Jun Xiang

Did a wiki scraper to obtain the constituencies that each Main Town in Singapore consists. (Eg. Ang Mo Kio – Nee Soon GRC and Ang Mo Kio GRC).

Did visualisation (line of best fit using seaborn) on the change in demographics of each race in a Town (X) against change in votes in a constituency (Y).

Did OLS to get the statistical summary of our results.

Realised that there was a bug in our data cleaning process. SOLVED.

Muskaan

As per our group meeting, we discussed that we would try to focus on two main areas for now:

- the demographics analysis

- the transcripts and newspapers analysis to find factors

For the newspapers, we discussed that we would analyse news from 6 months before elections began until election time for the past two election years. Then we will try to compare the election voting results to the sentiments from the newspapers and see if there is a pattern. If there exists a pattern, we will try to apply similar patterns for predicting the next election results for that particular factor.

Other possible factors that we have identified till now include income patterns (for which we hope to analyse number of private estates) and we also want to explore if the results depend on who is campaigning (considering years of political experience, number of previous wins, activity on social media and participation in parliament).

Moving on from the team meeting, for my part, I extracted sentences to get context for the most frequent words used and I also assigned a scoring weight to them. However, I realised this would not be meaningful if I had to do it for each individual article. So, I thought of adopting another approach, which would help with both the transcripts and newspaper analysis.

Firstly, I was thinking about what level of analysis I want to do. if the level is analysing individual transcripts to get individual word clouds, then I would go with simple NLP analysis which I have been working on till now. However, in this case, it would make more sense if we analyse multiple transcripts at once. For this, we can create a new dataset which has all the different sentences in different transcripts. The next step would be to perform clustering analysis on the new dataset. thereafter, each cluster would be related to a particular topic and each cluster would have a set of words related to the topic which is represented by the cluster. Therefore, we could investigate the sentences in which the words related to a particular cluster are located to identify the context (The same analysis could be applied to newspapers as well).

Sherman

Encountered a timeout error when accessing newslink. Unable to have multiple logins. Will create a scraper that retrieves the URL for articles then stores the <body> in a txt file

Newslink was a bust, worked on correlating housing resale values to votes. Turns out signpost holds all the postal code data. But allows one to one searches on addresses. Scraped singpost and retrieved postal codes and used [www.parliament.gov.sg/mps/find-my-mp](http://www.parliament.gov.sg/mps/find-my-mp) to link the various postal codes to the various addresses

Zexel

Next step would be figuring out how to combine the datasets, form our hypotheses regarding the household races and use linear regression as the first form of analysis.