Faculty of Computers and Information
Cairo University
NLP-2019

NLP
Assignment-2
Performing text classification

Objectives

- Practice how to perform text classification using a machine learning classification model and the results of tf-idf as a feature vector

Dataset

Movie reviews data set V2.0 contains 2000 text samples divided into 1000 positive reviews and 1000 negative reviews. Reference and download:
http://www.cs.cornell.edu/people/pabo/movie-review-data/

Task:

- (1) Load review samples (both positive and negative) and generate tf-idf for the samples
- (0.5) Generate labels vector for the dataset. For example 1 for positive review and 0 for negative review. So labels [234] = 1 means that sample 234 is a positive review
- (0.5) Randomly divide data to training and testing sets. Note that each set should contain samples of the two types
- (1) Train a classification model to predict the label of the review

Output

- (0.5) Print the accuracy of the model after testing it on the testing set
- (1.5) Your program should allow the user to input a new text review and then predict if it is positive or negative using the trained model

Bonus (2 grades)

Plot the testing samples with different colors for the positive and negative samples and draw the curve separating them

Teams

Form teems of 3 for this assignment