Meeting 2 Agenda

TODO list before Friday

1. Read all discussion slides related to project
2. Register for KDD account
   - https://www.biendata.xyz/kdd2024/#overview
   - Find project 1 (IND)
   - Click `Join this competition`
   - Click `Sign Up`
3. Download data and Run Baselines for IND
   - Download the dataset
     - `wget https://www.dropbox.com/scl/fi/o8du146aafl3vrb87tm45/IND-WhoIsWho.zip?rlke`
     - `unzip IND-WhoIsWho.zip`
   - Clone the baseline repository
     - `git clone https://github.com/THUDM/whoiswho-top-solutions.git`
     - `cd whoiswho-top-solutions/incorrect_assignment_detection`
   - Install required packages
     - `pip install -r requirements.txt`
   - Preprocess data
     - `python encoding.py --path pid_to_info_all.json --save_path roberta_embeddings.pkl`
     - `python build_graph.py --author_dir train_author.json --save_dir train.pkl --pub_dir pid_to_info_all.json --embeddings_dir roberta_embeddings.pkl`
     - `python build_graph.py --author_dir ind_valid_author.json --save_dir valid.pkl --pub_dir pid_to_info_all.json --embeddings_dir roberta_embeddings.pkl`
   - Train and test the model
     - `python train.py --train_dir train.pkl --test_dir valid.pkl`
4. Make dummy submission
   - https://www.biendata.xyz/kdd2024/#overview
   - Find project 1 (IND)
   - Click `Join this competition`
   - Find `Make a submission` in the sidebar
   - Keep an eye out for “add team members” button
5. Meet again Friday to discuss Project Proposal
   - Read some literature on this project

Dummy Submission

Dummy submission due last sunday

- Email professor and TA
- Read discussion slides to catch up on project
- Steps
  - Run the baseline code provided for your chosen task
  - Prepare the dummy submission file according to the specified format
  - Submit the dummy file to the contest portal
  - Verify that the submission was successful and meets the requirements
- All members must register for an account

WhoIsWho-IND

- Background
  - Increasing online publications make name ambiguity more complex
  - Inaccurate disambiguition results lead to invalid author rankings and award cheating
  - Competition aims to develop models to discover paper assignment errors for given authors
- Task
  - Given each author's profile (name and published papers)
  - Develop a model to detect incorrect paper assignments
- Dataset
  - Paper attributes provided:
    - Title, Abstract, Authors, Keywords, Venue, Publication year
  - Participants not allowed to use disambiguation results of existing academic search systems
- Evaluation
  - Weighted Area Under ROC Curve (AUC)
- Baselines
  - Graph-based anomaly detection methods
  - LLM-based methods

Project Proposal

- Due May 13th
- One submission per team

- Use NeurIPS LaTeX style files: 2 pages max excluding references

https://www.overleaf.com/latex/templates/neurips-2023/vstgtvjwgdng

- Include:
  - Problem statement
  - Literature review
  - Tentative schedule
  - Tentative approach
  - Division of workload per member
  - References
- Run official baselines
- Survey literature for improvement ideas
- Propose ≥ 1 method to improve baselines
- Discuss with TA/Professor to formalize idea
- Use proposal as blueprint for final report

Baselines - IND

- Download the dataset
- Clone the baseline repository
- Install required packages
- Preprocess data
- Train and test the model

Data Download

Find under "Files" in BruinLearn

- Go download IND dataset from BruinLearn