**CAPSTONE PROJECT**

**CSA1328 – Theory of Computation with Non Deterministic Problem**

**SAVEETHA SCHOOL OF ENGINEERING**

**SIMATS ENGINEERING**



**Supervisor**

**Dr. M Rajasekar**

**DONE BY**

**L Lava Kumar Naidu (192211006)**

**K Surya (192211028)**

**APRIL 2024**

**SUMMARY WRITTER USING NATURAL LANGUAGE PROCESSING**

**INTRODUTION**

* Text Summarization: Enhancing Information Retrieval
* Utilizes natural language processing to condense(reduce) large text volumes into concise (short and clear) summaries.
* Extracts essential sentences and phrases for brief overviews.
* Potentially improves information retrieval and productivity.
* Offers time-saving benefits for individuals and organizations dealing with large textual data.

**Natural Language Processing (NLP):**

* A branch of artificial intelligence enabling computers to understand, interpret, and generate meaningful human language.
* Techniques range from basic text processing to advanced machine learning models.
* Goal: bridge human communication and computer understanding for seamless interactions.  
   Addresses challenges like ambiguity, context, and semantics.

**OBJECTIVES**

* The project aims to improve information retrieval, facilitate quick comprehension, and enable efficient decision-making by condensing large volumes of text into key insights.
* The objective of automated text summarization using Natural Language Processing (NLP) can be multifaceted, but generally, it aims to condense a piece of text while preserving its key information and meaning.
* To save time and effort by automatically generating concise summaries of lengthy documents, articles, or texts.

**ALGORITHM**

* Clustering based model
* SVM algorithm
* TF-IDF

**TEXT SUMMARIZATION:**

To perform the text summarization, we have two types of techniques.

**1.**Abstractive summarization technique

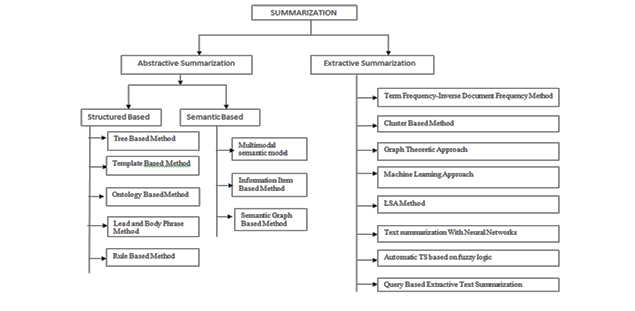
**2.**Extractive summarization technique

**EXTRACTIVE SUMMARIZATION**

* Involves selecting and extracting key sentences from original text.
* Doesn't involve rephrasing or rewriting.
* Relies on identifying key sentences encapsulating main ideas.
* Effective in maintaining original context but can lead to incoherent summaries.
* Commonly used in legal documents or news articles.
* Utilizes models like SVMs, Random Forests, and neural networks for sentence ranking.

**ABSTRACTIVE SUMMARIZATION**

* Involves interpreting original text and rephrasing it in a condensed form.
* Aims to capture source material's essence in a cohesive, readable manner.
* Involves natural language generation and requires deeper understanding.
* Produces more human-like summaries but poses challenges in factual accuracy and avoiding erroneous information generation.
* Relies on advanced architectures like Sequence-to-Sequence models and Transformer-based models.



**KEY STEPS**

**Preprocessing and Cleaning**

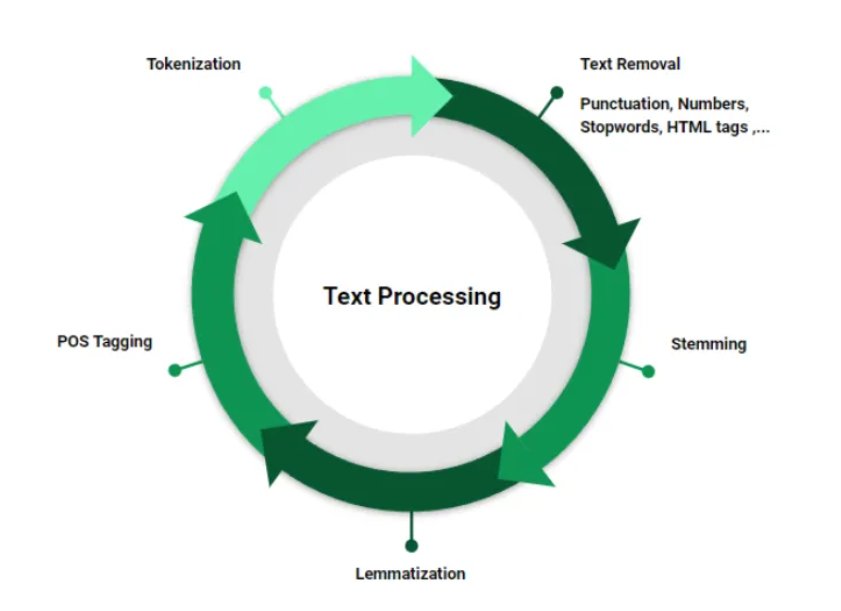
* Before creating a summary, the text data needs to be preprocessed and cleaned to remove any irrelevant information, such as stop words and punctuation. This step also involves tokenization, stemming, and lemmatization to ensure the text is in a format conducive to summarization.

**Sentence Scoring and Ranking Algorithms**

* After preprocessing, the sentences are scored and ranked based on their relevance to the overall content. Various algorithms, such as TF-IDF (term frequency-inverse document frequency) and Text Rank, are used to assign importance scores to each sentence for inclusion in the summary.

**Generating the Final Summary**

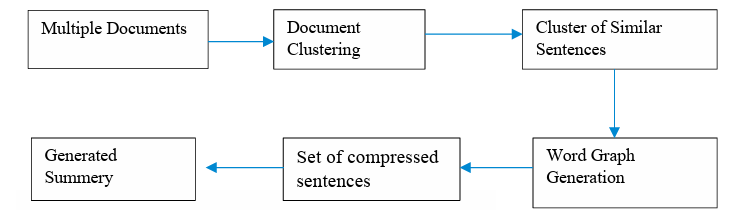
* Once the sentences are scored and ranked, the final summary is generated by selecting the most salient sentences that best represent the key points of the original text. This can be done using extractive or abstractive techniques based on the specific requirements.



**Unsupervised Text Summarization Approaches**

* Clustering techniques group similar documents.

• Latent Semantic Analysis (LSA) uncovers underlying semantic structures for significant sentence extraction.  
• Topic modelling, like Latent Dirichlet Allocation (LDA), assigns topics probabilistically to documents.  
• Word embeddings like Word2Vec or Doc2Vec capture semantic relationships for unsupervised summarization.  
• Text Rank algorithm applies graph-based ranking to sentences, identifying important content based on relationships.  
• Density-based clustering like DBSCAN identifies coherent themes for summarization.

****

**Evaluation metrics for text summarization**

* ROUGE (Recall-Oriented Understudy for Gisting Evaluation): A set of metrics used to evaluate the quality of a summary by comparing it to human-generated reference summaries.
* BLEU (Bilingual Evaluation Understudy): Originally designed for machine translation, it has been adapted to evaluate the effectiveness of text summarization.
* Pyramid Method: A manual evaluation approach that involves comparing different summaries and assigning scores based on specific criteria.
* Responsiveness: Measures the ability of a summary to accurately capture the most important and relevant information from the source text.

**CONCLUSION**

Summary writing in NLP plays a crucial role in extracting key information and condensing the content of large documents into concise summaries. It helps in improving the efficiency of information retrieval, document understanding, and decision-making processes. By summarizing text, we can save time and effort in reading and analyzing large volumes of information. It allows us to quickly grasp the main ideas, identify important details, and make informed decisions based on the summarized content.