CS 410 Project Proposal

1. The team, Zetabyte, has one member/captain Mohammad Arham Ahmed (mbahmed2).

2. The project topic is a recommendation system for open source (OS) projects based on OS projects a user has liked or contributed to.

   The task is mining the user's preferences and interests of OS projects then recommending similar projects. The planned approach is to apply content-based filtering from the lectures as the recommendation algorithm to find projects similar to one a user likes.

   This is an interesting problem to solve since it can be difficult to sift through the vast number of projects available and handpick the ones that look interesting. With this recommendation system there is a significant reduction in the amount of time and effort needed for such a task, giving the user more time to understand or contribute to the recommended projects.

   The system is planned to have some frontend component (command line interface, or website) and a backend engine that is exposed with a simple API. The design will be iteratively improved and it is expected to use several Python libraries for the implementation such as pandas, scikit, and numpy.

   The dataset to be used is from Kaggle where it contains information on the top-980 starred repositories on GitHub.

   The outcome is expected to be application of data mining theory to a real-world use-case where the end result are recommended projects the user finds useful or beneficial.

   Evaluation of this recommender system will be done empirically by asking peers and volunteers to anonymously use the system then provide explicit feedback on its utility.

3. Python will be the programming language of choice for this project.

4. There are several tasks involved to implement this system:
   a. Data preparation: cleaning, scraping for additional data, and enrichment 5 hours
   b. Recommendation engine: 10 hours
   c. Frontend interface: 2 hours
   d. Analysis and testing: 3 hours
   e. Documentation and report: 2 hours