School of Computer Science Engineering and Technology

Course- B. Tech	Type- Elective
Course Code- CSET346	Course Name: Natural language processing
Year- 2023	Semester- Even
Date: 17-02-2023	Batch- ALL

Lab Assignment 03 – Word Embedding (TF-IDF)

Objective:

The main goal of this assignment is to develop the TF-IDF vectorizer both from scratch and using a in-built Python library. This assignment includes facilitate learning word clustering in addition to word embedding.

Goal:

Social media analytics helps companies address these experiences and use them to do various tasks such as :

- Spot trends related to offerings and brands,
- Understand conversations — what is being said and how it is being received,
- Derive customer sentiment towards products and services,
- Gauge response to social media and other communications,
- Identify high-value features for a product or service,
- Uncover what competitors are saying and its effectiveness,
- Map how third-party partners and channels may affect performance etc.

Let, you are a social media data analyser. In your assignment, you need to take a raw online web article of your choice and find out the important key words in it. Moreover, you have to cluster the keywords to find out the related ones.

Input Dataset:

Take any single article from the link : https://tetw.org/Greats

Tasks:

Question 1:

Do the needful pre-processing to the input text.

a. Tokenization
b. Only consider characters.
c. Stopword Removal
d. Stemming / Lemmatiztion (Your choice)

Question 2:

Implement TF-IDF model from scratch to make the word embedding of the corpus.

Question 3:

Implement TF-IDF model with Sklearn package to generate word embeddings of the corpus.

Question 4:

Implement K-means clustering on the processed data with number of clusters given as user input.

Useful links:

1. https://docs.scipy.org/doc/scipy/reference/cluster.vq.html
2. https://scikitlearn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html