# CSE 184: Gendered Job Posting Analysis

Fall 2019 - Individual Contributions Report

**Kyle O’Brien**

Kyle was primarily responsible for the team's data wrangling challenges. In total, we had over 25,000 jobs from four data sources. Each raw data set had differing schemas, duplicate values, missing data, and varying formats for locations. It would have complicated our analysis work to attempt to work on multiple files with differing schemas. Kyle developed a data wrangling program using Python and Apache Spark to asynchronously format these various raw datasets into a single clean file ready for analysis.

Kyle worked closely with Bryan to consume the raw data he provided and with Jerico and Nikhil on how to best handoff the clean data to them for analysis.

Included in the pipeline is a sample Dockerfile allowing developers to deploy the model with minimal configuration as well as unit tests for the pipeline's utility functions.

Lastly, Kyle did the initial research on collecting the corpus on gendered terms from academic and industry research.

**Bryan Jimenez**

Bryan was responsible for scraping 5000 to 6000 jobs from Glassdoor and Indeed websites. Each website renders its content using JavaScript, content that wouldn’t be available using only Beautiful soup and requests. The web scraping scripts made use of Selenium, a tool to automate webpages navigation that mimics a human user. Randomized pauses to throttle requests, limiting requests 1 every second, and using tor requests to change IP addresses were a few of the techniques used that were mentioned in lectures to avoid getting errors from websites.

**Nikhil Dodd**

Nikhil mainly worked on visualization and data analysis for the jobs scraped. Using this data, Nikhil took each column of the entire dataset and found all the jobs that utilized any of the masculine words/phrases in the job description and created an updated dataset to use throughout the rest of the code. Using this data, a graph was created to show the most common words/phrases in job descriptions and find the number of jobs for each state in the US. Nikhil then used NLTK (word stemming) to find the most common stems in the job descriptions, and used the data picked up from that to create a second new dataframe that gives the state, the stem/phrase, and the number of job descriptions that use that use that phrase.

Using plotly and geopandas, Nikhil passed in the previously mentioned dataframe to the geopandas dataframe to help create the interactive map that would provide all the stats per state. Using the plotly library “graph\_objects”, Nikhil was able to create an interactive choropleth heat map that gave the correlation between the number of jobs and darkness of the state color.

**Jerico Factor**

Jerico worked on analyzing the data set and finding the percentages of gendered job listings and also broke them up by state. He found information about each of the phrases in each of the state and exported it as json so that we don't have to keep running through the entire data set all the time. He created bar graphs with the percentages