GU4243/GR5243 Spring 2017 Applied Data Science

Project 4 Entity Resoultation Algorithms Evaluation

In this project, working in teams, you will evaluate and compare a pair of algorithms for **Entity Resolution**.

Challenge

Entity Resolution refers to the process of identifying multiple references to the same object and distinguishing them from mentions of different objects. Entity resolution operates on natural language text; A special case of entity resoulation, *author name disambiguation*, operates primarily on metadata about authors and articles.

For this project, each team is assigned a pair of research papers from the *Entity Resolution* literature. You will study the papers carefully and implement the algorithms in R.

For submission, you will submit the GitHub repo of your codes, a *testing* report (must be a **reproducible** R notebook) on the algorithms in terms of a side-by-side comparison of their performance. For presentation, each team should briefly explain what each algorithm does, how the evaluation was carried out, and what are the main results.

All developments need to be carried out in group shared private repo on [https://www.github.com/TZstatsADS/] with clear project management log, taking advantage of GitHub issues.

Each week, we will give a tutorial in class and having live discussion and brainstorm sessions. The instruction team will join team discussions during class and online.

- week 1 [3/31]: Introduction and project description.
- week 2 [4/7]: An overview of entity resolution and Q&A.

Evaluation criteria

- Readabiity and reproducibility of codes
- Validity of evaluation
- Presentation (report, github and in-class presentation)

(More details will be posted as grading rubrics in courseoworks/canvas)

Suggested team workflow

1. [wk1] Week 1 is the **reading and coding** week. Read the papers assigned to you, understand the algorithms and start coding up the algorithms; Also load the data into `R` and understand its structure.
2. [wk1] Each team is strongly recommended to demonstrate project progress by posting a project plan with task assignments as issues on GitHub by April 3rd.
3. [wk1] As a team, brainstorm about your evaluation plan.
4. [wk2] Based on outcomes from week 1's reading and brainstorm sessions, continue coding and start evaluation.
5. [wk2] Week 2 is the **evaluation** week.

6. [wk2] It is ok to separate into two sub-teams, one working on one algorithm, as long as the two teams have the same criteria for evaluating the algorithms. The two sub-teams can also serve as others' validators.
7. By using R Notebook to carry out coding and evaluation, your final report can just be adding explanation and comments to your Notebook.

Working together

- Setup a GitHub project folder from joining the GitHub classroom link with everyone listed as contributors. Everyone clones the project locally via your GitHub desktop and create a local branch.
- The team can work in subgroups of 2-3, which might meet more frequently than the entire team. However, everyone should check in regularly on group discussion online and changes in the GitHub folder.
- Learn to work together is an important learning goal of this course.

Resources

Review papers
Algorithms