NLP Assignment 6

Q1) Given a three labels as positive, negative, and neural respectively. Illstrate a full-fledged sentiment analysis system that uses a non-parametric bootstrapping based ensemble methodology along with one metric for analysis.

→

sentiment analysis classify that data into categories: positive, negative and neural. A robust approach to enhance accuracy is using non-parametric bootstrapping along with ensemble leaning techniques:

a) Data collection: Gater a labeled dataset containing text samples with sentiments. potential source include social media platform, product review & news articles.

b) Data preprocessing:-
- Text cleaning: Remove punctuation special characters
- Tokenization:- split text into word or token ~~numerical format~~
- Vectorization:- convert text into numerical format using methods such as TF-IDF, Bow, word2vec.

c) Bootstraping & Ensemble learning:-
- Bootstraping:- Randomly sample the training dataset with replacement to create multiple subset.
- model training:- Train a seperate sentiment classification model or each bootstrap sample.
- Final prediction:- combine prediction from all models using majority voting or averaging for the final sentiment label.

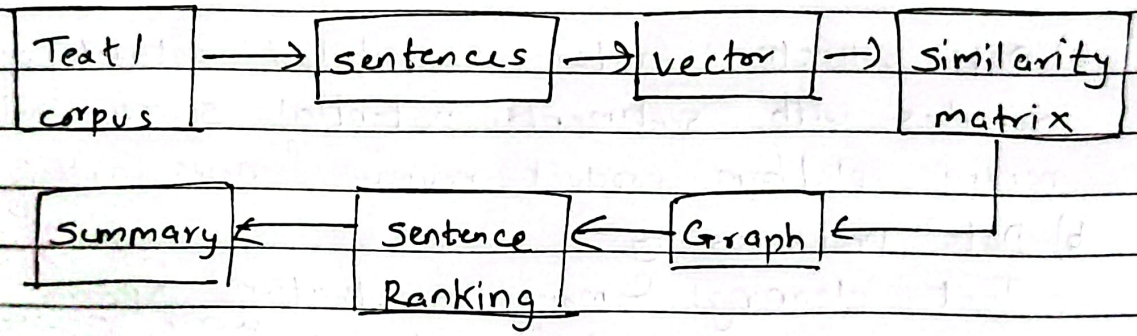
d) Evaluation matrix:- use the f1-score to evaluate model performance, balancing precision and recall

for multi-class classification. The f1-score is calculated as follows:-

$$\text{F1-score} = \frac{2 \times \text{precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Q2) Given a corpus of 100 lines stating current state of geo-political situation. illustrate a functional and meta-heuristic extractive text summarizing technique that (no) utilize text rank algorithm along with it's entire proces.

→



Extractive text summarization involves selecting the most important sentence from a document to create a consice summary. The textrank algorithm is powerful technique for extraction summarization, particularly useful for summarization large text, such as corpus of 100 lines describing the current geopolitical situation.

Process flow for performing:-

a) Data preparation:- corpus collection and cleaning: preprocessing the text by removing unnecessary characters, puntuations and stopwords to enhance quality of the analysis.

b) sentence tokenization:- split cleaned text into individual sentences. This can be done using libraries like nltk.

c) Graph construction:-
- Similarity matrix: create similarity graph where each sentence is a node. Calculate the similarity between each pain of sentence using co-sine similarity based on word embedding.
- edges: Add weighted edges between nodes, where the weights represents the similarity score

d) Text Rank Algorithm:-
- Assign each sentence an initial score
- update the score of each sentence based on neighbouring sentence in graph.

The update rule: $S_i = (1-d) + d \sum_{j \in N(i)} S_j / C_j$

S_i = score of sentence i
N_i = Neighbouring sentence
C_j = ~~Neighbouring~~ Number of edges
- Repeat until score stabilize

e) Sentence selection:-
- Ranking : Rank the sentence based on their final score select the top n sentences
- output: It compile the selected sentences into a coherent summary that reflects the main points at original text.