**DATA CURATION**

**Vedanti Vyawaharea, Rakshanda Mahajanb, Satvik Dandalec,**

**Prathamesh Dhawaled**

*a – UG student of Vishwkarma Institute of Technology, Pune 411037*

*b – UG student of Vishwkarma Institute of Technology, Pune 411037*

*c – UG student of Vishwkarma Institute of Technology, Pune 411037*

*d – UG student of Vishwkarma Institute of Technology, Pune 411037*

**Abstract**

In this project an attempt was made on Extract Technology based on research. Summarization Model is an agnostic content summarization technology that automatically parses news, information, documents and many more into relevant and contextually accurate abbreviated summaries. Summarization has been proven to be a useful and effective technique supporting data analysis of large amounts of data.

**Keywords**

Important words , frequency , score.

**INTRODUCTION**

Data Summarization is a simple term for a short conclusion of a big theory or a paragraph. This is something where you write the code and in the end, you declare the final result in the form of summarizing data. Summarization is exceptionally good in summarizing the text i.e. important part of the paragraph automatically without changing meaning of the paragraph. It will summarize text, emails, news, speech, etc into weighted lists of keywords and key phrases extracting the primary contextual sentence highlight of how the keyword / key phrase has been used . Uniquely positioned for web services, Auto Summarization is immediately capable of consuming documents of any length and subject matter, distilling the precise, contextual meaning of the content into keyword and key phrase summary formats.

**ALGORITHM**

1. Create a table named word frequency, word list and also include stop words. Transverse through the paragraph:

1. If current word is not a stop word then increment frequency by 1.
2. Finding the average frequency of words (total frequency/total no of words).
3. If frequency used>average then store the word in word list.

2. Create a table for sentence frequency and store each frequency =0.

3. Transverse through the paragraph:

1. If the sentence contains a number then freq= freq+1.
2. If the sentence contains imp word then freq=freq+1 for each word.
3. If the sentence contain wordlist then frequency=freq+1 for each word.

* These all parameters are used to give score to the sentence.

4. Find total words in paragraph.

5. Find average length=total no. of words/no. in sentence.

6. For each sentence: Final score = Initial score + length of the sentence

7. Find total frequency=add all final score of the each sentence.

8. Find average frequency= total frequency/no. of sentence.

9. If sentence frequency > avg frequency :

Extract that sentence.

10. Print the extract.

**CONCLUSION**

It is required to extract a paragraph rather than whole page. So, this model is very effective and efficient to extract a paragraph. This algorithm came after lot of permutations and combinations . This model is very effective in retrieving the correct information from the given huge information. This model is useful in saving ones time and energy. Summarization is responsible for summarize the textual information approximately one-third valuable information for further decision support system.

**LIMITATION**

When we are dealing with large amount of data , time complexity of program becomes very large because it parse the data more number of times.

**ACKNOWLEDGEMENT**

I would like to express my special thanks of gratitude to our guide, Prof. Pravin Khandare and our honourable head of department Prof. Deshpande who helped us in every aspect and problem that we had.

Special thanks to our team mates, who helped and supported us to make our work valuable and complete our project successfully.