Università
della
Svizzera
italiana

**Facoltà
di scienze
informatiche**

Data Analytics

Academic Year 2022-23

Course Assignment N. 20: Movie Reviews

Prof. Fabio Crestani

For this assignment you will work individually to carry out simple tasks of data analysis given a specific dataset. The goal of this assignment is to use Python and complementary libraries on a given dataset in order to *explore* and *analyze* the given data and *draw conclusions*.

Description

This data originates from snippets from the Rotten Tomatoes HTML files containing movie reviews. The corpus is based on the dataset introduced by Pang and Lee (2005) and consists of 11,855 single sentences extracted from movie reviews. Each sentence/phrase can be categorized into one of the five classes -- very negative, negative, neutral, positive, very positive depending on the probability threshold assigned by the sentiment analysis algorithm.

Consequently, your goal is to build models in order to predict the sentiment of a sentence/snippet based on the provided features (or a set of them). Your tasks are to:

- Explore and describe the data (*i.e.*, standard descriptive statistics, visualize the variables with different graphs, draw distributions and histograms of variables, are there outliers? Any interesting observation? Any correlations? Etc.)
- Pre-process the data (*i.e.*, handle and fill unknowns if there are any, etc.)
- Build one sentiment classification model for estimating the sentiment polarity (positive, negative, …) using the training data
- Evaluate the accuracy of the model using the test data

Submission procedure and evaluation

You should produce a report of your work and its evaluation along with the source code. It will be a concise explanation of how you tackled the different tasks, the reasons of your choices, successive conclusions, graphs you produced, results of the decisions and their accuracy *etc*.

Use Jupyter Notebook to produce results of the commands in a single .ipynb file. For more information check: https://jupyter.org/documentation

The report (max 5 pages) and the code of the project need to be submitted via iCorsi.

Please, upload all the required items in a single file and name it following the structure: **noProject_FirstnameLastname.[zip|tar.gz|7z]**. For instance, 05_ NameSurname.tar.gz The dataset regarding this project can be downloaded from the link provided on iCorsi.