School of Computer Science Engineering and Technology

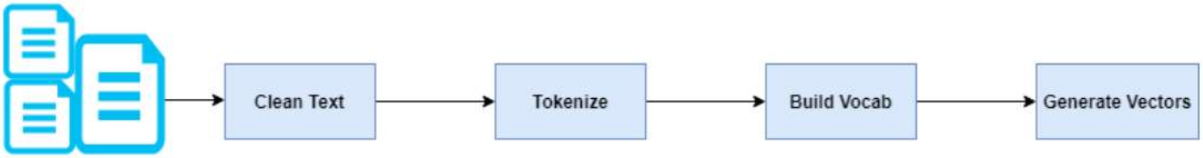
Course- B. Tech	Type- Elective
Course Code- CSET346	Course Name: Natural language processing
Year- 2022	Semester- odd
Date: 30-08-2022	Batch- ALL

Lab Assignment 03 – Word Vectorization

CO-Mapping

Exp. No.	Name	CO1	CO2	CO3
03	Word Vectorization	✓	✓	--

Objective: The main objective of this assignment is to know how to convert words into numerical vectors and its application.



Count Vectorizer: It creates a document term matrix, which is a set of dummy variables that indicates if a particular word appears in the document.

Problem1. Choose any text data of your choice and apply count vectorizer to convert the word into vectors after applying data cleaning and tokenization.

One-Hot Encoding: Represent each unique word in vocabulary by setting a unique token with value 1 and rest 0 at other positions in the vector

Problem2. Apply the OHE technique for the above problem.

Bag of Words: It takes a document from a corpus and converts it into a numeric vector by mapping each document word to a feature vector for the machine learning model.

Problem 3: Take an open-source text data, pre-process the data and apply Bag of Words techniques and display the vectors.