Project Proposal for CS410

Topic: Topic Mining
Title: Team Projects Topic Mining and Browsing
Team: Topic Thunder

1. **What are the names and NetIDs of all your team members? Who is the captain?**

- Creon Creonopoulos creonc2@illinois.edu
- Kurt Tuohy ktuohy@illinois.edu
- Suhas Ashok Bhat suhasb2@illinois.edu

Team Captain will be Creon Creonopoulos

2. **What is your free topic? Please give a detailed description. What is the task? Why is it important or interesting? What is your planned approach? What tools, systems or datasets are involved?**

Project Topic: Past CS410 Team Projects Topic Mining

Description: The objective of our project will be to run one or more topic mining algorithms (**Probabilistic Latent Semantic Analysis, Latent Dirichlet Allocation etc.**) through all the previous Team Project submissions and split them into thematic categories.

We will then create and host a UI Dashboard where students will be able to filter projects by topic instead of having to go through each project separately to figure out what its focus is. This will facilitate discovery of past projects on particular themes.

Dataset: We will use a cumulative collection of all previously submitted projects found on GitHub (https://github.com/CS410Assignments/CourseProject/network/members) and we will run our algorithms through each project's Final Reports, Progress Reports and Project Proposals.

Tools:

- Python libraries to scrape through project reports: GitPython (https://github.com/gitpython-developers/GitPython) and PyPDF2 (https://www.analyticsvidhya.com/blog/2021/09/pypdf2-library-for-working-with-pdf-files-in-python/)
- Python libraries to prepare and mine text: NLTK (https://www.nltk.org/) and GenSim

(https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/)
- ReactJS to build the frontend user facing dashboard (https://reactjs.org/)
- A database system (Firebase or SQLite) to store the results and source the dashboard.
- GitHub Projects for task tracking (https://github.com/kreonjr/CourseProject/projects/1)

3. What is the expected outcome? How are you going to evaluate your work?

1. We expect a user-friendly webpage that will contain a list of links to all the projects on GitHub, which will also contain a filter selector that will list all the mined topics and will be used to only show project links to the selected topic.
2. Our work can be evaluated by making sure that NO project link remains without a topic assigned to it, even for those that do not have the ability to be mined (video reports).
3. Observation-based, e.g., observing the top 'N' words in a topic.
4. Interpretation-based, e.g., 'word intrusion' and 'topic intrusion' to identify the words or topics that "don't belong" in a topic or document.

4. Which programming language do you plan to use?

- Python for Topic Mining
- JavaScript for UI construction

5. Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.

Main tasks:

- Scrape through GitHub links to find all project reports 8-10 hours
- Clean and analyze the data and create a corpus 20-25 hours
- Code the algorithms to mine topics and assign each doc to a topic 28-30 hours
- Evaluate Topic creation and document assignment 10-15 hours
- Code the database storing mechanism for the categorized data 8-10 hours
- Code the frontend to fetch and display the data either as a whole or as filtered 8-10 hours