Client Project Retrospective

Our primary objective for this project was to build a web-scraping program that can download separate stories from the health care review website (https://www.careopinion.org.au) and save various relevant sections of each page into separate files. As mentioned before, Care Opinion Australia is an independent site where individuals can share their experiences regarding the Australian health-care system. By using web-scraping our client will be able to study how Australians feel and think about their health-care experiences and perform further analysis on this data for research purposes.

As mentioned in our previous sprint documentation, our team has already demonstrated to the client the functionality of our web-scraping program to save different relevant sections of the pages from the Care Opinion website, specifically, story id and story, username, title, about (location), date, similar, progress, responses, updates, good tag, improved tag, feel tag, etc. A major additional component to this project was the database that was requested by the client; we have since created and refined a database that can store the downloaded information in a manner that it can be easily accessible using unique IDs. Our database has been specifically designed so that it can be filtered using tags that have been scraped or keywords. Our project has also incorporated a graphical user interface, created using the tkinter python package, to use in relation with our database, the features of which have been explained and demonstrated to the client who has responded positively. User guides, manuals and other support documents have all been created for use by the client to help aid with the installation and running of the web-scraping program and database.

For the duration of this sprint our team directed its focus towards the testing of our code and perfecting the format of the scraped output to cater specifically to our client's needs. Scraped file samples were continuously sent to the client to gather feedback and adjust the formatting of the output files. This subsequently led to specific detailed alterations to our code, some changes included removing certain whitespaces before words, removing the 'z' character from the date file and removing default tag words from the title of each scraped file. After finalising the format with the client, we were given the all clear to begin the final big scrape of all 10,000 files from the Care Opinion website, this has since been completed. Both the web-scraping code and the database itself have been through vigorous manual testing; numerous unit tests have also been written for varying functions of these components to ensure that the correct data is being scraped in each file, the data is correctly formatted and that the data being scraped and data being input into the database align with one another. Our group has gone through the individual tests and testing methods used for the web-scraper and database personally with the client to ensure and help build confidence in the product that we have created for him.

As with our previous sprints, our team has maintained a positive working relationship with the client through regular organised meetings and providing continuous status updates of our progress throughout the project. Our group has managed to incorporate all the clients requirements and additional features asked for throughout the duration of our CITS3200 project. Our client has repetitively conveyed his contentment regarding the overall progress of the project and has stated he would be more than happy to include the names of our group members in any research papers that may benefit from the data we have collected.