Elizabeth Oakes
014318490
Introduction to Methods in Digital Humanities
Final Project
05.01.2017

This project uses two corpora to compare language representing altered states (madness, dreams, drug trips) to the language of general fiction at the lexical level. The general fiction corpus includes 30 one thousand word text files drawn from Project Gutenberg English language fiction texts published after 1900. The altered states corpus includes 30 one thousand word text files assembled from texts I purchased. The general fiction corpus is available here, but copyright prevents me from sharing the altered states corpus.

To begin, both corpora were loaded into Yoshikoder, a text analysis program written by Will Lowe, which applies any suitably formatted categorical dictionary to any series of texts. For this project, Colin Martindale's Regressive Imagery Dictionary (RID) was used. This dictionary contains about 3000 words arranged in 52 nested categories designed to identify text written by a writer in an altered state or text representing altered mental processes. Both Yoshikoder and the RID are freely available through these links where you will also find documentation.

Yoshikoder output a .csv (provided here as DigiHum.csv) containing word counts for each text for each dictionary category. This .csv was processed with R to produce a series of visualizations that demonstrate the main lexical differences between texts representing altered states and general fiction texts.

I first used a simple regular expression to filtered the data into altered and fiction datasets. I then used the sturges function to establish break points for the Primary and Secondary summary categories in both these datasets. These break points were used to set sensible bin widths in the below histograms.

The stripcharts show the distribution of word counts across Primary (a collection of word categories characteristic of altered states) and Secondary (a collection of word categories characteristic of "normal" states) summary categories for both corpora. This visualization is helpful to obtain a quick overview of the general shape of the data though it does not say much on its own.

The histograms with overlaid kernel density estimation provide a suitable view for comparing Primary and Secondary summary category words between both corpora. The violin plots with added jitter layer show the relationship between Primary and Secondary words in each text in both corpora.

The provided R script contains a few more visualizations done to explore the data, but the above six included here are most useful for making broad comparisons between the corpora. Of course with 52 categories and subcategories at hand, the data resulting from applying the RID to two corpora could say much more. This brief exploration is meant to demonstrate an approach while outlining the main lexical differences between writing that represents altered states and writing that represents "normal" consciousness.