4/2/2021

Discussed:

Created and estimated Jira epics and stories

4/5/2021

Discussed:

Colin/Joe: Web Scraper

* Uses Python import requests to connect to web addresses
* Web addresses will be hardcoded for website layout consideration individually
* CSV saved in folders by website, and document title will be date/time

Nikki/Logan: Sentiment Analysis

* Categories
  + 1 -> positive
  + 2 -> neutral
  + 3 -> negative
  + 4 -> mixed

Brice: Reports

* Use line charts for historical patterns using Python’s Matplotlib package
* Use pi charts for one particular website to break down categories using Matplotlib
* Aggregate analysis of data

4/7/2021

Discussed:

Colin/Joe: Web Scraper

* Reddit is flagging our web scraper as a bot

Nikki/Logan: Sentiment Analysis

* Began researching the library NLTK and Scikit-Learn

Brice: Reports

* Discovered bar charts in Matplotlib and began testing results with faux data

Current Roadblocks:

* Scraping Reddit

4/9/2021

Discussed:

Joe/Colin: Web Scraper

* Reddit is still flagging our scraper as a bot

Nikki/Logan: Sentiment Analysis

* Decided on NLTK
  + We found a few tutorials that used NLTK and vader\_lexicon
  + Vader\_lexicon was created to analyze the English language and its jargon
  + We also found a data model called punkt which would separate the entire post into word lists

Brice: Reports

* Found the best way to get accurate results

Current Roadblocks:

* Reddit’s bot flag is still an issue, so for the next meeting they are going to try scraping another website

4/12/2021

Discussed:

Joe/Colin: Web Scraper

* They moved on from Reddit and found results with Facebook
* They also reviewed our list of websites and decided to build a crawler based on Reddit and Glassdoor because the other sites only needed scraped

Nikki/Logan: Sentiment Analysis

* Began experimenting with NLTK using copy/pasted posts from Reddit

Brice: Reports

* Began research in importing data to Python
  + Considered CSV files, plain text files, and pdf files

Current Roadblocks:

* While on the back burner, Reddit is still an issue
* Facebook is an issue as well because of dynamic loading/scrolling

4/16/2021

Discussed:

Joe/Colin: Web Scraper

* Had success with Facebook and began finding the tags needed in the other websites
* Colin went back to working through the Reddit issue

Nikki/Logan: Sentiment Analysis

* They have consistent success in the results from NLTK using copy/pasted posts
* Found that there are “trigger words” that will sway the analyzer
  + ‘Party’ is always positive, even in the post about someone having money stolen from them at a party
  + Words that don’t have sentiment will give exactly neutral results with no compound, such as last names and words that aren’t in English

Brice: Reports

* Can import csv and text files to Python from a local directory
* Can parse the files and break apart strings to find specific words

Current Roadblocks:

* Reddit
* Some websites are very difficult to find specific tags that are unique to only posts or only replies
  + This is an issue when posts are from an official Pitt-Johnstown account so only the replies should be analyzed

4/19/2021

Discussed:

Web Scraper

* After presenting our current project, Bilitski gave recommendations on how to bypass the Reddit bot issue
  + “Trick” the site into thinking we are a different computer every time
  + Use a VPN
    - Tried it, but it was inconvenient to change the locations incrementally
* Joe found a tagline we could include in our scraper to bypass the bot flag without having to use a VPN
* Even with the tagline, Reddit is only reading local posts

4/21/2021

Discussed:

Joe/Colin: Web Scraper

* With the Reddit bot issue cleared, they moved onto creating a crawler

Nikki/Logan: Sentiment Analysis

* Figured out how to take the text (decided to use text files instead of csv because they were easier to parse) and use it in the sentiment analysis engine

Current Roadblocks:

* Getting the crawler to detect a url from a scrape and open that url to find the next set of tags

4/23/2021

Discussed:

Colin/Joe: Web Scraper/Crawler

* Joe got the scraper to save the data in a text file that could be used in the sentiment analysis engine
* Colin was still getting the crawler to work correctly

Nikki/Logan: Sentiment Analysis

* Took Joe’s data and started testing small portions of the results at a time to ensure correct output
* Limited our results to positive and negative only because we determined the neutral results were not meaningful

Brice: Reports

* After the sentiment analysis has ran, he is testing that the results are accurate to certain posts that had “trigger words” from the analysis

Current Roadblocks:

* The crawler
* Not all posts will give correct output due to the trigger words, words that the analysis engine could not determine, and misspelled words or abbreviations

4/26/2021

Discussed:

* The crawler was finalized and data is able to flow between the scraper/crawler, the sentiment analysis, the results and result chart, and the output file
* There is still an issue where, when using multiple websites with multiple types of formatting, the output data is not formatted properly even though the sentiment results are correct

4/29/2021

Discussed:

* Reviewed the results given from running the entire process
* Figured out how to run the program on a Python scheduler library called scheduler
* Determined the order of our presentation and discussed what would be addressed by whom