School of Computing

Year 4 Project Proposal Form

SECTION A

Project Title Detecting brigading on social media with machine learning

Student Name Adam Pegman

Student ID 14351741 Stream CASE4

Project Supervisor Name Ray Walshe

[Note: It is the student's responsibility to ensure that the Supervisor accepts your project and this is only recognised once the Supervisor assigns herself/himself via the project dashboard. Project proposals without an assigned Supervisor will not be accepted for presentation to the Approval Panel.]

SECTION B

Area Covered

The aim of the project is to create a bot capable of reporting users automatically on Reddit, when they post a comment that belongs on a different section of the site.

Outline

Reddit is a social media and news aggregator site with over 250 million users. Users on reddit can create, moderate and subscribe to 'subreddits,' which are forms on which links and text can be posted, each with a thread of comments. Each subreddit has it's own topic, culture and rules.

The moderators of each subreddit remove dozens or hundreds of comments from their subreddits every day, and must keep track of frequent offenders so they can be banned from the subreddit. This task is made more difficult by the fact that creating a new account takes less than a minute, so moderator bans can be easily avoided.

Simple bots have been designed to help with this task via reddit's api. Some of these bots use wordlists to remove comments automatically, or watch rival subreddits for new users in order to ban them automatically from the subreddit. Some bots also automatically respond to comments with unit conversions or other messages.

My project idea is to create a Reddit moderator bot that can be configured via a web interface, that could either respond to a comment with a fixed message, remove the comment or ban the user, based on a manual flag by the moderator, a wordlist or a bayesian spam filter trained on comments submitted by the moderator.

A more advanced feature I may implement is to determine which of a set of subreddits a comment most likely came from via a bayesian or k-nearest classifier. This should allow the problem of excluding users from rival subreddits to be tackled more efficiently, as even if a user has multiple accounts they can still be detected automatically if they post a comment typical of another subreddit.

Programming Languages

Java

Tech Stack

I plan on using Java to write all parts of the project. I may use libraries such as spring for web serving, vavr, Lombok or Guice to avoid code clutter, and sqlite as a database if I need one.
I plan to write my own Reddit api client with just the parts I need, but I may use JRAW or another existing library if this proves problematic.
I plan to use Gradle as a build tool, and JUnit to test my Java.
I might use polymer or web components to create any components I need for the web interface.

Learning Challenges

I'll have to research and implement Baysian classifiers, k-nearest and possibly several other classifier algorithms to find the best approach. Challenges will include finding algorithms that work on text input that will not over fit and cause false positives.

Platform
Jvm, Web