HACETTEPE UNIVERSITY COMPUTER ENGINEERING DEPARTMENT

BBM 409 MACHINE LEARNING LABORATORY

ASSIGNMENT 2

Name: Hülya Şermin

Surname: Karakaş

Student ID: 21591198

E-mail: hulyasermin.karakas@gmail.com

1. Introduction

In this assignment, it is expected to predict the sentiment of tweets given in dataset. While doing prediction process, naive bayes and bag of words algorithm need to be used. In addition, unigram and/or bigram features are kindly recommended to add to the project. After that, accuracy computation of the given model should calculate by the formula given in the assignment information manual. [1]

2. Algorithm & Results

The following steps of my algorithm is;

- Importing the numpy and sklearn libraries
- Loading the inputs
- Initializing the tfid vectorizer to count the common word in tweets
- Categorizing the tweets by their properties which are negative, positive and neutral
- Initializing and declaration of the class priors
- Initializing the arrays that will store the sum of the classes' class priors with the smoothing. (The smoothing is done by adding the column number of the class array)
- Initializing and declaration of the likelihood arrays. The computation of the likely hood done by taking the logarithm of the column summation with adding 1 and dividing that value by classes' value summation.
- Categorizing the tweets in the validation data by their "0", "2" or "4" property.
- Initializing the classes' posterior arrays and declaring it to summation of substitution of negative likely hood matrix and the tweet data. Then, I added classes' class prior to that value.
- That value gives us the possibility of the test tweet to predict which class it belongs to. To find the highest possibility, I calculate the possibility with max number.
- Finally, I computed the accuracy of the program by the given formula in the lab manual.
- I tried my algorithm on unigram, bigram and both.

Results of the program;

Feature	Accuracy
Unigram	57.1999
Bigram	44.6
Both	55.7

The values observed with max_df = 0.8 and min_df=2 parameters for the bigram and both unigram and bigram. I did not give those parameters to unigram because program gives memory error on bigram feature.

3. Analysis of the Results

In the theory, the feature of the usage both unigram and bigram should give much high accuracy, but because of some reasons, I could not observe the expected result. I got the highest accuracy on unigram feature. The reasons could be;

- Noises in the tweets dataset
- Incorrectly labeled tweets
- Not having enough amount of train tweet.

4. References

1. https://d1b10bmlvqabco.cloudfront.net/attach/j7vgdoadakn3ce/iki7yi2bje7t8/j8yx1czm9d12/Assignment2.pdf