«Algorithms for massive datasets»

Master in Data Science for Economics

Projects for 2020-21

This document describes two projects for the «Algorithms for massive datasets» module taught at the Master in «Data Science for Economics» at the Università degli Studi di Milano. Submitting a successful project is compulsory before being admitted to the oral exam. The following sections describe the tasks and datasets associated with each project. The last section covers the details of project implementation and submission.

Note that this document does not refer to the analogous course taught at the Master in «Computer Science» at the Università degli Studi di Milano. Students from this master should refer to a [specific document](https://www.google.com/url?q=https://www.google.com/url?q%3Dhttps://docs.google.com/document/u/0/d/1aB-gpEdge7jscDwasyJhVwARdV_L0Ine_Jso6oviY7o/edit%26amp;sa%3DD%26amp;source%3Deditors%26amp;ust%3D1643228493365918%26amp;usg%3DAOvVaw3HjUBOKtf-IkVfH0wMbVvq&sa=D&source=docs&ust=1643228493376992&usg=AOvVaw1uq4fNyjFo_XmfNbXK4f2_).

Project 1: Finding similar items

The task is to implement a detector of pairs of similar items, analyzing one of the two datasets described here below.

Bird species

The «[250 Bird Species](https://www.google.com/url?q=https://www.google.com/url?q%3Dhttps://www.kaggle.com/gpiosenka/100-bird-species%26amp;sa%3DD%26amp;source%3Deditors%26amp;ust%3D1643228493366508%26amp;usg%3DAOvVaw1eJB_lIOb0AwJWyz1ujxfP&sa=D&source=docs&ust=1643228493377347&usg=AOvVaw0VZZYBSfsNtRLN_ys3E4uZ)» dataset is published on Kaggle. The detector must consider the images in the dataset and output the pairs inferred as similar.

StackSample

The «[StackSample](https://www.google.com/url?q=https://www.google.com/url?q%3Dhttps://www.kaggle.com/stackoverflow/stacksample%26amp;sa%3DD%26amp;source%3Deditors%26amp;ust%3D1643228493366834%26amp;usg%3DAOvVaw06JizwHkR4KZOj93rneYII&sa=D&source=docs&ust=1643228493377570&usg=AOvVaw11OUR6LoZtGaH85wkUxo6f)» dataset is published on Kaggle and released under the CC-BY-SA 3.0 license, with attribution required. The detector must consider the *Body* column of the Questions.csv file and output the pairs inferred as similar.

Project 2: Market-basket analysis

The task is to implement a system finding frequent itemsets (aka market-basket analysis), analyzing one of the two datasets described below.

IMDB

The «[IMDB](https://www.google.com/url?q=https://www.google.com/url?q%3Dhttps://www.kaggle.com/ashirwadsangwan/imdb-dataset%26amp;sa%3DD%26amp;source%3Deditors%26amp;ust%3D1643228493367370%26amp;usg%3DAOvVaw2RFetNE9_S_O4wQb51lC0W&sa=D&source=docs&ust=1643228493377818&usg=AOvVaw2TL56chBu49g-S-IaMqrLV)» dataset is published on Kaggle, under [IMDb non-commercial licensing](https://www.google.com/url?q=https://www.google.com/url?q%3Dhttps://www.imdb.com/conditions?pf_rd_m%253DA2FGELUUNOQJNL%2526pf_rd_p%253D3aefe545-f8d3-4562-976a-e5eb47d1bb18%2526pf_rd_r%253DK7VGSA5BY26HTH7KAGZV%2526pf_rd_s%253Dcenter-1%2526pf_rd_t%253D60601%2526pf_rd_i%253Dinterfaces%2526ref_%253Dfea_mn_lk2%26amp;sa%3DD%26amp;source%3Deditors%26amp;ust%3D1643228493367620%26amp;usg%3DAOvVaw0ss9ZQnTYQmpGi_N10Hq-Z&sa=D&source=docs&ust=1643228493377980&usg=AOvVaw32f44UjOmkPY5olt7OPvxQ). The analysis must be done considering movies as baskets and actors as items.

Old newspapers

The «[Old Newspapers](https://www.google.com/url?q=https://www.google.com/url?q%3Dhttps://www.kaggle.com/alvations/old-newspapers%26amp;sa%3DD%26amp;source%3Deditors%26amp;ust%3D1643228493367973%26amp;usg%3DAOvVaw36aerY5VthSDhs9MoW3qOu&sa=D&source=docs&ust=1643228493378184&usg=AOvVaw0z9sdx4bbGqZYLzt7yO7Wt)» dataset is published on Kaggle and released under the public domain license (CC0). The analysis must be done considering values of the *Text* attribute as baskets and words as items.

Project 3: Link analysis

The task is to implement a system ranking nodes in a graph using the PageRank index (or other approaches based on link analysis), processing the IMDB dataset described within Project 2. In this case, nodes in the graph will identify actors, and an edge will link two nodes if the corresponding authors played at least once in the same movie.

**Project implementation**

**Important:** the techniques used in order to analyze data have to scale up to larger datasets.

The project can be carried out individually, or in groups of two students. Code should be written in Python 3.

The project should be made available through a public github repository, containing code and a report describing the work done. The dataset should not be added to the repository, but downloaded during code execution, for instance via the kaggle API ([https://github.com/Kaggle/kaggle-api](https://www.google.com/url?q=https://www.google.com/url?q%3Dhttps://github.com/Kaggle/kaggle-api%26amp;sa%3DD%26amp;source%3Deditors%26amp;ust%3D1643228493368759%26amp;usg%3DAOvVaw24hOH3Z5x9zZq2QaDthbL5&sa=D&source=docs&ust=1643228493378527&usg=AOvVaw1la7SbF79TCW3wxirnLs1j)). Code should be implemented using a jupyter notebook executable on Google colab, possibly adding a badge/link directly from the repository to the colab version of the notebook.

The project report, preferably written in LaTeX, will be evaluated according to the following criteria:

* correctness of the general methodological approach,
* replicability of the experiments,
* correctness of the approach,
* scalability of the proposed solution,
* clarity of exposition.

The report should contain the following information:

1. the chosen dataset, and the parts of the latter which have been considered,
2. how data have been organized,
3. the applied pre-processing techniques,
4. the considered algorithms and their implementations,
5. how the proposed solution scales up with data size,
6. a description of the experiments,
7. comments and discussion on the experimental results.

The report must also contain the following declaration: *“I/We declare that this material, which I/We now submit for assessment, is entirely my/our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my/our work. I/We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study.“*

If the proposed solution is based on the ones published in Kaggle, this must be clearly stated, and the report should explain the differences and compare the experimental results.

Once the project has been finalized, students should send an email to Prof. Malchiodi (malchiodi AT di DOT unimi DOT it), specifying

* their names and student IDs,
* their enrollment in the Master in Data Science for Economics,
* a github link to the project.

After the project is evaluated, students will be able to schedule an appointment for the oral discussion.

These projects are valid for the academic year 2020/21.