AI607: Graph Mining and Social Network Analysis (Fall 2021)

Term Project: Same Author Detection

Release: Sep 17, 2021
Progress Report: November 5, 2021, 11:59 pm
Final Report: December 3, 2021, 11:59 pm
Presentation Video: December 5, 2021, 11:59 pm

This project's ultimate goal is to practice data mining research by addressing the same author detection problem using a paper-author relation dataset. In this project, you will design, implement, and evaluate your approach for finding the same authors in the paper-author network. Also, you will (a) write a progress report, (b) write a final report, and (c) present your approach. While details of the following steps will be announced later, tentative schedules are as follows:

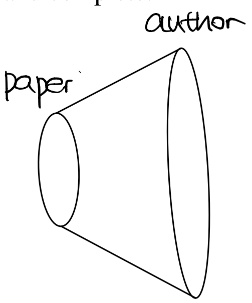
- Progress Report (Max Score: 20) - November 5, 2021, 11:59 pm
- Final Report (Max Score: 60) - December 3, 2021, 11:59 pm
- Presentation Video (Max Score: 20) - December 5, 2021, 11:59 pm
- Presentation - December 7 and 9, 2021

This is a team project, and each team should consist of two or three members. You can find your teammates 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 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.



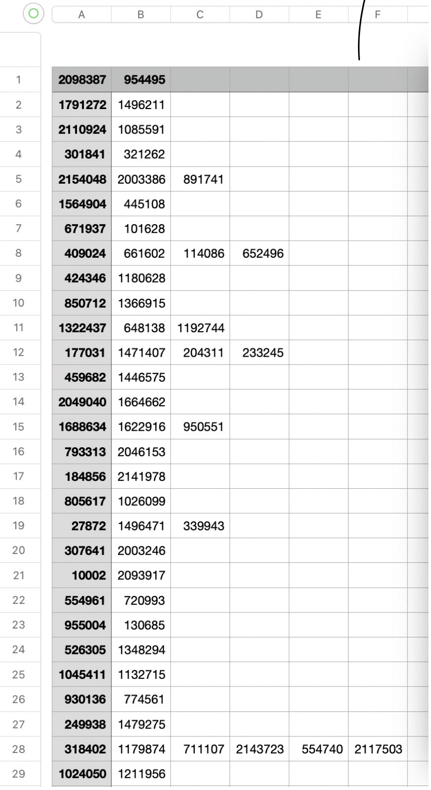
* BYOL Setting

⇒ Train에 있는 애들의 연결관계 제외 나머지 perturb..

Loss 2개 $\underbrace{\boxed{}}_{\text{BYOL's}}$ + prediction loss

* Start with little purturbation.

paper_ author_ relationship.csv



	A	B	C	D	E	F
1	2098387	954495				
2	1791272	1496211				
3	2110924	1085591				
4	301841	321262				
5	2154048	2003386	891741			
6	1564904	445108				
7	671937	101628				
8	409024	661602	114086	652496		
9	424346	1180628				
10	850712	1366915				
11	1322437	648138	1192744			
12	177031	1471407	204311	233245		
13	459682	1446575				
14	2049040	1664662				
15	1688634	1622916	950551			
16	793313	2046153				
17	184856	2141978				
18	805617	1026099				
19	27872	1496471	339943			
20	307641	2003246				
21	10002	2093917				
22	554961	720993				
23	955004	130685				
24	526305	1348294				
25	1045411	1132715				
26	930136	774561				
27	249938	1479275				
28	318402	1179874	711107	2143723	554740	2117503
29	1024050	1211956				

Only True.

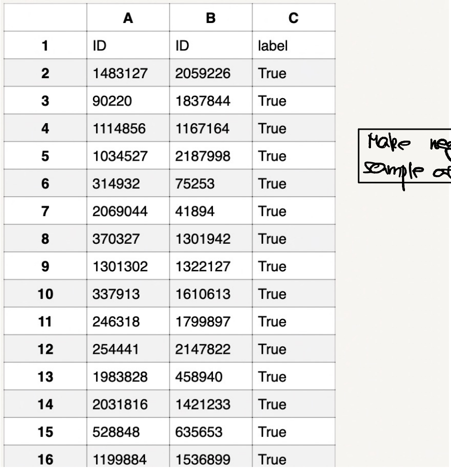
↑

train_ dataset.csv.

True + False

↑

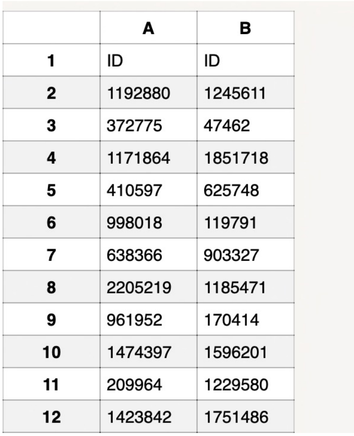
(valid dataset도 같은 Form)



	A	B	C
1	ID	ID	label
2	1483127	2059226	True
3	90220	1837844	True
4	1114856	1167164	True
5	1034527	2187998	True
6	314932	75253	True
7	2069044	41894	True
8	370327	1301942	True
9	1301302	1322127	True
10	337913	1610613	True
11	246318	1799897	True
12	254441	2147822	True
13	1983828	458940	True
14	2031816	1421233	True
15	528848	635653	True
16	1199884	1536899	True

Make negative sample of train set

query_ dataset. csv.



	A	B
1	ID	ID
2	1192880	1245611
3	372775	47462
4	1171864	1851718
5	410597	625748
6	998018	119791
7	638366	903327
8	2205219	1185471
9	961952	170414
10	1474397	1596201
11	209964	1229580
12	1423842	1751486

N = 100,000

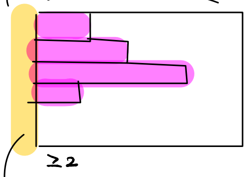
$\binom{100,000}{2}$

same author detection

paper id.

Authors id → ident matching

unique cnt : 61442



Paper # = 449006 : index

1 Problem: Same Author Detection

1.1 Provided Data

The provided dataset contains paper-author relationships, and author-ID pairs. Some of the pairs correspond to the same authors, while the others do not. In paper_author_relationship.csv, each line corresponds to one paper and contains the IDs of authors of the paper. This dataset consists of 61,442 authors and 449,006 papers. Note that each line of the paper_author_relationship.csv contains at least two author IDs, which means each paper contains more than two authors. Also note that the author IDs are (non-consecutive) integers.

데이터 결과좀 이상한데..?

unique authors

In this dataset, there are 3,000 author ID pairs that correspond to the same authors. Among them, 1000, 500, and 500 are included in the training (train_dataset.csv), validation (valid_dataset.csv), and query datasets (query_dataset.csv), respectively. Moreover, the validation and query datasets include 500 other author-ID pairs that do not correspond to the same authors. While the training and validation datasets contain labels, which indicate whether each pair corresponds to the same author or not, the query dataset does not contain labels. Each line of train_dataset.csv and valid_dataset.csv contains an author ID pair and a label, separated by commas. In addition, each line of query_dataset.csv contains only an author ID pair, also separated by commas.

2000

1000쌍이 남았다.

	# Same author ID pairs	# Different author ID pairs
Training dataset (train_dataset.csv)	1,000	0
Validation dataset (valid_dataset.csv)	500	500
query dataset (query_dataset.csv)	500	500

All True

True(500)+False(500)

but no " labels

Table 1: Statics of training, validation, and test datasets

Goal : 2개.

The project's goal is to find the same authors among the author ID pairs in the query dataset and to predict the remaining 1,000 same author ID pairs not included in the training, validation, and query datasets. The prediction of the same authors among the author ID pairs in the query dataset must be saved in the same format as the training dataset (train_dataset.csv) or validation dataset (valid_dataset.csv), and it should be named query_answer.csv. The prediction of the remaining 1,000 same author ID pairs also must be saved in the same format as the query dataset (query_dataset.csv) and it should be named same_author.csv. The evaluation metric is described in the next section.

1.2 Evaluation

You should submit query_answer.csv and same_author.csv. They are the same author ID pairs in the query dataset and the remaining 1,000 same author ID pairs which are not included in any of the training, validation, and test datasets. The format of query_answer.csv must be the same that of the training and validation datasets, and the format of same_author.csv must be the same as that of query_answer.csv.

Using these files, we will evaluate the performance of your approach. The performance will be evaluated using accuracy and precision. Accuracy will be measured on query_answer.csv, and precision will be measured on same_author.csv. The detailed formulas are given below,

$$Accuracy_{query\_answer} := \frac{\text{\# correctly predicted author ID pairs in the query dataset}}{1000},$$

and

$$Precision_{same\_author} := \frac{\text{\# correctly predicted author ID pairs in the remaining 1,000 same author ID pairs}}{1000}.$$

Note that we may run the submitted code on another query dataset 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 design and/or contact Hyeonsoo Jo (hsjo at kaist.ac.kr) and Taehyung Kwon (taehyung.kwon at 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 can use any programming language and any external libraries.

2 Presentation Video

The video should not be longer than 5 minutes; we recommend using PowerPoint to create a video. It should describe your approach with some intuition behind it, and it should discuss the accuracy of your approach (in terms of the accuracy) on the validation set of each dataset.

3 How to submit your project

3.1 Progress Report

Submit your progress report that is written using the attached template to KLMS by Nov 5, 2021, 11:59 pm. The file should be named report-[your student ids].pdf (e.g., report-20189000_20199000_20209000.pdf). Details will be announced soon.

3.2 Presentation Video

Submit your presentation video to KLMS by Dec 5, 2021, 11:59pm. The video should be named video-[your student ids].mp4 (e.g., video-20189000_20199000_20209000.mp4).

3.3 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
   - **test_answer.tar.gz**: this file should contain the `query_answer.csv` and `same_author.csv` files.

- **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.