Çağrı Çiftçi 2016400243
Öykü Yılmaz 2019400294
M. Erdinç Oğuz 2017400267

MACHINE LEARNING PROJECT REPORT
SENTIMENT ANALYSIS ON IMDB USER REVIEWS

STEP 1

DATA COLLECTION

Collection Method

The data necessary is collected with using Python. IMDbPY library is used with some modifications. Since the library was not enough to satisfy our needs. The changes made are listed below:

- In movieParser.py:

The tag that contains the title of the review was taken as if it was in a <div> tag. But currently, the site puts the title in an <a> tag (line 1603).

```
        extractor=Path('.//div[@class="title"]//text()')
```

was changed to:

```
        extractor=Path('.//a[@class="title"]//text()')
```

- In movieParser.py:

When retrieving the rating, library function was only returning the ratings only if they had length two. This corresponds to ratings with value 10. Besides, the function returned only the first digit of that rating, i.e. ‘1’. If the rating had only one digit, the output was “null”. The function is altered so that any rating is properly returned (line 1634-1635).

```
            if review.get('rating') and len(review['rating']) == 2:
                review['rating'] = int(review['rating'][0])
```

was changed to

```
            if review.get('rating') and len(review['rating']) <= 2:
                review['rating'] = int(review['rating'])
```

This method is chosen because in the future, if it is needed to increase the number of reviews it can be done by changing only one or two lines of code.

Details of Data Collection

The task was to retrieve reviews of the movies that start with the letter F. To do that, the top 250 IMDb movies and top 250 Indian movies are collected. Because these had high rankings, to avoid bias in data two words with initial 'f' (fail and fast) were searched among the movies. Maximum of twenty-five reviews were taken from a specific movie, also to prevent bias. Reviews taken are shuffled before being sorted by their ratings. Finally, they are written to files according to the instructions given. It has been realized that some (four) reviews are not encoded in python write function's default, hence we added a try catch check to 'write to file' part of code to validate the encoding of the review.

Summary of Work Done

Since the first part did not include heavy work, we met online for coding. First, Mehmet Erdinç Oğuz made a research about the library and started the initial project. Afterwards, someone shared their screen and coding was done simultaneously. Also, the report was written in the same way.

GITHUB REPOSITORY LINK

https://github.com/eridincu/cmpe462_project