Individual Project Proposal

Part 1. General Information

Name: Gowtham Goli

Project Title: Reddit data analysis

Project Description: The dataset consists of Reddit's entire publicly available comment dataset. It can be used for a wide range of experiments and analyses because it's a huge collection of timestamped events with interesting features such as comment, score, author, subreddit, position in comment tree and other fields. A few things that could be done are 1. Identify and track topics associated with every subreddit and username 2. Predict posts/subreddits a user will next engage in 3. Model flow of conversations (e.g. rate of replies compared to controversiality of comment/post) 4. Get the most controversial subreddits and what had redditors laughing, bickering 5. Create an advanced comments search tool

Describe who will benefit from your analytic: Reddit users and internet community in general

Data Sources - Use the table below to list and describe data sources.

Part 2. General Data Source Information

Data Sources	**Data Source Description**	**Data Size**
- Reddit comments	*The dataset consists of* Reddit's entire publicly available comments. It consists of 1.7 billion JSON objects complete with the comment, score, author, subreddit, position in comment tree and other fields which are all gathered using Redit's API and are available for public free download. There is also a small portion of the comments available on Kaggle which consists of only comments of May 2015	Full dataset is over 1TB Dataset of May 2015 is 8 GB
Data Source 1	https://mega.nz/#!ysBWXRqK!yPXLr25PgJi184pbJU3GtnqUY4wG7YvuPpxJjEmnb9A	
Data Source 2	https://www.kaggle.com/reddit/reddit-comments-may-2015	

Part 3. Detailed Data Source Information		
Data Sources - Reddit comments	**Data Characteristics** The data is static and will be loaded once. It consists of Reddit's entire comments starting from the year 2008 until June 2015	**Data Frequency** Data is static (one time loading)
Data Source 1	https://mega.nz/#!ysBWXRqK!yPXLr25PgJi184pbJU3GtnqUY4wG7YvuPpxJjEmnb9A	
Data Source 2	https://www.kaggle.com/reddit/reddit-comments-may-2015	