Group Assignment

Introduction

In this practical assignment we are going to put in practice all the theoretical concepts that we have been reviewing in class, as well as the practical materials that we have reviewed in the lab sessions.

I propose you to work on a typical recommendation system scenario but with large business impact: product recommendation in E-commerce, i.e., Amazon. We are going to make use of a real Amazon Dataset collected by Julian McAuley (More on the dataset in the related section), composed by 24 different product domains (e.g., cell phones, clothing, beauty, etc.), each one provided as a separate file.

For the assignment you only need to work on one of these domains, whichever you like. (i.e., I recommend you to not use the largest ones because your experimentation will take much more time to execute, but it is up to you)

After you select a particular dataset to work on, this project will mainly consist two required steps and another two optional steps:

1. Create the training and testing dataset from the original dataset (required)
2. Create a Collaborative Filtering RS based on the user ratings (required)
3. Create a Content-based Recommender system leveraging the textual content associated to the reviews (required)
4. Create a hybrid approach to combine both CF and Content-based RS (optional)

Dataset

Researchers from UC San Diego released a complete Amazon dataset that is publicly available online (http://jmcauley.ucsd.edu/data/amazon/). This dataset contains user reviews (numerical rating and textual comment) towards amazon products on 24 product categories (there is an independent dataset for each product category). **We will use the "Small subsets for experiment" (the 5-core dataset) on the website, which can be downloaded directly from the website.**

The structure of the dataset is explained on the website with detailed examples. Basically, each entry in a dataset is a user-item interaction record, including the following fields:

- **user-id**: which is denoted as "reviewerID" in the dataset
- **product-id**: which is denoted as "asin" in the dataset
- **rating**: a 1-5 integer star rating, which is the rating that the user rated on the product, it is denoted as "overall" in the dataset
- **review**: a piece of review text, which is the review content that the user commented about the product, it is denoted as "reviewText" in the dataset
- **title**: the title of the review, which is denoted as "summary" in the dataset
- **timestamp**: time that the user made the rating and review

- **helpfulness**: contains two numbers, i.e., [#users that think this review is *not* helpful, #users that think this review is helpful]
- **Image:** for each product, the dataset product the image of the product in a form of a 4096-dimensional vector that is learned by a CNN deep neural network (these vectors are provided in an independent dataset "Visual Features", also in the website, which is very large)
- **Metadata:** some metadata information for each product, including product title, price, image URL, brand, category, etc. It is also provided as an independent dataset ("Metadata"), which is also very large.

For the CF RS you may only use the user-id, item-id, and ratings. For the design of more advanced recommendation algorithms that may achieve better prediction and recommendation performance, you may use other information sources such as review text, timestamps, images, or metadata.

Tasks

- **Data selection and preprocessing (Mandatory)**
  First you need to select a product category (from the "Small subsets for experiment") and download the related file to create a training dataset and a testing dataset from therein for the experiment. A recommended standard pre-processing strategy is that: for each user, randomly select 80% of his/her ratings as the training ratings and use the remaining 20% ratings as testing ratings.
- **Collaborative Filtering Recommender System (Mandatory)**
  Based on the training dataset, you should develop a Collaborative Filtering model/algorithm to predict the ratings in the testing set. You may use any existing algorithm implemented in Surprise (or any other library) or develop new algorithms by yourself. After predicting the ratings in the testing set, evaluate your predictions by calculating the RMSE.
- **Content-based Recommender System (Optional)**
  You should leverage the textual information related to the reviews to create a Content-based RS to predict the ratings for the users in the test set. I do recommend you to make use of the lab session related to the topic. As in the previous case, you should evaluate the predictions by means of the RMSE
- **Hybrid RS (Optional)**
  As an extra, you can propose a hybrid recommender system joining the operation of the 2 previously developed systems. To that end, you can make use of any of the ideas explained in class.

Submission

You must upload to the campus before the deadline your submission including:

- The code of your project (Markdown, Notebook, Script…). **The code must work and to produce the same results that you are reporting** (If you think that some explanation is needed in order to execute the code or some additional library is required, please create a small user guide)**. You do not need to send the dataset**
- Report summarizing your project, results, findings and conclusions. Think about it as the presentation that you will do to the relevant stakeholders.

Evaluation

- The ML Pipeline in your markdown: 70%
- Report summarizing your work: 30%

Deadline

The deadline for the assignment is November 20th