Lab10-Natural Language Processing – Text Classification

Total Marks: **8 Marks + 2 Marks (individual assessment) = 10 Marks**

In this assignment you will use a provided musical dataset and by using natural language processing, build the classifiers and evaluate the performance of a system that assign positive (1) or negative (0) score by analyzing text based reviews of musical instruments. The dataset is a modified 1000 reviews of a dataset used in "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering by R. He, J. McAuley WWW, 2016 [cseweb.ucsd.edu]", which is attached with this assignment.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)Recall = TP / (TP + FN)

F1 Score = 2 * Precision * Recall / (Precision + Recall)

Using Python language, perform the followings NLP tasks to build the classifier for the given dataset:

1. Using NLTK word_tokenize function, tokenize the given dataset reviews
2. Using NLTK PorterStemmer, perform the stemming for the tokens of the reviews
3. Using NLTK WordNetLemmatizer, perform the lemmatization for the stemmed tokens
4. Build the Random Forest technique using sklearn library
5. Evaluate the model by finding its accuracy, precision and F1-score