Project Progress Report

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

- Creon Creonopoulos (team captain): creonc2@illinois.edu
- Suhas Ashok Bhat: suhasb2@illinois.edu
- Kurt Tuohy: ktuohy@illinois.edu

1) Progress made thus far

Our initial focus was surrounded on building an infrastructure and gather all the data we need to work on before we provide it to our algorithms for mining. More specifically:

- Using GitHub as our data source, we have been able to write the necessary scripts that gather all the documents containing the text we need to work on Topic Mining and Categorization.
- We have implemented a basic dashboard website, hosted it on firebase to make it publicly available, connected it to a backend (also firebase), that gets populated with the final output json that our algorithms will produce.
- We have decided the final structure of our data, in order to present the users the course project filtering. This will help with the frontend display and allow users to filter projects per topic in the dashboard
- We have completed some data cleaning, removing non-text bases content and we are in the process of deciding how much more cleaning we will do to our data (punctuation, lemmatization, stemming etc.)

2) Remaining tasks

We have completed the infrastructure of our project and we are now mostly focused on the algorithmic parts, trying to get a final output for our frontend to consume. In detail:

- Implementation of the algorithms to mine topics from the text and perform some clustering to allow for further filtering on the frontend

- Complete the frontend design by allowing users to filter on topics, allow for course project links to be interactable and potentially implement a very lightweight search method to search for the categories in our data
- Evaluate topic modeling and categorizing of our algorithms against a sample of empirically generated topics from a sample of projects

3) Any challenges/issues being faced

Although our focus initially was going to only be topic mining from the project submissions, we have decided to also consider some topic clustering and categorization. This will help with our final output data and its display on the frontend dashboard. Furthermore:

- At first, many of the python scraping libraries were not able to pull all the text data on GitHub correctly so there were a couple of different libraries that had to be tried out but we overcame that issue.
- We have also gone back and forth with our data cleaning implementation, especially surrounding what text symbols we should keep or not (like email symbols or symbols in URLs). We are still in discussion about this item.