Universidade Federal de Minas Gerais
Departamento de Ciência da Computação

TCC/TSI/TECC: Sistemas de Recomendação

Programming Assignment #2
Content-based Movie Recommendation

Deadline: Mar 1st, 2021 23:59 UTC-3 via Moodle and Kaggle

Overview The goal of this assignment is to implement a content-based movie recommender by exploiting movie metadata obtained from OMDB.[1] **Note that only personalized recommender implementations are acceptable.** As discussed in class, various implementation choices impact the quality of content-based recommendations, including choices for content representation (e.g., unigrams, n-grams, concepts, named entities, latent topics), user profiling (e.g., by aggregating positive and negative feedback), and user-item similarity estimation. As part of this assignment, you should try different instantiations of these components, and verify the resulting recommendation performance of your implementation by submitting your produced recommendations to Kaggle.[2]

Kaggle This assignment uses Kaggle in Class as a platform for automatically evaluating the quality of your produced recommendations. If you do not yet have a Kaggle account, you can register by clicking on the following link:

`https://www.kaggle.com/t/6fbe21f9887f42588ca54d1fd509e335`

Make sure to use your **matriculation number** as your "Team Name".

Teams This assignment must be performed individually. Any sign of plagiarism will be investigated and reported to the appropriate authorities.

Implementation You must use Python 3 for this assignment. Your code must run in a virtual environment using only the libraries included in the provided

[1] `http://omdbapi.com/`
[2] `http://www.kaggle.com`

`requirements.txt` file. Execution errors due to missing libraries or incompatible library versions will result in a zero grade. To make sure you have the correct setup, you can test it in one of the Linux machines provided by the Department of Computer Science[3] using the following commands:

```
$ python3 -m venv pa2
$ source pa2/bin/activate
$ pip3 install -r /path/to/requirements.txt
```

Execution Your implementation should include a `main.py` file, which will be executed in the same virtual environment described above as follows:

```
$ python3 main.py content.csv ratings.csv targets.csv
```

Input Your implementation must take three CSV files as input:

- `content.csv`, containing 22,080 $\langle item, content \rangle$ pairs[4]
- `ratings.csv`, containing 336,672 historical $\langle user, item, rating \rangle$ tuples
- `targets.csv`, containing 77,276 $\langle user, item \rangle$ pairs for prediction

Note that each of these input files contains a header line. These files can be downloaded from the data description page on Kaggle.[5]

Output For each of the $\langle user, item \rangle$ pairs in the `targets.csv` input file, your implementation should predict the corresponding numeric rating, by leveraging the item metadata available from `content.csv` and the historical user-item matrix available from `ratings.csv`. You **must not** use any historical information about the target item. The resulting prediction should be written to standard output[6] as a $\langle user, item, rating \rangle$ tuple, formatted as two CSV columns:

- `UserId:ItemId`, containing the $\langle user, item \rangle$ pair separated by a colon (:)
- `Prediction`, containing the predicted rating for the target pair

Submissions The predictions output by your implementation should be written to a submission file, to be uploaded to Kaggle. In total, a submission file must contain a header line plus one line for each of the n predictions. For this assignment, $n = 77{,}276$, meaning that your submission should have $n + 1 = 77{,}277$ lines. An example submission file is provided below:

[3] `https://www.crc.dcc.ufmg.br/infraestrutura/laboratorios/linux`
[4] The `content` column is formatted as a JSON document.
[5] `https://www.kaggle.com/c/recsys-20202-cbmr/data`
[6] `https://en.wikipedia.org/wiki/Standard_streams`

```
UserId:ItemId,Prediction
u0000039:i0060196,6.38666836618544
u0000039:i0099077,3.10922975975217
u0000039:i0102926,8.02790645781112
...
u0038723:i1951265,2.98912620664908
u0038723:i2395427,6.44312165951174
u0038723:i2413496,0.13009045472507
```

Your submission should be uploaded to Kaggle[7] to be automatically evaluated. Through the course of this assignment, you should try multiple instantiations of the various components of your implemented recommender, in the hope of further improving its effectiveness. To this end, you can upload a maximum of 20 submissions per day to Kaggle. The platform will maintain a live leaderboard indicating the relative performance of your submissions in comparison to those by your fellow classmates. Keep track of the performance of your submissions, so you can analyze what worked in your final assignment report.

Deliverables Before the deadline (Mar 1st, 2021 23:59 UTC-3), you must submit a package file (`zip` or `tar.gz`) via Moodle containing the following:

1. Source code of your implementation;
2. The last submission file (csv) uploaded to Kaggle;
3. Documentation file (pdf, max 2 pages).

Grading policy This assignment is worth a total of 15 points, with the possibility of attaining up to 5 extra points. These points are distributed as:

- 5 points for your *documentation*, assessed based on a short (pdf) report[8] describing your implemented data structures and algorithms, their computational complexity, as well as a discussion of your attained results (e.g., based on the various submissions you uploaded to Kaggle).
- 5 points for your *implementation*, assessed based on the quality of your source code, including its overall organization (modularity, readability, indentation, use of comments) and appropriate use of data structures.
- 5 points for your *performance*, assessed based on the RMSE score of your last submission on Kaggle's private leaderboard[9] relative to the performance of your fellow contestants.[10]

[7] `https://www.kaggle.com/c/recsys-20202-cbmr/submissions`

[8] Your documentation should be no longer than 2 pages and use the ACM LaTeX template (sample-sigconf.tex): `https://www.acm.org/binaries/content/assets/publications/consolidated-tex-template/acmart-primary.zip`

[9] Your performance on Kaggle's public leaderboard will normally reflect your performance on the private leaderboard, provided that your solution does not overfit.

[10] In a nutshell: stay within the average, and you get roughly all 5 awarded points; try your best to outperform the average, and you get up to 5 extra points.

To be eligible for the performance grades, you must satisfy the following:

1. You must upload at least one submission to Kaggle within the time-frame of this assignment;
2. The source code that you submit (via Moodle) by the deadline should be able to precisely generate your last submission to Kaggle;
3. Your implementation should be able to execute correctly in a Linux environment under 5 minutes.