Andrew Holmes

Mini Project 1: Text Mining and Analysis

Summary:

For this first Mini Project I decided to kill two birds with one stone and focus on the research I do with Professor Zhenya Zastavker on Discourse Analysis. Discourse Analysis is the process of analyzing language in naturally occurring environments and how it impacts (or can be impacted by) environments, genders, context, and other factors. At Olin we are analyzing teams of students in Design Nature and User Oriented Collaborative Design (UOCD). We film team meetings for a number of groups and outsource it to a company to create a transcript of the meeting. From there we perform a set of qualitative analyses to identify trends within teams and between teams. We have also followed several people from Design Nature to UOCD to see if their interactions have changed over time.

My goal with this project was to begin automating the analysis process and attempting a new process of analysis. I imported the transcripts, which have been saved as excel files, to csv files and did some cleaning up and sorting of the data. Then, I used the sentiment and modality functions in the pattern module for Python to produce a value that correlates to certainty, positivity or negativity, and subjectivity or objectivity. From there, I took the average for the output of each function by adding up the value for every line and dividing by the total number of lines. This provides some rudimentary quantitative data that we can begin to use in validation of our hypotheses.

I would like to point out that I cannot provide the full transcripts to people outside of our research group in compliance with the Institutional Review Board, however I can present small sections of the transcript as an example. The names have also been pseudonymised. I have included a small portion of a transcript at the end of this write up. Regardless, my project works with any csv file where the first column is names and the second column text.

Results:

We will now take a look at some of the results from running this program on different transcripts. Below is the output when we run an entire transcript for a UOCD team:

Philip Melter:
Average Modality: [0.6397517212189672], Average Polarity: [0.07259849236411742], Average Subjectivity[0.28806948452781767]

Elizabeth Homer:
Average Modality: [0.6823198417388558], Average Polarity: [0.08349982893996981], Average Subjectivity[0.2317629373321862]

Ken Roberts:
Average Modality: [0.7668320425758024], Average Polarity: [0.132764920112782], Average Subjectivity[0.27189077293929265]

Mary Young:
Average Modality: [0.7044270833333333], Average Polarity: [-0.02812499999999999], Average Subjectivity[0.12291666666666666]

Mike Lands:
Average Modality: [0.6932240445540618], Average Polarity: [0.053054238740950936], Average Subjectivity[0.28310748478684405]

While we have not completed qualitative analysis on this team, the results overall lineup with what we have observed. For example, we have noticed that Philip Melter and Mike Lands are both very talkative and tend to compete for control of the group. Their average subjectivity is nearly equal for both of them, and also higher than that of every other team member, suggesting they are the most opinionated of the group. Mary Young is a very quiet team member, and talked over for the majority of the transcript. Her average polarity (a rating for how positive or negative a person is) and average modality (rating of how certain a person is) are the lowest of the group, and her polarity actually dips toward the negative end of the spectrum. This team is also much more vulgar and off-topic than other teams we have studied, and this shows, as their average polarity is somewhat lower than that of other teams I have run the program on.

Reflection:

The focus of the Discourse Analysis research has actually shifted this year to focus more on software, so I was incredibly excited to hear that the first project in the class would be focused around text mining. My hope is to improve my software skills over the course of the semester so that I will be more useful as we move forward in our research, and this first project was the perfect fit. There were some limitations that I came across, but much of it stemmed from inconsistencies in our transcripts. I think the data will be more accurate if the transcripts are organized with a unified style. Another design choice that I have concerns about is that I am taking the average sentiment and modality over every line that each team member speaks. The issue with this is that some lines are really short and others are extremely long, giving simple lines like “yeah” and “okay” as much precedence as larger thoughts. I could not come up with a more elegant solution. Maybe next time I can attempt to parse lines into individual sentences and compute the average in terms of sentences? This was my first time creating a software project that isn’t just a simple toy program and actually has a real world application. I’m excited to create programs that are even more complex and intricate as the semester progresses!

Philip Melter	Like having some of this linear like, right. I mean we could put some information there and it'd maybe be useful, but…
Elizabeth Homer	True. I think it's so nice and concise the way that it is now, with…
Philip Melter	Yeah.
Elizabeth Homer	The before and after.
Philip Melter	Shouldn't we try to pack more info into it?
Elizabeth Homer	Right.
Mike Lands	I think you should zoom in. Like a lot of stuff… well, or I guess the least I'm saying is like a lot of our vignettes around regionals are all really interesting, right. There's the foghorn goes off while they're running in the middle of the woods and there's the like… we had to go around the lake, can anyone hear me on the radio? They're seeing someone through the scope like…
Philip Melter	Yeah. Yeah.
Mike Lands	There's not like really… they're really detailed, but I don't know if they belong in the broad interactions map... Or if any more details needs to be put in.
Philip Melter	Okay.
Mike Lands	Like I don't know if this is what Mary had in mind, but I think it works.

Outputs:

Philip Melter:
Average Modality: [0.7807692307692308], Average Polarity: [0.2182142857142857], Average Subjectivity[0.28154761904761905]

Mike Lands:
Average Modality: [0.5580808080808081], Average Polarity: [0.11755952380952382], Average Subjectivity[0.2967261904761905]

Elizabeth Homer:
Average Modality: [0.875], Average Polarity: [0.2119047619047619], Average Subjectivity[0.42857142857142855]