U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

IA - Group 85_3C

Artificial Intelligence - Natural Language Processing

Yelp Reviews

Diogo Câmara - up201905166
Mário Ferreira - up201907727
Pedro Moreira - up201904642

Definition of the machine learning problem

- Dataset of Yelp reviews
- Each review is defined by:
  - a review text
  - a score from 1 to 5 stars
- Our objective is to predict the score of any given review text using Natural Language Processing algorithms

Implementation work

- Observing the distribution of the review ratings in the train and test datasets
- Word clouds for each type of rating
- Sampling of the training dataset
- Filtering/Processing the samples
- Training the models:
  - Naive Bayes MultinomialNB
  - Naive Bayes ComplementNB
- Performance of the models and validation

Related work

- Work performed on the 7th practical class
- Classification Yelp Reviews
- A Guide to Automated Deep/Machine Learning for Natural Language Processing: Text Prediction
- Natural Language Processing - Python

Details on data pre-processing

- Removal of non alphabetic characters
- Lower casing every word
- Removal of stop words using the nltk library
- Stemming using the PorterStemmer from the nltk library

Developed models

- Bag of Words model
  - Using CountVectorizer from sklearn library
  - Using TfidfVectorizer from sklearn library

Tools and algorithms

- Anaconda
- Python libraries: pandas, matplotlib, re, nltk, wordcloud, sklearn
- Algorithms: Naive Bayes (MultinomialNB and ComplementNB)