I was recently inspired by [a paper](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2735096/) from computational biology, where given a data set of European individuals and their genomes, the researchers could use a technique called Principle Components Analysis (PCA) to reconstruct a map of Europe.

My question was whether we could reconstruct similar “maps,” but of literary texts instead of human individuals.

**What is PCA?**

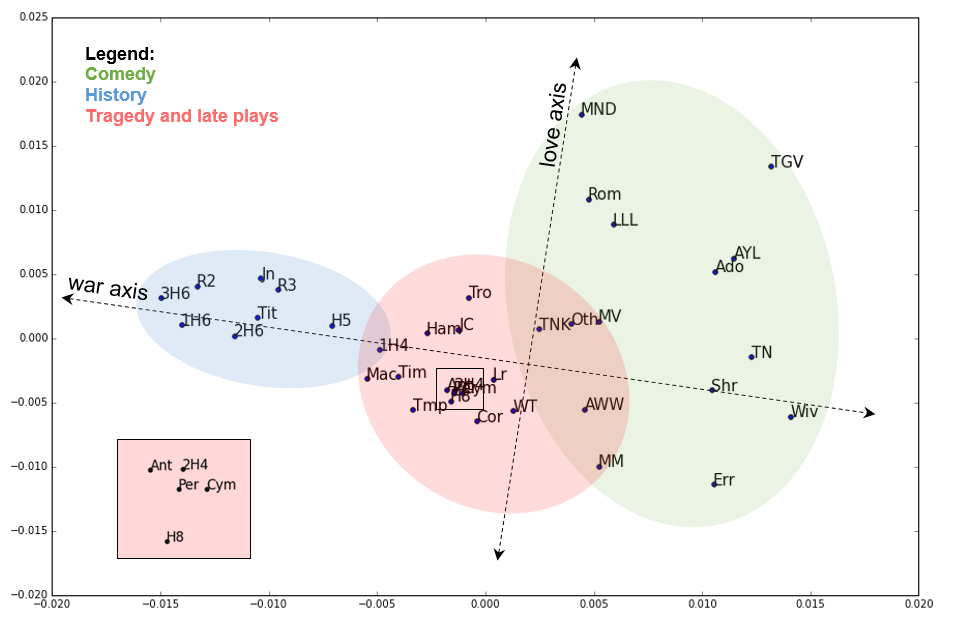
First, a little bit about PCA. I like to think of PCA as a way of understanding a group with a diverse set of members (whether those “members” are actual human individuals, or image files, or Shakespeare plays). Members can be similar or different from each other according to many different factors; for example, two people can have different eye color, hair color, and age, but maybe they have the same weight. PCA is a way to find the factors that most explain the variation within a group by calculating the “principle components” of a group. The first principle component is associated with the factor(s) that explains most of the variation in a group. The second principle component explains the most variation in the group after accounting for the variation caused by the first principle component. And so on.

So, if you plot members of the group as dots on a 2-dimensional graph using the first two principle components to spread out the dots, you can construct a “map” of the group like the map in the paper above.

# PCA Map of Shakespeare’s Plays

Instead of human genomes, I tried using PCA to construct a map of Shakespeare’s plays using [data from the Folger Shakespeare library](http://winedarksea.org/?attachment_id=2227). The metrics I used to distinguish one play from another were the frequencies of all the words in each play. It turns out that when we represent each play as a point on the graph and use PCA to cluster similar plays together, they tend to group according to the three genres traditionally associated with Shakespeare’s plays: history, comedy, and tragedy.

Here is the plot of the top two principal components for Shakespeare plays:



Interesting to consider in the above graph are the so-called “problem plays,” or plays that are difficult to characterize as fully comic or fully tragic. Most of such plays are also placed on the border between comedy and tragedy, including *All’s Well That Ends Well*, *Measure for Measure*, *Merchant of Venice*, *The Winter’s Tale* and *Two Noble Kinsmen* (a Shakespeare collaboration with playwright John Fletcher).

**The Comic Potential of *Romeo and Juliet***

I am especially interested, however, in the plays that may not be considered as “problematic” as those above, but that are still closely clustered with many plays of a different genre. For example, the famous tragedy *Romeo and Juliet* is clustered alongside comedies like *Love’s Labour’s Lost* and *A Midsummer Night’s Dream*. In a recent lecture I attended by Shakespeare scholar Stephen Orgel, he described *Romeo and Juliet* as a tragedy that constantly tempts us with a comic ending. It seems as if all this temptation was enough to fool PCA.

A nice thing about PCA is that it is somewhat interpretable, especially with the way our model is constructed. More specifically, we can see the words that are highly correlated with the two principle components. It turns out that, for the first principle component, the strongest associations are with the words are “arms” (as in weapons) and “allegiance.” For the second principle component, the most important word is “love.”  So much of the generic variance in Shakespeare plays can be explained by love and war!

*Romeo and Juliet* is classified as a comedy because words like “love” show up with such high frequency in the play. This supports Orgel’s claim about the comic potential of *Romeo and Juliet*; whenever there is a similar level of “love” in another Shakespeare play, that play ends in a happy marriage. Only in *Romeo and Juliet* does Shakespeare devote so much time developing a (successful) marriage plot before forcing the play in a tragic direction. It almost seems as if a Shakespeare play is allowed to continue beyond a marriage, tragedy will inevitably ensue.

Still, a relatively simple model like PCA on word frequencies does remarkably well at clustering most plays according to genre. It would be absurd to think that Shakespeare’s oeuvre can be reduced to words like “love” and “war,” and so based on these results, it should seem equally absurd to think of Shakespeare’s plays as being classified as a single genre. Perhaps when teaching and thinking about any play, we should actively try to reason beyond the traditional notion of genre.