Reddit Analytics Engine (RAE)

Software Process

Development Team:

Heather Bradfield, Jesus Bamford, Matt Huber,

Phillip Porter, Preston Peterson, Zoltan Batoczki

Development Model

The overall development structure is incremental and plan-driven. We chose incremental because each phase specified builds upon the phases before it and represents a distinct step in the design process. However, Phase 3 and 4 will require an Agile development process because we will not know exactly what our engine is capable of and because of time constraints, we will not be able to fully implement all the features we hoped for. Therefore, the features implemented in phase 3 will have to be built in short-term sprints requiring constant testing and re-evaluation.

There are 5 weeks left before the deadline and to assess the progress of our group we expect to spend about a week on each of the following 4 phases with the last remaining week to finalize everything and prepare documentation and reports.

Phase 1

* Create barebones bot written in python (Reddit Username: ci\_rae)
* Allow the bot to respond to basic user requests. Example: The command **/u/ci\_rae test** will reply with a comment “Hello!”. This will help us verify that the bot is capable of reading commands and writing responses on Reddit.
* Allow the bot to retrieve data about a Reddit user in JSON format, using the Reddit API.
* Implement a module that will convert Reddit user data from JSON format into a format that we can easily manipulate with python.

Phase 2

* Build foundation of remaining software infrastructures (namely the database and user CLI/GUI).
* Implement a barebones analytics engine.
* Implement our first bot command: **/u/ci\_rae activity /u/example\_user**
* Set up a server with a database using MySQL.
* Expand our Reddit bot to be able to access the Imgur API, so the bot will be able to upload statistical photos.

Phase 3

* Implement data extrapolation to our analytics engine using information stored on the database.
* Add more commands, with time permitting (this could also become part of the Evolution phase)
  + Popularity
  + Suggested Subreddits
  + WordCloud
* Implement user cross-referencing if time allows.

Phase 4

* Unite modules together to test bot capabilities.
* Repeat phase 3 if we have time to implement more commands or more powerful analytics algorithms.
* Testing and release of the final product.

Requirements:

User:

1. User can make data generation requests directly from a gui/cli in addition to from a Reddit comment.
2. Engine will have at least 3 different types of data requests a user can make.
3. Engine only processes data that is publicly available.

System:

1. System only processes one request at a time, to avoid failure.
2. System must not support requests taking up too much time, so that it is able to respond to new requests as quickly as possible.
3. System must be executed on a server, so that it is always running and available for use.

End User specification:

Making a data request from Reddit:

Login to a Reddit account and submit a comment to any thread in the [forthcoming] list of supported Subreddits. A data-request comment to the Reddit Analytics Engine looks like this: “ci\_rae [USER\_NAME] REQUESTS…”

Ex 1. “ci\_rae wordcloud“ → generates a wordcloud in image form for the user account that owns the comment (since no discrete user was specified).

Ex 2. “ci\_rae /u/bilbobaggins sub\_suggest popularity” → generates a list of subreddits that may be of interest to the user account bilbobaggins as well as a list of the subreddits that the user account bilbobaggins is most popular in (in terms of user’s overall upvote ratios).

Making a data request directly from the engine interface follows the same syntax except that it does not require the user to qualify commands with the prefix “ci\_rae”.

Modules of the software:

* Data Processing: Takes in data pulled from reddit, cleans it up and prepares it for the rest of the modules.
* Analytics Engine: Performs analytics based on the command given, pulls required data from the database, and generates reports of the results.
* User Interface: A user interface/client to interact to allow more robust commands to be performed without having to directly interact with the bot on reddit. Calls the analytics engine but displays the resulting reports within the UI.
* Database: Stores all cleaned up data and results from previous queries to keep track of the data collected and to allow for quick generation of reports.
* Bot: Autonomous software that runs continuously and scans reddit comments until a call to the bot is found. Detects which user called the bot and and which command was given, replies to the user once the report is finished.

Data Structures involved:

Data pulled from reddit will be in Javascript Object Notation (JSON) format, which will be parsed into the following classes:

* A comment class will store the comment body, the date, the unique identifier for the comment, the user who made the comment, the comment score (upvotes subtracted by downvotes), what subreddit the comment was in, and what comment this comment was a reply to (if applicable).
* A post class will contain the title of the post, the date, the URL that the post links to, the post description (if applicable), the unique identifier for the post, the user who made the post, the post score (upvotes subtracted by downvotes), which subreddit the post was in, and a linked list of all the unique identifiers of the comments in the post.
* A subreddit class will store the name of the subreddit, how many subscribers, how many users are currently present, the date of subreddit creation, the description of the subreddit (called the sidebar), and a linked list of the unique identifiers of the 10 latest posts in that particular subreddit.
* A user class will have the username, the age of the user account, the amount of post upvotes the user has received, the amount of comment upvotes the user has received, the amount of gold the user has been gifted by other users for superior content, what trophies the user has received, a linked list of all the unique identifiers of all the users comments and posts.
* A MySQL database will store all data that needs to be kept track of. This includes but is not limited to: comments, user accounts, and subreddits that have already been processed as well as any reports that were generated with these objects.The MySQL database is a relational database management system and uses a B+ tree structure to store its data.

The Following is an entity-relationship model of the database, showing the user table, the comment table, the post table, and the subreddit table and the relationships between them.

