Overview

Jun Xiang

Week 0/1:

Tried using Twitter API and Facebook API. However, not effective as Twitter API requires premium account whereas Facebook GraphAPI as researched only allows for requests of personal account/page.

Data cleaning of csv files. Grouping sub-towns into their main towns to get a better idea/estimates of the overall numbers.

Week 1-2:

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).

Muskaan

Week 0/1:

So, according to our thought process as a team, we wanted to figure out what factors affect the election results in Singapore and what is the extent of that effect. We broke this problem down into a few parts. One of them was to analyse the speeches given by the Prime Minister and Members of Parliament before and after elections. The main purpose of doing this was to understand what they talk about in their speeches and how they try to influence voters. So, for this, I took a sample speech transcript and used tokenizing, removing stop words, parts of speech tagging, word lemmatizing and then finally got a word cloud frequency distribution for the final list of words used during the speech.

This process helped us identify a few commonly used words, but it didn’t help us identify the context in which they were used. However, it did help us solidify our previous hypothesis with the household data. So, one of my next few steps would be to find the context in which these words were used and to scale the word cloud to multiple transcripts. Now, the problem that I encountered while looking for transcripts was that they’re not many of those available in text format. However, there’s lots of YouTube videos with speeches. So, to scale it further I would like to investigate converting YouTube speeches to text format for further analysis.

Week 1/2:

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

Zexel

I swapped the key value pairs of the dictionary to match it with our other data.

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.