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: 27-02-2023	Batch- ALL

Lab Assignment 04 –Apply NLP to real life data (Word2Vec Word Embedding)

CO-Mapping

Exp. No.	Name	CO1	CO2	CO3
06	Apply NLP to real life data	✓	✓	✓

Objective:

The main goal of this assignment is to implement the Word2Vec embedding and analyze its different characteristics. Moreover, create a text classification model with real time NLP dataset.

Tasks 1:

Question 1:

Find the datasets available for word vectorization in ***genism***.

Question 2:

Use ***glove-twitter-50*** for training your model.

Question 3:

Take any ten random words of your choice. (Say, ***chosen_words***)

Question 4:

Find the three most similar words for the ***chosen_words***

Question 5:

Find the similarity value between each ***chosen_word*** with its most similar words.

Question 6:

Find all these words embeddings.

Question 7:

Reduce their dimension to 2 using a dimension reduction algorithm (eg. t-SNE or PCA) and plot the results in a 2d-scatterplot

Question 8:

Show that the Semantic regularities captured in word embeddings.
Ex. queen = king + woman - man
(using gensim *most_similar* with *positive* and *negative*)

Task 2:

Implement a Text Classification Model using Word2Vec. For implementing the task, consider any real time text classification dataset and any classification model of your choice.