AI506: Data Mining and Search (Spring 2020)

Term Project: Co-authorship Prediction

Release: Apr 08, 2020
Progress Report: May 15, 2020, 11:59pm
Final Presentation: June 30 or July 2, 2020
Final Report: July 3, 2020, 11:59pm

The ultimate goal of this project is a practice of data mining research by addressing the co-authorship prediction problem. In this project, you will design, implement, and evaluate your own approach for predicting future collaboration based on past collaboration history. Also, you will (a) write a progress report (b) present your approach, and (c) write a final report. While details of the following steps will be announced later, tentative schedules are as follows:

- Progress Report (Max Score: 20) - May 15, 2020, 11:59pm
- Presentation (Max Score: 20) - June 30 or July 2, 2020
- Final Report (Max Score: 60) - Jul 3, 2020, 11:59pm.

This is a team project, and a team should consists of two or three members. You can find your team mates by all means (e.g., Classum), and one progress report should be submitted per team.

Your submission will be evaluated based on

- Presentation of your reports & presentation - 40%,
- Novelty of your proposed approach - 20%,
- validity of your proposed approach - 20%,
- **Accuracy - 20%**.

Note that the accuracy is not our only concern. Instead of spending all your time optimizing the accuracy, we recommend spending more time developing novel and valid approaches and making your presentation clear and complete.

1 Problem: Co-authorship Prediction

1.1 Provided Data

You can use the provided training data. Below, we provide some details of the training data.

1. Paper-Author dataset (`paper_author.txt`): Author information of past publications is included in the file. The first line of the file contains the number of authors $|N|$ and the number of publications $|M|$. Each $i+1$-th line provides the ids of the authors' of the i-th publication. **There is no fake collaboration injected in this dataset.**

2. Public and private query sets (`query_public.txt, query_private.txt`): For each query set, the first line of the file contains the number of queries $|Q|$. Each of the next $|Q|$ lines contains several integers, which denotes the ids of the authors of future collaboration. **Unlike the paper-author dataset, these datasets contain some fake collaboration as well as real collaboration.**

3. Answers for the public query set (`answer_public.txt`): This file contains the ground-truth labels of each collaboration in the public query set. Each i-th line provides "True" or "False", which indicates whether the collaboration in the $(i+1)$-th line of `query_public.txt` is real or fake.

1.2 Evaluation

You should submit `answer_private.txt`, which contains the predicted label of each collaboration in the private query set. Using this file, we will evaluate the accuracy of your approach. The **format of the file must be exactly the same as** `answer_public.txt`.

Note that we may ask you to run your submitted code on another query set if your answer is suspiciously similar to any other group's answer.

1.3 Notes

- You may encounter some subtleties when it comes to implementation, please come up with your own design and/or contact Jihoon Ko (jihoonko@kaist.ac.kr) and Hojoon Lee (joonleesky@kaist.ac.kr) for discussion. Any ideas can be taken into consideration when grading if they are written in the *readme* file.

- Unlike the other assignments, you are allowed to use any programming language and any external libraries.

2 How to submit your project

2.1 Progress Report

Submit your progress report that is written using the attached template to KLMS by May 15, 2020, 11:59pm. The file should be named report-[your student ids].pdf (e.g., report-20189000_20199000_20209000.pdf).

2.2 Final Submission

1. Submit project-[your student ids].tar.gz (e.g., project-20189000_20199000_20209000.tar.gz) to KLMS. Your submission should contain the following files:

   - **final_report.pdf**: a final report that is written using the attached template written in LaTeX
   - **slides.pdf**: slides used for final presentation
   - **answer_private.txt**: the predicted label of each collaboration in the private query set, which we provide.
   - **readme.txt**: this file should contain the names of any individuals from whom you received help, and the nature of the help that you received. That includes help from friends, classmates, lab TAs, course staff members, etc. In this file, you are also welcome to write any comments that can help us grade your assignment better, your evaluation of this assignment, and your ideas. This file also should describe how to run your code.
   - **code.tar.gz**: your implementation

2. Make sure that no other files are included in the tar.gz file.