Project Instructions

Data Mining

Stephen Spengler, Georgios Panayiotou, Bahri Uzunoglu

OVERVIEW

The objective of the project is to test the knowledge and skills acquired during the course on a real Data Mining process. This is, by design, an *independent* activity, where you are expected to identify relevant questions that can be answered using Data Mining methods, autonomously reason about the encountered problems and identify appropriate solutions. The project is performed on *real data*: nothing has been simplified to force some educational concepts to emerge, as done instead in the preparatory assignments.

GROUPS

The project is performed by the groups formed on Studium. However, grades are individual: You will be able to acquire 10% of the points as a group, and the remaining 60% will be determined by your individual performance during the final oral presentation. For this, it is crucial that every single group member knows the whole project and can answer all questions.

PROJECT STRUCTURE

- Interpret your results and be able to convince your examiner that your results can be trusted

SUBMISSIONS

Over the course of the project, you will have to submit the following documents in two stages:

PROPOSAL: A one-page project proposal, specifying the data and article that you plan to use:

1. If you chose to implement a single algorithm, please submit the PDF of the article describing it. In case the article contains multiple algorithms and you only want to focus on one/some of them, please indicate this as a note to the submission.
2. If you chose to perform a comparison of different methods, please submit a PDF document with a link to the survey paper (if any), to the library-ies you are planning to use, and a short description of how you plan to compare the different methods.
3. If we have feedback you will hear from us latest one week after the submission deadline. Otherwise you can consider the selection accepted and start looking into it in more detail.

Additionally, you must give a (brief) explanation on how you will be using Data Mining to answer your research question.

FINAL: A data exploration report, containing a description of the main parameters with basic benchmarking to the article chosen. The final presentation. It should contain an introduction into your research question and a *brief* overview over the article. The main part should focus on the algorithms that you have used and your results, validations, evaluations, benchmark and conclusion.

All submissions must be done on Studium before the deadlines stated there.

EXAMINATION

Each submission will be graded and counts towards the final grade of the project. To make the examination transparent, we provide detailed examination criteria as a checklist.

1. Project Proposal [10 points]
   - [5] The research question is clearly stated at the beginning, in non-technical terms.
   - [5] The research question has the relevant validation and evaluation data and documentation.

Note: After submission and grading, your project coordinator will accept your proposal or request changes. In the latter case, you will need to resubmit, but the grade will stay the same. If the requested changes are not addressed in the resubmission, the project will be graded as failed.

2. Modelling [90 points]

If clustering, evaluate the article cluster accuracy by investigating the following:

1. Evaluation methodology relevance: [15points]
2. Evaluation methodology results: [15points]
3. Evaluation methodology conclusion: [15points]

If association, evaluate the article association accuracy by investigating the following:

4. Evaluation methodology relevance: [15points]
5. Evaluation methodology results: [15points]
6. Evaluation methodology conclusion: [15points]

If classification, evaluate the article modelling accuracy by investigating the following:

7. Evaluation methodology relevance: [15points]
8. Evaluation methodology results: [15points]
9. Evaluation methodology conclusion: [15points]

Note: This part of the project will be assessed in an oral clarification of the written project documents after the submission to assess the written project results. You will be able to book an examination time for your group after the groups have been finalised on Studium. During the examination, each group member will individually answer exactly one of the above questions, as chosen by the examiner. Therefore, everyone needs to be prepared to answer each question.

The grading will be as follows. Each question is worth 15 points. For each question answered by one of your group colleagues, you will get their scored points, which amounts to a maximum of 30 points. For each question answered by yourself, you will get four times the score, for a maximum of 60 points. This totals to 90 points.
Note that if a student achieves less than the required individual grade to pass the last assignment, their scores will not be counted towards the final grade.
For example: Assume a group consisting of members A, B, C, each answering two questions
The final individual grade for each member will look like the table below:

Question [ans. by, grade]	**A**	**B**	**C**
Q1 [A, 4/5]	60	15	15
Q2 [B, 5/5]	10	40	10
Q3 [C, 3/5]	8	8	32
Final individual grade (sum)	78	63	57

GRADING DETAILS

The project is on the 5 / 4 / 3 / U grading scale:

5 [100-90) Pass with distinction:	Outstanding performance with only minor errors.
4 [90-70) Pass with credit:	Generally sound work with a number of notable errors.
3 [70-50) Pass:	Fair but with significant shortcomings.
U [50-0) Fail:	Fail – considerable further work is required.

During the course, grading will be out of 100 internally and will be converted to Swedish System of U, 3, 4 or 5. Submissions are final, no changes are allowed after submission. If the task is delayed due to sickness or emergency, documentation needs to be provided. Otherwise project assignments submitted after the deadline will lose 20 points out of 100 for every day of late submission for 2 consecutive days. If the projects are submitted later than 2 days, grading will be done over 60 points out of 100 for that project.
Plagiarism is not allowed and suspected cheating leads to notification. Cheating will by default make you fail the assignment and/or exam and will cause more penalties.
If a student fails, a retake will be provided for the failed items that can make the student reach a passing grade for the late submission.

TOOLS

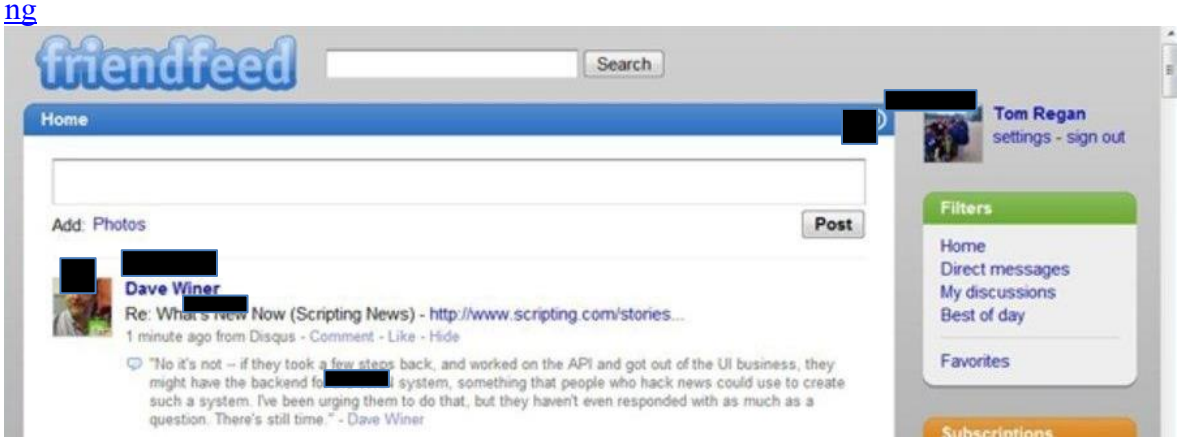
You are free to choose any tools to perform your analysis. A recommended combination is MySQL or any other relational DBMS (if you need a lot of initial preprocessing power and you want to exploit your knowledge of SQL) and RapidMiner, or only RapidMiner if you do not need database methods, but you are free to choose other tools/languages if you prefer, e.g., R, Python, etc. You do not need any approval for this.

You are expected to bring your process/code to the presentation, because we may ask you to execute some parts of it live, to change parameters, etc.

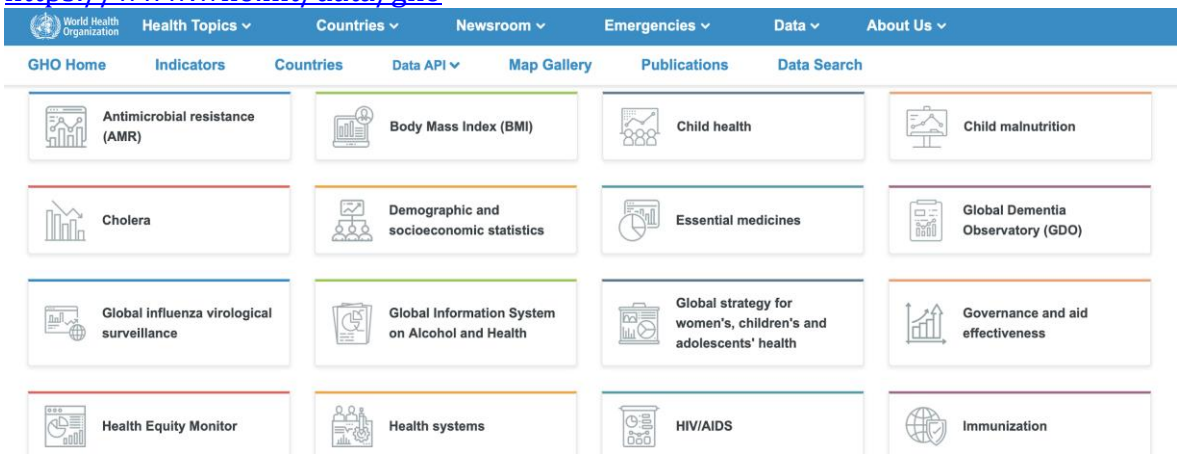
DATA

For your analysis, you can use a dataset of your choice and interest (that must be approved by your tutor) or one of the following data sources (that do not need approval):

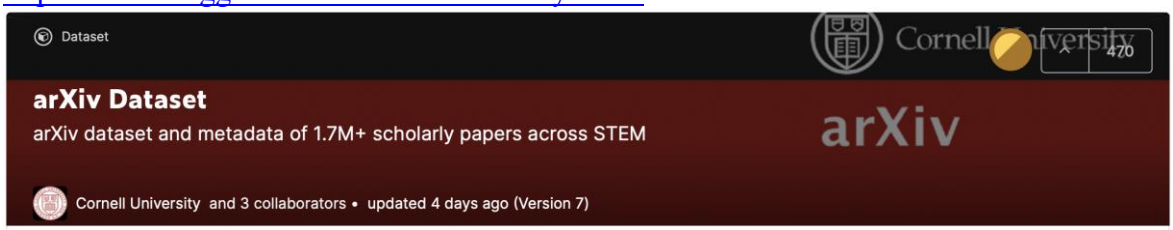
- Data from the Friendfeed study, obtained by monitoring an Online Social Network, with user posts, likes, following/followers, etc.: https://drive.google.com/folderview?id=0B_D5tuT1vDQtckFGWkk1aTh5VlE&usp=sharing



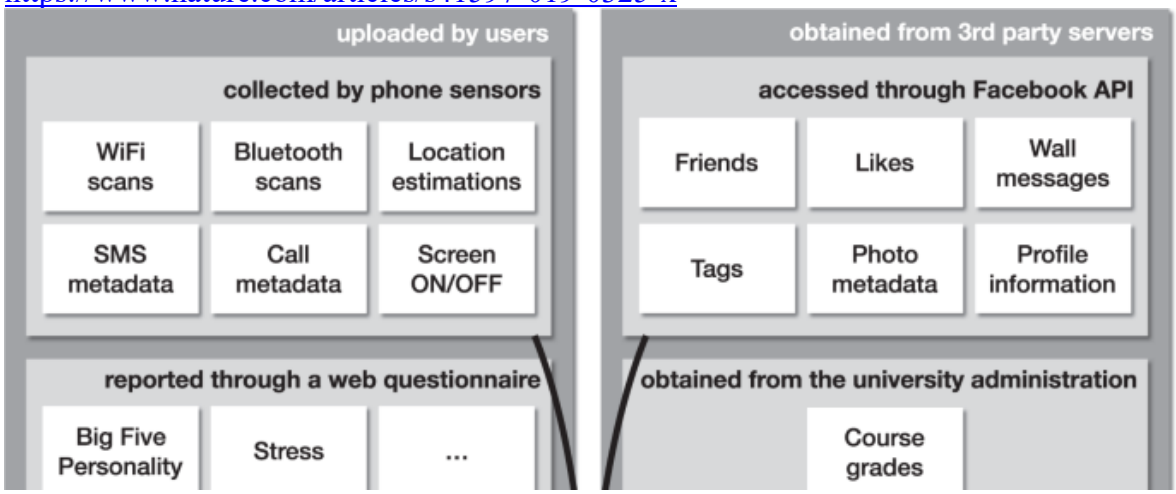
- Data from the Global Health Observatory Data Repository: https://www.who.int/data/gho



- Data from the arXiv repository, with information about research papers in STEM: https://www.kaggle.com/Cornell-University/arxiv



- Data from the Copenhagen Networks Study, with interactions between university students: https://www.nature.com/articles/s41597-019-0325-x



These data sources should give you an indication of what we expect if you choose your own data. In general, we will *not* accept simple data sources that have already been prepared for a specific analysis, as is the case for many Kaggle datasets. Some Kaggle datasets (like the one listed above) can still be ok, if you choose an original question requiring some preprocessing and non-trivial analysis.

Independently of the chosen dataset, consider that you will need to spend most of the project time understanding, retrieving, and preprocessing the data. This is one of the intended learning outcomes of the project, that cannot be obtained with the lectures.

SUPPORT

This project tests your ability to independently design and execute a knowledge discovery process on real data validate and benchmark the existing results.

Therefore, you should work independently on this project. Nevertheless, if you have further questions, feel free to contact your project coordinator at any time.

We hope you enjoy this experience,
and we look forward to hearing your presentations!