Assignment Recommendations Systems

Philipp Borchert

Case description

Introduction

LastFM is a well-known music streaming service that has open-sourced some of their data sets. Three of these data sets are, together with a readme file, included the folder 'Group Assignment Data' and serve as data sources for this assignment. For more information about the data sets I refer to the readme file.

- LastFM, a well-known music streaming service, is improving their recommendation system to provide their customers with a better experience and increase the amount time spent on their platform.
- You, as part of the internal data science department, are asked to improve the current system, which recommends the 10 most popular artists to all users.
- The company expects a benchmark including multiple recommendation models, as well as a written report containing your findings and important decisions. You should also elaborate on pros and cons of the benchmarked techniques.
- Apart from the general evaluation metrics (defined in the specific requirements), LastFM wants to encourage users to engage with a **variety** of different artists.

Deliverable

A detailed report:

- **Report**: PDF File.
- **Code:**
  - **Dashboard**: Python Script (.py file)
  - Optional: Additional Code in a Jupyter Notebook (.ipynb file)
- Presenting the benchmark, final model and a discussion of relevant decisions.
- The general guideline here is that you should provide all the code required for a colleague to be able to replicate the model and sufficient documentation so as for your colleague to understand the steps and choices you made.
- There is no '*page limit*' to your report but you are advised to be both **concise** and to provide sufficient insight and detail.

Hand-in

- **Deadline: March 8th, 2021, 20:00h.**
- The report must be uploaded to **IESEG-online**
  - In case there are issues with the website (e.g. file size) you can submit your assignment by email to **p.borchert@ieseg.fr**
  - Include yourself in cc of your email to us so that you can verify whether your email was sent correctly.

Guidelines

- The datasets are provided on the course platform (IESEG-Online – Communication Tools):
  - Artists.dat
  - Tags.dat

    - User_artists.dat
    - User_taggedartists.dat

- A detailed description of the variables in the datasets is provided below.
- The code examples presented in class can be used for all parts of the assignment but may require some adjustments for this specific business case.
- Try to display in the report both technical mastering of the various steps in the model development process, maturity in making choices with respect to understanding the business problem and importantly also communicating your results to the **manager of the data science department** of LastFM.

Specific requirements

The data is provided in separate files. Read and combine them in Python and transform the data to create input matrices for the recommendation systems.

- To create a base matrix for **content-based** recommendation systems, you should merge 'user_taggedartists.dat' and 'tags.dat' data.
- To create a base matrix for **collaborative filtering** you should use the 'user_artists.dat' data. Watch out, the data in the weights column is continuous and skewed. It might be a good idea to categorize the data.

Finally, make sure your base matrices for both content-based and CF recommendation systems contain the same set of users and items.

Benchmark

- Apply at least 4 different recommendation systems.
- Apply at least 2 different hybrid recommendation systems.
- Use **cross-validation** to evaluate and compare your models.
- Evaluate your model based on **RMSE, MAE, NDCG and F1-Score.**
- Qualitative assessment of your recommendations (only for the final model):
    - Display an example of how your recommendations would look like for a specific user
    - Explain why the selected recommendations are highest rated for this user (E.g. “Users who bought this also bought this…”, “You viewed X, how about this…”, etc.)
- Propose a recommendation strategy that encourages users to engage with a **variety** of different artists.

Note: The specific requirements above leave a lot of freedom in terms of how exactly to develop your storyline and report your findings. As mentioned before, carefully consider all the choices you make and explain them in your report.

Dataset description

1. **Artists.dat**

Variable	Name	Meaning	Type	Remarks
1	id	Unique identifier	Numeric	
2	name	Name of the artist	String	
3	url	Link to the LastFM artist profile	String	
4	pictureURL	Link to the artists' profile photo	String	

2. **tags.dat**

Variable	Name	Meaning	Type	Remarks
1	tagID	Unique identifier	Numeric	
2	tagValue	Tag description	String	

3. **user_artists.dat**

Variable	Name	Meaning	Type	Remarks
1	userID	Unique identifier	Numeric	
2	artistID	Unique identifier	Numeric	
3	weight	Play count	Numeric	

4. **user_taggedartists.dat**

Variable	Name	Meaning	Type	Remarks
1	userID	Unique identifier	Numeric	
2	artistID	Unique identifier	Numeric	
3	tagID	Tag user assigned to the artist	Numeric	
4	day	Day the user tagged the artist	Numeric	
5	month	Month the user tagged the artist	Numeric	
6	year	Year the user tagged the artist	Numeric	