VIVEKANANDA INSTITUTE OF TECHNOLOGY

HYBRID DOCUMENT SUMMERIZATION

DONE BY,

NAVEEN PANDURANGI(1VK15CS034)

NIKHIL CHANDRAN(1VK15CS038)

SRIVATSA V JAMADAGNI(1VK15CS054)

MOHAMMED ABU TALHA AHMED(1VK15CS030)

UNDER THE GUIDANCE OF

MRS. CHANDRAMMA R

ASSOC.PROFESSOR ,DEPT OF CSE,VKIT

TABLE OF CONTENTS

INTRODUCTION :

- Hybrid Document Summarization is the technique by which the huge parts of content are retrieved.
- The Document Summarization plays out the summarization task by unsupervised learning system.
- The significance of a sentence in info content is assessed by the assistance of 3 algorithms.
- 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).
- The Hybrid Document Summarization undertaking makes the users simpler for various Natural Language applications, like, Data Recovery, Question Answering or content decreasing etc.

LITEARTURE SURVEY

- Document Summarization is the technique by which the huge parts of content are retrieved.
- The Document 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.
- A specific word may have distinctive significance in various setting. So, the principle task of word sense disambiguation is to decide.

SL No	Title	Methodology	Algorithm
01	A Neural Attention Model for Abstractive Sentence Summarization	local attention-based model	Beam Search
02	Abstractive Text Summarization using Sequence-to-sequence RNN and Beyond	RNN	-
03	Automatic Keyword Extraction for Text Summarization : survey	-	-
04	Summarization with Pointer-Generator Networks	Pointer generator network and CNN	-
05	Text Summarization Techniques: A Brief Survey	Frequency Driven Approaches	-
06	Ranking Sentences for Extractive Summarization with Reinforcement Learning	CNN/RNN	REINFORCE algorithm

07	Neural Document Summarization by Jointly Learning to Score and Select Sentences	RNN	-
08	Machine Learning Techniques for Document Summarization: A Survey	Talks about all the methods	-
09	Ranking with Recursive Neural Networks and Its Application to Multi-Document Summarization	R2N2/RNN	-
10	A Neural Attention Model for Sentence Summarization	local attention-based model	Beam Search
11	A Review Paper on Text Summarization	Natural Language Processing	-
12	Automatic Text Summarization Using Natural Language Processing	Natural Language Processing	LESK Algorithm
13	A Novel Technique for Efficient Text Document Summarization as a Service	Natural Language Processing	LESK Algorithm

14	Automatic Text Summarization and its methods : A Review	Term frequency based method/ Graph based method/ Time based method/Clustering based method	-
15	A Review on Automatic Text Summarization Approaches	Frequency Based Approach/Term Frequency–Inverse Document Frequency/Machine Learning Approach/Discourse Based Method	-

EXISTING SYSTEM

- Usually summarization is done by hand by the editor or user.
- Most of the methods currently in use do summarization for text .
- Summarization of document is hardly even possible
- Most methods utilize supervised learning.
- This method is tedious, time consuming and very inefficient.

PROPOSED SYSTEM

• A solitary or single input content is going to be outlined utilizing unsupervised learning.

•Sentences with induced weights are composed in sliding solicitation concerning their weights.

• Certain quantities of sentences are chosen as an outline.

•The proposed computations, abridges solitary or single report content utilizing unsupervised learning approach.

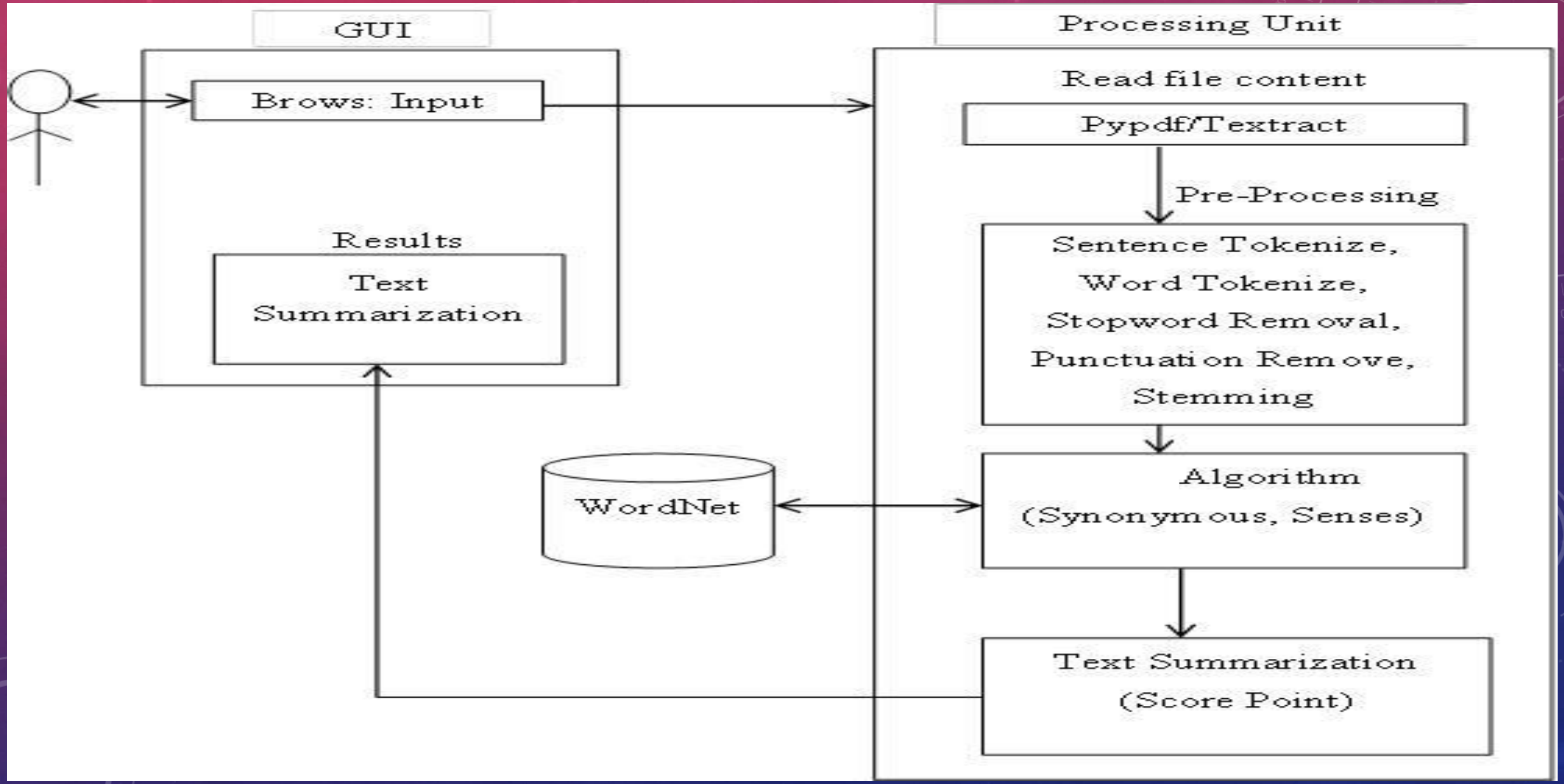
•Heaviness of every sentence in a substance is resolved using streamlined Lesk's computation and wordnet.

•Summarization procedure is performed as indicated by the given rate of synopsis.

•Input document will be in the form of a .txt file.

- The pre-processing includes the data cleaning and data abstraction.
- The data will be the input to the lesk algorithm with the weights given to the words.
- Wordnet acts as a dictionary for comparing the importance of the word given is the input.
- Once the data is processed in the lesk algorithm it gives the output values which will be further converted into the summarized document format.

SYSTEM ARCHITECTURE OF PROPOSED SYSTEM



METHODOLOGY

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 be performing sentence part, tokenization, empty stop word, clear accentuation and perform stemming.

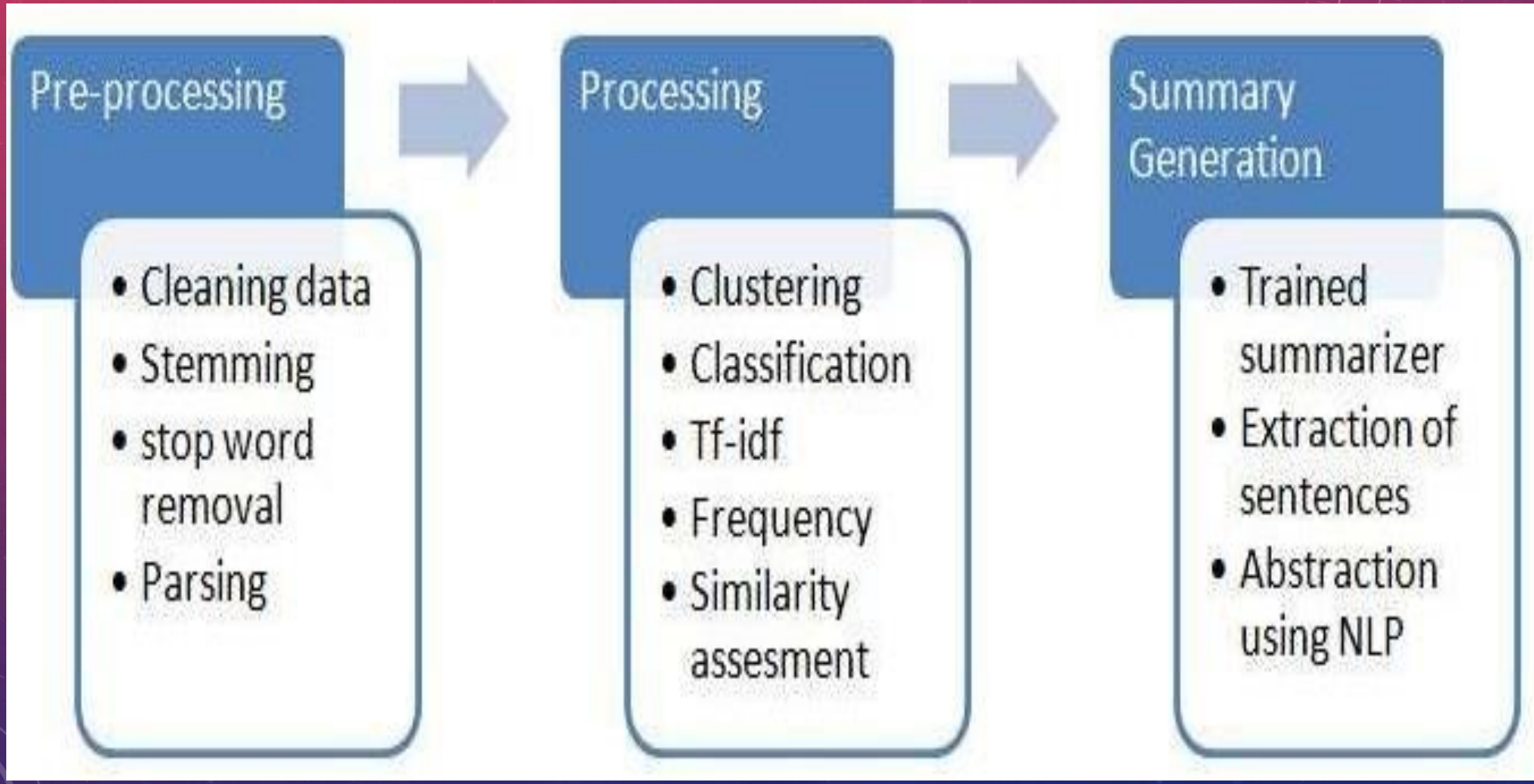
STAGE 2: EVALUATION OF WEIGHTS

This is the next phase of document summarization. After the Text data is cleaned and pre-processed, the processing techniques are applied in which the scores, frequency of words are calculated. The data is grouped on the basis of similarity, dissimilarity so that the summary generation can be made efficiently. For this different clustering techniques are used.

STAGE 3: SUMMARIZATION

After the data is processed, the summary is to be generated based on the requirement. Extraction of sentences is done from the processed data and the summary is processed. The words to added to sentences, reducing the sentences based on score etc. is done in this step to produce summary. The representation of the summary can be in the form of words sentences, paragraphs, graphs etc. for the summary to be generated different summarizers are used.

MODEL PROCESSING



ALGORITHMS

1. AUTOMATIC SUMMARIZATION BASED ON USER QUERY

The algorithm of this approach is as follows:

1. Generate a Document Graph (DG) for each sentence in the input documents,
2. Generate a DG for the query (topic),
3. Measure the similarity between each sentence and the query (topic),
4. Search for and add the best sentence to the summary,
5. If the summary's length restriction is met or there are no more sentences to add then finish and report the target summary; otherwise add the DG for the chosen sentence to the query graph,
6. Repeat from step 3 until no more sentences can be added to the summary.

2. AUTOMATIC SUMMARIZATION BASED ON WEIGHTING

This calculation compresses multiple 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.

Info: Multiple-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 sequence in the info content.
Step 11: Stop.

3. INFORMATION EXTRACTION ALGORITHM

1. Introduce a method to extract the merited keyphrases from the source document. For example, you can use part-of-speech tagging, words sequences, or other linguistic patterns to identify the keyphrases.
2. Gather text documents with positively labeled keyphrases. The keyphrases should be compatible to the stipulated extraction technique. To increase accuracy, you can also create negatively-labeled keyphrases.
3. Train a binary machine learning classifier to make the text summarization. Some of the features you can use include:

- Length of the keyphrase
- Frequency of the keyphrase
- The most recurring word in the keyphrase
- Number of characters in the keyphrase

4. Finally, in the test phrase, create all the keyphrase words and sentences and carry out classification for them.

WORDNET

- **WordNet** is a lexical database for the English language.
- It groups English words into sets of synonyms called *synsets*.
- WordNet can is a combination of dictionary, and thesaurus.
- Primarily used in automatic text analysis and artificial intelligence applications.
- The database and software tools are freely available for download from the WordNet website.
- Both the lexicographic data (*lexicographer files*) and the compiler (called *grind*) for producing the distributed database are available.

NATURAL LANGUAGE TOOL KIT

•The **Natural Language Toolkit** is a suite of libraries and programs for symbolic and statistical natural language processing (NLP).

• It was developed by Steven Bird and Edward Loper in the Department of Computer and Information Science at the University of Pennsylvania.

• NLTK includes graphical demonstrations and sample data.

• It is accompanied by a book that explains the underlying concepts behind the language processing tasks supported by the toolkit.

Summarization using query-based algorithm

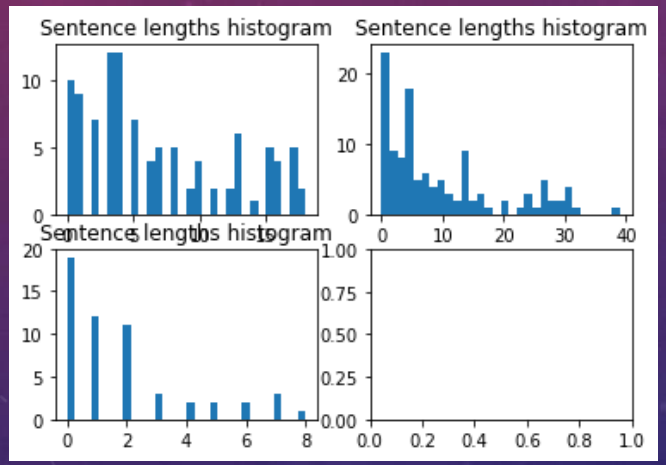
Input:

Please enter a query: friend

OUTPUT:

PROCESSED QUERY: [U'FRIEND']

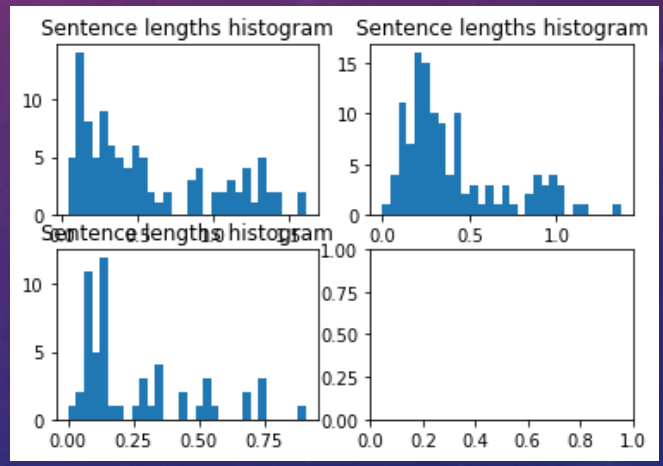
[("SHE WAS RANSACKING THE STORES FOR JIM'S PRESENT.", 1.324313323278029), ('JIM STEPPED INSIDE THE DOOR, AS IMMOVABLE AS A SETTER AT THE SCENT OF QUAIL.', 1.3260330166909116), ('"GIVE IT TO ME QUICK" SAID DELLA.', 1.337516806722689), ('JIM HAD NOT YET SEEN HIS BEAUTIFUL PRESENT.', 1.3436049382716049), ('"YOU NEEDN\'T LOOK FOR IT," SAID DELLA.', 1.3593439153439153), ('AND THEN DELLA LEAPED UP LIKE A LITTLE SINGED CAT AND CRIED, "OH, OH!"', 1.3629012660542072), ('ONLY $1.87 TO BUY A PRESENT FOR JIM.', 1.4183846153846156), ('IT SURELY HAD BEEN MADE FOR JIM AND NO ONE ELSE.', 1.444810744810745), ('"JIM, DARLING," SHE CRIED, "DON\'T LOOK AT ME THAT WAY.', 1.5813651999874911), ("ONE WAS JIM'S GOLD WATCH THAT HAD BEEN HIS FATHER'S AND HIS GRANDFATHER'S.", 1.6168950552674828)]



Sentence length histogram for Weight-based output

Automatic summarization based on weighting

[('Jim drew a package from his overcoat pocket and threw it upon the table.', 16.0), ('"You needn\'t look for it," said Della.', 16.0), ('The door opened and Jim stepped in and closed it.', 16.0), ('It surely had been made for Jim and no one else.', 17.0), ('Jim had not yet seen his beautiful present.', 17.0), ('And then Della leaped up like a little singed cat and cried, "Oh, oh!"', 17.0), ('Only $1.87 to buy a present for Jim.', 17.0), ('Jim looked about the room curiously.', 17.0), ('"Jim, darling," she cried, "don\'t look at me that way.', 18.0), ("One was Jim's gold watch that had been his father's and his grandfather's.", 18.0)]



Sentence length histogram for query-based output

Summarization Information Extraction

Output: Separation between Subject verb Object

(NX ONE/CD DOLLAR/NN AND/CC EIGHTY-SEVEN/NNP CENTS/NNP NX) ./.
(NX THAT/WDT NX) (VX WAS/VBD VX) ALL/PDT ./.
(NX AND/CC SIXTY/CD CENTS/NNP NX) of/IN (NX it/PRP NX) (VX was/VBD VX) in/IN (NX pennies/NNS NX) ./.
(NX Pennies/NNP NX) (VX saved/VBD VX) (NX one/CD and/CC two/CD NX) at/IN (NX a/DT time/NN NX) by/IN (VX bulldozing/VBG VX) (NX the/DT grocer/NN NX) and/CC (NX the/DT vegetable/NN man/NN NX) and/CC (NX the/DT butcher/NN NX) until/IN (NX one/CD 's/POS cheeks/NNS NX) (VX burned/VBN VX) with/IN (NX the/DT silent/JJ imputation/NN NX) of/IN (NX parsimony/NN NX) (NX that/WDT such/JJ close/NN NX) (VX dealing/VBG implied/VBN VX) ./.
(NX Three/CD times/NNS Della/NNP NX) (VX counted/VBD VX) (NX it/PRP NX) ./.
(NX One/CD dollar/NN and/CC eighty-seven/JJ cents/NNS NX) ./.
And/CC (NX the/DT next/JJ day/NN NX) (VX would/MD be/VB VX) (NX Christmas/NNP NX)

CONCLUSIONS

- We have worked on three algorithms and by comparing those algorithms, query-based algorithm gives the least accurate summary.
- The weight-based algorithm gives the summery according to the score of the sentence but it is not accurate in the formation of the sentences.
- Sometimes the sentences might not a right meaning. The information extraction algorithm uses parts of speech tagging which gives the summary having proper meaning for the sentences.
- Hence, we like to conclude that information extraction algorithm is more accurate than query and weight-based algorithms.

THANK YOU!