DATA MINING CUP 2021

Scenario

Before the pandemic, Johannes Gutenberg managed a flourishing little book shop in the historic city center of Mainz. He took great joy in building personal relationships with each of his clients, recommending books catered to their personal taste and assisting in widening their literary palette. In his city and beyond he developed a formidable reputation with a respectable base of loyal customers, who considered him more of a connoisseur than a traditional salesman.

Unfortunately, this loyal base of customers is not enough to make his business profitable. And so, like many traditional retailers, Johannes also relies on walk-in customers.

At the beginning of the pandemic, this source of revenue vanished. To keep his employees and cover ongoing costs, Johannes had to find an alternative form of revenue.

With great initial reservation, he decided to expand his business by launching an online shop, which he believed would save his beloved business from imminent bankruptcy.

At first, Johannes and his employees tried their best to provide suitable recommendations for every product manually. But as the number of products increased and associates worked to keep at least some personal contact to clients via phone and email, this manual process was just not feasible.

Today, Johannes is looking for a reliable recommendation system to provide a targeted recommendation to every product page. This solution should meet his high personalization standards and only require a small amount of manual support to implement.

Since Johannes is only interested in the best process possible, he decides to organize a contest to identify the best recommendation solution.

Data

In order to create a recommender model, the participants are provided with historical transaction and descriptive item data in the form of structured text files (.csv)

The data is provided in three individual files. One file containing the transactions (“transactions.csv”) one the descriptive item data (“items.csv”) and the final one (“evaluation.csv”) containing the template for the result submission.

Here are some points to note about the files:

1. The first line (top line) has the same structure as the data sets but contains the names of the respective columns (data fields).
2. A list of all the column names, which occur in the appropriate order, can be found in the “features.pdf” file as well as brief descriptions and value ranges of the associated fields.
3. The top row and each data set contain several fields, which are separated from each other by the “|” symbol.
4. Floating point numbers are not rounded. “.” is used as the decimal separator.
5. There is no escape character: quotes are not used.
6. The encoding is “utf-8”.

The “items.csv” file is a master data set that contains descriptive features. The features may be categorical or numerical. The list of features is explained in the “features.pdf” file. Each data line contains the description for one single item.

The “transactions.csv” file contains information about clicks, baskets and orders over a period of three months. Each line displays one transaction for one single item. All the attributes are described in the “features.pdf” file.

The “evalutation.csv” file contains a list of product IDs. This list is a subset of the products from the “items.csv” and the reference for the submission.

Entries

Participants may submit their results by **2 p.m. on 29 June 2020 (UTC+2 or CEST)**. The task description below explains how to submit entries.

Task

The goal for each participating team is to create a recommendation model based on historical transactions and item features. For any given product, the model should return its five best recommendations.

One file containing the following information should be used to send the solution data:

Column name	Description	Data type
ItemID	Unique identifier for each product	String
rec_1	Item identifier of the product with the best recommendation for the product in the "itemID" column	String
rec_2	... second best recommendation	String
rec_3	… third best recommendation	String
rec_4	… fourth best recommendation	String
rec_5	… fifth best recommendation	String

The key attribute for every prediction is the "itemID". The prepared subset of "itemID"s for the evaluation can be found in the "evaluation.csv" file.

Possible values for the "rec_1" – "rec_5" recommendation columns are the available values from the "itemID" column in the "items.csv" file. Note, the predicted recommendations shall not be limited to the items from the evaluation file.

The different columns are separated by the "|" symbol. A possible extract from the solution file might look like this:

itemID|rec_1|rec_2|rec_3|rec_4|rec_5
737039|774491|447268|787699|265939|860516
812772|842080|276953|482172|474665|489790
512879|976538|222918|394752|689554|251347
...

The solution file must match the specifications described in the Data section. Incorrect or incomplete submissions cannot be assessed.

The solution file must be uploaded as a structured text file (csv) to the Data Mining Cup website: **https://www.data-mining-cup.com/dmc-2021/.**

Please make sure that the mandatory boxes on the form are correctly and fully completed before uploading the data.

The name of the text file consists of the team's name and the file type:

"<Teamname>.csv" (e.g. TU_Gutenberg_1.csv)

The team's name was communicated to the team leaders when their registration was confirmed.

Evaluation

To compare the different recommender systems, we've launched a special website, through which participants and the general public can vote for the best recommendations.

Each participating team is asked to provide their solution for the above specified subset of books.

By entering the evaluation page, a book from that set will be drawn randomly and displayed with its cover and some descriptive information. For this product, three of the submitted solutions are selected and displayed underneath.

The website visitor can then decide which recommendations are most appropriate. The best recommendations receive points. The team with the highest normalized score wins. Since the submitted results of different participants appear by chance, a completely equal number of appearances is very unlikely.