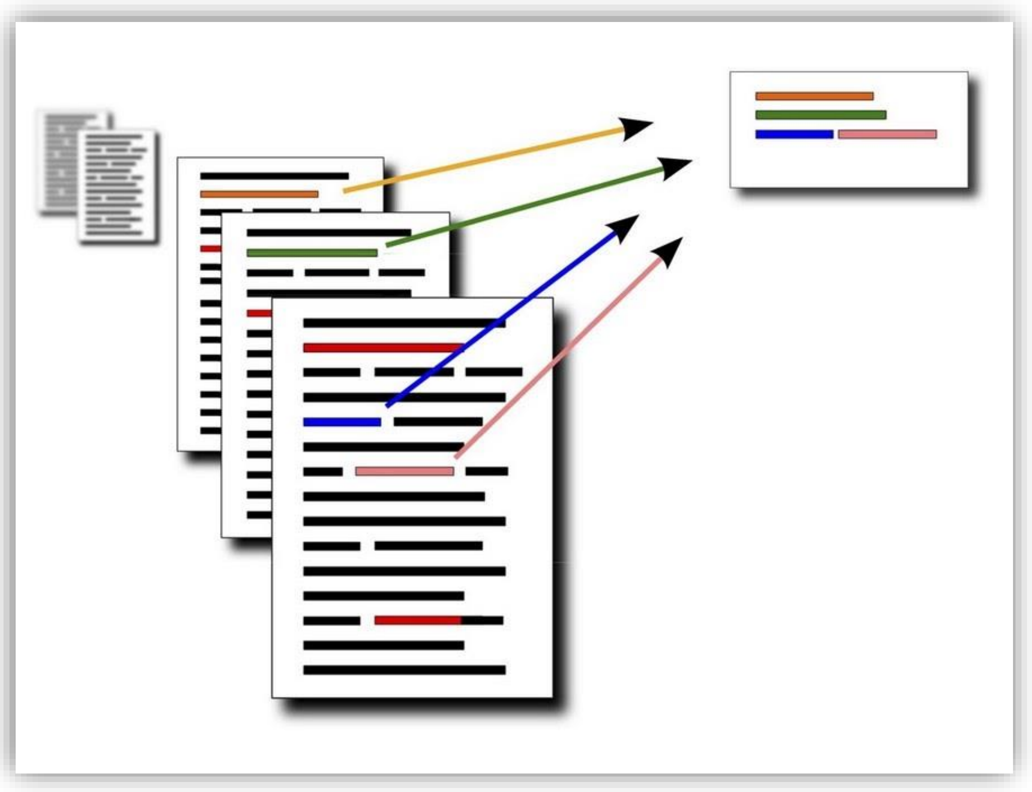
DOCUMENT SUMMARIZATION

ABSTRACT

The process of obtaining important information from any text document os known as text summarization. Text Summarization has been an area that attracts a large number of pursuits and works. Here, the retrieved information is collected by squeezing the content and absorbing useful information. It is very important that the summary we collect is understandable and useful to humans otherwise it is all vague. Two types of approaches are used for text summarization: Extractive and abstractive summarization. In this era of the internet, where there are so many documents out there it is important that we have the essential tool to extract the required important sentences from the long documents.



INTRODUCTION

Before seeing further in Text summarization, first, we need to know what a summary is. A summary is a text that is produced from one or more texts, that shows only germane and crucial information of the original text, and it is of a precise form. The main goal of text summarization is to present the long documents into a petite version with the main meaning in it. The most crucial advantage of reading a summary is, it reduces the reading time. In this world, the importance of text summarization catches more attention because of the abundance of data available on the web. Hence it is necessary to cut down to essential points only. Text summarization has many applications such as summaries of books, stock market, news, highlights of any events/sport/meetings, etc. And due to such vast importance, many universities are consistently working on their improvements.

EXTRACTIVE CONTENT RUNDOWN:

This procedure can be isolated into two stages: Pre-Processing step and Processing step. Pre-Processing is organized portrayal which is recognized as first content. It generally incorporates:

a) Sentences limit recognizable proof. In English, sentence limit is related to nearness of speck toward the finish of sentence

b) Stop-Word Elimination—Common words with no semantics

c) Stemming—the reason for stemming is to get the stem or radix of each word, which underscore its semantics. In Processing step, highlights impacting the importance of sentences are chosen and determined and afterward

Loads are doled out to these highlights utilizing weight learning technique. Last score of each sentence is resolved utilizing Feature-weight condition. Top positioned sentences are chosen for conclusive rundown

Rundown assessment is a significant perspective for content outline. By and large, rundowns can be assessed utilizing inborn or outward measures. While characteristic techniques endeavor to gauge rundown quality utilizing human assessment and extraneous strategies measure the equivalent through an undertaking based execution measure such the data recovery situated errand.

HIGHLIGHTS FOR EXTRACTIVE TEXT SUMMARIZATION

Most of the current robotized content outline frameworks use extraction strategy to deliver a synopsis. Sentence extraction systems are generally used to create extraction rundowns. One of the techniques to acquire reasonable sentences is to relegate some numerical proportion of a sentence for the synopsis called sentence scoring and afterward select the best sentences to shape report rundown dependent on the pressure rate. In the extraction technique, pressure rate is a significant factor used to characterize the proportion between the length of the synopsis and the source content. As the pressure rate expands, the outline will be bigger, and progressively irrelevant substance is contained. While the pressure rate diminishes the rundown to be short, more data is lost. Indeed, when the pressure rate is 5-30%, the nature of rundown is adequate

EXTRACTIVE SUMMARIZATION METHODS:

- Term Frequency-Inverse Document Frequency (TF-IDF) method:
- Group based method
- Chart theoretic approach:
- AI approach:
- Content outline with neural systems:
- Programmed content synopsis dependent on fluffy rationale:

Term Frequency-Inverse Document Frequency (TF-IDF) method:

It is a numerical measurement which reflects how significant a word is in a given document. The TF-IDF esteem expands relatively to the occasions a word shows up in the document. This strategy essentially works in the weighted term-recurrence and backwards sentence recurrence worldview. where sentence-recurrence is the quantity of sentences in the archive that contain that term. These sentence vectors are then scored by likeness to the question and the most noteworthy scoring sentences are picked to be a piece of the summary. Summarization is inquiry explicit:

$$TFIDF\ score\ for\ term\ i\ in\ document\ j = TF(i,j) * IDF(i)$$

$$where$$

$$IDF = Inverse\ Document\ Frequency$$

$$TF = Term\ Frequency$$

$$TF(i,j) = \frac{\text{Term i frequency in document j}}{\text{Total words in document j}}$$

$$IDF(i) = \log_2 \left(\frac{\text{Total documents}}{\text{documents with term i}}\right)$$

$$and$$

$$t = Term$$

$$j = Document$$

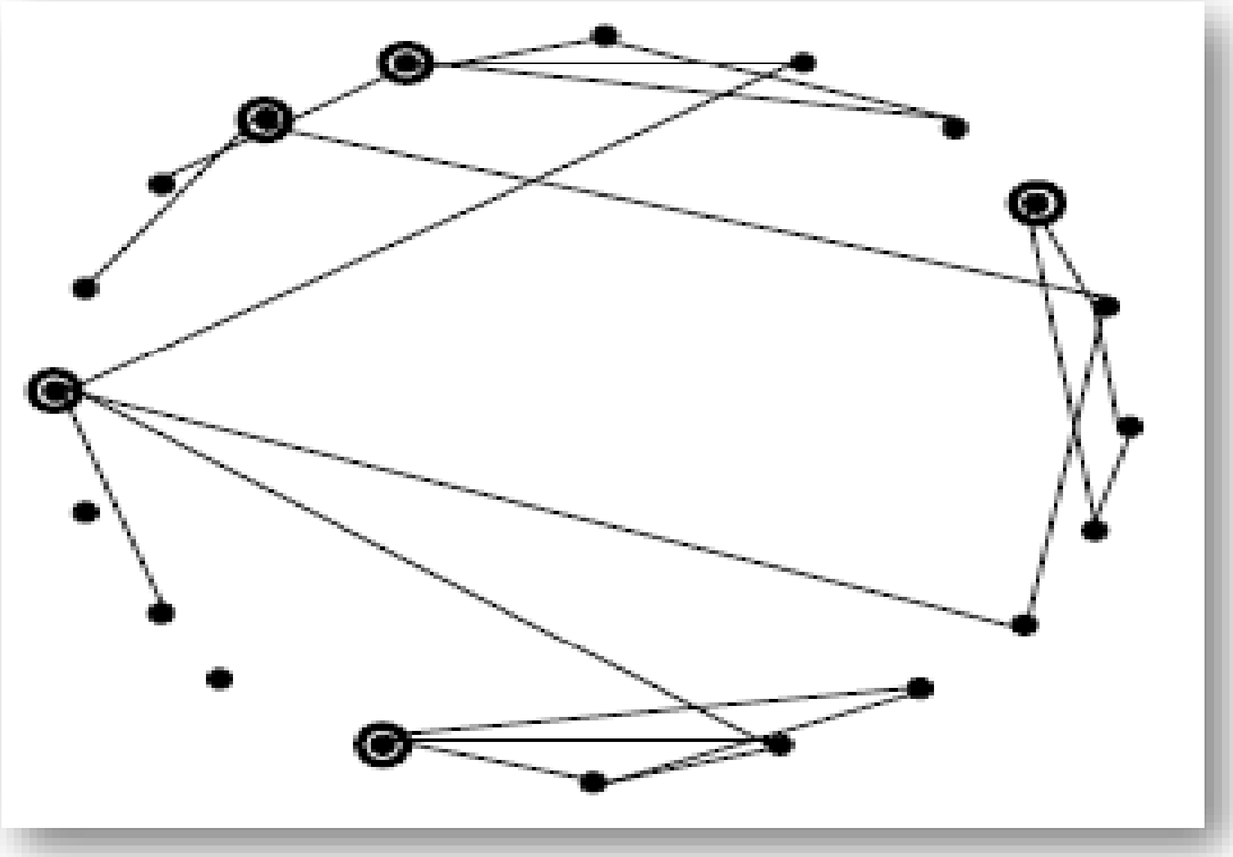
The speculation expected by this methodology is that on the off chance that there are "increasingly explicit words" in a given sentence, at that point the sentence is generally progressively significant. The objective words are generally things .This strategy plays out a correlation between the term recurrence (tf) in a record - right now sentence is dealt with as a record and the report recurrence (df), which implies the occasions that the word happens along all archives. The TF/IDF score is determined as follows

Group based method:

In this strategy, the semantic idea of a given record is caught and communicated in characteristic language by a lot of triplets (subjects, action words, objects identified with each sentence).Cluster these triplets utilizing comparable data. The triplets explanations are considered as the essential unit during the time spent summarization. More comparable the triplets are, the more the data is futile rehashed; in this way, a rundown might be developed utilizing an arrangement of sentences related the registered bunches.

Diagram theoretic approach:

In this strategy, there is a hub for each sentence. Two sentences are associated with an edge if the two sentences share some normal words, as such, their closeness is over some edge. This portrayal gives two outcomes: The segments contained in the chart (that is those sub-diagrams that are detached to the next sub diagrams), structure unmistakable subjects canvassed in the reports. The second outcome by the chart theoretic technique is the ID of the significant sentences the record.



The hubs with high cardinality (number of edges associated with that hub), are the significant sentences in the parcel, and thus convey higher inclination to be remembered for the summary. Figure shows a model diagram for an archive. It very well may be seen that there are around 3-4 themes in the archive; the hubs that are surrounded can be believed to be educational sentences in the record, since they share data with numerous different sentences in the report. The diagram theoretic technique may likewise be adjusted effectively for representation of entomb and intra record likeness.

AI approach:

In this method, the preparing dataset is utilized for reference and the synopsis procedure is displayed as a characterization issue: sentences are named rundown sentences and non-outline sentences dependent on the highlights that they have. The grouping probabilities are found out measurably from the preparation information, utilizing Bayes' standard: where, s is a sentence from the report assortment, F1, F2...FN are highlights utilized in order. S is the synopsis to be created, and P (s∈< S | F1, F2, FN) is the likelihood that sentence s will be picked to frame the outline given that it has highlights F1, F2...FN.

Content rundown with neural networks:

In this strategy, each record is changed over into a rundown of sentences. Each sentence is spoken to as a vector [f1,f2,...,f7], made out of 7 features. Seven Features of a Document

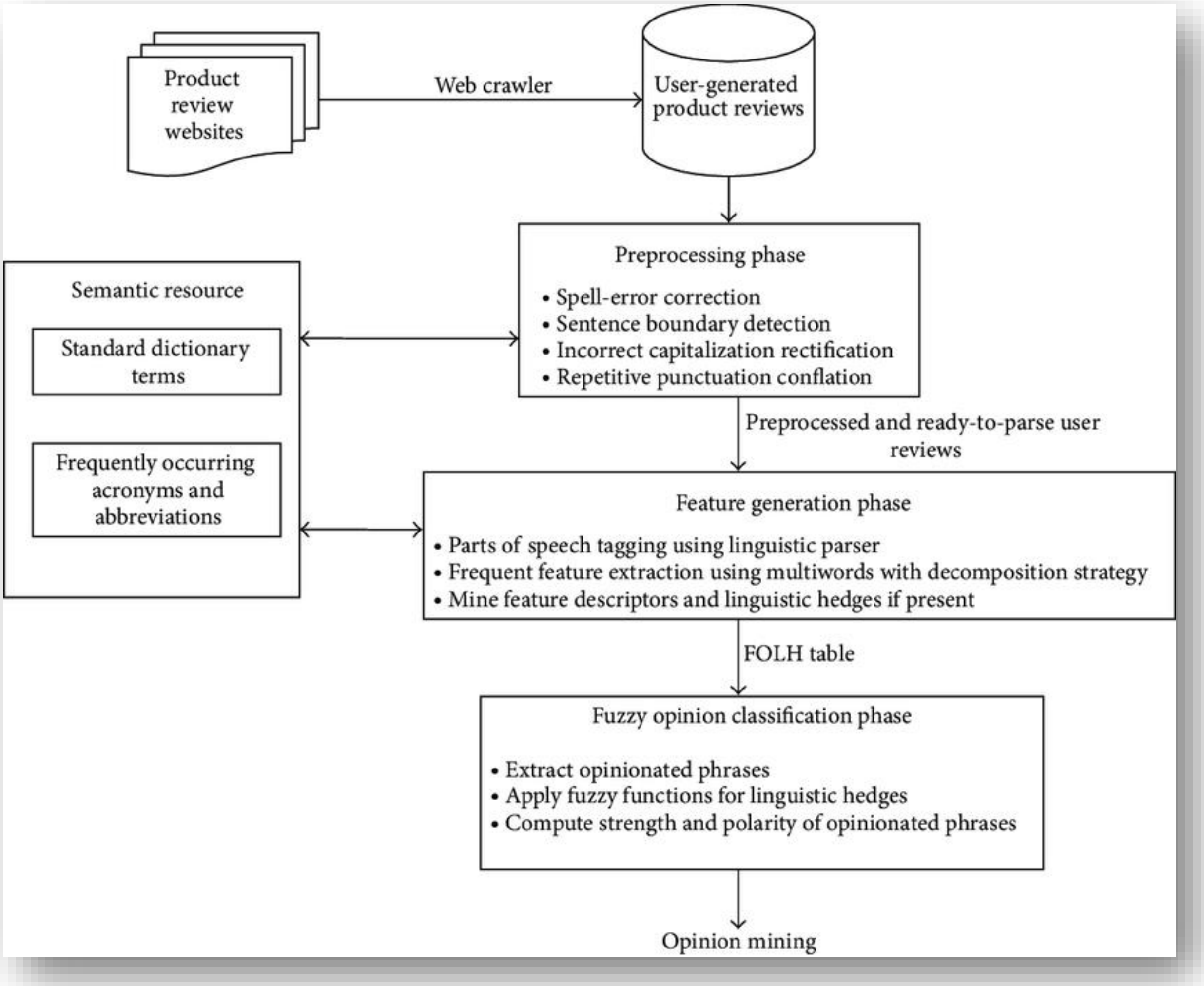
1) F1 Paragraph follows title
2) F2 Paragraph area in document
3) F3 Sentence area in paragraph
4) F4 First sentence in paragraph
5) F5 Sentence length
6) F6 Number of topical words in the sentence
7) F7 Number of title words in the sentence

The first period of the procedure includes preparing the neural systems to gain proficiency with the kinds of sentences that ought to be remembered for the rundown. When the system has taken in the highlights that must exist in outline sentences, we have to find the patterns and connections among the highlights that are inalienable in most of sentences. This is practiced by the element combination stage, which comprises of two stages: 1) taking out unprecedented highlights; and 2) falling the impacts of basic highlights.

Programmed content rundown dependent on fluffy logic:

This technique considers every trait of a book, for example, sentence length, similitude too little, comparability to catchphrase and so on as the contribution of fluffy framework

.Then, it enters all the standards required for synopsis, in the information base of structure.


Product review websites
Web crawler
User-generated product reviews
Preprocessing phase
• Spell-error correction
• Sentence boundary detection
• Incorrect capitalization rectification
• Repetitive punctuation conflation
Semantic resource
Standard dictionary terms
Frequently occurring acronyms and abbreviations
Preprocessed and ready-to-parse user reviews
Feature generation phase
• Parts of speech tagging using linguistic parser
• Frequent feature extraction using multiwords with decomposition strategy
• Mine feature descriptors and linguistic hedges if present
FOLH table
Fuzzy opinion classification phase
• Extract opinionated phrases
• Apply fuzzy functions for linguistic hedges
• Compute strength and polarity of opinionated phrases
Opinion mining

Afterward, an incentive from zero to one is gotten for each sentence in the yield dependent on sentence attributes and the accessible standards in the information base. The got an incentive in the yield decides the level of the significance of the sentence in the last outline.

The information participation work for each element is separated into three enrollment capacities which are made out of unimportant qualities (low L), exceptionally low (VL), medium (M), noteworthy qualities (High h) and high (VH). The significant sentences are extricated utilizing IF-THEN guidelines as indicated by the component criteria.

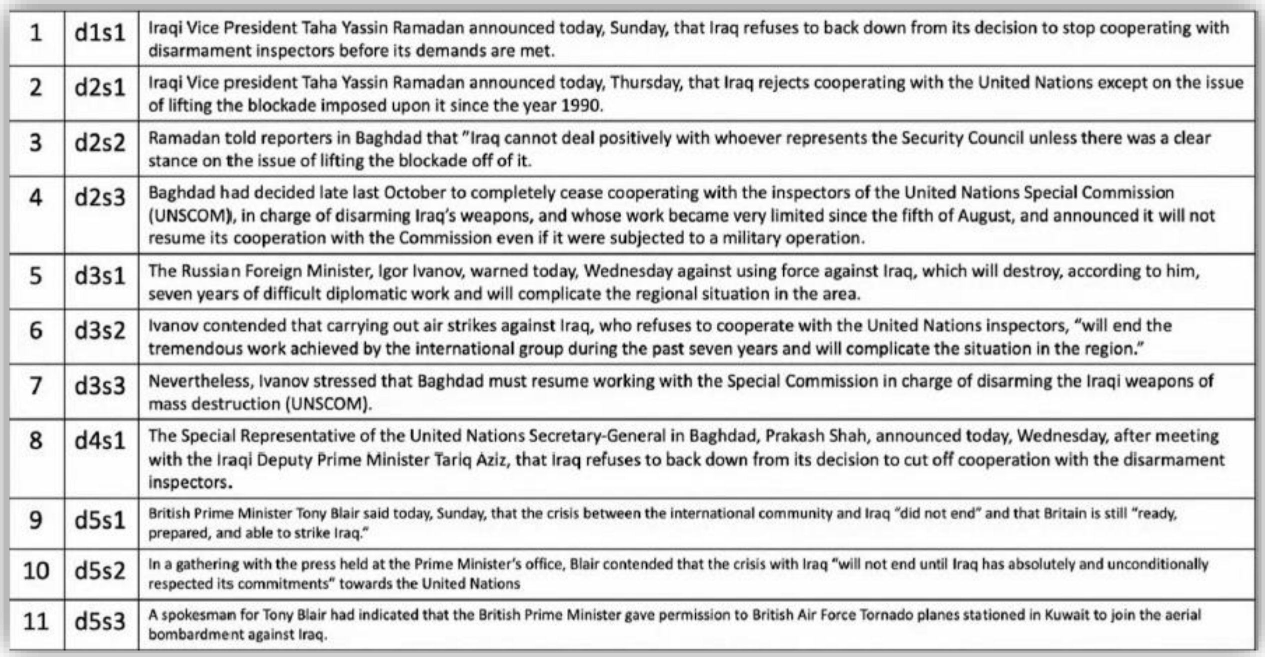
The fluffy rationale framework comprises of four segments: fuzzifier, induction motor, fuzzifier, and the fluffy information base. In the fuzzifier, fresh information sources are converted into semantic qualities utilizing an enrollment capacity to be utilized to the info etymological factors. In the last advance, the yield semantic factors from the derivation are changed over to the last fresh qualities by the fuzzifier utilizing participation work for speaking to the last sentence score.

Extractive Text Summarization Techniques

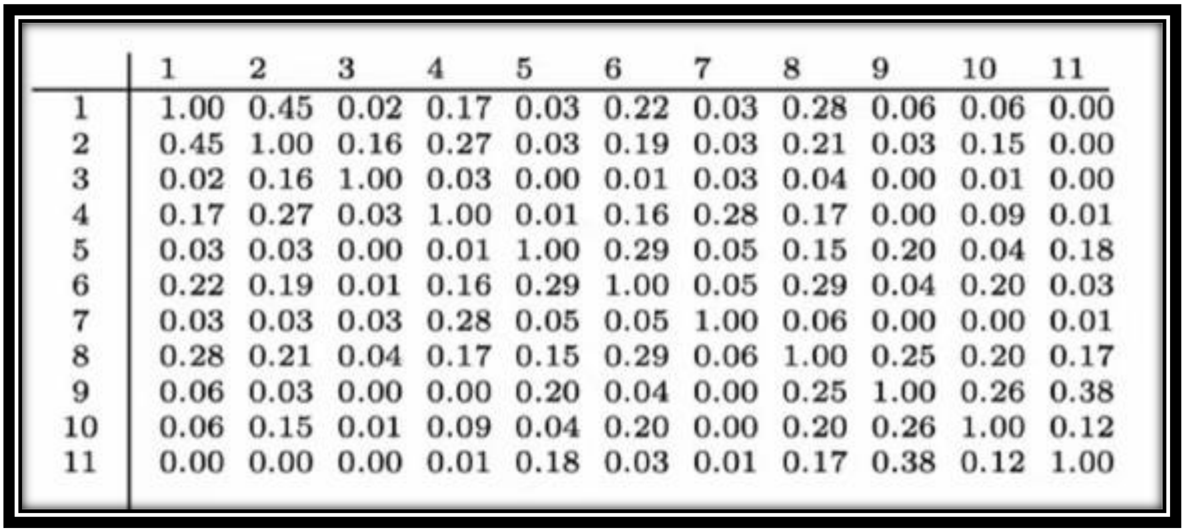
They can be majorly divided as Unsupervised Learning and Supervised Learning.

1. Unsupervised Learning Methods: We do not supervise the model in this case and the model discovers on its own, using algorithms. There is higher level of automation. They are classified as –

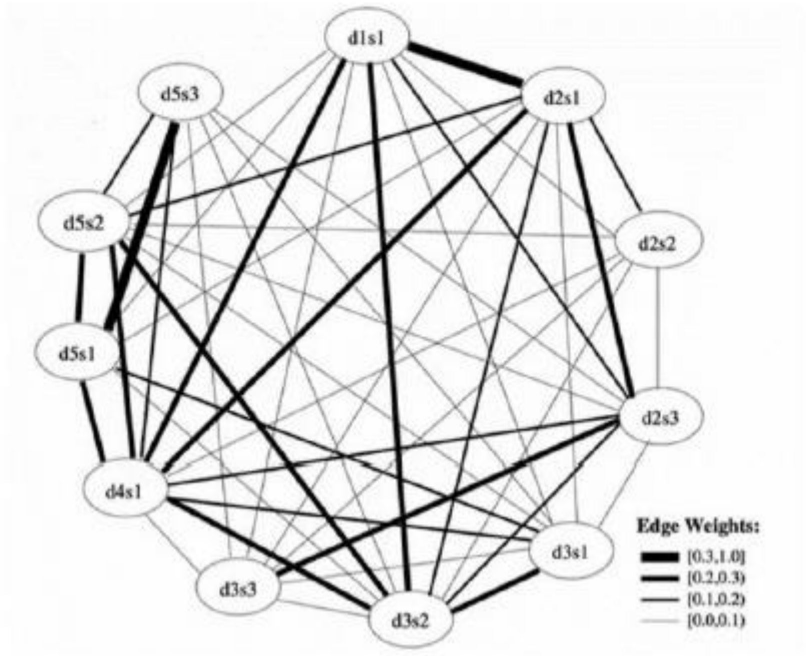
- Graph based method (LexRank): Let’s say we have some document or group of sentences and then we create a similarity matrix. Well similarity matrix sees how similar are the 2 sentences in a document. Each sentence is represented as nodes in the graph. Then we connect the sentences based on similarity matrix. If 2 sentences have a lot in common like, number of words, etc. then there is a high value in similarity graph. Consider 11 sentences from different documents and their LexRank also the graph:



1	d1s1	Iraqi Vice President Taha Yassin Ramadan announced today, Sunday, that Iraq refuses to back down from its decision to stop cooperating with disarmament inspectors before its demands are met.
2	d2s1	Iraqi Vice president Taha Yassin Ramadan announced today, Thursday, that Iraq rejects cooperating with the United Nations except on the issue of lifting the blockade imposed upon it since the year 1990.
3	d2s2	Ramadan told reporters in Baghdad that "Iraq cannot deal positively with whoever represents the Security Council unless there was a clear stance on the issue of lifting the blockade off of it.
4	d2s3	Baghdad had decided late last October to completely cease cooperating with the inspectors of the United Nations Special Commission (UNSCOM), in charge of disarming Iraq's weapons, and whose work became very limited since the fifth of August, and announced it will not resume its cooperation with the Commission even if it were subjected to a military operation.
5	d3s1	The Russian Foreign Minister, Igor Ivanov, warned today, Wednesday against using force against Iraq, which will destroy, according to him, seven years of difficult diplomatic work and will complicate the regional situation in the area.
6	d3s2	Ivanov contended that carrying out air strikes against Iraq, who refuses to cooperate with the United Nations inspectors, "will end the tremendous work achieved by the international group during the past seven years and will complicate the situation in the region."
7	d3s3	Nevertheless, Ivanov stressed that Baghdad must resume working with the Special Commission in charge of disarming the Iraqi weapons of mass destruction (UNSCOM).
8	d4s1	The Special Representative of the United Nations Secretary-General in Baghdad, Prakash Shah, announced today, Wednesday, after meeting with the Iraqi Deputy Prime Minister Tariq Aziz, that Iraq refuses to back down from its decision to cut off cooperation with the disarmament inspectors.
9	d5s1	British Prime Minister Tony Blair said today, Sunday, that the crisis between the international community and Iraq "did not end" and that Britain is still "ready, prepared, and able to strike Iraq."
10	d5s2	In a gathering with the press held at the Prime Minister's office, Blair contended that the crisis with Iraq "will not end until Iraq has absolutely and unconditionally respected its commitments" towards the United Nations
11	d5s3	A spokesman for Tony Blair had indicated that the British Prime Minister gave permission to British Air Force Tornado planes stationed in Kuwait to join the aerial bombardment against Iraq.



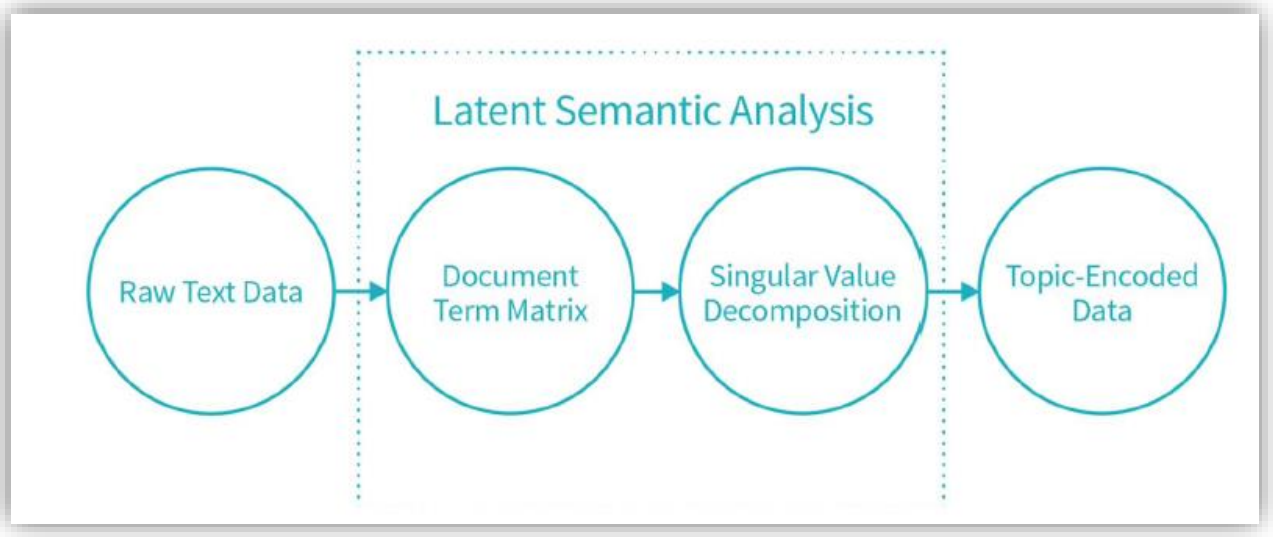
	1	2	3	4	5	6	7	8	9	10	11
1	1.00	0.45	0.02	0.17	0.03	0.22	0.03	0.28	0.06	0.06	0.00
2	0.45	1.00	0.16	0.27	0.03	0.19	0.03	0.21	0.03	0.15	0.00
3	0.02	0.16	1.00	0.03	0.00	0.01	0.03	0.04	0.00	0.01	0.00
4	0.17	0.27	0.03	1.00	0.01	0.16	0.28	0.17	0.00	0.09	0.01
5	0.03	0.03	0.00	0.01	1.00	0.29	0.05	0.15	0.20	0.04	0.18
6	0.22	0.19	0.01	0.16	0.29	1.00	0.05	0.29	0.04	0.20	0.03
7	0.03	0.03	0.03	0.28	0.05	0.05	1.00	0.06	0.00	0.00	0.01
8	0.28	0.21	0.04	0.17	0.15	0.29	0.06	1.00	0.25	0.20	0.17
9	0.06	0.03	0.00	0.00	0.20	0.04	0.00	0.25	1.00	0.26	0.38
10	0.06	0.15	0.01	0.09	0.04	0.20	0.00	0.20	0.26	1.00	0.12
11	0.00	0.00	0.00	0.01	0.18	0.03	0.01	0.17	0.38	0.12	1.00



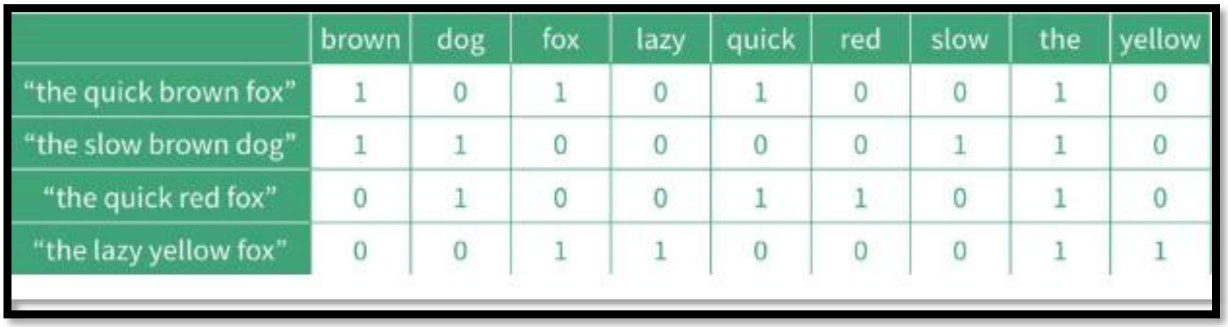
So basically if a sentence is connected to many sentences then that statement should be into our summary.

● **Latent Semantic Analysis Method (LSA):**

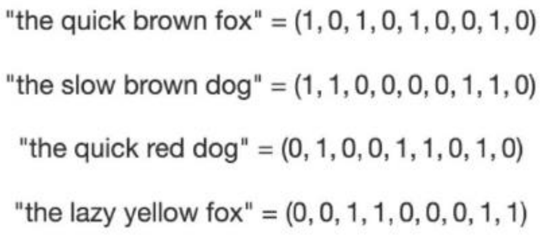
The word latent means “hidden”. Features that are hidden in the data which cannot be directly measured. They are essential to data, but are not original features of data set. LSA is a NLP technique. The aim of LSA is to create representations of text data in terms of these topics or latent features. We decrease the dimensionality of the original data set.



In **Document Term Matrix** , text documents can be represented as points also known as vectors. Example



	brown	dog	fox	lazy	quick	red	slow	the	yellow
“the quick brown fox”	1	0	1	0	1	0	0	1	0
“the slow brown dog”	1	1	0	0	0	0	1	1	0
“the quick red fox”	0	1	0	0	1	1	0	1	0
“the lazy yellow fox”	0	0	1	1	0	0	0	1	1



"the quick brown fox" = (1, 0, 1, 0, 1, 0, 0, 1, 0)

"the slow brown dog" = (1, 1, 0, 0, 0, 0, 1, 1, 0)

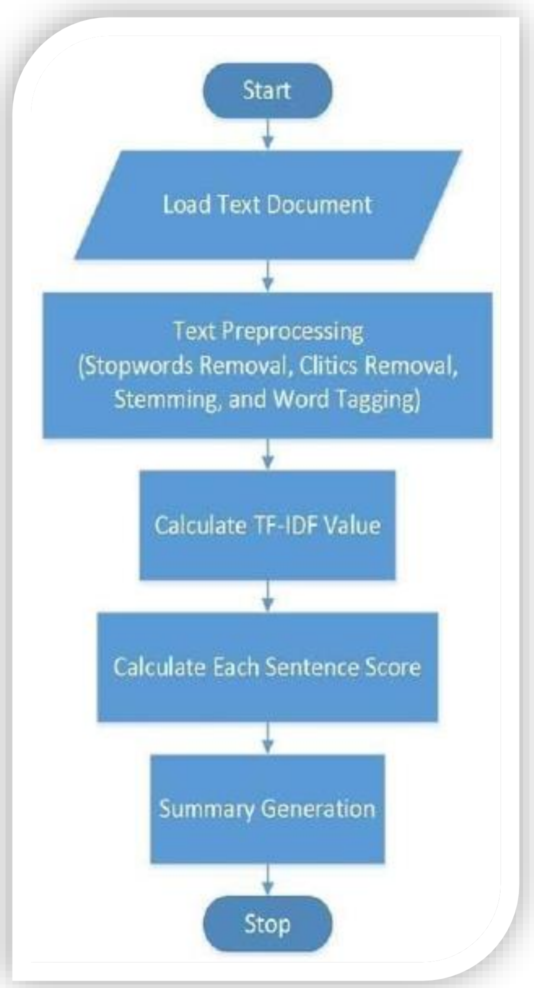
"the quick red dog" = (0, 1, 0, 0, 1, 1, 0, 1, 0)

"the lazy yellow fox" = (0, 0, 1, 1, 0, 0, 0, 1, 1)

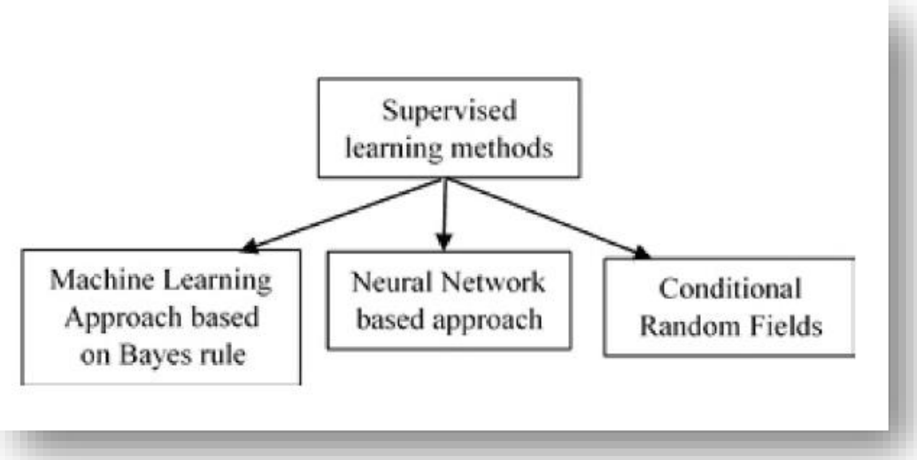
- **Singular Value Decomposition** reduces the dimensionality of original data set by encoding using these latent features. These features represent topics in the original text data. Then the topic encoded data is formed.

● **Frequency Based Approach:**

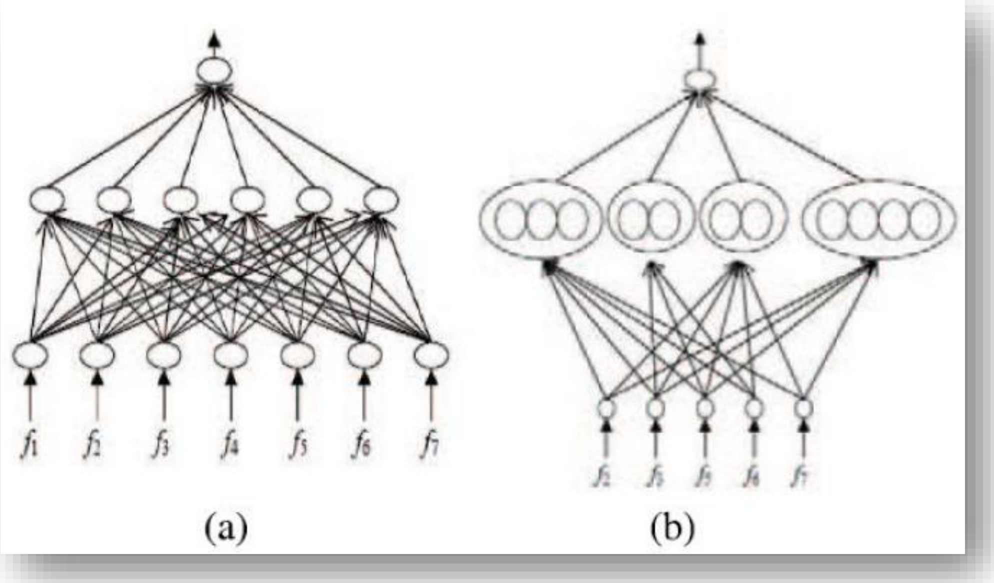
The occurrence of words is compared and counted. There are 2 ways to do it viz. word probability and TFIDF (Term Frequency Inverse Document Frequency) . Word probability counts the occurrence of a word in a sentence and then divides it by the total number of words in a document. The words' whose probability turns out to be maximum is finalized in the summary. TFIDF searches for common words that should not be a part of the final summary. The same topics or words with same meaning using various features are clustered.


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

2. Supervised Learning Methods: The model is trained here and they know the difference between a summary sentence and a non-summary sentence. Therefore we can perform automatic text summarization using various methods.


Supervised learning methods
Machine Learning Approach based on Bayes rule
Neural Network based approach
Conditional Random Fields

1) **Machine Learning Approach based on Bayes rule**:
   We train our model and the sentences are classified in different categories. If a new statement is taken then its probability is calculated according to different categories
2) **Neural Network based approach:**



(a) (b)

In this approach, important sentences are identified through neural networks. A two-level neural network is used along with back propagation. Initial step involves recognizing the data using a ML approach and then collect features of a sentence in test sets as well as train sets which is then passed on to neural networks to rank the sentences. The disadvantage with this method is that usage of words and sentences are ignored while giving weights which leads to low accuracy.

3) Conditional Random Field

This approach is a statistical modeling approach. Conditional random fields focus on ML to provide a structured outcome. The main benefit of using this approach is that it identifies correct features and gives a better summary. And one of the main disadvantages of the method is that it requires domain specific training, a training cannot be applied to any general document.

CHALLENGES AND FURTHER RESEARCH

To evaluate the summaries of any document is very tedious task. It takes a lot of time to evaluate a large document. The main problem comes because there is no possibility to build a standard against which results can be tested.People have completely different kind of thinking and can develop different thinking approaches to reach upto a summarization.There already exists an approach for automatically summarizing using paraphrases.Nowadays most text summarization takes the extractive text summarization approach.This approach performs by selective extensive and important sentences from an lengthy document. It will create an concise document containing the gist and relevant information only. Though the human beings can copy and paste relevant information from the extensive document but still sometimes it can miss out various information or on the other side there is high probability that it may join disparate data into one sentence.Also further developments and research can improve the various methods and can also help in creating novel techniques.

NEW APPLICATION AREAS

There are various areas where this summarization techniques are becoming very beneficial. Amongst them the main four areas include multiple languages, hybrid sources, multiple documents and multimedia. In all these four areas summarizers must be able to deal with disparate documents such as HTML and XML. They must be able to deploy the relevant and concise information into the summarized document. For further developments in this summarization containing multiple languages and hybrid sources is less mature.

- **MULTIPLE LANGUAGES**

The high-quality translation of machine where input is unrestricted and is out of reach. The feasible and logical solution to this kind of problem is filtering mechanism. Users can apply a filter to produce a single summarized document from multiple documents available. After getting a view on that document they can then decide to further include translations or not**.**

- **HYBRID SOURCES**

For this application, the summarization contains information from both the formatted data as well as unformatted text. An example of that summary is the information of a baseball player related to matches, scores, average, best run rate to the data of news stories related to that player. This application is recently developed and is novel, very few researches is done under this application**.**

- **MULTIPLE DOCUMENTS**

The collection of data here ranges from gigabytes to bytes so various different approaches are taken into consideration for various data. Whereas each method includes the analysis of the document and then fusing various information from the documents in synthesis phase. Summarizers still carry out various approaches such as segregation, elimination and generalization approaches. Simply aggregation of various documents would not be suffice because there can too many summaries with various redundant data.

Summarizers are having great acumen and can easily discern the differences among the documents by simply comparing them.They can easily find out what’s common, what’s different and how the documents discern.For example the same news may appear in various newspapers or newsletters, summarizers perform their work by discerning the differences and eliminate the redundant information to provide a concise document**.**

- **MULTIMEDIA**

Although the research is in very rudimentary stage , the application of multimedia makes this most important novel application in summarization.Techniques have their gamut from cross media information during transformation or analysis or during synthesis. The techniques which are used currently for obtaining the salient information includes the audio, video, content analysis and various other.For example the current goal is to identify the content of videos using pattern recognition to predict the important and salient events such as appearances, incidents, fights, main characters and so on.

CONCLUSION

Overall, the research on this summarization is still young. There is some general agreement that there is more need for the evaluation but many challenges still remain untouched. This review has shown the various methods which can be implemented for the extractive text summarization. The extractive text summarization is less redundant, very much cohesive. The aim is to give a complete understanding of the usefulness of the process, various approaches, further research's and different approaches of the extractive text summarization. Although the research in this field started many years back but still various unprecedented and novel researches are impending. Providing the concise view of the document is very useful and meticulous as it saves the time. Nevertheless, many of the techniques discussed here there exists still vast knowledge which is to be debunked to conquer the huge information universes looming.

REFERENCE

https://www.researchgate.net/publication/317420253_A_survey_on_extractive_text_summarization

https://towardsdatascience.com/a-quick-introduction-to-text-summarization-in-machine-learning-3d27ccf18a9f

https://blog.floydhub.com/gentle-introduction-to-text-summarization-in-machine-learning/