and later banned the use of cattle brain, spleen,thymus, intestines and spinal cord in food for humans. Sheep offalis still used in pig and poultry feed. Earlier this month, the government announced it would payfarmers 100 percent of market value or average market price,whichever is less, for each animal diagnosed with BSE. ``I think it is a recognition _ not just of pressure fromfarmers _ but that the public would feel more confident that noBSE-infected animal would ever be likely to go anywhere near thefood chain if there was 100 percent compensation,'' said Sir SimonGourlay, president of the National Farmers Union. The disease struck one of his own cows, Gourlay said. ``In thecourse of 24 hours, the animal went from being ostensibly quitenormal to very vicious and totally disoriented.'' As of Feb. 9, the Ministry of Agriculture, Fisheries and Foodsaid that 9,998 cattle have been destroyed after being diagnosedwith BSE. The government has paid $6.1 million in compensation, and isbudgeting $16 million for 1990. Ireland's Department of Agriculture and Food said about 20 caseshave been confirmed there, all of them near the border with theBritish province of Northern Ireland. Because of the disease, the U.S. Department of Agriculture'sAnimal and Plant Health Inspection Service banned imports ofcattle, embryos and bull semen from Great Britain in July, saidMargaret Webb, a USDA spokeswoman in Washington. Similar embargoes have been imposed by Australia, Finland,Israel, Sweden, West Germany and New Zealand, according to theagriculture ministry, and the European Community has proposed a banon exports of British cattle older than 6 months. David Maclean, a junior agriculture minister, has complained of `BSE hysteria'' in the media and has insisted that the risk of thedisease passing to humans is ``remote.'' The government has committed $19 million to finding the cause ofthe disease. A commission chaired by Professor Sir Richard Southwood ofOxford University reported last year that the cause of BSE ``isquite unlike any bacteria or known viruses.'' The report said the disease was impossible to detect inapparently healthy animals because it did not prompt the immunesystem to produce antibodies. The Southwood report said it was ``most unlikely'' that thedisease was a threat to humans. But the report added: ``If ourassessments of these likelihoods are incorrect, the implicationswould be extremely serious.'' There is a human variant of spongiform encephalopathy, known asCreutzfeldt-Jakob disease. About two dozen cases were reported inBritain last year. Another form, known as kuru, had been found cannibals in NewGuinea. According to a report in the British Medical Journal, theincidence of Creutzfeldt-Jakob disease is no higher in Britain thanit is in countries free of scrapie. ``It is urgent that the same reassurance can be given about thelack of effect of BSE on human health,'' a consultative committeereported to the agriculture ministry. The committee's report,released early this year, said it is only a ``shrewd guess'' thatBSE was transmitted through sheep offal in cattle feed.

generated summary

``Mad cow disease'' has killed 10,000 cattle,restricted the export market for Britain's cattle industry andraised fears about the safety of eating beef. Ireland's Department of Agriculture and Food said about 20 caseshave been confirmed there, all of them near the border with theBritish province of Northern Ireland. ``It is urgent that the same reassurance can be given about thelack of effect of BSE on human health,'' a consultative committeereported to the agriculture ministry. According to a report in the British Medical Journal, theincidence of Creutzfeldt-Jakob disease is no higher in Britain thanit is in countries free of scrapie.

golden summary

"Mad cow disease" (bovine spongiform encephalopathy or BSE) has killed10,000 cattle, restricted the export market for British beef andraised doubts about the safety of eating beef. Although the Britishgovernment insists that the incurable disease poses only a remote riskto human health, scientists are not sure what causes the disease orhow it is transmitted. It has symptoms similar to scrapie, a diseaselong observed in sheep, eating holes in the brains

Figure 6: Summary Page

7. Result Analysis of Algorithms:

Result Using Sentence Scoring Methods:

Method	Rough - 1			Rough - 2			Rough - L		
	Recall	**Precision**	**F-Score**	**Recall**	**Precision**	**F-Score**	**Recall**	**Precision**	**F-Score**
Proper Noun	0.258	0.244	0.248	0.083	0.080	0.080	0.238	0.226	0.229
Sentence Length Character	0.258	0.249	0.250	0.080	0.079	0.078	0.236	0.228	0.230
Sentence Length Word	0.261	0.246	0.251	0.075	0.072	0.073	0.239	0.225	0.230
Sentence Position	0.287	0.275	0.278	0.101	0.098	0.099	0.268	0.256	0.259
Word Frequency	0.279	0.271	0.272	0.093	0.090	0.091	0.255	0.248	0.250
Numerical Value	0.265	0.252	0.256	0.086	0.082	0.083	0.244	0.232	0.236
Named Entity	0.264	0.248	0.254	0.084	0.080	0.081	0.242	0.227	0.232
Iterative Query Score	0.299	0.291	0.293	0.111	0.106	0.108	0.276	0.269	0.270
Cue Word	0.279	0.266	0.269	0.092	0.089	0.090	0.259	0.247	0.250

Table 3: Result Using Sentence Scoring Method

Result After Combination:

Rouge	**Recall**	**Precision**	**F-Score**
Rouge-1	0.297	0.285	0.288
Rouge-2	0.109	0.104	0.106
Rouge-L	0.276	0.264	0.268

Table 4:Result After Combination

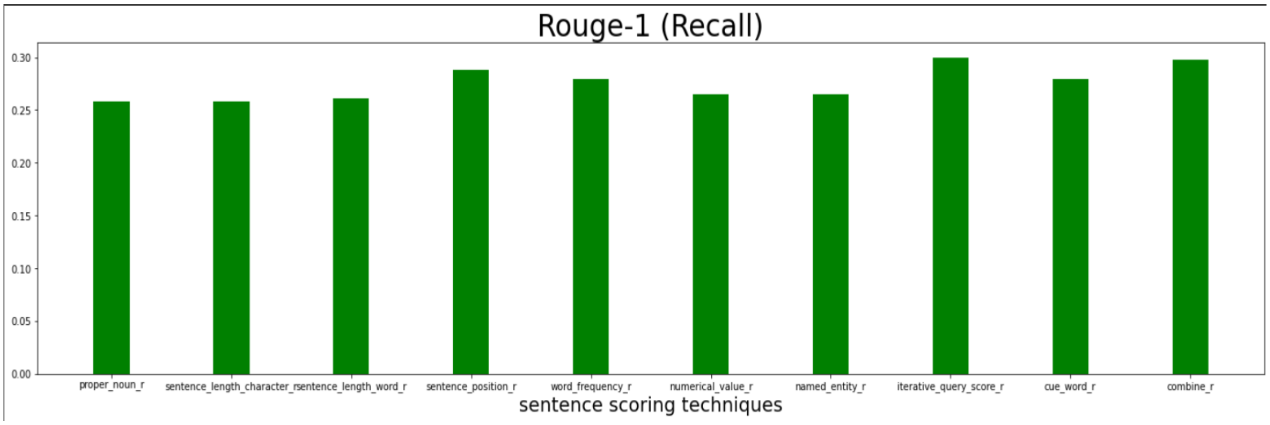
Result After Cuckoo Search Algorithm:

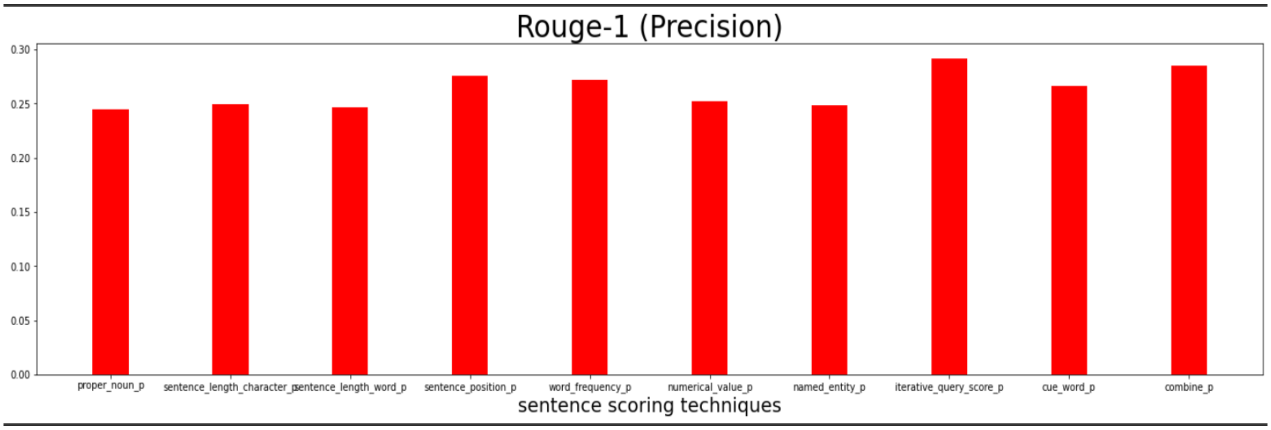
Rouge	**Recall**	**Precision**	**F-Score**
Rouge-1	0.323	0.231	0.265
Rouge-2	0.102	0.069	0.081
Rouge-L	0.2921	0.2083	0.292

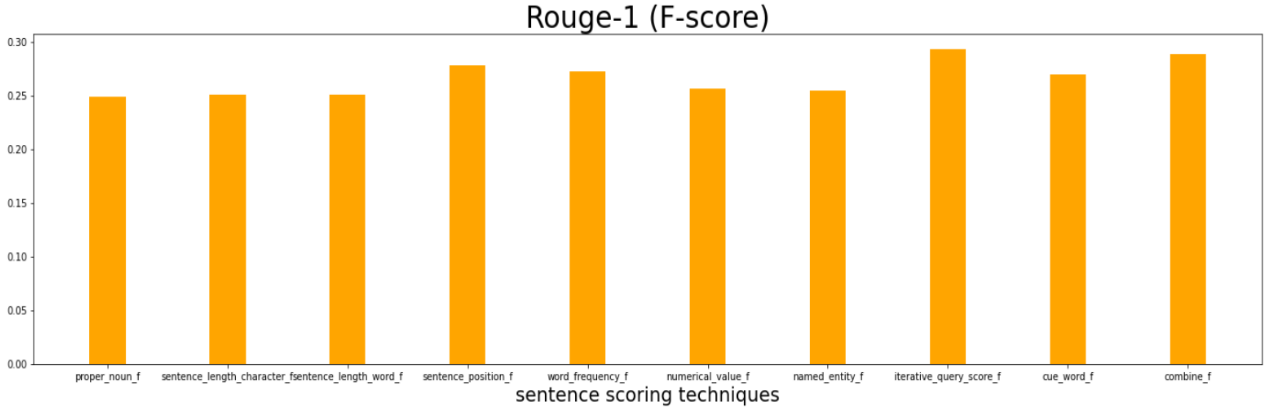
Table 5: Result After Cuckoo Search Algorithm

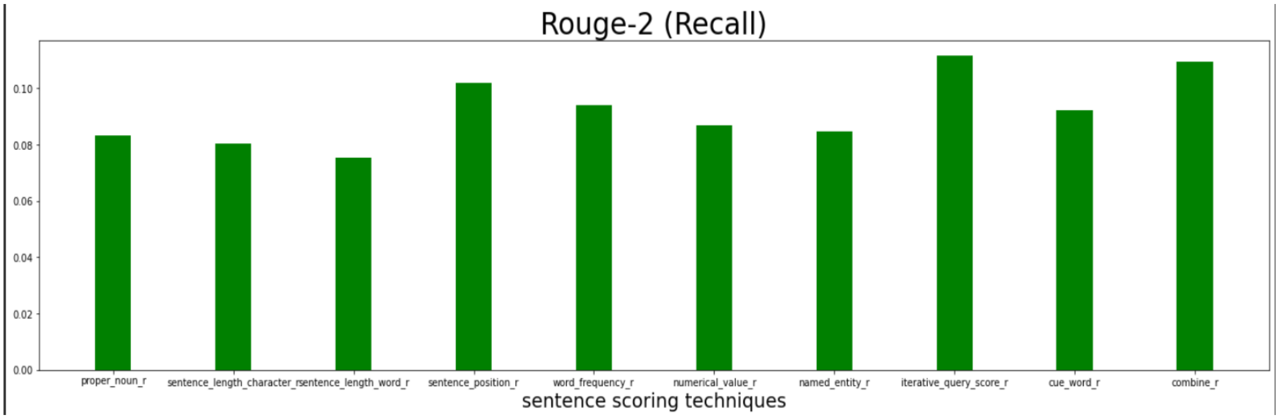
Result of Sentence Scoring Techniques:

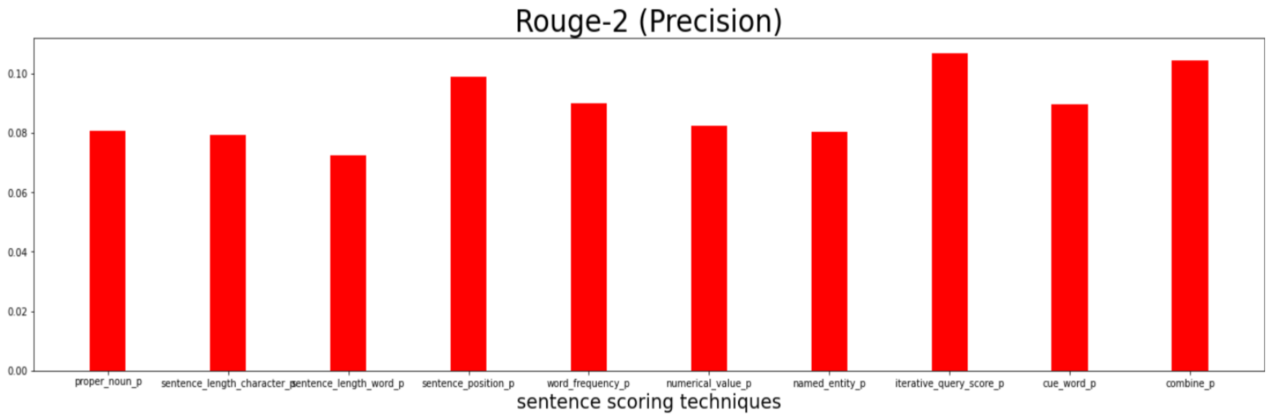
Figure 7 : Graphs

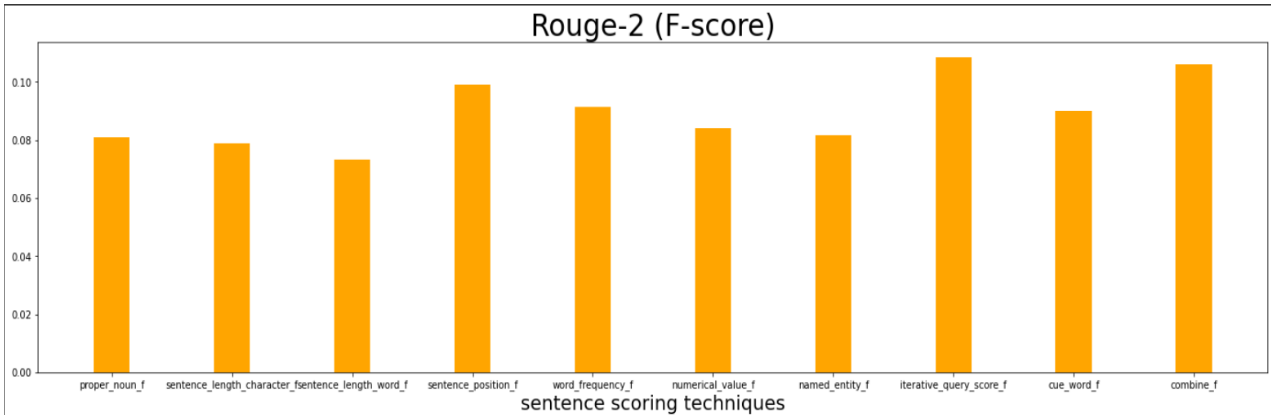


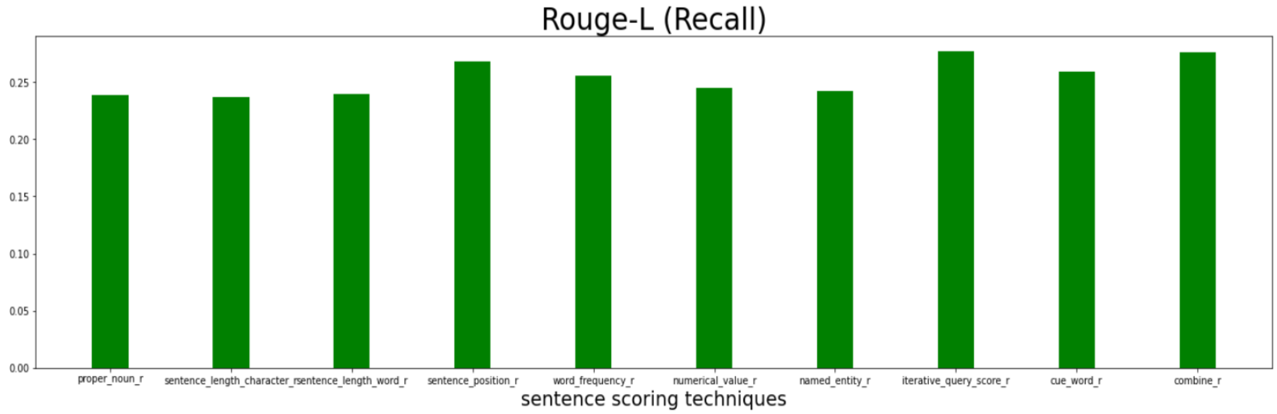

Rouge-1 (Precision)
0.30
0.25
0.20
0.15
0.10
0.05
0.00
proper_noun_p
sentence_length_character_p
sentence_length_word_p
sentence_position_p
word_frequency_p
numerical_value_p
named_entity_p
iterative_query_score_p
cue_word_p
combine_p
sentence scoring techniques

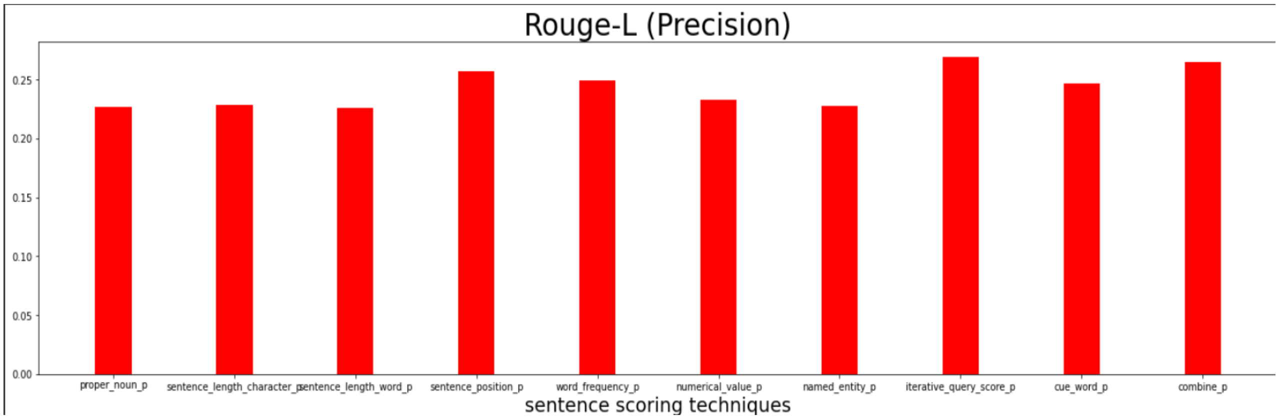

Rouge-1 (F-score)
0.30
0.25
0.20
0.15
0.10
0.05
0.00
proper_noun_f
sentence_length_character_f
sentence_length_word_f
sentence_position_f
word_frequency_f
numerical_value_f
named_entity_f
iterative_query_score_f
cue_word_f
combine_f
sentence scoring techniques

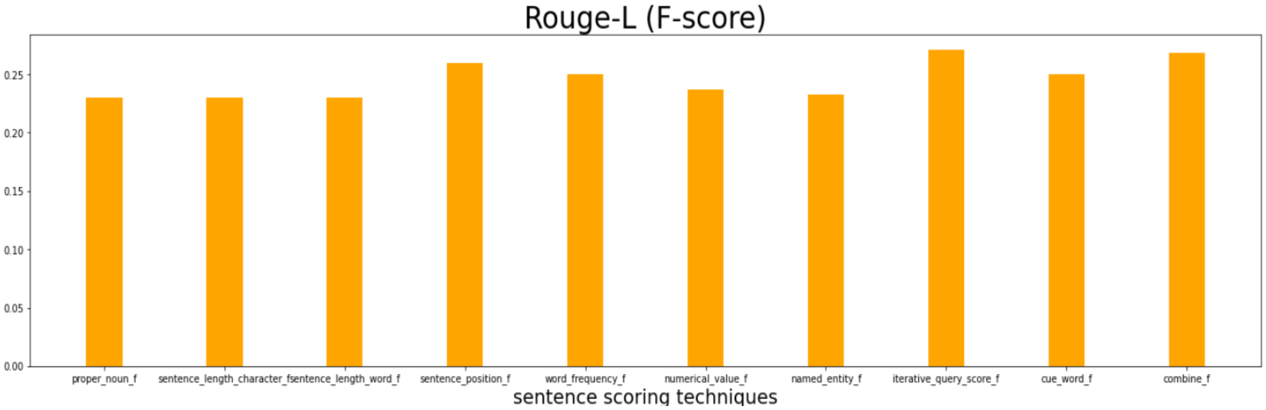

Rouge-2 (Recall)
0.10
0.08
0.06
0.04
0.02
0.00
proper_noun_r
sentence_length_character_r
sentence_length_word_r
sentence_position_r
word_frequency_r
numerical_value_r
named_entity_r
iterative_query_score_r
cue_word_r
combine_r
sentence scoring techniques

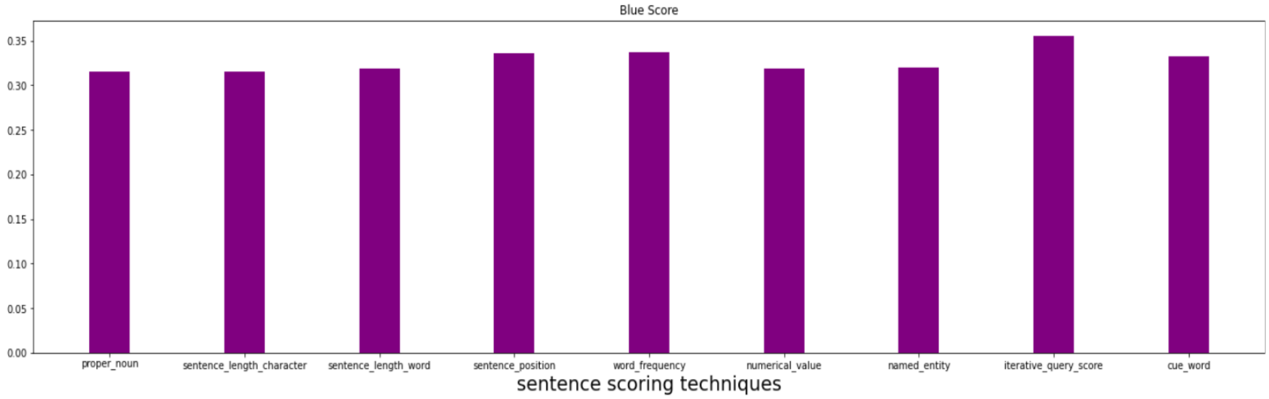

Rouge-2 (Precision)
0.10
0.08
0.06
0.04
0.02
0.00
proper_noun_p
sentence_length_character_p
sentence_length_word_p
sentence_position_p
word_frequency_p
numerical_value_p
named_entity_p
iterative_query_score_p
cue_word_p
combine_p
sentence scoring techniques


Rouge-2 (F-score)
0.10
0.08
0.06
0.04
0.02
0.00
proper_noun_f
sentence_length_character_f
sentence_length_word_f
sentence_position_f
word_frequency_f
numerical_value_f
named_entity_f
iterative_query_score_f
cue_word_f
combine_f
sentence scoring techniques


Rouge-L (Recall)
0.25
0.20
0.15
0.10
0.05
0.00
proper_noun_r
sentence_length_character_r
sentence_length_word_r
sentence_position_r
word_frequency_r
numerical_value_r
named_entity_r
iterative_query_score_r
cue_word_r
combine_r
sentence scoring techniques


Rouge-L (Precision)
0.25
0.20
0.15
0.10
0.05
0.00
proper_noun_p
sentence_length_character_p
sentence_length_word_p
sentence_position_p
word_frequency_p
numerical_value_p
named_entity_p
iterative_query_score_p
cue_word_p
combine_p
sentence scoring techniques


Rouge-L (F-score)
0.25
0.20
0.15
0.10
0.05
0.00
proper_noun_f
sentence_length_character_f
sentence_length_word_f
sentence_position_f
word_frequency_f
numerical_value_f
named_entity_f
iterative_query_score_f
cue_word_f
combine_f
sentence scoring techniques


Blue Score
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
proper_noun
sentence_length_character
sentence_length_word
sentence_position
word_frequency
numerical_value
named_entity
iterative_query_score
cue_word
sentence scoring techniques

8. Conclusion and Future Work:

Hence in this project we have done some of the works in the area of text summarization. Summarization has always been a necessity for many years as there is a huge amount of information being released on the internet every day. This project described all the major summarizations techniques and the prominent works that are being done on each technique. There has always been an improvement to the earlier technique which improved the accuracy like a single document summarization.

This project will be widely used in the future by the all types of users. This project presents algorithm based single document summarizer with the help of sentence scoring method. This project discusses the simple and easy extractive technique of text summarization. This project is done mostly with python and any number of extensions for scoring techniques can easily be added.

Future Work

As for the future work we will be more focusing on improving the sentence score's by optimizing the scoring techniques. We will also try to implement different optimization's algorithms like FireFly, ABC algorithm. Currently we are working on DUC dataset for single document text summarization and now We will now start to work on multi document text summarization.

9. Reference:

- Python Documentation

  https://www.python.org
- https://www.researchgate.com
- https://www.towardsdatascience.com
- https://sci-hub.com
- Assessing sentence scoring techniques for extractive text summarization
  www.elsevier.com/locate/eswa
- Single document text summarization technique using optimal
  https://doi.org/10.1007/s41870-021-00739-2
- Text Summarization using Sentence Scoring Method
  www.irjet.net