Appendix 6.3
Text Mining and Analysis Paper (two-three pages)
Robert Davis

Overview

Working with a partner, you will create a corpus of plays, books, or articles to analyze using the tools we have covered in class. Each student will write a 500–750-word paper discussing the process and what you learned from this reading of the texts.

Finding Data

There are many ways to get full-text sources. Here are a few:

1. Go to *Project Gutenberg*: http://www.gutenberg.org/browse/authors/ and find the author/play you are looking for.
    a. You will see several file formats: HTML is for online viewing, EPUB for Nooks and Kindles, and Plain Text. We want Plain Text. Open the link.
    b. *Project Gutenberg* has a lot of headers and footers, i.e., there is a ton of licensing material at the start and bottom. Select the text (if you don't know about SHIFT + Select, you should!) without the headers and footers.
    c. Paste it into another document
    d. Save that document as a plain text (.txt) file. You are reading to work with it!
2. If you know an author or play title you are looking for, search for it on Google and add "filetype:txt," and it will only return results that are in plain text. Some PDF files will let you copy and paste them into a .txt format.
3. *Literature Online* has LOTS of plays in full-text as does *Internet Archive*.

A Note on Preparing Texts

The peskiest thing about analyzing plays is the character names. Depending on the tool you are using, you can either type them in later as stop words (a set of words

that should be excluded from the results of the tool), or adjust the text before you analyze it. I would recommend the latter, so you don't have to remove character names each time you analyze. The Find/Replace function on Word and most TextEdit applications works quite well for this. For example, for *The Hunchback*, where every character's name when they spoke was abbreviated [i.e., "_Wal_" or "_Julia_" I ran Find/Replace to find the abbreviation (with the "_") and replaced it with a space, which effectively took them out].

Tools

You are free to use any of the tools that we have discussed in class (Word Clouds, Voyant, AntConc, Excel, etc.). I made a spreadsheet of available tools and their uses.

Organizing the Paper

Your paper should be divided into the following sections:

1. Research Topic/Hypothesis/Question
2. What tools you are going to use and why.
3. Your findings: this is the bulk of your paper and should include details on the tools that you used (what worked, what didn't), the work, and your overall conclusions. Please include screenshots or graphs!

Possible Topics

You can work on any topic you want and should feel free to choose a range of texts. Possible topics include comparing gender language, comparing the language of plays from two different parts of the century, comparing several plays by different playwrights, visualizing how many speaking roles or lines are given to different types of characters, etc.