Homework

Daniel García Hernández
Statistical Programming

February 29, 2020

The dataset with which you'll be working is a subset of the dataset `title.basics.tsv.gz` that contains the basic information of movies, shorts, tv episodes, etc released during 2019.

The whole dataset can be found in the imdb dataset repository with explanations of what it contains: **link**

Return a Jupyter notebook witht the following name `NameAndLastName_imdb.ipynb` **on Friday 6th March** containing the results and answers to the following points and questions.

One (1) point per correct answer/result.

1. **Open the dataset as a `pandas` DataFrame named `imdb`:** please note that the file is .tsv type. Investigate what this is, and how to pass a different separator value `sep=""` when using `pd.read_csv`
2. **How many types of titles are there in the column titleType? No for loops allowed!** Check `pandas.unique, pandas.Series.value_counts, set`
3. **Create a slice of `imdb` that only contains the following columns:**
   - `titleType, primaryTitle, startYear, runtimeMinutes`
4. **Create a subset of `imdb` named `tvEpisodes_2019` that only includes the type `tvEpisodes`**
5. **Percentage of adult films over total releases in 2019.** Check `pd.Series.mean`
6. **Create a column named `words_in_title` that contains the total number of words in the title**. Use `map` alongside a function or a lambda function.
7. **What's the average value of `runtimeMinutes` for the type `short`?**
8. **Filter `imdb` to return `tvMovie` type with 3 or more words in the title, and less than 75 minutes of `runTimeMinutes`**