Project Report

Computational Intelligence

Guilan University

Submitted by:

Mahta Amraji(940122611101)

Mojan Jamal omidi(940122611103)

Azar bagheri(9501226810l6)

Under Supervision of:

a. Tourani

January 2020

Overview:

This report discusses the result of the work done in development of "sentiment analysis" with python on Jupyter notebook as the final project of computational intelligence course.

Case study:

The objective of our case study is to let the program find the polarity of the words retrieved from sentences.

The dataset we are using on this project consists of 50000 movie reviews taken from IMDb. The data is split evenly in half as training and testing sets.Moreover each set consists of half positive and half negative reviews.

Step one:

The first step is loading the data as training data as it is done in the first code box. There are two columns in our dataset, the first one is the sentimentText or the comments and the second one provides the comments sentiment.

Step two:

the second step in our project is data pre processing, at this step

the goal is to remove all the irrelevant and redundant information present or noisy and unreliable data to make the process easier and faster.

Some of the preprocessing we have done here in this project is removing the extra white spaces,stopwords,rare words,frequent words,punctuations and etc.

Step three:

our next step is splitting the dataset to make our training and testing sets.this task is done with sklearn train_test_split.

Step four:

On this step we have used the wordcload data visualization technique to show the frequency or importance of the positive and negative words in our set.

Step five:

On this step where the main work is done we use two functions ,the first one builds a pipeline of tfidf vectorizer and two methods as the classifier "logistic regression" and "Ridge classifier" and the accuracy score of the methods are computed in the second function where it first fits the model and predicts the polarity of the test set.

Step six:

This is the last step where we test some of our own input

sentences to try to challenge the program.

Conclusion:

we get the best result using Ridge classifier with three_gram .

References:

www.towardsdatascience.com

www.datasciencetoday.net

www.nlpforhackers.io