Project proposals consist of the names of the team members, project topic, a brief summary of the project, and a plan for completing the project including specifications of tools/libraries that will be used. The proposals must be submitted in writing and presented to the class on Feb 7, 2019. Each team will get 5 minutes for proposal presentation.

**Team Members:**

Bill Fisher, Ron Richardson, Christina Eusanio

**Project Topic:**

Predict which Reddit posts (subreddits) will receive the highest score based on features extracted from the Reddit API.

**Project Plan:**

Predicting page popularity, no matter what the platform, has many practical uses. Namely, advertisers and marketers can leverage this information to target certain posts, pages, etc. Reddit is the fifth most popular site on the internet according to the data captured by Amazon’s Alexa. Reddit is a social message board where users can create profiles, create topics, and interact with others in a variety of ways. When users create posts, they receive upvotes and downvotes. The aggregate of those votes constitutes as a post’s score. By extracting various features from the Reddit API, our goal is to create a machine learning model that will classify posts as popular or unpopular, as defined below.

Our plan is to use the framework from *Popularity Prediction of Reddit Texts*, a thesis paper by Tracy Rohlin. The author narrowed down her dataset to 12,000 posts from six subreddits. We are going to focus on the following subreddits: r/aww, r/fitness, r/conspiracy, r/knitting, r/askmen, and r/askwomen. Instead of splitting the entire dataset into train, validation, and test sets, the author deliberately broke up each subreddit into train, validation, and test sets that were stratified to include the same ratio of popular and unpopular posts within each. This way each subreddits’ posts would be evaluated by the model equally. The author also purposely uses posts that are roughly six months old because at that point voting has ended. Finally, the author labeled the data as popular or unpopular. Posts were deemed popular if they fell above the 75th percentile of voting scores from the collected dataset.

The author preprocessed the data by removing stop words, low frequency words (those that only show up in <= two posts), and, punctuation. The author also chose not to use stemming or lemmatization. The posts were then fed to various preprocessing algorithms before being fed into predictor algorithms. Finally, the author uses F1 scores and accuracy to evaluate the algorithms.

**Data we want to scrape for possible features:**

* Post title
* Post content
* User that made the post
* User data
* Sub-Reddit
* Sub-Reddit data
* Create date/time

**Possible Libraries and Packages:**

* Reddit API
* Scikit Learn
* PRAW (Python Reddit API Wrapper)

<https://www.alexa.com/topsites/countries/US>