

## **TABLE OF CONTENT**

|                   |     |
|-------------------|-----|
| Title page        | i   |
| Certification     | ii  |
| Dedication        | iii |
| Acknowledgement   | iv  |
| Table of contents | v - |
| vi                |     |

### **CHAPTER ONE: INTRODUCTION**

|                                |   |
|--------------------------------|---|
| 1.1. Background of the study   | 1 |
| 1.2. Statement of the problem  | 2 |
| 1.3. Aim of the study          | 2 |
| 1.4. Objectives of the study   | 2 |
| 1.5. Research Methodology      | 2 |
| 1.6. Scope of the study        | 3 |
| 1.7. Significance of the study | 3 |
| 1.8. Definition of terms       | 3 |

### **CHAPTER TWO: LITERATURE REVIEW**

|  |     |
|--|-----|
| 2.1. Introduction                                      | 4   |
| 2.2. What is an article                                | 4   |
| 2.3. What is summarization                             | 4   |
| 2.3.1. Automatic Summarization                         | 5   |
| 2.3.2. Automatic Article /Text Summarization           | 5-7 |
| 2.4. Single Vs Multiple Documentation summarization    | 8   |
| 2.5. Historical Development of Automatic Summarization | 9   |
| 2.6. Concept of Automatic Summarization                | 10  |
| 2.7. Techniques of achieving automatic summarization   | 10  |
| 2.7.1. Abstractive Approach                            | 11  |

|  |       |
|--|-------|
| 2.7.1.1 Structured based Abstractive Summarization methods |       |
| 11-14  |       |
| 2.7.2 Extractive Approach                                  | 15    |
| 2.7.2.1 Intermediate Representation                        | 16    |
| 2.7.2.1.1 Topic Representation Approach                    | 17-21 |
| 2.7.2.1.2 Indicator Representation Approach                |       |
| 22-23  |       |
| 2.7.2.2 Sentence Score                                     | 24    |
| 2.7.2.3 Summary Sentence selection                         | 25    |
| 2.8 Techniques of Extractive Approach                      | 25    |
| 2.9 Impact of context in summarization                     | 26    |
| 2.10 Application of Automatic summarization                | 27    |
| 2.11 Evaluation of Automatic summarization                 |       |
| 28-29  |       |
| 2.12 Review of related works                               |       |
| 30-32  |       |

## **CHAPTER THREE: RESEARCH METHODOLOGY**

|   |    |
|---|----|
| 3.1 Introduction  | 33 |
| 3.2 Overview of existing system                                   | 33 |
| 3.3 Description of the proposed system                            | 33 |
| 3.4 Methodology for article summarization                         | 34 |
| 3.4.1 Extractive summarization using TextRank                     | 34 |
| 3.5 Framework of the system                                       | 35 |
| 3.5.1 High level system framework for previous text summarization | 35 |
| 3.5.2 Proposed framework for article summarization                | 36 |
| 3.6 Sample Interface for Article summarization                    | 37 |
| 3.6.1 First Interface   | 37 |
| 3.6.2 Second Interface  | 38 |
| 3.6.3 Third Interface   | 39 |
| 3.7 System Modelling  | 40 |
| 3.7.1 Use case Diagram  | 41 |

## REFERENCES

42-43

## CHAPTER ONE

### INTRODUCTION

#### 1.1 BACKGROUND OF THE STUDY

As the web continues developing, individuals are getting overpowered by the measure of computerized data and archives. This accessible has requested serious research in the zone of programmed content outline. An outline is a book that is begat from at least one records that depicts significant data in the source report, and that is no longer than half of the source archive Radef,1950.

Programmed rundown is the regular language preparation of AI and information mining, the principle thought of outline is to discover portions of information which contains the data of the whole source archive. Such systems are broadly utilized in the business today, web search tools are a model, and others incorporate synopsis of records, picture assortment and recordings. It has become significant that this examination ought to and must be paid attention to thus the reason for this exploration work.

At a certain point in time everyone has had the undertaking of condensing an article or record so as to make short and brief data from the entire article or archive given, this procedure supposedly is a distressing errand, as you start to worry over which sentences are significant, what key expressions ought not be surrendered or more all, the rundown should convey all the significant data required without composing back the entire article. Here comes the errand of programmed article synopsis framework, this framework is made to dispense with the above expressed issues, programmed outline framework does this by taking a solitary article or groups of article and delivers a succinct and familiar rundown of the most significant data.

Throughout the years synopsis framework has been worked to help individuals in different fields, flawed as they seem to be, they have just appeared to support clients and furthermore offer approach to other programmed outline application. This task targets making a programmed article outline framework which will likewise empower clients transfer records, look into papers and get a summarized version of it. Automatic Summarization receives two

methodology which is the extractive and abstractive methodology, the Extractive synopsis is basically selecting sentences from the content that can best speak to its outline. Extractive outline procedures have been pervasive for a long while now, attributable to its starting point in the 1950s. It's progressively about figuring out how to comprehend the significance of each sentence and their relations with one another instead of attempting to comprehend the substance of the content. While the Abstractive synopsis, then again, is tied in with attempting to comprehend the substance of the content and afterward giving an outline dependent on that, which could conceivably have indistinguishable sentences from present in the first content. Abstractive outline attempts to make its own sentences and is unquestionably a stage towards progressively human-like synopses.

Over the years the summarization system has been built to help people in various fields, imperfect as they are, they have already shown to be very useful to users and also give way to other automatic summarization applications. This project aims at creating an automatic article summarization system which will also enable users to upload documents, research papers and get a summarized version of it.

In the early days, text summarization was done exclusively using rule-based algorithms. It was called “importance evaluator”, this was worked based on ranking different parts of a text according to their importance. Two important knowledge bases were used by the evaluator; one of them being the “importance rule base” which made use of the ‘IF-THEN’ rules and the other being the “encyclopaedia” which contained domain specific world knowledge represented using a network of frames. The importance rule based method makes use of a theory called Hierarchical Propositional Network (HPN), where numerical representations are given to the conceptual units of extended linear representations (ELR) of sentences to constitute the importance of it. Afterwards, a referential structure is then applied by goal interpreter to interpret the importance of each sentence.

In 1984, a method called “Production rule system for summarization” was used. It primarily works on three steps;

- Inferencing
- Scoring the format rows for their importance
- Selecting the appropriate ones as summary.

As researchers in the field of summarization technique progressed, a lot of developments were made to interpret the importance of sentences in a textual data corpus.

One of such methods is to calculate the relatedness of a piece of text to other texts in the corpus and then decide on the importance by the degree/quantity by which this text is related to other texts.

## **1.2 STATEMENT OF THE PROBLEM**

- Due to the large amount of data available digitally, it has become difficult to sort through this data to get the particular information you need.
- Summarizing articles and documents manually is seen to be a very difficult task as you have to stress over picking out the most important information.

## **1.3 AIM OF THE STUDY**

The aim of this project is to create an automatic summarization system that allows users to request from a pool of data and summarize it and also have the ability to upload documents and research papers and get a summarized version of it. It is also aimed at allowing users to be able to discover, consume and digest relevant information faster.

## **1.4 OBJECTIVES OF THE STUDY**

- To review the operations of existing systems like Extract, Resoonmer and ResearchGate.
- To create and develop an automatic article summarization system using Html, css, javascript, and python.
- To give users the ability to upload data and extract the summarized version of the uploaded data.
- To validate and evaluate performance of the system research.

## **1.5 RESEARCH METHODOLOGY**

To develop an efficient automatic article summarization system, various methodologies used is as follows:

- Literature review : Various manuals, journal papers, documents and reports on past works are being reviewed in the course of research. This helps to get an insight on the research work, limitations and lapses identified so as to provide a better system.
- Internet search: Further research is being made on the internet through the use of search engines like Google Scholar, Google, Wikipedia, medium articles.

- Study of existing system: Close observation and study of how the existing system works, the interfaces used is done to get additional knowledge on how to develop an improved system.

## **1.6. SCOPE OF THE STUDY**

This project focuses on developing an automatic article summarization system using the TextRank algorithm. It will be limited to articles, documents, research and journal papers only.

## **1.7 SIGNIFICANCE OF THE STUDY**

The significance of this project is to help users extract summarized data from a large pool of data without having to read through very large amount of articles. It allows users to read less data but still receive the most relevant information to make solid conclusion.

## **1.8 DEFINITION OF TERMS**

Automatic summarization: The process of shortening a text document with software in order to get the major points of the original documents.

Data mining: Practice of examining large pre-existing databases in order to generate new information.

Algorithm: A process to be followed in order to solve a problem.

TextRank Algorithm: An automatic summarization technique that ranks text chunks in order of their importance in the text document.

Interface: The shared boundary between users and the computer.

Extractive method: Involves the selection of phrases and sentences from the source document to make up the new summary.

Cluster: A group of similar things/occurring closely together.

Imperative: of vital importance, crucial.

## **CHAPTER TWO**

### **LITERATURE REVIEW**

#### **2.1 INTRODUCTION**

In this chapter insight will be given into various studies conducted by researchers, as well as explained terminologies in regards to automatic article summarization systems.

This chapter will also give insight to the application areas and usefulness, history and present state of automatic article summarization.

#### **2.2 WHAT IS AN ARTICLE?**

Article is a piece of writing included with others in a newspaper, magazine or other publications. An article can be an essay, reports, accounts, column and composition e.t.c.

#### **2.3 WHAT IS SUMMARIZATION?**

Summarization can be described as the demonstration of communicating the most significant reality or thought regarding a person or thing in a brief and familiar structure. In basic words outline is to summarize the most significant piece of something, it recognizes the most significant thought in a book, and how to disregard the superfluous parts and how to combine the focal thought in a pertinent manner. Rundown is a significant part of article assessment, it is likewise imperative to understudies to figure out how to outline in light of the fact that an understudy will experience a circumstance where the person needs to condense a given exposition, article, examine paper and creation. Synopsis causes understudy to:

- It helps decide key ideas and combine those important ideas that support them.
- It empowers understudies to concentrate on catchphrases and expressions of a doled out content that are important and recalling.
- It shows understudy how to take huge determination of content and lessen it to the primary concern for increasingly concise comprehension

### **2.3.1 AUTOMATIC SUMMARIZATION**

Automatic Summarization is the way toward shortening a book record, recordings and pictures with programming, so as to make a rundown with the purposes of the first archive.

This procedure is broadly utilized in businesses today, web search tools and news altering organizations are a model. The restorative field has broadly embraced the utilization of programmed outline, on a few events programmed rundown framework has being worked to help information extraction from full content in precise audit advancement. "Deliberate surveys are significant data hotspots for medicinal services suppliers, research and arrangement creator. A SR endeavors to completely distinguish, assess and orchestrate the best accessible proof to discover solid responses to look into questions". As information extraction is one of the means in SR improvement whose objective is to gather important data from distributed reports to perform quality evaluation, yet examines have indicated that manual information extraction has a high danger of blunders PC techniques have been proposed throughout the years to upgrade the profitability and decrease mistake in SR information with the utilization of automatic summarization. ([https://en.wikipedia.org/wiki/Automatic\\_summarization](https://en.wikipedia.org/wiki/Automatic_summarization))

Automatic Summarization is a piece of information mining and AI. Programmed outline can be applied in various structures, from abridging archives/articles to condensing pictures and condensing recordings. Note that the universe of automatic summarization is a quickly developing and required field as it enables chopping down so a lot of exertion is spent experiencing huge articles to get the significant ones. Programmed rundown for the most part has two ways to deal with it "the extractive and abstractive methodology", the extractive methodology will be utilized with the end goal of this research work.

### **2.3.2 AUTOMATIC ARTICLE/TEXT SUMARIZATION**

Automatic content/article summarization is the information science and natural language processing issue of making a short, succinct and familiar from an enormous report. Rundown strategies are extraordinarily expected to expand the regularly developing measure of content accessible on the web, generally synopsis lets us devour data quicker. Filtering through heaps



of records can be tedious and troublesome, without a dynamic or outline it can take minutes before you make sense of what a specific paper or report is discussing, simply envision experiencing several archives.

Representatives, examiners, legitimate assistants, understudies and specialists need to sift through gigantic quantities of reports each day to move forward, and an enormous piece of their time is spent simply making sense of what record is applicable and what isn't. By expelling significant sentences and making comprehensive outlines, it's conceivable to rapidly check whether a record merits perusing.

Programmed content outline is likewise helpful for understudies and creators. Envision having the option to consequently produce a dynamic based for your examination paper or part in a book in a reasonable and brief way that is devoted to the first source material.

Programmed rundown of content works by first examining the word thickness for the whole content record. At that point, the most widely recognized words are put away and arranged. Each sentence is then scored dependent on what number of high-thickness words it contains, the higher the word thickness the higher its value. At long last, the top N sentences are then taken, and arranged depending on their situation in the first content.

Programmed content outline makes things straightforward and broadly useful, the programmed content synopsis calculation can work in various circumstances that different usage may battle with, for example, reports containing unknown dialects or one of a kind word affiliations that aren't found in Standard English language corpuses. Collin (2015).

In content synopsis, perhaps the best issue is to extricate significant data from given basic sources including website pages, any archive, and database. A decent rundown must be delivered by content synopsis systems utilizing less time and less repetition. Outlines produced by the Rule based strategy are of high data thickness however it is extremely dull work since every one of the guidelines and examples are composed physically. In the strategy for Ontology, treatment of questionable information is conceivable which is preposterous in basic space metaphysics. Issue with this strategy is that solitary area specialists can characterize the cosmology of the space which is tedious. In the Tree based strategy, the nature of outline gets improved in light of the utilization of language generators. The Only issue with this technique is that the fundamental setting of the sentences gets dismissed while catching the crossing point of expressions. The Multimodal semantic model technique

produces a theoretical rundown in which it incorporates printed information just as graphical information and thus, gives great outcome. Issue with this strategy is that assessment is to be done physically. In the Information thing-based technique, the determination of valuable data is finished. Based on the chosen data thing, the sentences and outlines are created. This methodology gives a little, cognizant and data rich outline.

Issue with this technique is that occasionally helpful data things get dismissed while the development of important and syntactically right sentences diminishes the phonetic nature of rundown. The Semantic diagram strategy, Sentences framed are less repetitive just as linguistically right. In any case, this strategy is restricted to just a single record. In spite of the fact that the system of programmed synopsis is an old test, the specialists these days are getting more slanted towards abstractive rundown methods as opposed to extractive outline strategies. This is on the grounds that, abstractive rundown strategies produce progressively intelligible, less repetitive and data rich summary. Creating dynamic utilizing abstractive synopsis strategies is a troublesome assignment since it requires increasingly semantic and etymological investigation. Because of the above reasons the investigation of abstractive outline procedures demonstrates to be increasingly helpful.

Automatic text/article summarization tools are needed because:

- Summaries decrease understanding time.
- When searching for records, outlines make the choice procedure simpler.
- Automatic outline calculations are less one-sided than human summarizers.
- Personalized outlines are valuable being referred to by noting frameworks as they give customized data.
- Using programmed or self-loader outline frameworks empowers business theoretical administrations to build the quantity of writings they can process.

These challenges can be achieved because automatic text/article summarization:

- Determines which sentences are most important or salient.
- It makes the synopsis firm and coherent.
- It limits the quantity of references to thought and substances not referenced in the synopsis.

Automatic text/article summarization can be seen as a single or multiple document challenge.

## **2.4. SINGLE VS. MULTIPLE DOCUMENT SUMMARIZATION**

Single archive outline manages abridging single record instead of various report synopsis which targets removing rundown from different content or articles expounded on a similar point. The subsequent synopsis report enables singular clients to rapidly acquaint themselves with data contained in an enormous group of records. In a manner various report synopsis framework assists clients with adapting to data over-burden. The significant advantage of different archive rundowns is that it makes data reports that are both brief and comprehensive with unmistakable suppositions being assembled and characterized, each theme is depicted from numerous points of view inside a solitary record. While the objective of a rundown is to make data search basic and cut the time by choosing the most significant source report, exhaustive various archive outlines should itself contain the required data, subsequently decreasing the requirement for approaching the first records to situations when refinement is required. Programmed outlines present data removed from various sources algorithmically with no article contact subsequently making it impartial. The numerous record outline is more mind boggling than abridging a single archive. The trouble originated from related randomness inside an enormous arrangement of records, a great outline innovation intended to choose and consolidate the primary topic with culmination, intelligibility and quickness. The ideal numerous record synopsis framework makes the first content short as well as presents data sorted out around the key angle to speak to the assorted view.

A multiple document summary is successful if it has the following quality:

- Good readability
- Clear structure, including a framework of the primary substance, from which it is anything but difficult to explore the full content segments.
- Gradual changes from increasingly broad to progressive explicit related angles.
- Text inside segments is partitioned into important sections.

## **2.5 HISTORICAL DEVELOPMENT OF AUTOMATIC SUMMARIZATION**

Luhn in the 1950s had the primary work on automatic summarization, he utilized the techniques sentence extraction, his methodology was executed to chip away at specialized papers and magazine articles. This basic thought that Luhn set forward later formed the future

research of programmed rundown, to be specific that a few words in an archive are graphic and the sentences that draws out the most significant data in the record are the ones that contain numerous such engaging words near one another. After the arrival of Luhn take a shot at programmed synopsis, this examination increased more fascination which was significantly utilized for condensing logical research, Luhn proceeded to acquaint a technique with remove remarkable sentences from the content utilizing highlights, for example, word and expression recurrence, he proposed to gauge the sentences of a report as an element of high thickness normal words, and further went to depict an example dependent on key expressions notwithstanding standard recurrence depending loads, which utilized the accompanying strategies to decide sentence weight;

**Cue Method:** in this technique the pertinence of a sentence is determined dependent on the existences of a specific sign word in the prompt lexicon

**Title Method:** The heaviness of a sentence is figured as an entirety of all substance words showing up in the title and heading of a book.

**Location Method:** This technique accepts that sentences showing up in the Start of individual passages has a higher likelihood of being significant.

Over the years many works have been published to address the problems of automatic summarization with newer methods being introduced with the rise of advanced technology like Machine learning, Natural language processes and data mining.

## **2.6 CONCEPT OF AUTOMATIC SUMMARIZATION**

An automatic summarization system takes many documents as input and attempts to produce a brief and inclusive summary of the most important information in the input. Writing a concise and inclusive summary requires the ability to identify, alter, and join information expressed in diverse sentences in the input. The automatic summarization system is built with all these in mind.

## **2.7 TECHNIQUES TO ACHIEVING AUTOMATIC SUMMARIZATION**

There are two general approaches to achieve an automatic summarization system, which are the Abstractive and Extractive approach. The two approaches are explained in brief below:

- **Abstractive Approach:** The abstractive approach will generate new phrases or use new words that were not in the original text to generate the summary. Naturally abstractive approaches are harder, as the model has to first understand the documents and then try to express that understanding using new words and phrases to create a perfect summary.

The abstractive approach is barely used as it deals with semantic problems and produces less effective summary than the extractive approach. This method uses advanced natural language processing and deep learning which is a growing field in itself.

- **Extractive Approach:** In this approach the automatic system takes out sentences from the entire collection, without altering the sentences themselves.

This approach will check out the important sentences and use these sentences to form the summary. There are different techniques and algorithms that are used to check out the weight of the sentences and then rank those sentences depending on how similar and important they are. Extractive outlines frequently give preferable outcome over abstractive rundowns, this is on the grounds that abstractive synopsis techniques adapt to issues like semantic portrayal, surmising and characteristic language age which is generally harder than information driven methodologies like sentence extraction.

There are numerous systems accessible to produce extractive synopsis, with the end goal of this exploration work we will utilize a solo learning approach of the extractive methodology.

### **2.7.1 ABSTRACTIVE APPROACH**

The abstractive approach of automatic summarization takes the document given and uses entirely new words to form the summary needed, an abstractive approach can be seen as a human-like approach to summarizing. Abstractive summarization understands the main concept and relevant information in a document and provides a short and clear format of the summary, it doesn't necessarily use the same words in the document it is smart enough to coin its own word after understanding the document. This methodology can be arranged into two classifications. Which are in particular organized based and semantic based techniques. Organized based methodologies decide the most significant data through archives by utilizing layouts, extraction rules and different structures, for example, trees, metaphysics and so forth.

### **2.7.1.1        STRUCTURED        BASED        ABSTRACTIVE        SUMMARIZATION METHODS**

#### **• Rule Based Method:**

The rule based method comprises of three steps:

Right off the bat, the reports to be gathered are spoken to as far as their classifications, the classes can be from different areas. Subsequently the main errand is to sort these. The subsequent advance is to pose inquiries dependent on these classes E.g among the different classifications like crowds, fiascos, wellbeing and so forth taking the case of a horde class various inquiries can be postponed out like:- what was the deal? When did it occur? Who got influenced? What were the outcomes? And so on

At the point when these inquiries have been posed, rules will be produced. A few action words and things which have comparable implications will be utilized and their positions will be accurately recognized. The context selection module selects the best candidate amongst these.

Generation patterns will then be used for the creation of summary sentences.

#### **Ontology Method**

In this strategy area metaphysics for news occasions is characterized by the space specialists. Next stage is archive preparing stage, important terms from corpus are created in this stage, and the significant terms are characterized by classifier on premise of occasions of news. Participation degree related with different occasions of area metaphysics enrollment degree is produced by fluffy surmising. The confinement of philosophy strategy is that space cosmology must be characterized by area specialists and this makes it tedious.

#### **Tree Based Method**

In the tree-based strategy the pre-preparing of the same sentences is finished utilizing shallow parser, after we append these the same sentences to the predicate-contention structure, distinctive calculation can be utilized for choosing the basic expression from the sentences, for example, topic calculation. The expression depicting a similar significance is chosen, extra data will be included and it will be organized in a legitimate request. Toward the end

FUF/SURGE language generator will be utilized for making new synopses by joining and sorting out the chosen normal expression.

Utilizing language generators will build the expert articulation of the language and diminish syntactic blunders. This trait is the fundamental nature of this technique. The principal issue with this technique is that the states of the sentences doesn't get included while determining basic expression and it is a significant piece of the sentences regardless of whether it isn't a piece of the basic expression.

### **2.7.1.2 SEMANTIC BASED ABSTRACTIVE SUMMARIZATION**

Multimodal semantic model depicts the idea and shape the connection among these thoughts. These chose ideas are communicated as sentences. This model acknowledges content archive just as picture record.

Multi modal consist of three phases:-

- **Semantic Modal**

Ideas are only words which speaks to significant data. Thoughts are developed utilizing information portrayal dependent on objects. Hubs speak to ideas and connections between these thoughts speak to connection between them. Ibrahim .F (2010). Utilizing this semantic model are built as appeared in Figure 2.1

- **Rated Concepts**

Thoughts are assessed using information thickness (ID) structure. This structure is used to survey importance of thoughts. Parts for choosing the noteworthiness – Completeness of characteristics it is just the extent of filled properties in the semantic model to without a doubt the quantity of qualities in semantic. This gives sentences which contain more information. One thought is related with various thoughts and these relations are counted. Observing this count makes us know how critical this thought is.

- **Sentence Generation**

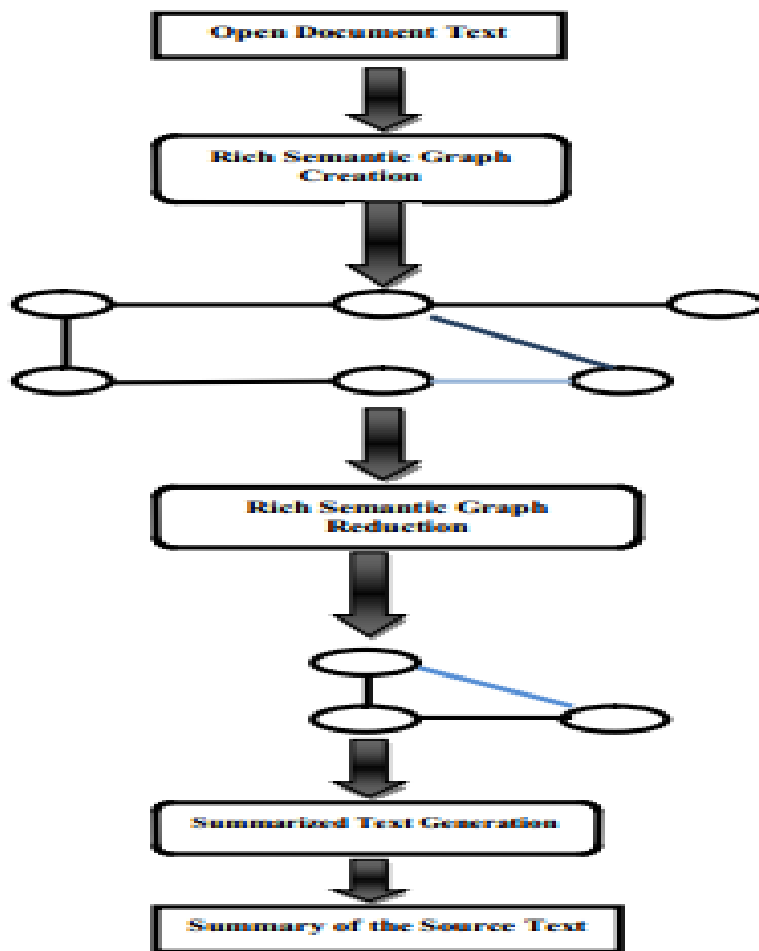
When the ideas are evaluated utilizing ID grid the subsequent stage is to produce sentences utilizing parsing methods.

## **INFORMATION ITEM BASED METHOD**

In this strategy, rather than producing theoretical sentences from the info document, it is created from a unique portrayal of the information record. The dynamic portrayal is only a data thing which is the littlest component of data in a book. The system utilized in his technique was proposed with regards to the Text Analysis Conference (TAC) 2010 for multi-record synopsis of news. The modules of this system are: Information thing recovery, sentence age, sentence choice and synopsis age. In the Information Item (INIT) recovery stage, subject-action word object significantly increases are shaped by grammatical investigation of content finished with the assistance of parser. During the linguistic examination, action word's subject and item are separated. In sentence age stage, the sentences are created utilizing a language generator. In the following stage for example sentence determination stage, ranking of each sentence is done based on the normal archive recurrence (DF) score. Finally, in the rundown age stage, exceptionally positioned sentences are organized and conceptual is created with appropriate arranging. From this strategy, a short, intelligent, data rich and less repetitive outline can be framed. Regardless of such a large number of favorable circumstances, this strategy has additionally numerous constraints. While making syntactic and significant sentences, numerous significant data things get dismissed. Because of which, the etymological nature of resultant synopsis gets diminished.



## Semantic Graph Based Method



Source: (vijay Mathur 2009)

**Figure 2.1: Semantic Graph Reduction**

Figure 2.1 shows the method of generating a summary by creating a semantic graph called rich semantic graph (RSG).

The semantic graph approach consists of three phases:-

- The first stage speaks to include an archive utilizing rich semantic charts (RSG). In RSG, the action words and things of the info report are spoken to as diagram hubs and the edges compare to semantic and topological relations between them.

- The second stage diminishes the first chart to an increasingly decreased diagram utilizing heuristic guidelines.
- The third stage creates an abstractive summary. The upside of this strategy is that it creates less repetitive and syntactically right sentences. The inconvenience of this technique is that it is restricted to a solitary archive and not various records.

### **2.7.2 EXTRACTIVE APPROACH**

The extractive methodology of building a programmed synopsis framework accepts a few sentences as they show up in the materials and link them to create an outline. Early works in the synopsis framework managed a single report rundown where the framework delivered an outline of that archive.

The extractive outline is of two kinds relying upon what your synopsis motor spotlights on right off the bat the conventional rundown whose significant point is to separate unique from the whole set and besides the question pertinent which will abridge record explicit to an inquiry. Contingent upon what the client may require, the synopsis framework ought to make nonexclusive and inquiry significant rundowns.

An example of outline issue will be synopsis of a report which endeavors to make a short rundown from a given archive, once in a while a client perhaps keen on producing a rundown from a solitary source record or various source record. At an elevated level synopsis calculation attempts to discover subsets of sentences which spread data of the whole set, these calculation models thoughts like decent variety, inclusion, data and representativeness of the outline. A few procedures or calculations which normally model rundown issues are TextRank, Submodular set capacity, Cosine similitude, Matrix likeness, Marginal significance and so on.

The extractive methodology can take a managed or unaided learning approach towards its rundown procedure. The administered learning approach endeavors to dissect and prepare the model to deliver a deduced capacity which can be utilized for mapping the sentences, it produces interpretable standards for what highlights portray a key expression however they will in general require huge measure of preparing information and reports with realized key

expressions are expected to execute this methodology. While the unaided learning approach expels the requirement for preparing information, it moves toward the issue from an alternate point as opposed to attempting to learn express highlights that portray key expressions, it misuses the structure of the content itself. With the end goal of this examination work we will embrace the solo learning methodology. So as to see how this synopsis framework functions, we will portray the three autonomous undertakings that a summarizer experiences.

### **2.7.2.1. INTERMEDIATE REPRESENTATION**

Each synopsis framework makes some transitional outline of the printed substance it plans to condense and finds remarkable substance situated in this portrayal. There are two styles of procedures dependent on the portrayal: topic representation and marker outline. Theme outline forms redesign the printed substance into a transitional portrayal and decipher the subject matter(s) talked about inside the content. Point portrayal based rundown procedures differ regarding their multifaceted nature and portrayal model, and are partitioned into recurrence driven systems, topic express strategies, and inactive semantic examination and Bayesian topic models. We complex topic portrayal strategies in the accompanying segments. Marker delineation systems depict each sentence as a rundown of highlights (pointers) of hugeness together with sentence length, position inside the archive, having positive expressions, and so on.

#### **2.7.2.1.1 TOPIC REPRESENTATION APPROACHES**

In this section, I would describe some of the most widely used topic representation approaches.

- **TOPIC WORDS:**

The theme phrases system is totally one of the customary point depiction methodology which wants to see words that portray the subject of the data record. It became one the most timely

works that used this strategy by the utilization of repeat points of confinement to locate the connection with articulations in the archive and address the subject of the record. An extra predominant model of Luhn's thought was shown in. In which they used a log-probability extent check to discover sensible words which in outline composing are suggested as the "topic signature". Utilizing subject mark communicates as point depiction changed into effective and copied the exactness of different archive rundowns inside the news zone. There are two unique approaches to process the centrality of a sentence: as a characteristic of the wide grouping of point marks it contains, or as the degree of the topic checks inside the sentence. Both sentence scoring features relate to the proportional subject depiction, nevertheless, they may consign different scores to sentences. The first system may consign better scores to longer sentences, in view of the truth they have extra words. The ensuing framework measures the thickness of the subject expressions.

## **FREQUENCY-DRIVEN APPROACHES**

When doling out heaps of words in point depictions, we can think about twofold (0 or 1) or authentic worth (tireless) stacks and pick which words are progressively related to the subject. The two most fundamental frameworks in this characterization are: word probability and TFIDF (Term Frequency Inverse Document Frequency).

### **Word Probability:**

The least troublesome method to use repetition of words as markers of centrality is word probability. The probability of a word  $w$  is settled as the amount of occasions of the word,  $f(w)$ , isolated by the amount of all words in the data (which can be a single document or different reports):

$$P(w) = \frac{f(w)}{n}$$

Vanderwende(1980). proposed the SumBasic system which uses only the word probability approach to manage choice sentence importance. For each sentence,  $S_j$ , in the data, it does out a weight proportional to the typical probability of the words in the sentence:

$$g(S_j) = \frac{\sum_{w \in S_j} P(w)}{|\{w \mid w \in S_j\}|}$$

Where  $g(s_i)$  is the weight of the sentence  $s_i$ . In the subsequent stage, it picks the best scoring sentence that contains the most noteworthy likelihood word. This progression guarantees that the most noteworthy likelihood word, which speaks to the theme of the report by then, is remembered for the synopsis. At that point for each word in the picked sentence, the weight is refreshed:

$$P_{new}(w_i) = P_{old}(w_i) P_{old}(w_i)$$

This word weight update shows that the probability of a word appearing in the once-over is lower than a word happening once. The recently referenced decision advances will go over until the perfect length plot is come to. The sentence decision approach used by SumBasic relies upon the greedy technique. Yih et al (1950). used a streamlining approach (as sentence assurance framework) to grow the occasion of the huge words comprehensive over the entire summary. is another instance of using an upgrade approach.

#### **TFIDF (Term Frequency Inverse Document Frequency):**

Since word probability methodologies depend upon a stop word list in order to not consider them in the diagram and in light of the fact that picking which words to put in the stop list isn't particularly straightforward progress, there is a prerequisite for additional created methodology. One of the further created and common methodologies to offer burden to words is TFIDF (Term Frequency Inverse Document Frequency). This weighting framework assesses the criticalness of words and identifies incredibly normal words (that should be barred from thought) in the document(s) by giving low loads to words appearing in numerous chronicles. The weight of each word  $w$  in document  $d$  is enlisted as seeks after:

$$q(w) = fd(w) * \log \frac{|d|}{fD(w)}$$

where  $fd(w)$  is term repeat of word  $w$  in the report  $d$ ,  $fD(w)$  is the amount of chronicles that contain word  $w$  and  $|D|$  is the amount of records in the collection  $D$ . For more information about TFIDF and other term weighting plans, TFIDF loads rush to enroll and moreover are extraordinary measures for choosing the hugeness of sentences, consequently many existing summarizers have utilized this technique (or some kind of it). Centroid-based layout, another plan of systems which has become an ordinary benchmark, relies upon TFIDF subject

depiction. This kind of method positions sentences by enrolling their prominence using a ton of features. We will briefly design the basic idea. The first step is to distinguish proof and documents that depict a comparable subject assembled. To achieve this target, TFIDF vector depictions of the documents are made and those words whose TFIDF scores are underneath an utmost are emptied. By then, a gathering estimation is run over the TFIDF vectors, consecutively adding records to bundles and preparing the centroids as demonstrated by:

Where is the centroid of the  $j$ th gathering and is the course of action of reports that have a spot with that bundle. Centroids can be considered as pseudo-records that include those words whose TFIDF scores are higher than the edge and structure the gathering. The resulting advance is using centroids to perceive sentences in each bundle that are essential to the subject of the entire gathering. To accomplish this target, two estimations are defined: bunch based relative utility (CBRU) and cross-sentence enlightening subsumption(CSIS).CBRU picks how relevant a particular sentence is to the general purpose of the entire pack and CSIS measures reiteration among sentences. To assess two estimations, three features (for instance central worth, positional worth and first- sentence spread) are used. Next, the final score of each sentence is prepared and the selection of sentences is settled.

## **LATENT SEMANTIC ANALYSIS**

Inert Semantic Analysis (LSA) , which is exhibited, is an independent system for evacuating a depiction of substance semantics subject to watched words. Gong and Liu from the start proposed a technique using LSA to pick incredibly situated sentences for single and multi-report plots in the news space. The LSA system first creates a term- sentence grid ( $n$  by  $m$  arrange), where each line identifies with a word from the information ( $n$  words) and each portion looks at a sentence ( $m$  sentences). Each section  $a_{ij}$  of the system is the greatness of the word  $I$  in sentence  $j$ . The heaps of the words are prepared by the TFIDF framework and if a sentence doesn't have a word the weight of that word in the sentence is zero. By then specific worth crumbling (SVD) is used on the structure and changes the cross section  $A$  into three frameworks:  $A = U \Sigma V^T$ .

System  $U$  ( $n \times m$ ) addresses a term-subject matrix having heaps of words. Matrix  $\Sigma$  is a corner to corner grid ( $m \times m$ ) where every segment  $I$  identify with the largeness of a point  $I$ . Lattice

VT is the subject sentence system. The network  $D = \Sigma VT$  delineates how much a sentence address a topic, as such,  $d_{ij}$  shows the greatness of the point I in sentence j.

Gong and Liu's system was to pick one sentence for each topic, thus, considering the length of summary to the extent sentences, they held the amount of focus. This system has a drawback due to the manner in which a subject may require more than one sentence to pass on its information. In this manner, elective courses of action were proposed to improve the presentation of LSA-based systems for layout. One redesign was to utilize the weight of each subject to pick the general size of the overview that should cover the topic, which gives the flexibility of having a variable number of sentences. Steinberger et al. displayed a LSA-based strategy which achieves a significantly favored introduction over the main work. They comprehended that the sentences that look at some of noteworthy subjects are incredible contender for once-overs, thusly, to discover those sentences they defined the largeness of the sentence as seeks after:

$$g(s_j) = \sqrt{\sum_{i=1}^m d_{ij}^2}$$

## • BAYESIAN TOPIC MODELS

Many of the existing multiple-document summarization methods have two limitations

- ☐ They consider the sentences as autonomous of one another, so points inserted in the reports are ignored.
- ☐ Sentence scores registered by most existing methodologies regularly don't have exceptionally clear probabilistic elucidations, and a significant number of the sentence scores are determined utilizing heuristics.

Bayesian theme models are probabilistic models that reveal and speak to the points of archives. They are very incredible and engaging, in light of the fact that they speak to the data (for example points) that is lost in different methodologies. Their bit of leeway in portraying and speaking to points in detail empowers the improvement of summarizer frameworks which can decide the likenesses and differences between records to be utilized in synopsis. Aside from the upgrade of theme and record portrayal, point models frequently use an unmistakable measure for scoring the sentence called Kullbak-Liebler (KL). The KL is a

proportion of difference (disparity) between two likelihood disseminations P and Q. In outline where we have likelihood of words, the KL dissimilarity of Q from P over the words w is defined as:

$$DKL(P||Q) = \sum_w P(w) \log \frac{P(w)}{Q(w)}$$

where  $P(w)$  and  $Q(w)$  are probabilities of w in P and Q. KL uniqueness is an intriguing strategy for scoring sentences with regards to the rundown, since it shows the way that great synopses are instinctively like the information records. It depicts how the significance of words adjusts in the synopsis in correlation with the info, for example the KL disparity of a decent rundown and the information will be low. Probabilistic point models have increased emotional consideration as of late in different areas. Inactive Dirichlet assignment (LDA) model is the cutting edge unaided procedure for removing topical data (subjects) of an assortment of reports. However, the primary thought is that archives are spoken to as an arbitrary blend of inactive themes, where every subject is a likelihood dissemination over words. LDA has been widely utilized for multi-archive synopsis as of late. For instance, Daume et al (1960). Proposed BayeSum, a Bayesian synopsis model for inquiry centered rundown. Wang et al(1950's). presented a Bayesian sentence-based theme model for synopsis which utilized both term-archive and term-sentence affiliations. Their framework accomplished significant execution and outflanked numerous other synopsis techniques. Celikyilmaz(1960). portray multi-report outline as an expectation issue dependent on a two-stage half and half model. To start with, they propose a various leveled subject model to find the theme structures all things considered. At that point, they figure the similarities of up- and-comer sentences with human-if synopses utilizing a novel tree-based sentence scoring capacity. In the second step they utilize these scores and train a relapse model agreeing the lexical and auxiliary attributes of the sentences, and utilize the model to score sentences of new reports (concealed records) to frame a rundown.

#### **2.7.2.1.2 INDICATOR REPRESENTATION APPROACHES**

Pointer portrayal approaches intend to display the portrayal of the content dependent on a lot of highlights and use them to legitimately rank the sentences as opposed to speaking to the



subjects of the information content. Chart based strategies and AI systems are regularly utilized to decide the significant sentences to be remembered for the outline.

- **GRAPH METHODS FOR SUMMARIZATION**

Chart strategies, that are influenced through PageRank calculation, comprise the records as a related diagram. Sentences shape the vertices of the chart and edges between the sentences propose how comparable the two sentences are. A typical strategy utilized to associate vertices is to degree the similitude of two sentences and if it's miles extra than a limit they might be associated. The most generally utilized strategy for likeness measure is cosine similitude with TFIDF loads for words. This chart portrayal results in two results. Initially, the segments (sub- diagrams) remembered for the chart, make discrete subjects secured inside the archives. The subsequent result is the identification of the fundamental sentences in the record. Sentences which can be identified with numerous different sentences in the segment are plausible the focal point of the chart and substantially more prone to be remembered for the synopsis. Diagram based strategies can be utilized for unmarried just as various report synopses. Since they needn't bother with language specific phonetic preparation other than sentence and expression limit location, they additionally can be actualized to various dialects. Regardless, the use of TFIDF weighting plan for comparability degree has constraints, since it is the least difficult jelly recurrence of words and does not now consider the syntactic and semantic records. In this manner, likeness is basically dependent on syntactic and semantic information that supplements the exhibition of the outline framework.

- **MACHINE LEARNING FOR SUMMARIZATION**

AI approaches model the synopsis as a classification issue. is an early research endeavor at applying AI strategies for rundown. Kupiec et al (1970's) built up a classification work, guileless Bayes classifier, to characterize the sentences as outline sentences and non-synopsis sentences dependent on the highlights they have, given a preparation set of records and their extractive rundowns. The classification probabilities are found out factually from the preparation information utilizing Bayes' standard:

$$P(s \in S | F_1 F_2 \dots F_k)$$

Where  $s$  is a sentence from the archive assortment,  $F_1, F_2, \dots, F_k$  are highlights utilized in classification and  $S$  is the synopsis to be created. Expecting the contingent autonomy between the highlights:

$$P(s \in S | F_1 F_2 \dots F_k) = \frac{\prod_{i=1}^k P(s \in S | F_i)}{\prod_{i=1}^k P(F_i)}$$

The likelihood of a sentence to have a place with the outline is the rating of the sentence. The settled on classifier plays out the job of a sentence scoring capacity. A portion of the successive capacities used in rundown include the situation of sentences inside the record, sentence length, nearness of capitalized words, closeness of the sentence to the archive title, and so forth. AI systems were broadly used in rundown to give some examples. Guileless Bayes, decision trees, help vector machines, Hidden Markov models and Conditional Random Fields are probably the most widely recognized framework becoming acquainted with procedures utilized for synopsis. One basic difference between classifiers is that sentences to be incorporated inside the summary should be resolved freely. It appears that techniques expressly accepting the reliance among sentences which incorporate Hidden Markov form and Conditional Random Fields often outflank different systems. One of the essential issues in utilizing managed acting techniques for outline is that they need a lot of tutoring reports (ordered information) to prepare the classifier, which may not be persistently accessible without trouble. Specialists have proposed a few choices to address this issue:

- **Annotated corpora creation:** Making comments on corpus for synopsis fundamentally benefits the analysts, since additional open benchmarks will be to be had which makes it less convoluted to assess different rundown systems together. It likewise brings down the opportunity of overfitting with restricted information. Ulrich(1990's). conveyed freely to be explained by email corpus and its creation procedure. In any case, making clarified corpus could be very tedious and all the more basically, there is no standard settlement on picking the sentences, and different individuals may pick various sentences to collect the outline

- **Semi-supervised approaches:** Utilizing a semi-managed procedure to prepare a classifier. In semi-administered learning we use the unlabeled information in preparation. There is normally a limited quantity of marked information alongside a lot of unlabeled information.

Olivier et al(2006). proposed a semi-directed technique for extractive outline. They co-prepared two classifiers iteratively to misuse unlabeled information. In every cycle, the unlabeled preparing models (sentences) with top scores are remembered for the marked preparing set, and the two classifiers are prepared on the new preparing information. AI techniques have been demonstrated to be effective and fruitful in single and multi-record synopsis, specifically in class specific rundown where classifiers are prepared to find specific kinds of data, for example, scientific paper outline and true to life outlines.

#### **2.7.2.2 SENTENCE SCORE**

At the point when the intermediate representation is created, we dole out a significance score to each sentence. In point portrayal draws near, the score of a sentence speaks to how well the sentence clarifies the absolute most significant themes of the content. In the greater part of the pointer portrayal strategies, the score is registered by amassing the proof from different markers. AI procedures are frequently used to find marker loads.

#### **2.7.2.3 SUMMARY SENTENCE SELECTION**

Over the long haul, the summarizer system picks the top k most huge sentences to convey a framework. A couple of strategies use voracious computations to pick the huge sentences and a couple of procedures may change over the assurance of sentences into a streamlining issue where a combination of sentences is picked, considering the confinement that it should enhance all things considered criticalness and coherency and cut off the abundance. There are various segments that should be considered while picking the noteworthy sentences. For example, setting in which the summary is made may be helpful in picking the criticalness. Sort of the report (for instance news story, email, scientific paper) is another factor which may influence picking the sentences.

### **2.8 TECHNIQUES OF EXTRACTIVE APPROACH**

There are several techniques towards the extractive approach, few of which will be discussed briefly.

**TextRank Algorithm:** TextRank is a universally useful chart based positioning calculation for NLP, it is like PageRank calculation and is an unaided learning approach towards

extractive synopsis. The essential thought of TextRank is to give a score of each sentence in a book and take the top-n sentence and sort them as they show up in the content to fabricate a programmed rundown. A chart is developed by making a vertex for each sentence in the record; the edges between sentences depend on some type of semantic closeness.

**Cosine Similarity:** Cosine likeness is a proportion of closeness between two non-zero vectors of an internal item space that estimates the cosine of the edge between them,

sentences are spoken to as a pack of vectors so the cosine similitude discovers comparability among sentences.

**Sub-modular set function:** In arithmetic, a submodular set capacity is a set capacity whose worth casually has the property that the distinction in the gradual estimation of the capacity that a solitary component makes when added to an information set increments.

In rundown submodular works normally model ideas of inclusion, data portrayal and decent variety. For instance in a report outline one might want the rundown to cover exceedingly significant data in the record, this is an occurrence of set spread in submodular capacities.

**LexRank Algorithm:** Lexrank is a more principled way to estimate sentence importance using random walks and eigenvector centrality, lexrank is very similar to textrank.

## 2.9 THE IMPACT OF CONTEXT IN SUMMARIZATION

Synopsis structures regularly have additional confirmation they can utilize in order to decide the most noteworthy purposes of document(s). For example, when delineating locales, there are discourses or comments coming after the blog section that are extraordinary wellsprings of information to make sense of which parts of the blog are essential and fascinating. In scientific paper summary, there is a great deal of information, for instance, referred to papers and assembling information which can be used to perceive noteworthy sentences in the principal paper. In the going with, we portray some of the settings in more nuances.

- **Web Summarization**

Pages contain loads of components which can't be abridged, for example, pictures. The printed data they have is frequently rare, which makes applying content rundown strategies constrained. In any case, we can consider the setting of a site page, for example snippets of data separated from the substance of the considerable number of pages connecting to it, as extra material to improve outline. The soonest look into in such a manner is the place they enquiry web search tools and bring the pages having connections to the specified site page. At that point they break down the up-and-comer pages and select the best sentences containing connections to the page heuristically. Delort et al(2003).extended and improved this methodology by utilizing a calculation attempting to choose a sentence about a similar point that spreads however many parts of the site page as could reasonably be expected. For blog synopsis, propose a strategy that first gets agent words from remarks and afterward chooses significant sentences from the blog entry containing delegate words.

### **Scientific Articles Summarization**

A valuable wellspring of data while abridging a scientific paper (for example reference based rundown) is to find different papers that refer to the objective paper and concentrate the sentences where the references occur so as to distinguish the significant parts of the objective paper. Amjhad(2011). propose a language model that gives a likelihood to each word in the reference setting sentences. They at that point score the significance of sentences in the first paper utilizing the KL dissimilarity technique (for example finding the comparability between a sentence and the language model).

### **Email Summarization**

Email has some particular attributes that show the parts of both spoken discussion and composed content. For instance, rundown strategies must think about the intelligent idea of the exchange as in spoken discussions. Nenkova et al (1970's). introduced early research in such a manner, by proposing a strategy to produce a rundown for the first two degrees of the string talk. A string comprises at least one discussion between at least two members after some time. They select a message from the root message and from every reaction to the root, thinking about the cover with root setting. Rambow et al (1970's). utilized an AI method and included highlights identified with the string just as highlights of the email structure, for example, position of the sentence in the track, number of beneficiaries, and so on. Newman et

al(1970's). depict a framework to condense a full letter box instead of a solitary string by bunching messages into topical gatherings and afterward extricating synopses for each group.

## 2.10 APPLICATION AREA OF AUTOMATIC SUMMARIZATION

- **Media Monitoring:** The issue of data over-burden and substance stun has been generally talked about. Programmed rundown gives a chance to consolidate the persistent downpour of data into littler snippets of data.
- **Newsletters:** Numerous week by week bulletins take the type of a presentation pursued by a curated choice of applicable articles. Synopsis would enable news associations to additionally enhance bulletins with a flood of outlines which can be a specific helpful configuration for versatile.
- **Search marketing and SEO:** When assessing scan inquiries for SEO, it is basic to have a
  - balanced comprehension of what your rivals are discussing in their substance. This has gotten especially significant since Google refreshed its calculation and moved concentration towards topical power.
- **Internal Document Workflow:** Large organizations are continually creating inward information, which often gets put away and under utilized in databases as unstructured information. These organizations should grasp devices that let them reuse these proficiencies. Outline can empower investigators to rapidly comprehend everything the organization has just done in a given subject, and rapidly collect reports that join various perspectives.

## 2.11 EVALUATION OF AUTOMATIC SUMMARIZATION

Assessment of an outline is a difficult task in light of the fact that there is no perfect rundown for an archive or an assortment of reports and the meaning of a decent synopsis is an open inquiry to enormous degree. It has been discovered that human summarizers have low

understanding for assessing and delivering outlines. Moreover, pervasive utilization of different measurements and the absence of a standard assessment metric have additionally made outline assessment be difficult and testing.

### **Evaluation of Automatically Produced Summaries**

There have been a few assessment battles since the late 1990s in the US. They incorporate SUMMAC (1996-1998), DUC (the Document Understanding Conference, 2000-2007), and all the more as of late TAC (the Text Analysis Conference, 2008-present). These gatherings have essential job in structure of assessment models and assess the outlines dependent on human just as programmed scoring of the rundowns. So as to have the option to do programmed outline assessment, we have to vanquish three significant difficulties:

- i) It is basic to choose and indicate the most significant pieces of the first content to save
- ii) Evaluators need to naturally recognize these bits of significant data in the applicant rundown, since this data can be spoken to utilizing dissimilar articulations.
- iii) The readability of the summary in terms of grammaticality and coherence has to be evaluated.

### **Human Evaluation**

The most straightforward approach to assess a synopsis is to have a human evaluate its quality. For instance, in DUC, the judges would assess the inclusion of the rundown, for example how much the applicant rundown secured the first given info. In later ideal models, specifically TAC, question based rundowns have been made. At that point judges assess to what degree an outline answers the given question. The components that human specialists must think about when offering scores to every competitor outline are grammaticality, non-excess, combination of most significant snippets of data, structure and lucidness.

### **Automatic Evaluation Methods**

There have been a lot of measurements to consequently assess rundowns since the mid 2000s. ROUGE is the most generally utilized measurement for programmed assessment.

## ROUGE.

Lin presented a lot of measurements called Recall Oriented Understudy for Gisting Evaluation (ROUGE) to consequently decide the nature of a synopsis by contrasting it with human (reference) rundowns. There are a few varieties of ROUGE, and here we simply notice the most extensively utilized ones

- **ROUGE-n:** This measurement is review put together measure and based with respect to examination of n-grams. a progression of n-grams (for the most part two and three and once in a while four) is evoked from the reference outlines and the competitor rundown (naturally created synopsis). Let  $p$  be "the quantity of regular n-grams among up-and- comer and reference rundown", and  $q$  be "the quantity of n-grams separated from the reference outline as it were". The score is registered as:

$$\text{ROUGE-n} = \frac{p}{q}$$

- **ROUGE-L:** This measure utilizes the idea of longest regular subsequence (LCS) between the two groupings of content. The instinct is that the more drawn out the LCS between two synopsis sentences, the more comparative they are. Despite the fact that this measurement is more flexible than the past one, it has a downside that all n-grams must be sequential.
- **ROUGE-SU:** This measurement is called skip bi-gram and uni-gram ROUGE and considers bi-grams just as uni-grams. This measurement permits addition of words between the first and the final expressions of the bi-grams, so they shouldn't be continuous arrangements of words.

The expanding development of the Internet has made an enormous measure of data accessible. It is difficult for people to outline a lot of content. Accordingly, there is a gigantic requirement for programmed rundown devices in this period of data over-burden. In this paper, we underscored different extractive methodologies for single and multi-record synopsis. We depicted the absolute most broadly utilized strategies, for example, subject portrayal draws near, recurrence driven strategies, chart based and AI systems. In spite of the fact that it isn't possible to clarify every single differing calculation and approach exhaustively in this paper, we think it gives a decent knowledge into late patterns and



advances in programmed rundown strategies and portrays the cutting edge in this exploration region.

## 2.12 REVIEW OF RELATED WORKS

Many other research works have contributed to the development of an automated article summarization system and the systems have been yielding good results. These related works together with the techniques used and their limitations are listed below

| Author & Year              | Article Title   | Techniques                                  | Limitation   |
|----------------------------|---|---|--|
| Collins, (2015)            | What Automatic Summarization is all about                                 | Application Area of Automatic summarization | The areas where automatic summarization were applied showed some challenges  |
| Kathleen(August 2014)      | A survey of text summarization techniques applying the textRank algorithm | The TextRank algorithm                      | Although the textRank algorithm can be seen as one of the best algorithm when dealing with text summarization its limitation are grammar, while concatenating it does not fully eradicate grammar errors |
| Amjad, (January 13, 2011). | Coherent citation based summarization of scientific paper                 | Topic words, pageRank, LexRank              | Technique was limited to the use of keyphrase, and it couldn't go further to summarize other type of data  |

|                           |   |   |  |
|---------------------------|---|---|--|
| John, (march 2011)        | Personalized and automatic social summarization of events in videos   | Topic words, PageRank                           | The technique gave lots of issues has it didn't drop some redundant words  |
| Ibrahim F. (January 2010) | "Semantic Graph Reduction Approach for Abstractive Text Summarization | Semantic graph based approach                   | This technique is limited to single document summarization it cannot be used for multiple document summarization   |
| Atif (February 2008)      | A Review on Abstractive Summarization Methods"                        | Structure based abstractive method (rule based) | The abstractive approach can be seen has the best technique for summarization but it is very much limited as it has to deal with a whole lot of limitation like semantics and the study of advanced natural language processing which in itself is still a growing field |
| Ani, (January 2004).      | Facilitating email thread access by extractive summary generation     | Topic words                                     | The technique gave lots of issues has it didn't drop some redundant words  |

## **CHAPTER THREE**

### **METHODOLOGY**

#### **3.1 INTRODUCTION**

This chapter focuses on the design, development and system modeling. The system to be developed will help in automatic text summarization that is presenting the source document into a shorter version with semantics.

To achieve this aim, we have carried out a proper review of various techniques and algorithms used for summarization.

#### **3.2 OVERVIEW OF THE EXISTING SYSTEM.**

Summarization as we know is the act of writing concise and fluent summary from a document or multiple documents. Since time immemorial summarization has been done majorly by hand i.e. humans, the problem with this system is that humans are very prone to errors and the possibility of leaving out the most important sentences that should be included in the summary compared to a computer generated.

#### **3.3 DESCRIPTION OF THE PROPOSED SYSTEM**

The system at a high level of abstraction simply allows the visitors to the site view summary of previously uploaded documents by signed up users. This implements text summarization to automatically create an abstract of the uploaded document which the visitors to the website can read. Interested visitors may want to get the full document. This is only possible for signed up users hence the system compels such visitors to sign up and hence has the privileges of other signed up users such as ability to download full document and ability to upload your own documents. A handy tool for signed up users is also an ability to automatically create your document abstract using the summarization engine embedded within the system.

### 3.4 METHODOLOGY FOR ARTICLE SUMMARIZATION

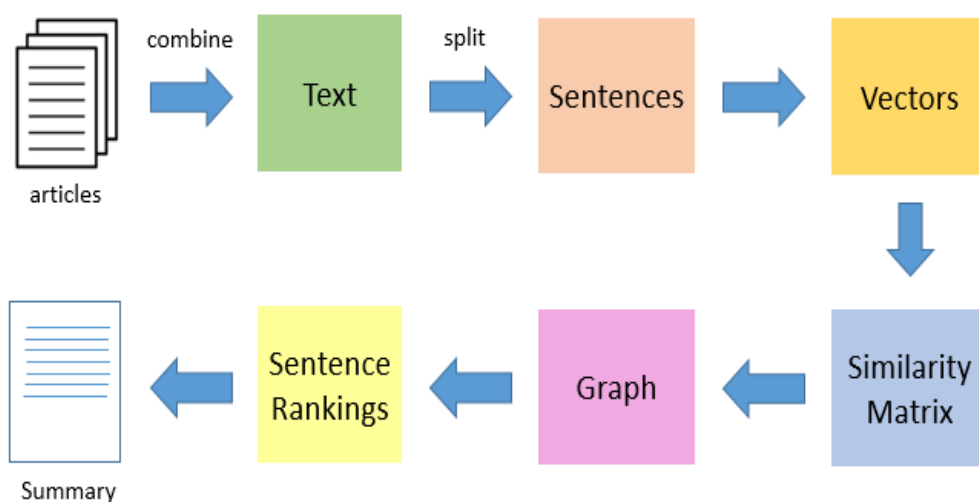
Extractive summarization approach for abstractive summarization approach can be used in summarizing a set of documents but I considered the extractive approach using the TextRank algorithm.

#### 3.4.1 EXTRACTIVE SUMMARIZATION USING TEXTRANK

The extractive approach of building an automatic summarization system takes several sentences or phrases and stacks them to produce a summary.

TextRank algorithm is an extractive and unsupervised text summarization technique. To generate the required summary the following steps would be considered using TextRank algorithm:

- All the texts contained in an article will be concatenated.
- The texts will be split into individual sentences.
- Vector representation for each sentences will be found
- Similarities between sentence vectors are then calculated and stored in a matrix
- The similarity matrix is then converted into a graph for sentence rank calculation
- Finally, a certain number of top ranked sentences form the final summary



**Figure 3.1**

### 3.5 FRAMEWORK OF THE SYSTEM

#### 3.5.1 HIGH LEVEL SYSTEM FRAMEWORK FOR PREVIOUS TEXT SUMMARIZATION

The system at a high level of abstraction basically allows the visitors to the site view summary of previously uploaded documents by signed up users. This implements text summarization to automatically create an abstract of the uploaded document which the visitors to the website can read. Interested visitors may want to get the full document. This is only possible for signed up users hence the system compels such visitor to sign up and hence has the privileges of other signed up users such as ability to download full document and ability to upload your own documents. A handy tool for signed up users is also an ability to automatically create your document abstract using the summarization engine embedded within the system.

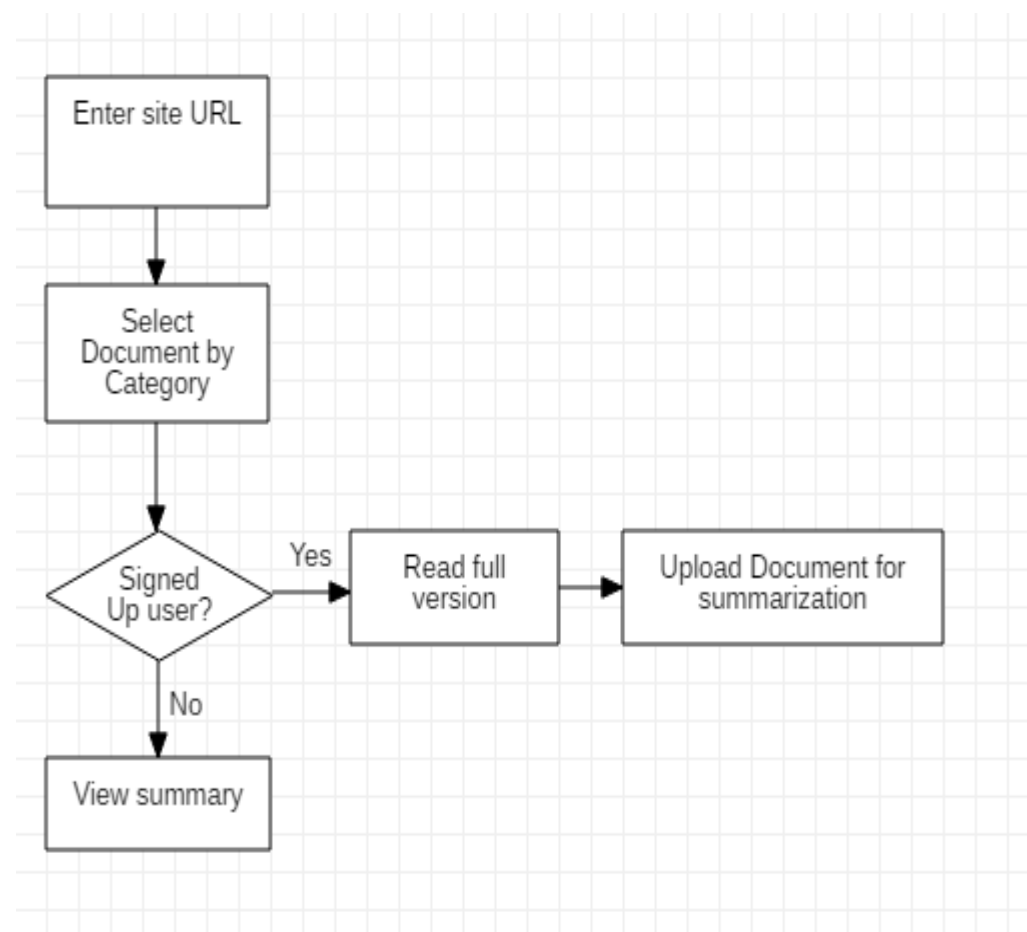
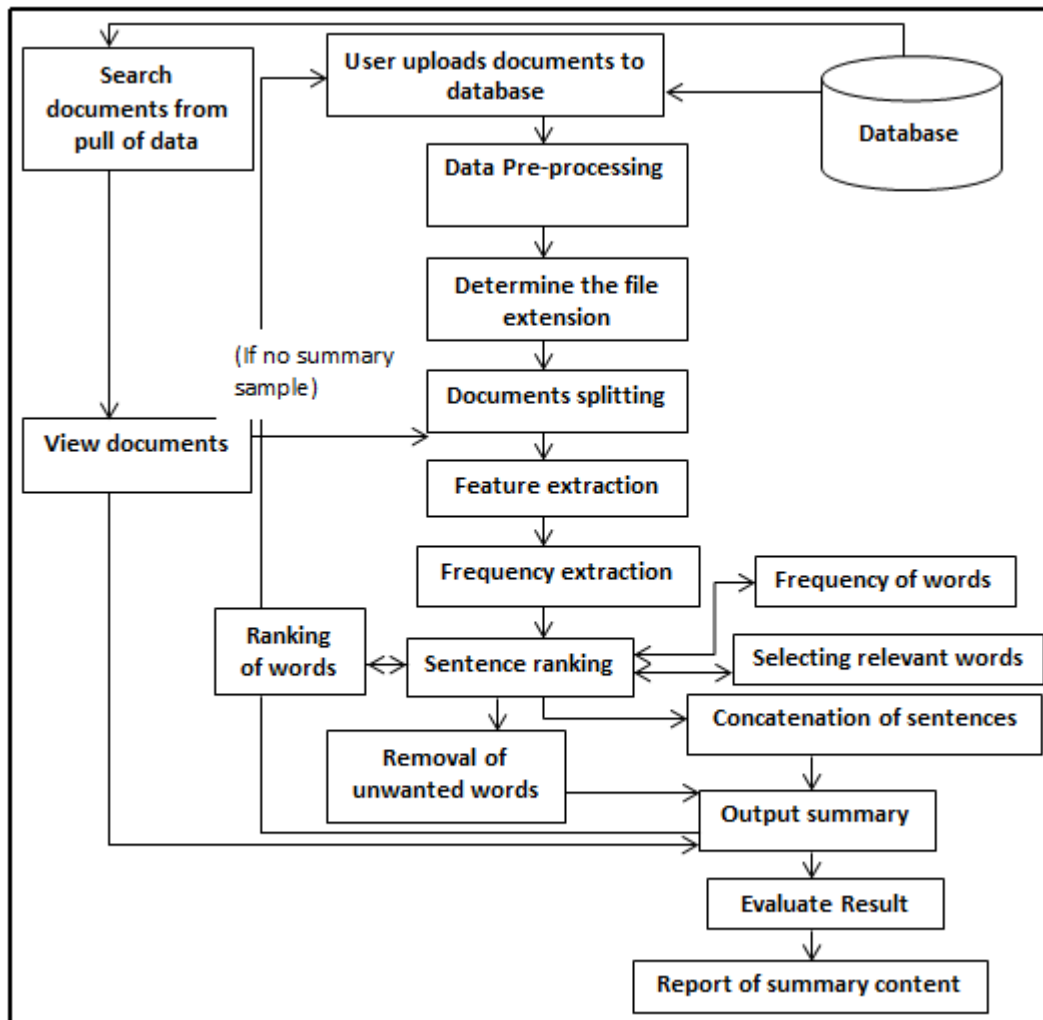


Figure 3.2: (Source; Abraham, January 2016)

### 3.5.2 PROPOSED FRAMEWORK FOR ARTICLE SUMMARIZATION SYSTEM

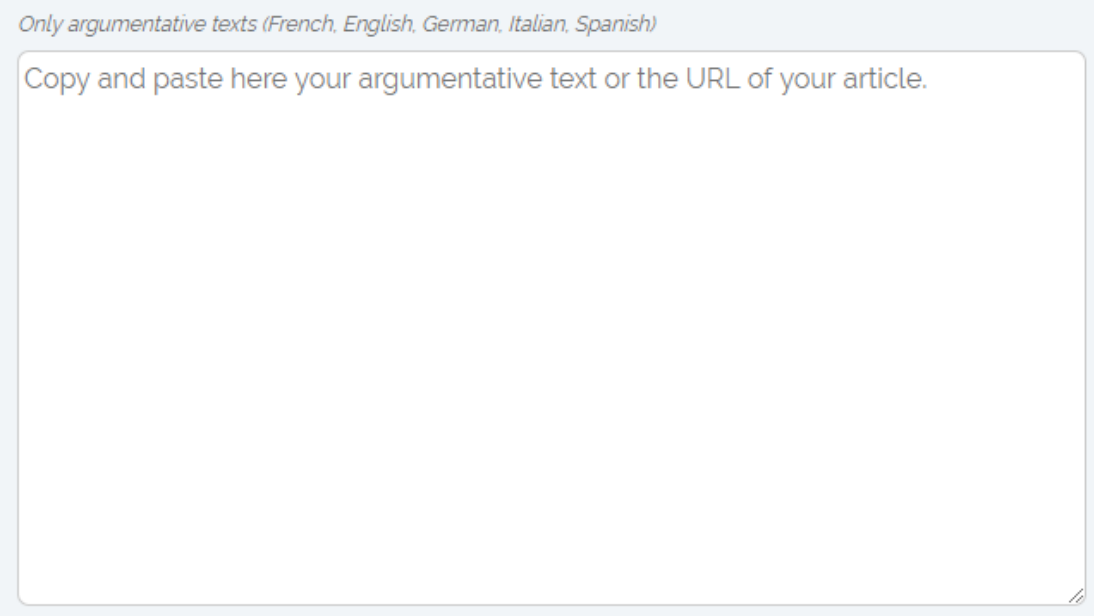
Below is the architecture for the summarization engine already discussed in the chapter two of this work:



**Figure 3.3:** Adapted from Source: (Rakesh et al, 2007)

### 3.6 SAMPLE INTERFACE FOR ARTICLE SUMMARIZATION

#### 3.6.1 FIRST INTERFACE

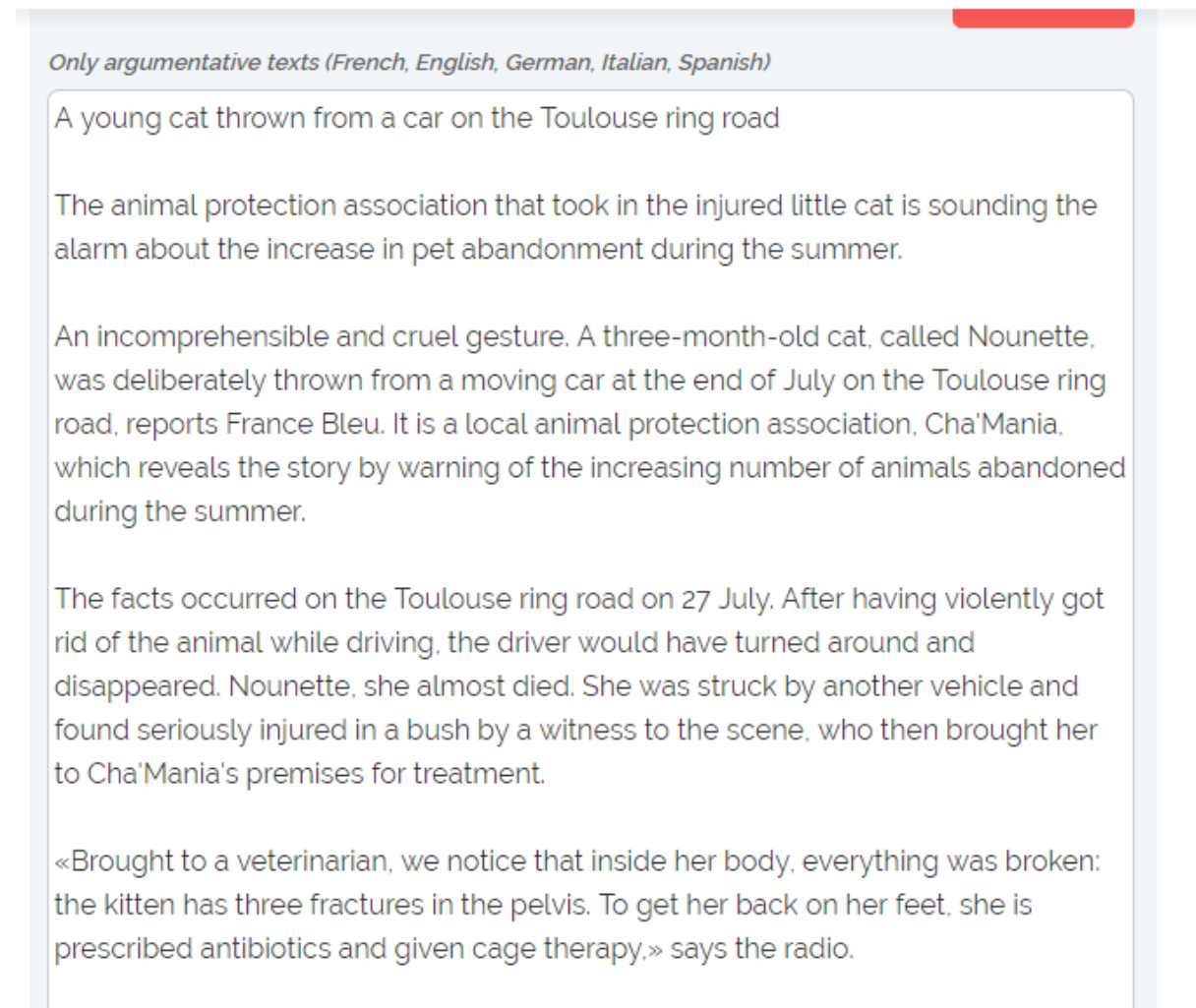


*Only argumentative texts (French, English, German, Italian, Spanish)*

Copy and paste here your argumentative text or the URL of your article.

**Figure 3.4: Interface to Upload document for summarization**

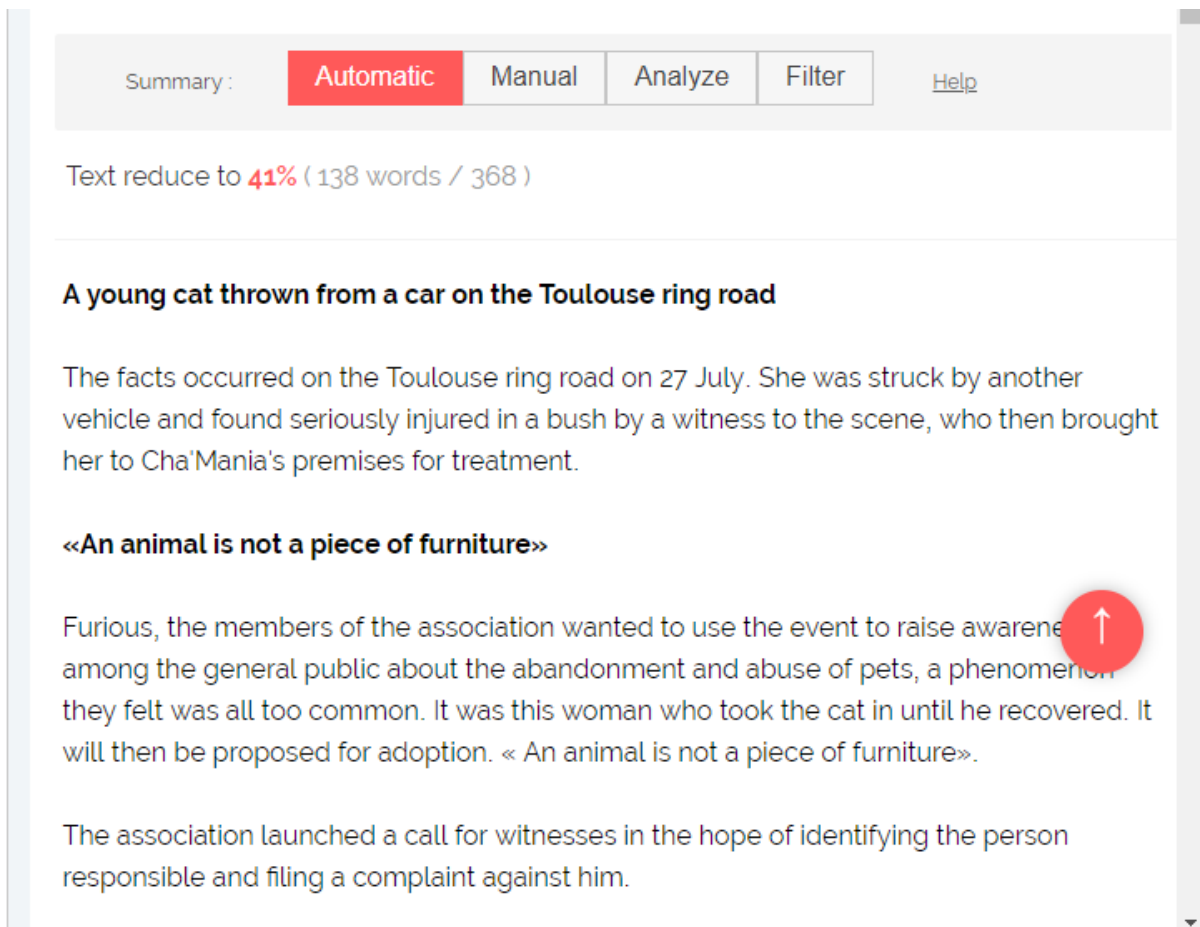
### 3.6.2 SECOND INTERFACE



**Figure 3.5:** Sample document uploaded for summarization



### 3.6.3 THIRD INTERFACE



The screenshot shows a web-based text summarization interface. At the top, there is a navigation bar with the following elements: a label 'Summary :', a red button labeled 'Automatic', and three buttons labeled 'Manual', 'Analyze', and 'Filter'. To the right of these buttons is a link labeled 'Help'. Below the navigation bar, the interface displays the result of the summarization process. It states 'Text reduce to 41% ( 138 words / 368 )'. The main content area contains a bolded title 'A young cat thrown from a car on the Toulouse ring road'. Below the title, there are three paragraphs of text. The first paragraph describes the incident on July 27th. The second paragraph is a quote: '«An animal is not a piece of furniture»'. The third paragraph describes the association's reaction and the call for witnesses. A red circular icon with a white upward-pointing arrow is positioned to the right of the second paragraph. The interface has a light gray background and a vertical scrollbar on the right side.

Summary : **Automatic** Manual Analyze Filter [Help](#)

Text reduce to **41%** ( 138 words / 368 )

**A young cat thrown from a car on the Toulouse ring road**

The facts occurred on the Toulouse ring road on 27 July. She was struck by another vehicle and found seriously injured in a bush by a witness to the scene, who then brought her to Cha'Mania's premises for treatment.

**«An animal is not a piece of furniture»**

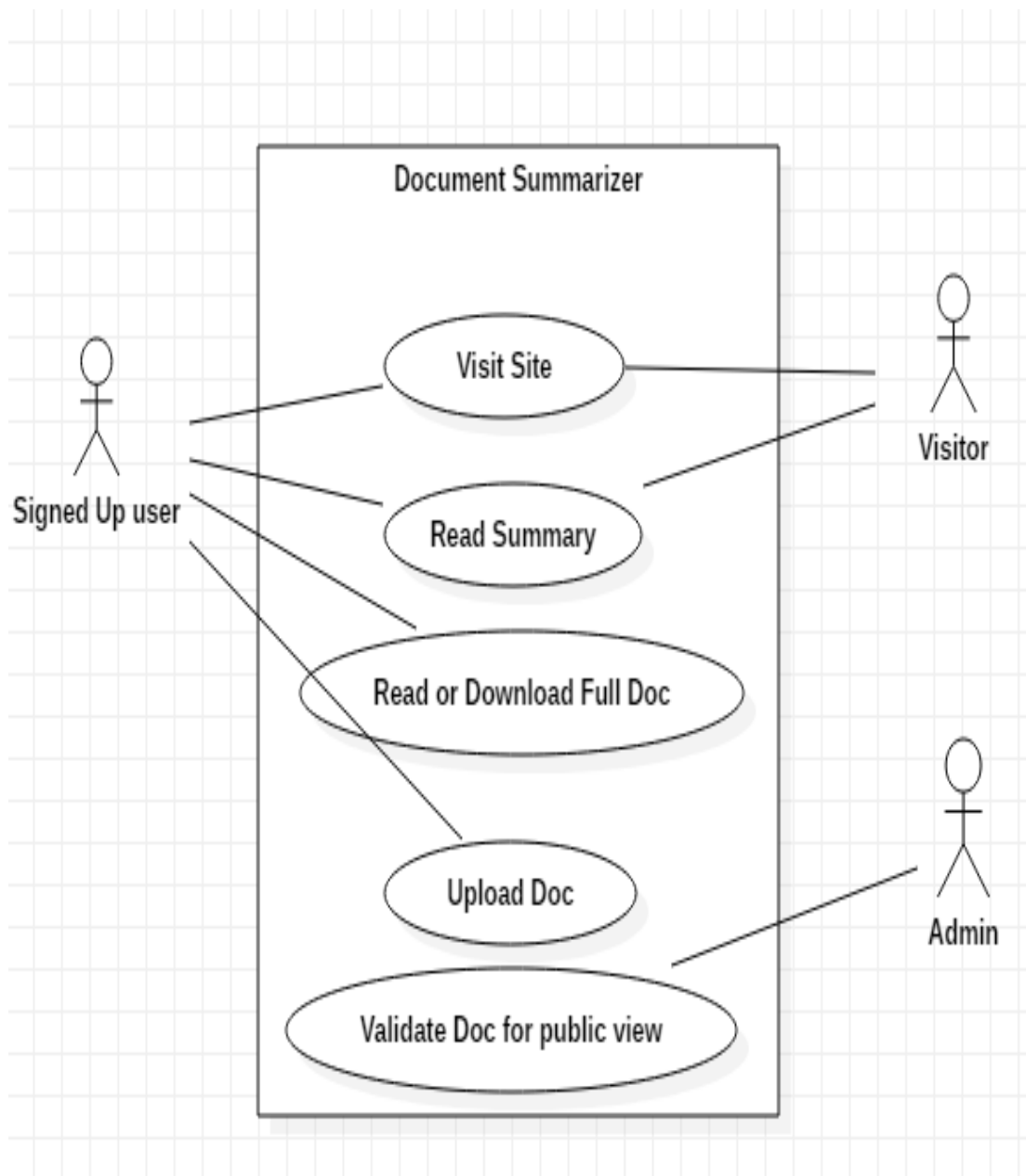
Furious, the members of the association wanted to use the event to raise awareness among the general public about the abandonment and abuse of pets, a phenomenon they felt was all too common. It was this woman who took the cat in until he recovered. It will then be proposed for adoption. « An animal is not a piece of furniture».

The association launched a call for witnesses in the hope of identifying the person responsible and filing a complaint against him.

**Figure 3.6: Result of Summarization using textRank algorithm**

### 3.7 SYSTEM MODELLING

#### 3.7.1 USE CASE DIAGRAM



**Figure 3.7.1: use case diagram for automated article summarization**

## REFERENCES

- Olivier Chapelle, Bernhard Schölkopf, Alexander Zien, and others. (2006). Semisupervised learning. Vol. 2. MIT press Cambridge.
- Ping Chen and Rakesh Verma. (2006). A query-based medical information summarization system using ontology knowledge. In Computer-Based Medical Systems, 2006. CBMS 2006. 19th IEEE International Symposium on. IEEE, 37- 42.
- Freddy Chong Tat Chua and Sitaram Asur. (2013). Automatic Summarization of Events from Social Media.. In ICWSM.
- John M Conroy and Dianne P O'leary. (2001). Text summarization via hidden markov models. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 406- 407.
- Hal Daumé III and Daniel Marcu. (2006). Bayesian query-focused summarization. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 305- 312.
- Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. (1990). Indexing by latent semantic analysis. JASIS 416 (1990), 391- 407.
- J-Y Delort, Bernadette Bouchon-Meunier, and Maria Rifqi. (2003). Enhanced web document summarization using hyperlinks. In Proceedings of the fourteenth ACM conference on Hypertext and hypermedia. ACM, 208- 215.
- Ted Dunning. (1993). Accurate methods for the statistics of surprise and coincidence. Computational linguistics 19, 1 (1993), 61- 74.
- Harold P Edmundson. (1969). New methods in automatic extracting. Journal of the ACM (JACM) 16, 2 (1969), 264- 285.
- Günes Erkan and Dragomir R Radev. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res.(JAIR) 22, 1 (2004), 457- 479.

- Yihong Gong and Xin Liu. (2001). Generic text summarization using relevance measure and latent semantic analysis. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 19- 25.
- Vishal Gupta and Gurpreet Singh Lehal. (2010). A survey of text summarization extractive techniques. Journal of Emerging Technologies in Web Intelligence 2, 3 (2010), 258- 268.
- Ben Hachey, Gabriel Murray, and David Reitter. (2006). Dimensionality reduction aids term co-occurrence based multi-document summarization. In Proceedings of the workshop on task-focused summarization and question answering. Association for Computational Linguistics, 1- 7.