Report - TP SD214

Renata PORCIUNCULA BAPTISTA

May 30, 2018

Answers

Part I - Implementation of the classifier

1-

Code in anex.

2-

They classify if a review is positive or negative, extracting the actual value of the rate, where the maximum rating is specified explicitly. They were able to recognize numbers e.g, 8/10 , four out of five or stars ****:*** and letters systems. After that they choose two threshold from which greater than: positive or less than : negative.

3-

Code in anex.

4-

To do that, we had to adapt the function to be able to use the function cross_validation. The modifications was mainly due to the functions fit and predict had to return an array and only receive this parameters (X,(y)). The performance is worst, because /

5-

Part II - Scikit-learn use

1-

Scikit-learn implementation is better, in the term of count_vectorizer.

2-

Logistic Regression is better.

3-

Stemmer improved the results once that reduces the number of features and agrupates in same class.

4-