Group Assignment

One member of your group can hand in the group assignment until Friday, 20 May, 17:00 via `https://filesender.surf.nl`. Send this to your tutorial teacher: Include n.m.i.a.fatahelrahmanabdulqadir@uva.nl as recipient. **Please compress all files into a single .zip or .tar.gz file and use GroupAssignment as subject line.**

Please include the following files:

File formats

- A set of scripts used to preprocess and analyse the data, and to build the recommender system
- The output files (if you work in Jupyter Notebooks, code and output can be integrated).
- The research report in .pdf format

Note: You may, but do **not** have to create a shared (public) github repository where you store all code and documentation. In that case, please include the link to the github repository together with the written assignment to your tutorial teacher.
Either way, make sure your code is well-documented!

Hand in the answer to tasks 1 and 3 as PDF file. You can hand in the answer to task 2 either *as one Jupyter Notebook-file, integrating code, output, and explanations* or *as one .py file containing the Python code and one PDF file with output and explanations.*

Compress all files into one single .zip or .tar.gz file with your name! If you want to compress your files on Linux, you can do so as follows. Imagine you have a folder called '/home/damian/takehome' in which you have everything you want to hand in, you can do

```
cd /home/damian
tar -czf /home/damian/Desktop/takehome-damian.tar.gz takehome
```

to create a file `takehome-damian.tar.gz` on your Desktop.

Then go to `https://filesender.surf.nl` and upload the file. You get a mail that confirms that you have handed it.

Good luck!!!

Teams

Form groups of around 4 students (5 is the maximum). Your tutorial teacher will help you with this.

Datasets

Together with your group, you can select from one of two datasets

1. Dataset on podcasts: `https://www.kaggle.com/listennotes/all-podcast-episodes-published-in-december-2017`. This link will lead you to two .csv files. You and your group may decide yourself whether you want to work with one of the files, or combine both of them.

2. Dataset on books. More specifically `google_books_1299.csv`: `https://www.kaggle.com/bilalyussef/google-books-dataset?select=google_books_1299.csv`.

For both datasets, it is important to note that you do not need to consider all the columns. Make a selection of relevant variables yourself, and argue in your assignment why you have decided to include or exclude specific information.

This dataset will form the basis of the assignment. Your task is to explore this dataset, describe it in a meaningful, data-scientific way, and ultimately, to build a recommender system.

The group assignment consists of three tasks: Writing a research report, exploring a dataset, and building a simple knowledge-based or content-based recommender system. Specifically, the following tasks are part of this assignment:

1 Write a research report (30% of final grade)

The research report consists of...

Method section

- A description of the steps you took, which type of variables were selected and how they were transformed.
- Explain your analytical strategy;
- What techniques (and why) will you be using to describe the dataset?
- What type of recommender system are you building? Why?

Result section

- A description of the dataset (how many observations, what type of variables)
- Results of the inductive analysis (e.g., description of the topics you've found).
- Demonstration of the recommender system; explanation of how it works, and some examples from the type of recommendations you get for different types of input.

2 Explorative data analysis (30% of final grade)

Explore, pre-process, and clean the dataset, and provide some descriptive analyses.

- Explore the dataset, and inspect what type of relevant variables are present, what data can be used. Select which variables might be of interest and can be used later on.
- Feature engineering is an important step here (keeping in mind the type of descriptive analysis you want to conduct in step 2). The literature and code examples from week 1 and week 2 should help you here.
- Describe the dataset using an inductive analysis.
- Provide a clear description of data you will be working with. E.g., describe the most interesting variables in terms of data 'type', number of unique observations, mean, distribution, etc.
- Plotting the data, to visualise some of the relations in the dataset, is appreciated.
- Describe the dataset using some of the techniques as discussed in week 3 and week 4. For example, apply LDA to describe the number of topics present in the dataset.

3 Recommender system (20% of final grade)

Build a simple knowledge-based or content-based recommender system.

- Build a recommender system, based on the insights from week 6. It's up to you to decide whether you build a knowledge-based or content-based recommender system.
- Think about relevant features that you want to use in your algorithm design. Based on which features do you want to recommend content?

4 Quality of the code and documentation (20% of final grade)

Make sure your code is well documented and understandable for people that see your code for the first time.