WEB-SCRAPER – PROJECT OVERVIEW

1.0 OVERVIEW AND SCOPE

The project overview gives the user an understanding of the project as a whole. Specifics are not explained here.

2.0 PROJECT OVERVIEW

2.1 Tableau

Tableau is used as the final deliverable's front-end display. From here, the user can compare Seattle Reddit posts to what the Seattle weather was at the time the post was made. The user can access the Tableau dashboard in one of two ways:

1. The user has or will create a Tableau Online account. Click this link and sign in: https://us-west-2b.online.tableau.com/#/site/webscraping/views/Web-Scraper/rSeattlevs_Weather
2. The user doesn't need an account and can access it through Tableau public here: https://public.tableau.com/profile/john.fitzgerald7009#!/vizhome/Web-Scraper/rSeattlevs_Weather

Instead of a word map (which would have required an expansive database table holding individual words, word counts, and numerous foreign keys to the original post), a simpler visual was created using the relationship of the Reddit Post's title and the corelating weather (temperature and phrase – e.g. cloudy, sunny, etc.) as seen in fig. 1.

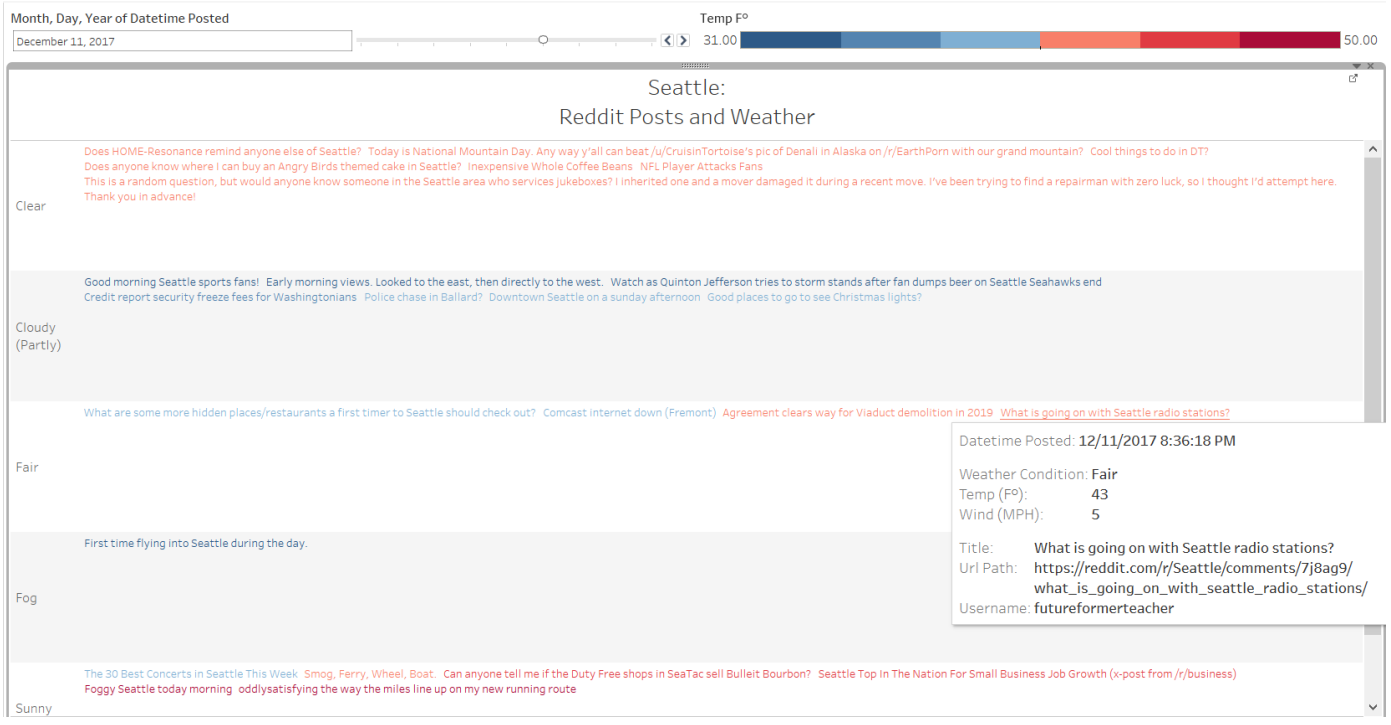

Fig. 1 Example of Tableau dashboard. Shows a single day, grouped by weather type, and ordered by temperature. Highlighting a title with the mouse will show context data.

2.2 Python

Python, BeautifulSoup, and Pyodbc are used as the back end to scrape (get HTML, parse for relevant lines, and use regular expressions to extract only the relevant data). Robust sanity checks aren't in place to detect invalid strings (e.g. currently can't insert emojis). When an invalid database insert is attempted using Pyodbc, and email goes out to the user indicating the title and time of the error. This requires a manual restart of the program. Fig. 2 is a screenshot of the scrapers running on an Ubuntu virtual machine.

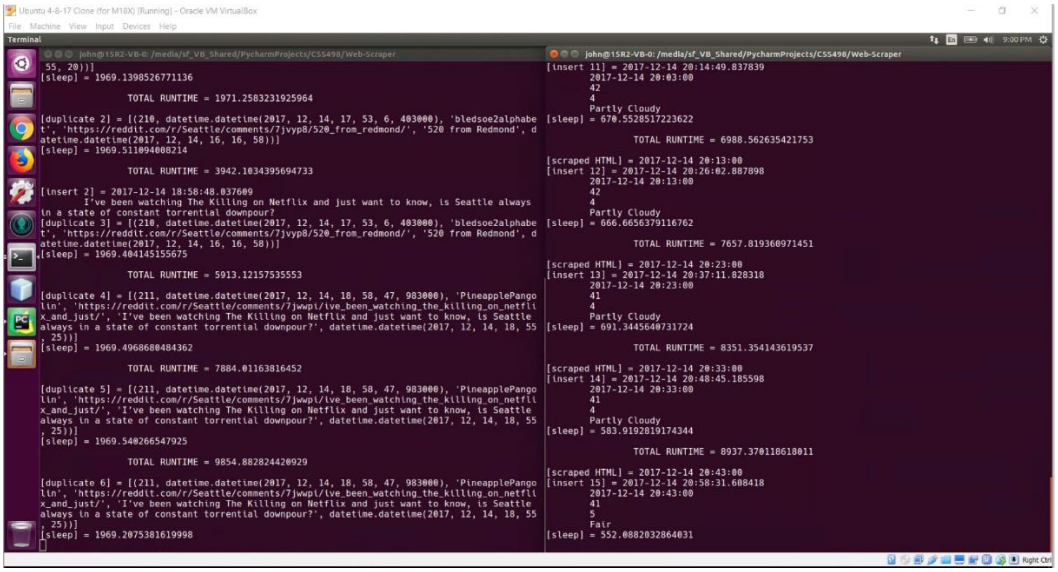

Fig. 2 Reddit (left) and weather (right) collector running on an Ubuntu VM.

2.3 Amazon Web Services Relational Database (AWS RDS)

A free AWS RDS instance was created to store the scraped data. There is 19.9 GB of free space which, under the current load of inserting a single weather update every ten minutes and less than 50 Seattle Reddit posts a day, it won't fill up for many years. Fig. 3 shows details regarding the instance on AWS' dashboard.

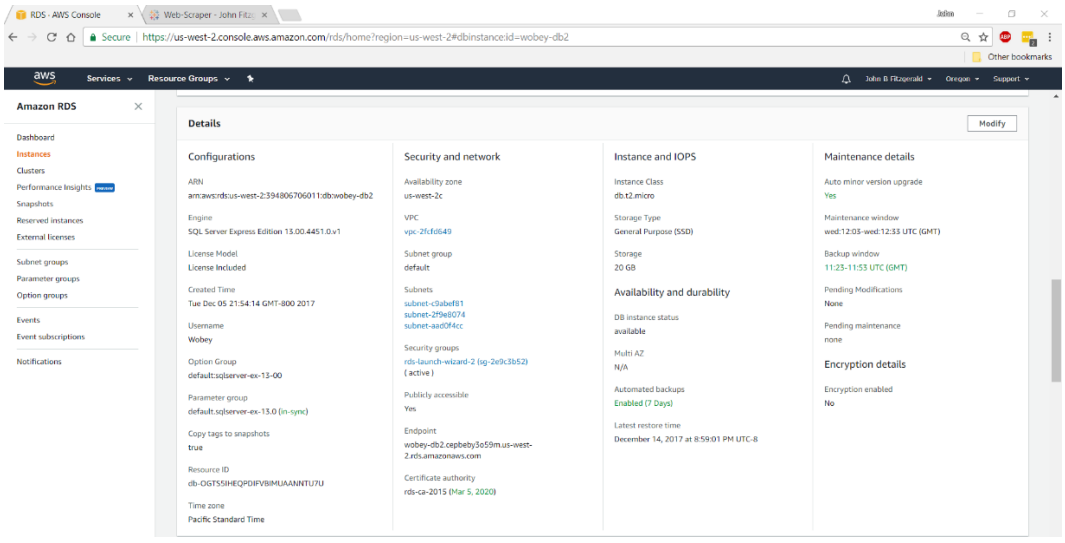

Fig. 3 AWS RDS instance that holds the database as seen from the AWS dashboard.