Automatic Text Summarization Using Natural Language Processing

Pratibha Devihosur[1], Naseer R[2]

[1] *M.Tech. student, Dept. of Computer Science and Engineering, B.I.E.T College, Karnataka, India*
[2] *Assistant Professor, Dept. of Computer Science and Engineering, B.I.E.T College, Karnataka, India*

---***---

Abstract - *Automatic Text Summarization is the technique by which the huge parts of content are retrieved. In this paper The Automatic Text Summarization plays out the summarization task by unsupervised learning system. The significance of a sentence in info content is assessed by the assistance of Simplified Lesk calculation. As an online semantic lexicon WordNet is utilized. Word Sense Disambiguation (WSD) is a critical and testing system in the territory of characteristic dialect handling (NLP). A specific word may have distinctive significance in various setting. So the principle task of word sense disambiguation is to decide the right feeling of a word utilized as a part of a specific setting. To begin with, Automatic Text Summarization assesses the weights of the considerable number of sentences of a content independently utilizing the Simplified Lesk calculation and orchestrates them in diminishing request as indicated by their weights. Next, as indicated by the given level of rundown, a specific number of sentences are chosen from that requested rundown. The proposed approach gives best outcomes up to 50% summarization of the first content and gives attractive outcome even up to 25% outline of the first content.*

Key Words: Automatic Text Summarization, wordnet, Streamlined lesk Calculation, Word Sense Disambiguation

1. INTRODUCTION

Automatic Text Summarization [1] H. Dalianis, [2] M. Hassel, is the plan to get an important data from a huge amount of information. The amount of data accessible on internet is increasing every day so it turns space and time expanding matter to deal with such huge amount of information. So, managing that large amount of data is makes a major problem in different and real data taking care of uses. The Automatic Text Summarization undertaking makes the users simpler for various Natural Language applications, like, Data Recovery, Question Answering or content decreasing etc. Automatic Text Summarization assumes an inescapable part by creating significant and particular data from a lot of information.

Filtering from heaps of reports can be troublesome and tedious. Without a summary or rundown, it can take minutes just to make sense of what the people will discuss in a paper or report. So the Automatic Text Summarization that concentrates a sentence from a content record, figures out which are the most imperative, and returns them in a readable and organized way. Automatic Text Summarization is a piece of the field natural language processing, which is the manner by which the PCs can break down, and get importance from human dialect.

Automatic Text Summarization that uses the classifier structure and its rundown modules to look over huge amount of reports and returns the sentences that are helpful for producing a summary. Programmed outline of content works by taking the overlapping sentences and synonymous or sense from wordnet most overlapping sentences are considered as high score words [3] H. Seo, H. Chung, H. Rim, S. H., Myaeng, S. Kim, [4] A. J. Cañas , A. Valerio, J. Lalinde-Pulido, M. Carvalho, M. Arguedas. The higher recurrence words are considering most worth. And the top most worth words and are taking from the content and sorted according to its recurrence and generate a summary.

Lesk algorithm [5] S. Banerjee, T. Pedersen, [6]M. Lesk, is used for evaluating the waits for the input text using online semantic dictionary wordnet and it also uses the word sense disambiguation to identifying the most overlapping sentences in the input content that type of sentences are called equivocal words. Those types of words or sentences are having higher recurrences during the summarization.

In numerous normal dialects, a word can speaks to numerous implications/sense, and such type of word is called a homograph. WSD is the route toward making sense of which sentiment a homograph is used as a piece of given setting. WSD is a long-standing issue in computational linguistics, and has a come bonafide application including machine elucidation, information extraction, and information recuperation. Gener-accomplice, WSD use the setting of a word for its sense disambiguation, and setting information can begin from either clarified/unannotated content or other learning resources, for instance, responsive view point word expert, parallel corpora.

1.1 Natural Language Processing

Natural Language Processing technique using the nltk for building a main stage for python projects to work with human dialect information. This gives the easier to-utilize by giving the interfaces to one or more than 40 corpora and lexicon assets, for libraries for characterization, for splitting paragraphs sentences, to get its original form of words, labeling, parsing, and vocabulary thinking, and wrappers for modern thinking quality common dialect handling libraries, and for dynamic discourse discussion.

The NLTK is going to use an enormous tool compartment, and is going for make a favour for people with the entire common dialect handling procedure. This will going to help people with all thing from part sentences from passages, to part up words, seeing the syntactic components of those words, marking the essential topics, doing this is helps to your machine b appreciating what really matters to the substance.

1.2 Streamlined Lesk Calculation

Calculation 1: This calculation compresses a single report content utilizing unsupervised learning approach. In This approach , the heaviness of each sentence in a content is determined utilizing Improved Lesk calculation and WordNet. The summarization procedure is performed as indicated by the given level of summarization [4]A. J. Cañas , A. Valerio, J. Lalinde-Pulido, M. Carvalho, M. Arguedas.

Info: Single-report input content.

Yield: Summarized content.

Step 1: The list of distinct sentences of the content is prepared.
Step 2: Repeat steps 3 to 7 for each of the sentences.
Step 3: A sentence is gotten from the list.
Step 4: Stop words are expelled from the sentence as they don't take an interest straightforwardly in sense assessment system.
Step 5: Glosses(dictionary definitions) of all the important words are extricated utilizing the WordNet.
Step 6: Intersection is performed between the sparkles and the information content itself.
Step 7: Summation of all the crossing point comes about speaks to the heaviness of the sentence.
Step 8: Weight appointed sentences are arranged in descending request concerning their weights.
Step 9: Desired number of sentences are chosen by the level of summarization.
Step 10: Selected sentences are re-orchestrated by their real sequency in the info content.
Step 11: Stop.

1.3 Advantages

- Reading the whole document, dismembering it and isolating the critical thoughts from the crude content require some serious energy and exertion. Perusing a document of 600 words can take no less than 10 minutes. Programmed outline programming condense writings of 500-5000 words in a brief instant. This enables the client to peruse less information yet get the most essential data and make strong conclusion.
- It reduces the human effort while creating a synopsis. A few vital products compress records as well as website pages.
- The persons quickly determine which points are imported for reading.

2. PROPOSED SYSTEM

In the Automatic Text summarization, we are using a solitary or single input content is going to outlined by the given rate of summarization utilizing unsupervised learning. In any case, the streamlined lesk's computation is associated with each of the sentences to find the guarantees of each sentence. After that, sentences with induced weights are composed in sliding solicitation concerning their weights. Presently as per a particular rate of summarization at a specific occurrence, certain quantities of sentences are chosen as an outline.

The proposed computations, abridges solitary or single report content utilizing unsupervised learning approach. Here, the heaviness of every sentence in a substance is resolved using streamlined Lesk's computation and wordnet. After that, summarization procedure is performed as indicated by the given rate of synopsis. In which, we are taking solitary info content and display summarization as yield. First info content is passed, to the lesk' computation and wordnet, where the weights of each sentences of the content are inferred utilizing and semantic investigation of the concentrates are performed. Next, weight doled out sentences is passed to derive the final summary according to the percentage of synopsis, where the last abridged outcome is assessed as and showed.

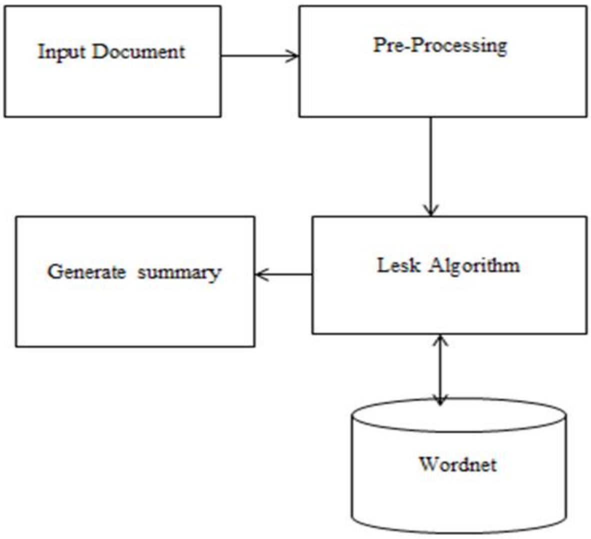


Fig -1: Overall Representation for Automatic Text Summarization Using Natural Language Processing.

1.2 System Architecture Of The Proposed System

The proposed system depicts the three stages for Automatic Text Summarization and they are listed below.

Stage 1: Data Pre-Processing

Stage 2: Evaluation of weights

Stage 3: Summarization

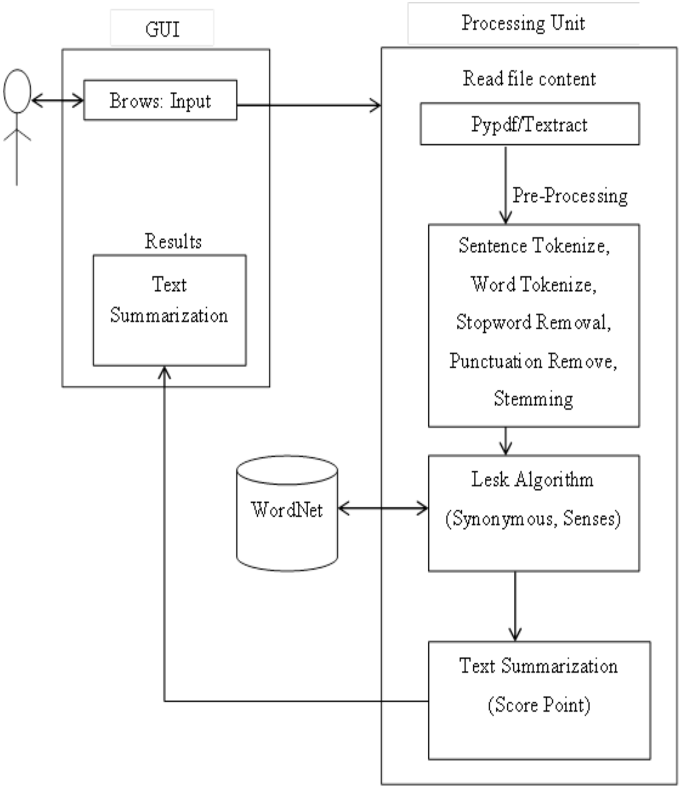


Fig -2: System Architecture For Automatic Text Summarization Using Common Handling Dialect.

Stage 1: Data Pre-Processing

Programmed record outline generator is for clearing the undesirable things which exist in the substance. Henceforth it will additionally process it will performing sentence part, tokenisation, empty stopword, clear accentuation and perform stemming.

Stage 2: Evaluation of weights

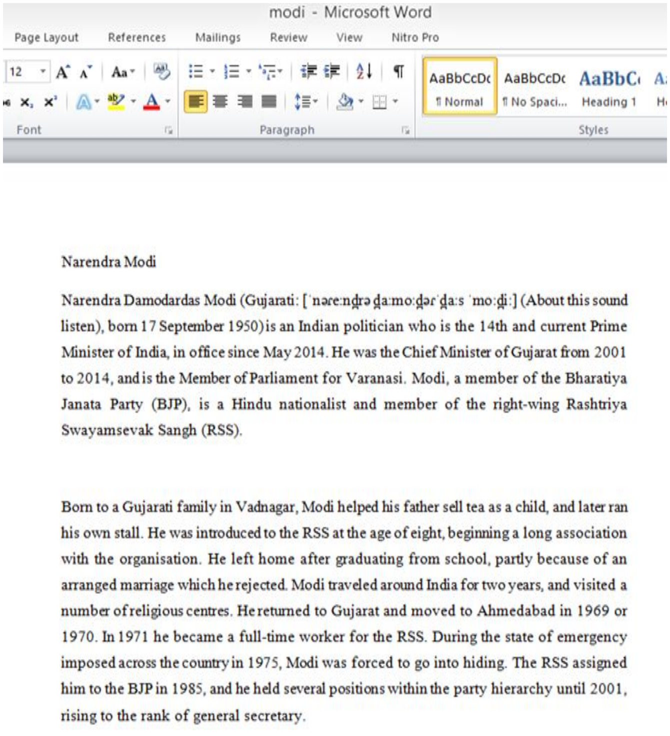
This stage processes the repeat of the sentences of a substance utilizing lesk count and wordnet. In the first place finding the total number of spreads between a particular and the radiance this philosophy is performed for the all n number of sentences. By then once-over a particular sentence of the substance is set up for each of the sentences. A sentence is snatched from the once-over. Stopwords are removing from the sentence as they don't take an intrigue particularly in sense task method. Sparkles of each vital word removed using wordnet. Union is performed between the sparkles and the data content itself. Once-over of all the intersection guide comes to fruition talks toward the largeness of the sentence.

Stage 3: Summarization

This stage evaluates the last outline of a substance and the introductions the yield, which is surveyed at the period of arranging the sentences. In the first place it select the once-over of weight named sentences are planned in jumping demand concerning their weights. Pined for number of sentences is picked by the rate of summary. Picked sentences are re-composed by their genuine gathering in the information content. The modified substance summary will gathers a substance without depending upon the association of the substance, rather than the semantic information lying in the sentence. Modified substance once-over is without vernacular. To remove the semantic information from a sentence, only a semantic word reference in the last vernacular is required.

3. OUTPUT AND DISCUSSION

Trial consequences of the venture for pre-preparing, assessment of the weights and showing the outline stage are executed. The results of following of these stages are represented in roar figure. In this approach we are using the word document and pdf document as input source.

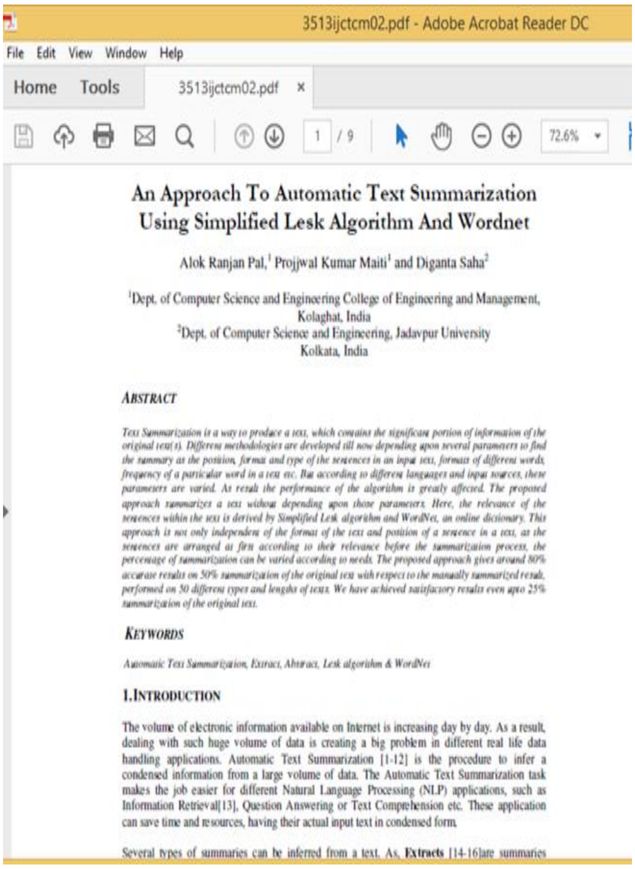


Narendra Modi

Narendra Damodardas Modi (Gujarati: [ˈnəreːnd̪rə d̪aːmoːd̪ərˈd̪aːs ˈmoːd̪iː] (About this sound listen), born 17 September 1950) is an Indian politician who is the 14th and current Prime Minister of India, in office since May 2014. He was the Chief Minister of Gujarat from 2001 to 2014, and is the Member of Parliament for Varanasi. Modi, a member of the Bharatiya Janata Party (BJP), is a Hindu nationalist and member of the right-wing Rashtriya Swayamsevak Sangh (RSS).

Born to a Gujarati family in Vadnagar, Modi helped his father sell tea as a child, and later ran his own stall. He was introduced to the RSS at the age of eight, beginning a long association with the organisation. He left home after graduating from school, partly because of an arranged marriage which he rejected. Modi traveled around India for two years, and visited a number of religious centres. He returned to Gujarat and moved to Ahmedabad in 1969 or 1970. In 1971 he became a full-time worker for the RSS. During the state of emergency imposed across the country in 1975, Modi was forced to go into hiding. The RSS assigned him to the BJP in 1985, and he held several positions within the party hierarchy until 2001, rising to the rank of general secretary.

Fig -3: Input File for Word Document.


3513ijctcm02.pdf - Adobe Acrobat Reader DC
File Edit View Window Help
Home Tools 3513ijctcm02.pdf ×
1 / 9
72.6%

An Approach To Automatic Text Summarization Using Simplified Lesk Algorithm And Wordnet

Alok Ranjan Pal,[1] Projjwal Kumar Maiti[1] and Diganta Saha[2]

[1]Dept. of Computer Science and Engineering College of Engineering and Management, Kolaghat, India
[2]Dept. of Computer Science and Engineering, Jadavpur University
Kolkata, India

ABSTRACT

Text Summarization is a way to produce a text, which contains the significant portion of information of the original text(s). Different methodologies are developed till now depending upon several parameters to find the summary as the position, format and type of the sentences in an input text, formats of different words, frequency of a particular word in a text etc. But according to different languages and input sources, these parameters are varied. As result the performance of the algorithm is greatly affected. The proposed approach summarizes a text without depending upon those parameters. Here, the relevance of the sentences within the text is derived by Simplified Lesk algorithm and WordNet, an online dictionary. This approach is not only independent of the format of the text and position of a sentence in a text, as the sentences are arranged at first according to their relevance before the summarization process, the percentage of summarization can be varied according to needs. The proposed approach gives around 80% accurate results on 50% summarization of the original text with respect to the manually summarized result, performed on 50 different types and lengths of texts. We have achieved satisfactory results even upto 25% summarization of the original text.

KEYWORDS

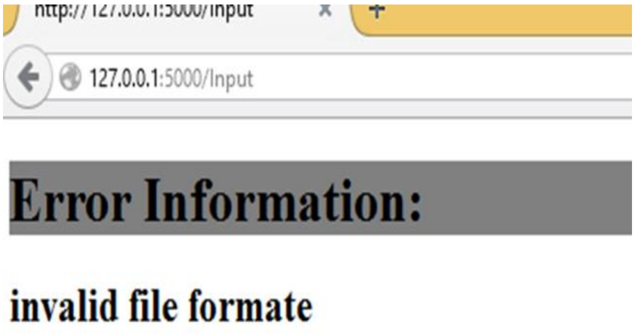
Automatic Text Summarization, Extract, Abstract, Lesk algorithm & WordNet

1.INTRODUCTION

The volume of electronic information available on Internet is increasing day by day. As a result, dealing with such huge volume of data is creating a big problem in different real life data handling applications. Automatic Text Summarization [1-12] is the procedure to infer a condensed information from a large volume of data. The Automatic Text Summarization task makes the job easier for different Natural Language Processing (NLP) applications, such as Information Retrieval[13], Question Answering or Text Comprehension etc. These application can save time and resources, having their actual input text in condensed form.

Several types of summaries can be inferred from a text. As, **Extracts** [14-16]are summaries

Fig -4: Input File For pdf Document.



Error Information:

invalid file formate

Fig -5: Input File For Other than pdf or Word Document.

If info record is other than .pdf or .docx organize blunder will show like invalid data and invalid document design

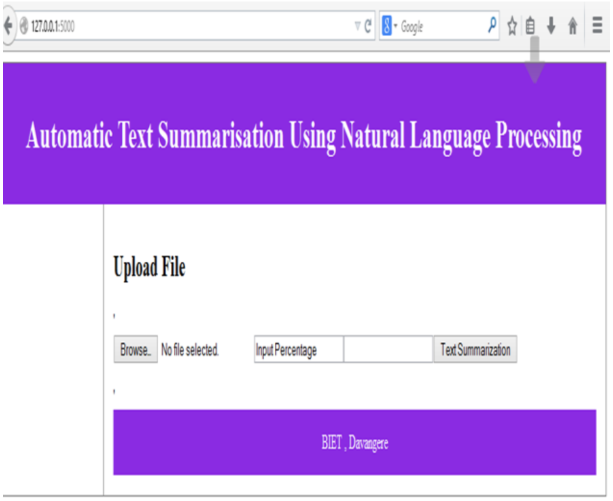

127.0.0.1:5000
Google
Automatic Text Summarisation Using Natural Language Processing
Upload File
Browse... No file selected.
Input Percentage
Text Summarization
BIET , Davangere

Fig -6: User Interface Form.

The User interface shape comprises of 2 catches, Browse and Text Summarization. The Brows catch will open a document to compress and Text Summarization is to begin procedure of the summarization.

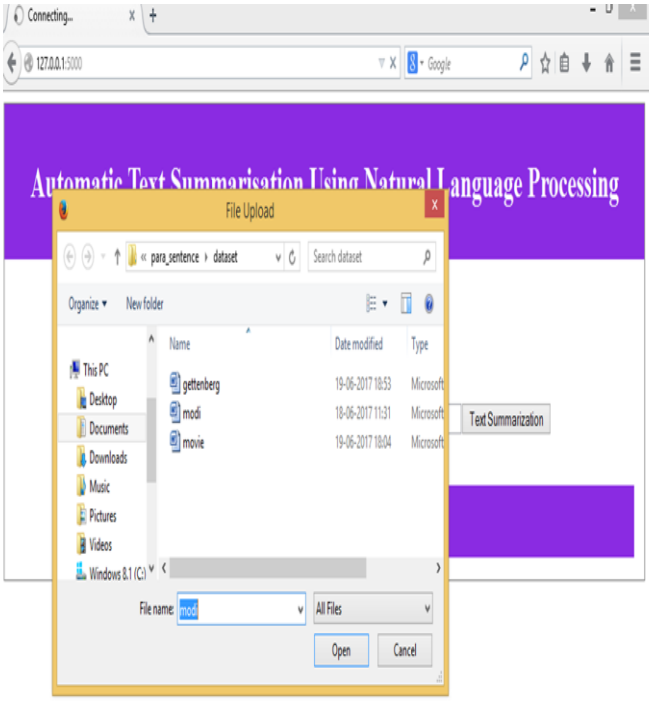
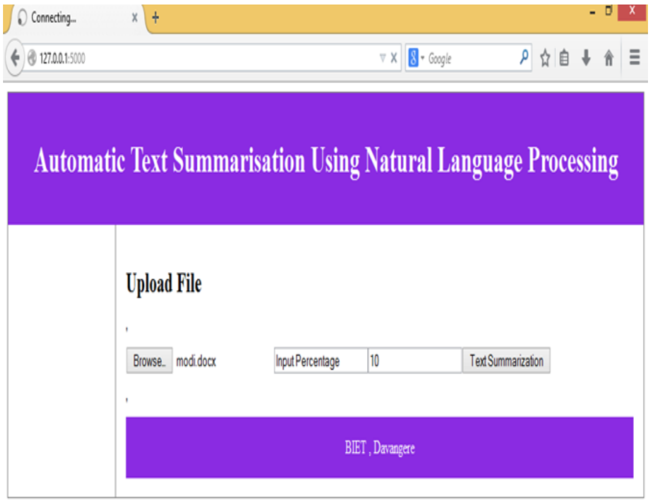


Fig -7: Brows Catch will Brows the file.

The brows catch will select the input file to give summarization process



Automatic Text Summarisation Using Natural Language Processing

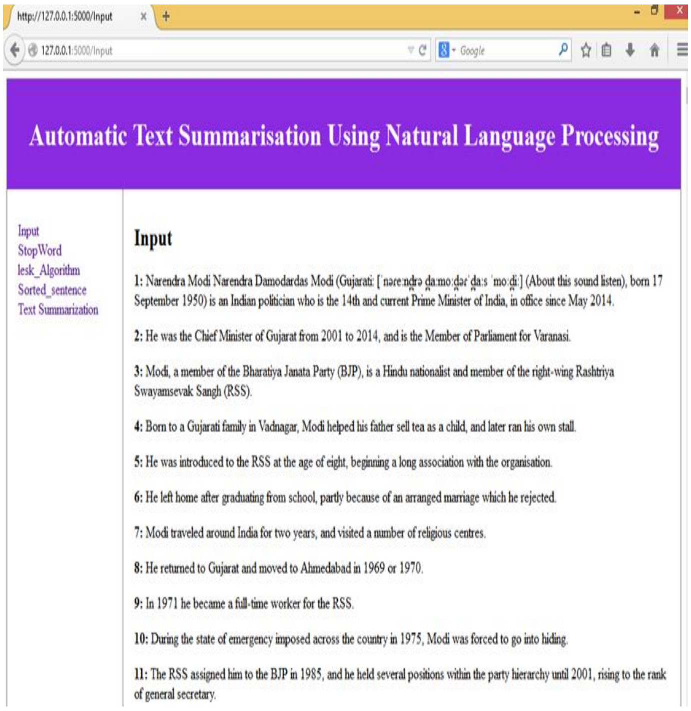
Upload File

Browse... modi.docx Input Percentage 10 Text Summarization

BIET , Davangere

Fig -8: Input Percentage.

After that client needs to give rate, how much summary need to show.



http://127.0.0.1:5000/Input
127.0.0.1:5000/Input
Google
Automatic Text Summarisation Using Natural Language Processing
Input
StopWord
lesk_Algorithm
Sorted_sentence
Text Summarization

Input

1: Narendra Modi Narendra Damodardas Modi (Gujarati: [ˈnəreːnd̪rə d̪aːmoːd̪ər ˈd̪aːs ˈmoːd̪iː] (About this sound listen), born 17 September 1950) is an Indian politician who is the 14th and current Prime Minister of India, in office since May 2014.

2: He was the Chief Minister of Gujarat from 2001 to 2014, and is the Member of Parliament for Varanasi.

3: Modi, a member of the Bharatiya Janata Party (BJP), is a Hindu nationalist and member of the right-wing Rashtriya Swayamsevak Sangh (RSS).

4: Born to a Gujarati family in Vadnagar, Modi helped his father sell tea as a child, and later ran his own stall.

5: He was introduced to the RSS at the age of eight, beginning a long association with the organisation.

6: He left home after graduating from school, partly because of an arranged marriage which he rejected.

7: Modi traveled around India for two years, and visited a number of religious centres.

8: He returned to Gujarat and moved to Ahmedabad in 1969 or 1970.

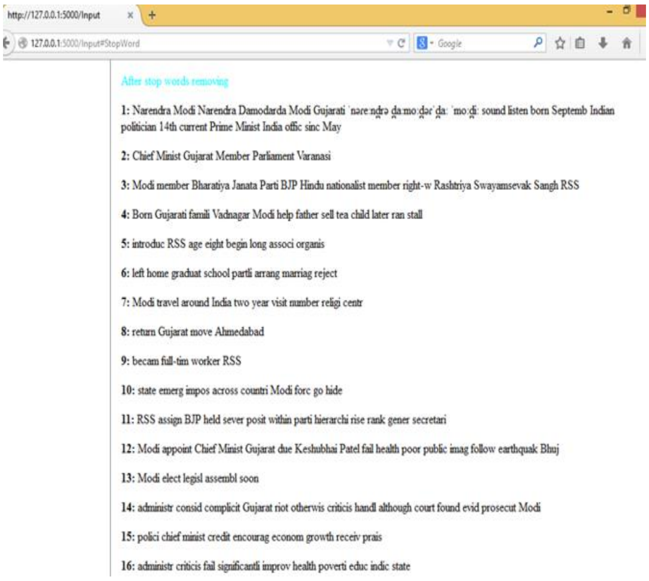
9: In 1971 he became a full-time worker for the RSS.

10: During the state of emergency imposed across the country in 1975, Modi was forced to go into hiding.

11: The RSS assigned him to the BJP in 1985, and he held several positions within the party hierarchy until 2001, rising to the rank of general secretary.

Fig -9: Brows Catch will Brows the file.

Therefore In Pre-handling the tokenization is parts the contribution as sentences or words.



After stop words removing

1: Narendra Modi Narendra Damodarda Modi Gujarati ˈnəreːnd̪rə d̪aːmoːd̪ərˈd̪aː ˈmoːd̪iː sound listen born Septemb Indian politician 14th current Prime Minist India offic sinc May

2: Chief Minist Gujarat Member Parliament Varanasi

3: Modi member Bharatiya Janata Parti BJP Hindu nationalist member right-w Rashtriya Swayamsevak Sangh RSS

4: Born Gujarati famili Vadnagar Modi help father sell tea child later ran stall

5: introduc RSS age eight begin long associ organis

6: left home graduat school partli arrang marriag reject

7: Modi travel around India two year visit number religi centr

8: return Gujarat move Ahmedabad

9: becam full-tim worker RSS

10: state emerg impos across countri Modi forc go hide

11: RSS assign BJP held sever posit within parti hierarchi rise rank gener secretari

12: Modi appoint Chief Minist Gujarat due Keshubhai Patel fail health poor public imag follow earthquak Bhuj

13: Modi elect legisl assembl soon

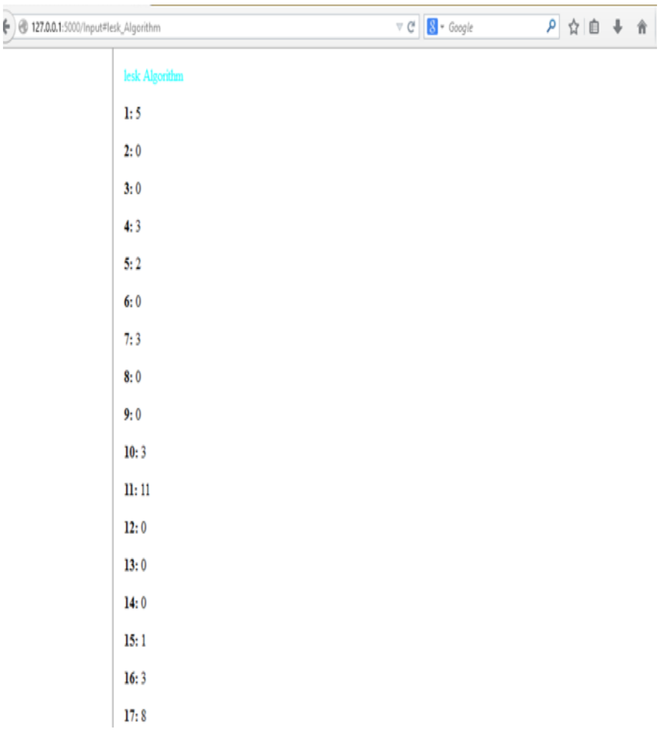
14: administr consid complicit Gujarat riot otherwis criticis handl although court found evid prosecut Modi

15: polici chief minist credit encourag econom growth receiv prais

16: administr criticis fail significantli improv health poverti educ indic state

Fig -9: Brows Catch will Brows the file.

After it will list the sentences in the wake of evacuating the stopwords.



lesk Algorithm

1: 5

2: 0

3: 0

4: 3

5: 2

6: 0

7: 3

8: 0

9: 0

10: 3

11: 11

12: 0

13: 0

14: 0

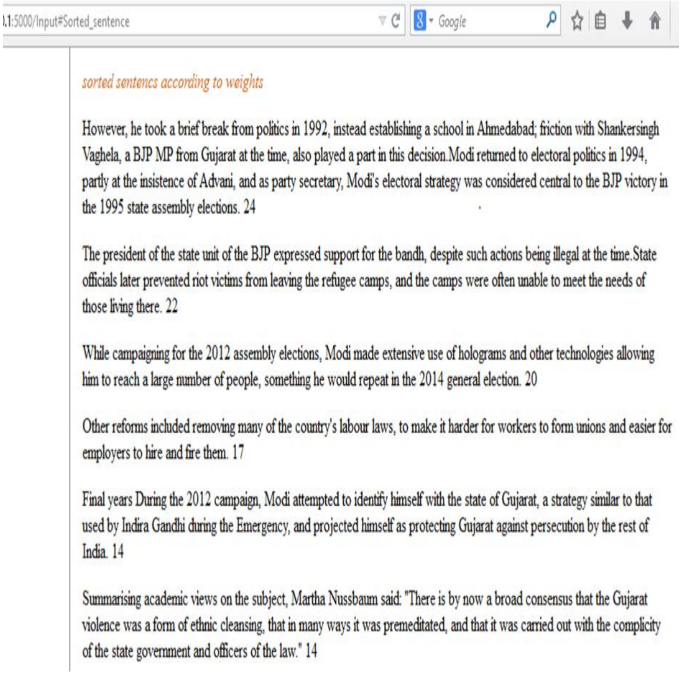
15: 1

16: 3

17: 8

Fig -10: Lesk Calculation.

It will show weights for the input sentences according to its most important sentences



sorted sentencs according to weights

However, he took a brief break from politics in 1992, instead establishing a school in Ahmedabad; friction with Shankersingh Vaghela, a BJP MP from Gujarat at the time, also played a part in this decision.Modi returned to electoral politics in 1994, partly at the insistence of Advani, and as party secretary, Modi's electoral strategy was considered central to the BJP victory in the 1995 state assembly elections. 24

The president of the state unit of the BJP expressed support for the bandh, despite such actions being illegal at the time.State officials later prevented riot victims from leaving the refugee camps, and the camps were often unable to meet the needs of those living there. 22

While campaigning for the 2012 assembly elections, Modi made extensive use of holograms and other technologies allowing him to reach a large number of people, something he would repeat in the 2014 general election. 20

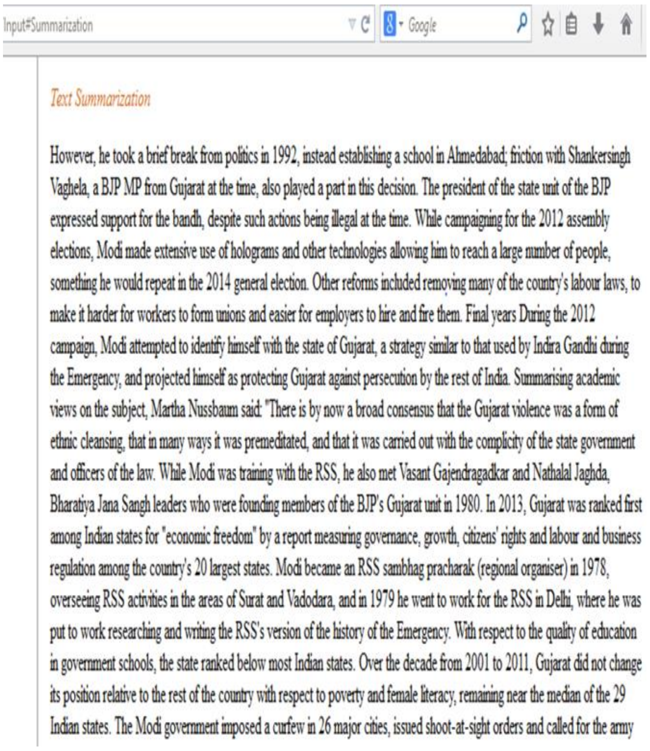
Other reforms included removing many of the country's labour laws, to make it harder for workers to form unions and easier for employers to hire and fire them. 17

Final years During the 2012 campaign, Modi attempted to identify himself with the state of Gujarat, a strategy similar to that used by Indira Gandhi during the Emergency, and projected himself as protecting Gujarat against persecution by the rest of India. 14

Summarising academic views on the subject, Martha Nussbaum said: "There is by now a broad consensus that the Gujarat violence was a form of ethnic cleansing, that in many ways it was premeditated, and that it was carried out with the complicity of the state government and officers of the law." 14

Fig -11: Brows Catch will Brows the file.

After it demonstrates the arranged sentences According to weights.



Text Summarization

However, he took a brief break from politics in 1992, instead establishing a school in Ahmedabad; friction with Shankersingh Vaghela, a BJP MP from Gujarat at the time, also played a part in this decision. The president of the state unit of the BJP expressed support for the bandh, despite such actions being illegal at the time. While campaigning for the 2012 assembly elections, Modi made extensive use of holograms and other technologies allowing him to reach a large number of people, something he would repeat in the 2014 general election. Other reforms included removing many of the country's labour laws, to make it harder for workers to form unions and easier for employers to hire and fire them. Final years During the 2012 campaign, Modi attempted to identify himself with the state of Gujarat, a strategy similar to that used by Indira Gandhi during the Emergency, and projected himself as protecting Gujarat against persecution by the rest of India. Summarising academic views on the subject, Martha Nussbaum said: "There is by now a broad consensus that the Gujarat violence was a form of ethnic cleansing, that in many ways it was premeditated, and that it was carried out with the complicity of the state government and officers of the law. While Modi was training with the RSS, he also met Vasant Gajendragadkar and Nathalal Jaghda, Bharatiya Jana Sangh leaders who were founding members of the BJP's Gujarat unit in 1980. In 2013, Gujarat was ranked first among Indian states for "economic freedom" by a report measuring governance, growth, citizens' rights and labour and business regulation among the country's 20 largest states. Modi became an RSS sambhag pracharak (regional organiser) in 1978, overseeing RSS activities in the areas of Surat and Vadodara, and in 1979 he went to work for the RSS in Delhi, where he was put to work researching and writing the RSS's version of the history of the Emergency. With respect to the quality of education in government schools, the state ranked below most Indian states. Over the decade from 2001 to 2011, Gujarat did not change its position relative to the rest of the country with respect to poverty and female literacy, remaining near the median of the 29 Indian states. The Modi government imposed a curfew in 26 major cities, issued shoot-at-sight orders and called for the army

Fig -12: Brows Catch will Brows the file.

Finally it will show the section of sentences constrained by rate.

4. CONCLUSION AND FUTURE SCOPE

Automatic Text Summarization approach depends on upon the semantic data of the concentration in a substance. So this way, gathered parameters like approaches, spots of different substances are not considered. In this recommendation, Lesk mean for word sense disambiguation by utilizing the vocabulary definitions to the electronic dictionary information base on utilizing wordnet. This goal is clear from covering sentence, couple of fusing words that give the setting of the word, in this not utilizing the late using the definitional shines of those words, other than those of words related to them through with the unmistakable relations portrayed in wordnet. So furthermore we are endeavoring to use other enlightening record away by wordnet for each word. For example, design sentences and identical words et cetera.

Among future work is the use of all the more balanced gathering to upgrade occurs additionally. Attempting diverse things with more tongue specific segments for instance, morphological parsers, printed entailment and anaphoric assurance is an open research for more updates later on. Programmed content summarisations should be possible for various archives. Client can be given an office to print the record from the interface specifically. A point of confinement to re-synopsis alternative perhaps included for record Shorter long. Additional line hole acquired in the outline can be evacuated. Spare as choice can be added to the application for the client to spare the synopsis in various arrangement.

REFERENCES

[1] H. Dalianis, "SweSum – A Text Summarizer for Swedish," Technical report TRITA-NA-P0015,IPLab-174, NADA, KTH, October 2000.D.

[2] M. Hassel,"Resource Lean and Portable Automatic Text Summarization. PhD thesis, Department of Numerical Analysis and Computer Science," Royal Institute of Technology, Stockholm, Sweden 2007.

[3] H. Seo, H. Chung, H. Rim, S. H., Myaeng, S. Kim, "Unsupervised word sense disambiguation using WordNet relatives," Computer Speech and Language, Vol. 18, No. 3, pp. 253-273, 2004.

[4] A. J. Cañas , A. Valerio, J. Lalinde-Pulido, M. Carvalho, M. Arguedas, "Using WordNet for Word Sense Disambiguation to Support Concept Map Construction," String Processing and Information Retrieval, pp. 350-359, 2003.

[5] S. Banerjee, T. Pedersen,"An adapted Lesk algorithm for word sense disambiguation usingWordNet," In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, February, 2002.

[6] M. Lesk,"Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone," Proceedings of SIGDOC, 1986.

BIOGRAPHIES

Pratibha Devihosur (M.Tech). student, Dept. of Computer Science and Engineering, B.I.E.T College, Karnataka, India.

Naseer R Assistant Professor, Dept. of Computer Science and Engineering, B.I.E.T College, Karnataka, India.