CS 5615: Information Retrieval – Assignment 1

Overview

In this assignment, you will:

1. Implement basic text pre-processing techniques using available software tools

Data

Refer to the data file (assignment_data.txt) given in Moodle

Task

assignment_data.txt contains sample text from a twitter feed, student course feedback, and a research paper.

1. For each text type in assignment_data.txt, tokenize the text using an available tokenizer. Discuss about the accuracy of the tokenizer on each type of text.
2. Carry out isolated word correction and context sensitive word correction on the text. Based on your observations on the processed text, discuss the impact of each type of spell corrections on the three types of text.
3. Stem and lemmatize the text using a suitable stemmer and a lemmatizer. Based on your observations on the processed text, discuss the suitability of stemming and lemmatizing for retrieving base forms of words.

Deliverables

1. Source Code

Source code and a readme file with CLEAR instructions on how to execute the code.

2. Report

The report should include the following:

1. Brief description of the tokenizer, spell correctors, stemmer, and lemmatizer you used.
2. Discussion as given in the task.