CSCI – 630 Homework Lab3

Submitted By- Prakhar Gupta

**Data Scraping**- Data was scraped using selenium from Wikipedia and dutch Wikipedia. The folder data scraping contains code for scraping Wikipedia.

Around 300 random articles of each language was chosen to make train600.dat

file . The data was also cleaned to remove hyperlinks and other special characters.

Chrome driver -<https://chromedriver.storage.googleapis.com/index.html?path=101.0.4951.15/>

**Running code**

To run the code pass inline arguments to main.py

For training – train <examples> <hypothesisOut> <learning-type>

For testing- predict <hypothesis> <file>

**Answers**

1. a description of your features and how you chose them

I did my research on semantics and word use of both languages Dutch and English. I was able to get a list of most used words in both Dutch and English.

It turn out all the languages follow a distribution called ZIPF distribution.

(<https://www.youtube.com/watch?v=fCn8zs912OE>) video reference.

(<https://en.wikipedia.org/wiki/Zipf%27s_law>) wiki link

Thus I took features

Count of frequent dutch words in sentence

Count of frequent english words in sentence

Average word length

Largest words length

Vowel to consonants ratio (VC ratio)

Based on the train data cut-off was decided to make true/false flags

Count of frequent dutch words in sentence <2

Count of frequent english words in sentence <2

Average word length =5.1

Largest words length< 11

VC ratio <0.4

Other thresholds were also tried. These give best results.

1. a description of the decision tree learning, how you came up with the best parameters (max depth, etc.) and your own testing results

2 level dtree turned out to give best performance below is the performance on 20 percent data

Accuracy on test of size (support)- 120

1 level dtree -0.933

2 level dtree -0.983

3 level dtree -0.983

4 level dtree- 0.983

1. a description of the boosting, how many trees turned out to be useful, and your own testing

3 dtree turned out to give best performance below is the performance on 20 percent data

Accuracy on test of size(support) - 120

3 round boosting adaboost - 0.983

5 round boosting adaboost - 0.966

10 rounds boosting adaboost - 0.925

1. anything else you think we might like to know-

2 level decision tree was chosen as the best model as it’s always better to choose a simpler model if the complex model has a similar performance

We have used accuracy metric as data classes are balanced if they were not precision, recall and F1 score would have been better metrics