1. Any surprises from your domain from these data?
   1. No major surprises other than how much info there was out there about spam classification. It seems to be a hot button topic these days. I thought it was interesting that on the whole the spam text messages were longer than the non-spam messages. It is like the spam authors decided to use every one of their 160 characters to build credibility. Yes, phones these days happily concatenate texts messages giving you up to 1600 characters, but I remember back when you had 160 and if you used more than that it was another 10 cents. This was why we looked at the length of the text messages as well.
2. The dataset is what you thought it was?
   1. Yes, it really was! I did have to strip out a lot of special characters, which was not difficult, and removing stopwords did take the bulk of the processing time. So much that I ended up having to write a function that essentially gave a completion timer for the processing step. I am actually rather proud of it! It was interesting working with the data as well – since most of the texts originated from Singapore, they were crafted a little bit differently from American text messages so it was a fun little exercise in multiculturalism which I rather enjoyed.
3. Have you had to adjust your approach or research questions?
   1. A little bit. I wanted to do my own thing with the dataset, so I took a couple of different methods. One of the really pretty visualizations I created was a word cloud. It worked really nicely with the spam text, as much like robocalls you start to see the same types of text message spam as well. Many offer free things, or tell you to act now, so it was a nice visualization that put the spam texts into perspective and showed a little bit of what stood out to the human reader that was then taught to the machine learning algorithm.
4. Is your method working?
   1. Seems to be working so far! I was able to utilize two different methods of classification that actually gave fairly accurate results! I believe in both cases I did not have any spam messages that were falsely identified as ham (or not spam) messages, which is amazing! There were a few spam recognized as ham but still, two very accurate models. I wonder if that implies that the method used to generate the spam messages is somewhat automated as well? It would make sense, because sometimes the sheer amount of output required to make a few hits would require a ton of manpower that could be saved via automation.
5. What challenges are you having?
   1. Some of the feedback I got regarding my first step was the conclusion and the next steps could have been a little bit stronger. That is something I want to work on in this project – being able to project beyond the project’s scope to identify the next steps or what would appeal to prospective employers. As an associate analyst, this is my wheelhouse – our senior actuaries are charged with some sort of outcome and they have me implement it. Part of what makes an actuaries role so difficult is that they truly are the bridge between the busines world and the analyst world. We need to be comfortable in both realms and acting as a laison between the two, and I feel like if I can get better at that it would serve me well in the future as well.