Spark (Scala)

Data: Movie Lens dataset

Loading and parsing the file

Q1: Download the ratings file, parse it and load it in an RDD named `data_rdd

Q2 : How many lines does the `data_rdd` RDD contain?

Q3 : Count how many times the rating '1' has been given.

Q4: Count how many unique movies have been rated.

Q5: Which user gave most ratings? Return the `userID` and number of ratings.

```
Array((userid,number))
```

Q6: Which user gave most '5' ratings? Return the `userID` and number of ratings.

Out[12]:

```
Array((userid,number))
```

Q7: Which movie was rated most times? Return the `movieID` and number of ratings.

Out[13]:

```
Array((movieid,number))
```

Q8: Read the `movies` and `users` files into RDDs. How many records are there in each RDD?

Q9: How many of the movies are a comedy?

Q10: Which comedy has the most ratings? Return the title and the number of rankings. Answer this question by joining two datasets.

Out[19]:

```
Array((American Beauty (yr),number))
```