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: 06-02-2023	Batch- ALL

Lab Assignment 01 – Initial Steps towards data pre-processing

Learning Object

The main objective of this assignment is to know pre-processing of text data through different steps using natural language tool kit(nltk). Tokenization, Stop word removal, Stemming, Lemmatiation are the parts of the initial text data pre-processing.

Goal

Amit, a journalism student, wants to analyze the BBC news data in his assignment. Help him to achieve his goal.

Input Dataset

Download the BBC news dataset from the following link: http://mlg.ucd.ie/datasets/bbc.html

There are five different topics in the dataset: business, entertainment, politics, sport, tech

Output

Create vocabulary of pre-processed words in the documents.

Tasks

1. Take all documents from each topic.
2. Read the documents line by line
3. Tokenize the lines
4. Apply stemming (Print the result for `PorterStemmer and LancasterStemmer of nltk`). Consider the results for any of the stemmers for further processing.
5. Perform lemmatization and stop-word removal.
6. Write the vocabulary of each topic within a file along with the frequency of each word.
7. Repeat the previous steps for each topic and create the complete vocabulary for the BBC news dataset and write it in a separate file.

Useful links:

1. https://www.datacamp.com/tutorial/stemming-lemmatization-python
2. https://www.geeksforgeeks.org/python-lemmatization-with-nltk/
3. https://www.machinelearningplus.com/nlp/lemmatization-examples-python/