Reddit Analytics Software Reuse

Mattias Huber, Preston Peterson, Phillip Porter

Heather Bradfield, Zoltan Batoczki, Jesus Bamford

Over the course of this project we have utilized many different software tools in order to speed up development. We started originally with a certain set of tools, but as the project evolved we found much better resources. The cost of using these tools was having to conduct extensive research on how to use them, which resulted in us ditching some tools for others. We had to search for tools such as wordcloud, which saved us time in not having to implement this functionality from scratch. We also had to learn how to interact with the PRAW and ImgurPython API by reading the documentation.

We utilized the abstract design of other reddit bots, such as having a listener and a request handler. Other reddit bots were utilizing the ‘mentioned’ technique as well which drastically sped up response time of the bot. This technique required the user to directly mention the bot in their comment by typing /u/ci\_rae, which causes reddit to send a message to the bot’s inbox. This allows the bot to not need to scan all of reddit looking for commands. Instead it just responds to the message it receives in its inbox and then marks the message as read once the job has been completed.

Objects from various libraries were directly reused. These include: Numpy arrays for easy manipulation of our data, Matplotlib graph axes to generate report graphs, WordCloud’s generate to generate word clouds, ImgurClient from the ImgurPython API to log into and submit images to imgur, PRAW’s get\_subreddit, get\_mentions, and get\_redditor objects. These objects were used to get all of our data about subreddits, mentions, and redditors.

We also reused many software components. The Python Reddit API Wrapper (PRAW) was used to retrieve data and post data to reddit. ImgurPython was used to upload our generated reports to imgur.com, an image hosting site. Matplotlib was used to generate the graphs of our reports. Numpy was used in our analytics modules to perform calculations. WordCloud was used to generate the word cloud of a user's comment history. Tornado was used as dependency for the Oauth2 authentication. The python threading package was used to utilize multithreading. Jupyter Notebook was used to build the analytics engine dashboard which will be used in our presentation.

The largest system that we reused, and the one that is core to our entire engine, is the Reddit website itself. The Reddit servers perform the role of our database, as they store all the data upon which our analytics are performed. The Reddit website performs as the graphical user interface which lets users interact with our analytics engine. Our engine also relies on the Prawoauth2 system, which handles authentication with Reddit and maintains a stable connection to Reddit’s servers. As for the images generated by our analytics, the Imgur website/servers are used for public hosting of our generated reports. Through the use of these systems, our engine has very little overhead. It can run on a small server, with very low storage and memory requirements, and yet access and process vast amounts of data.

Overall, the things we would have done differently would be conducting more research at the beginning so that we wouldn’t have wasted time trying to use tools that ultimately were not needed. An example of this is when we planned a whole part of the bot to scan reddit. This ended up not being required because of the versatility provided by PRAW. Another example is how we planned on utilizing a database. This proved to not be needed as we were fetching new data from reddit and our database would continually be outdated. We learned a lot about the advantages of software reuse on all levels - abstraction, object, component, and system - and together they allowed us to minimize the ultimate cost and time needed to implement and evolve our Reddit Analytics Engine.