CMPE 462 - Spring 2021
Machine Learning Project
Sentiment Analysis on IMDB User Reviews

Introduction

In this project, you are going to apply machine learning techniques on text data. You will propose and implement a feature extraction/selection and classification model for sentiment analysis on IMDB user reviews. This is a group project and will be submitted in 3 steps. Details of each step, submission and grading are found in the following sections. In the implementation phase, you will be using a conda virtual environment for Phyton. Details of this environment can also be found in the following sections. Set your environment as soon as possible in order to keep up with the project timeline.

Sentiment analysis is the process of understanding the underlying sentiment of a given text, which can be positive, negative or neutral. Sentiment analysis is actually applied and popularly used for analyzing social media messages in order to get an understanding of what the general opinion about a specific subject is. Elections and marketing are two main example areas.

You will implement sentiment analysis on IMDB user reviews. In a film main page on IMDB, you can reach user reviews about that film from the top menu (Figure 1). This page lists the reviews (Figure 2). Each review has a title, text and a corresponding rating score. There is also a user ratings page where you can observe the distribution of the ratings (Figure 3). In this project, you will design and implement a machine learning system that takes a user review as input and decides its sentiment. The target classes are Positive (P, for ratings 7,8,9,10), Negative (N, for ratings 1,2,3) and Neutral (Z, for ratings 4,5,6).

Step1: Data Collection

The first step of the project is data collection. You will select reviews with the criteria given, save them and check them.

Each project group will :

- be assigned an English letter, say S.
- collect user reviews for films having (English) names starting with your assigned letter, say Scarface, Star Wars.
- collect 150 user reviews in English: 50 Positive (with rating 7,8,9,10), 50 Negative (with rating 1,2,3), 50 Neutral (with rating 4,5,6)

The IMDB user reviews have three main parts that are important for the project. The first one is the header, the second one is the review text and the third one is the rating. You will save each review in a separate txt file. The first line of the file will be the header of the review. The rest of the file is the review text. There is a naming convention for the file. There are three parts of the file name connected with "_": 1. starting letter, say S, 2. your index of the file, 1:150, 3. class label for the user rating given with the review. For the first review in Figure 2, the corresponding file name will be S_1_P.txt, stating that this is a review for a film starting with S, this is the first of reviews for films starting with S and the review rating is Positive.

Before submitting the review files for Step1 of the project, you will check your files if there exists any special characters that will cause error while reading with Python. You can use the below code and make sure that all your files can be processed without errors.

```
with open(<filename>, 'r') as f:
        lines = f.readlines()
        print(f)
```

If there are errors for some files, contact me and we will analyze and resolve the errors together. There can be situations to ignore the selected review and select a new one. Therefore, apply your checks while you are building your dataset.

Step2: First Run

Details will be announced later.

Step3: Second Run

Details will be announced later.

Project Base Environment

You will be implementing your code with Python 3.6.
You need to create a python virtual environment with Anaconda for your project. After installing Anaconda, a base environment can be created with below commands:

```
conda create -n 462project python=3.6
conda activate 462project
```

While you keep working on your models, you will need to import additional libraries. List these libraries in a requirements.txt file. State any special versions if needed. A sample requirements file can be as below:

```
scikit-learn >= 0.22.2
scipy
pandas
sentencepiece==0.1.91
```

For grading, we will load your requirements with the command below:

```
python3 -m pip install -r requirements.txt
```

Before submission, test your code on a clear new conda environment by installing additional libraries from your requirements file. Because, there will be penalty if your code doesn't run like this.

Grading Details

The project will be graded over 100 points. You will be graded for your code and project reports.

- 20 points for Step 1
- 40 points for Step 2
- 40 points for Step 3

Additional details will be announced later.

Submission Details

This is a group project. Your code should be original. Any similarity between submitted projects or to a source from the web will be accepted as cheating.

If you have any further questions, send an e-mail to the course assistant: ozlem.simsek@boun.edu.tr

Step1

- The deadline for submitting Step 1 is **April 20, 2021 - 23:59**.
- For txt files: You should compress all your txt data files in a zip file with name as the assigned capital letter, say S.zip
- For project report:
  - You should submit a detailed project report in pdf format.
  - Clearly state group members and which member did what for this step.
  - You should name your report as step1_report_<teamname>.pdf, say step1_report_TeamA.pdf
- Submit max 2 items in a big zip file: txt files zip and report pdf. Name your submission zip file as step1_<teamname>.zip, say step1_TeamA.zip
- The final zip will be submitted on Moodle. Only one member of each group will make the submission.

Figures

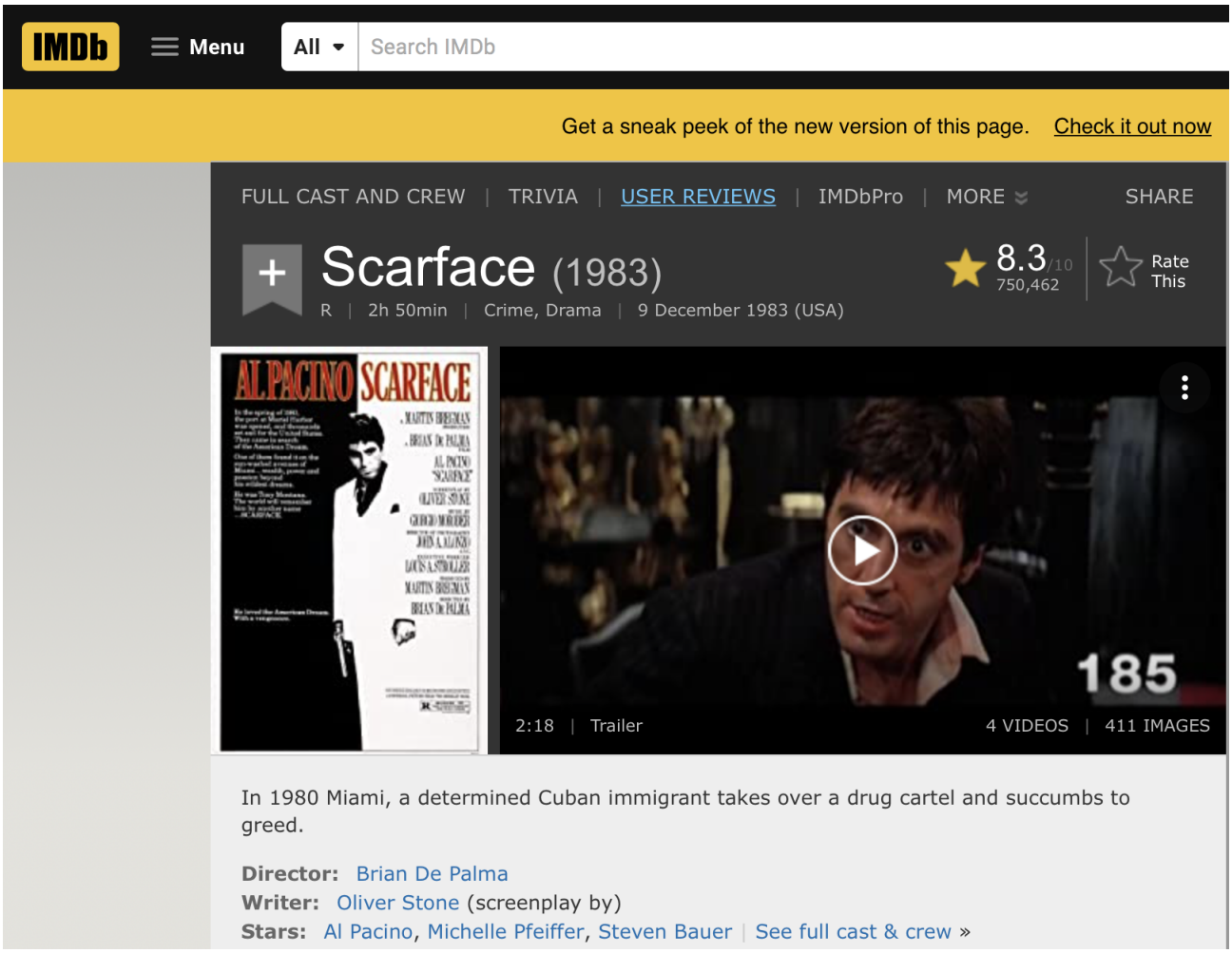

Figure 1: IMDB - film main page

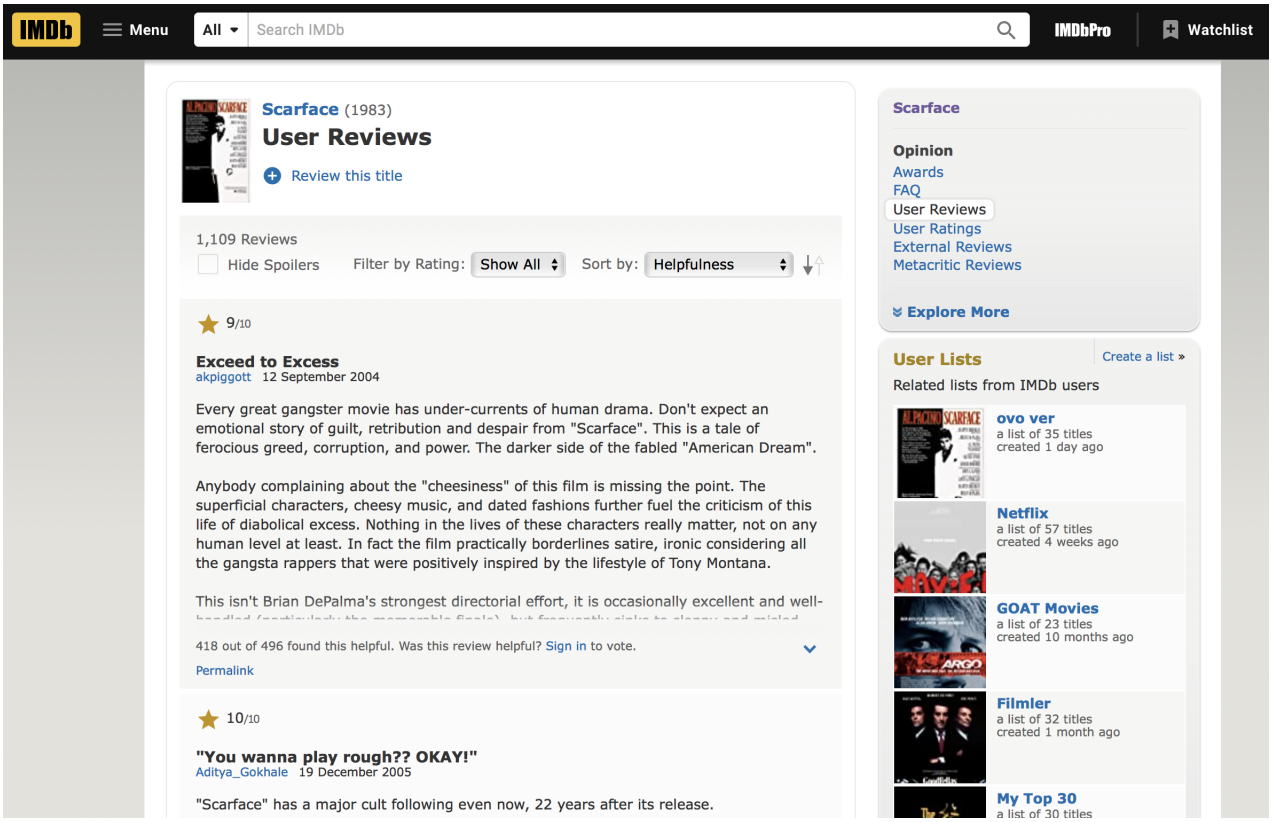

Figure 2: IMDB - user reviews

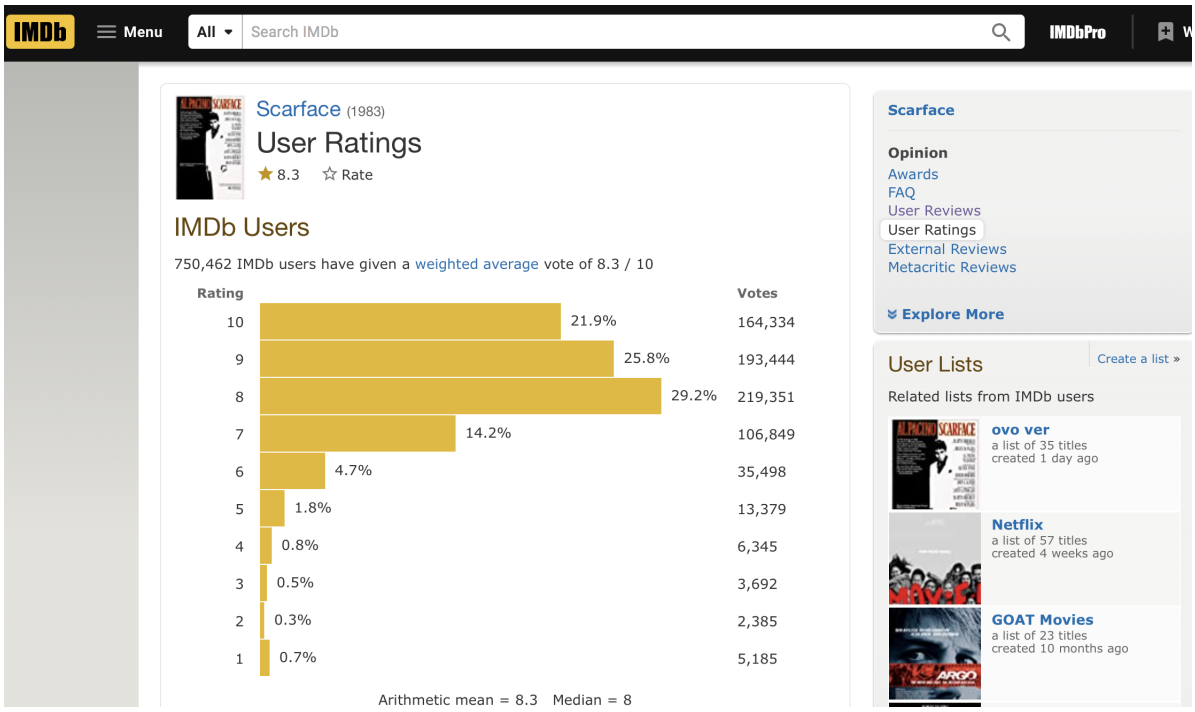

Figure 3: IMDB - user ratings