180128228

Natural language processing report lab 1

Introduction

The purpose of this assignment was to implement a binary perceptron and apply it against a sentiment analysis focusing on movies reviews to predict the sentiment of movie reviews. This is based upon 2000 documents split into 1000 documents for the positive folder and 1000 documents for the negative folder. Furthermore, 1600 documents were used for training purposes and the other 400 was used for testing purposes.

Implementation stages

- Reading all the files into a list which removes any relevant spaces
- I have gone through all the relevant files by using the derive function
- I have implemented a count unigram function which gets all the relevant unique keys and implements a dictionary
- I have divided the data into two segments which are valid data set and the training set and a valid data set
- I have implemented and tested the training weight
- Another implementation is that of regarding the iteration, I have set the maximum iteration to ten moreover, it updates it by using the number of iterations as I have used random permutation function
- I have implemented the binary perception standard which can be considered in correlation with the training data set
- The weight has been taken into consideration only after the iteration for the sole purpose of taken into account the average for all the documents
- The program that has been implemented prints out the following:
- Reading the relevant files
- Time to extract relevant characteristics
- Time to build the matrix
- Time of dividing the training data set and the valid data set
- The amount of time it takes for training
- Positive
- Negative
- Precision
- Recall

Results

When running the program, it will print out the following demonstrated in a table form:

Features that the program outputs - unigram	Relevant files	Extract relevant characteristics	Building matrix	Training and valid data	Time to train	Positive	Negative	Precision	Recall
Results/time	0.04 seconds	1.27 seconds	15.96 seconds	0.61 seconds	126.71 seconds	86	84	0.86	0.84

Recall is the portion of relevant documents returned and on the other hand precision is the portion of retrieved documents that are relevant. Moreover, both precision and recall take into consideration the relationship between the retrieved and relevant files.