Katherine Roe

Spring 2015

For my 6.UAP project, I will examine the works of Shakespeare and his contemporaries using natural language toolkits and a variety of data analysis techniques. In order to perform this analysis, I have converted all of Shakespeare’s plays into python objects. Each object corresponds to a single line from a play, and includes the text of the line, the speaker, the play it is from, which line number it is, and which act and scene it appears in. I plan also to convert select plays by Christopher Marlowe, Robert Greene, and Ben Jonson into this format.

A fair amount of work has already been done with natural language processing in order to identify the author of Shakespeare’s plays, and to identify potential collaborators. Therefore, I will not spend time trying to identify the authorship of the plays (which is, in my mind, not open to reasonable dispute in any case). However, very little quantitative analysis has been done on the text itself, and computer science could, I believe, shed new light on these plays. To achieve this end, I have come up with a series of questions that have not been answered through quantitative analysis to date, and which data analysis could answer. A sampling of questions I’d aim to answer are presented below:

* Are there any linguistic differences among the comedies, tragedies, and histories? For instance, are the comedies likely to exhibit different sentence structure or make use of a different vocabulary?
* Is it possible to determine, from the language, whether a speech is delivered by a male or female character?
* How did the plays of Shakespeare’s contemporaries change in language or structure after Shakespeare became popular? Is it possible to pinpoint how Shakespeare changed Elizabethan theatre conventions?

I will perform these analyses in Python using natural language packages such as NLTK (the Python Natural Language Toolkit) and Textblob. I will likely augment the methods available through these packages with my own that I will write as the need for them becomes more apparent. These toolkits will allow me to analyze the linguistic characteristics of the plays by determining the parts of speech of each word, by determining the sentence structures of monologues, and by assigning each sentence a score related to how “polarized” (ie. Positively or negatively emotional) it is.

In order to verify that any trends I find are not due to overfitting the data, I will select 12 acts from each category (history, tragedy, and comedy) to use as controls. I chose to use acts instead of entire plays as my control set because there are only 37 plays, and if a single play in the control group were particularly anomalous for some reason, it might skew the results considerably. Having analyzed the other acts, I will test my hypotheses against these 12 control acts. For instance, if I had determined that women did speak differently from men in a measurable way in Shakespeare’s plays, I could check whether my model correctly determined whether a speaker of a control monologue was a male or a female. If the model correctly predicted the gender of a monologue that it had never seen before, that would strongly suggest that the model was revealing something inherent in the data rather than being accurate due to overfitting.

Overfitting will still be a danger due to the limited amount of data available. This is especially true of the plays of Shakespeare’s contemporaries, who tended to be less prolific and less well-documented than Shakespeare himself. Although I think that this risk can be avoided with careful selection of control groups and categorization criteria, I will be sure to monitor whether overfitting is becoming an issue as the project progresses.