PSG College of Technology

Department of Applied Mathematics and Computational Sciences

MSc SS / TCS / DS

Natural Language Processing Lab

1. Given a dataset for drug sentiment with Medicine, disease, tweet about the adverse effect of medicine for a disease and the sentiment, do the following tasks:
    a. Preprocessing steps for sentiment feature:
        i. Tokenization in which convert the tweet into words
        ii. Remove stop words
        iii. Do stemming
        iv. Formulate bag of words
    b. Formulate training set using features and class label (Drug, disease, bag of words, sentiment)
    c. Do visualization using different plots which depict
        i. Frequency distribution of words disease wise / sentiment wise / medicine iwse
        ii. Adverse effects Vs Drug for disease wise
        iii. Drugs Vs disease Vs sentiment
        iv. Etc many more (Innovative themes and plots are welcome)
    d. Develop a UI
    e. Apply the following ML algorithms on the data and choose the best algorithm
        i. SVM
        ii. Logistic regression
        iii. KNN
        iv. Naïve Bayes