Information Retrieval (CS4051)
Programming Assignment No. 3
Spring 2023

Submission Date: May 11, 2023

Assignment Objective

This assignment is for the task of clustering. You need to perform document clustering for the given dataset. You are required to use K-Mean algorithm for the clustering. The evaluation is based on external measure Purity for the given task. Its calculation can be done as follows: For each cluster, count the number of documents from the most common class in said cluster. Now take the sum over all clusters and divide by the total number of documents. For a complete reference read the chapter 16 on partition clustering.

Datasets

For this assignment the dataset we have selected a subset of dataset of NewsGroup20. There are 50 documents from 5 different classes. We are providing you dataset with ground truth of all clusters.

Feature Selection for Clustering

Baseline – I - You need to first create a baseline for the clustering by using all TF based features using any filtering criteria like TF scores for the selected term (feature).

Baseline – II – You can create a slightly string baseline using TF*IDF feature selection using all TF based features using any filtering criteria like TF scores for the selected term (feature) and DF filtering criteria based on DF.

Word2Vec embedding of selected TF can be used as a semantic feature to perform clustering. You need to evaluate the clustering result using purity.

Coding can be done in either Java, Python, C/C++ or C# programming language. For K-Mean algorithm you can use SKLearn library implementation of it.

Evaluation/ Grading Criteria

The grading will be done as per the scheme of implementations and clustering evaluation metric

Grading Criteria:
Preprocessing (3 marks)
Feature Selection (3 marks)
K-Mean Implementation (1 marks)
Evaluation (2 marks)
Code Clarity (1 mark)

<The End>