ENG110: Introduction to Professional Writing
Textual and Sentiment Analysis of Company Documents

Computational Text Analysis

Scholars who have pioneered and pushed computational analyses of digitized texts have made it possible for undergraduate scholars of any discipline to engage in a new kind of close reading, a computational sweep for word frequencies and other measures of a single or library of documents. This hands-on exercise can be a fun probe of any collection of text files (raw .txt, .html, or .xml files, Unicode welcome), but it also reinforces effective practices when working with digitized and/or large sets of files, yes: practice too. Practice with working with tools and texts and data and analyses and that generate new and that generate new questions; repeat.

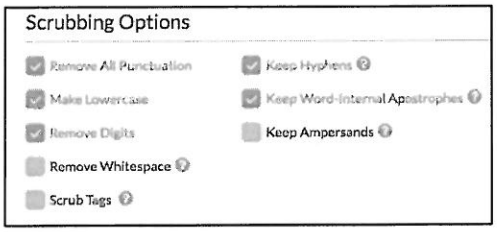
(1) Using the Chrome or Firefox browser, open the Lexos tools: http://lexos.wheatoncollege.edu

Workflow Example 1. CLUSTERING by Author
Upload -> Scrub -> Cut -> Cluster -> Kmeans

Example Uses:
(i) students use Lexos to "close read" their own writing samples in small groups;
(ii) students examine novels or papers by multiple authors;
(iii) students design authorship attribution experiment, including controls
(iv) students examine positive or negative sentiment embedded within documents

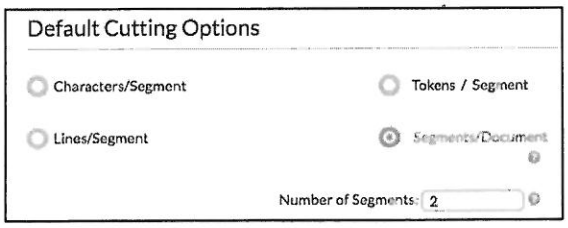
(2) **Upload** two (raw) text files from your group:

(3) Under the Prepare menu, select **Scrub**; select the settings shown



and **Apply Scrubbing** Apply Scrubbing

(4) Under the Prepare menu, select **Cut**; cut the four documents each into two (2) Segments

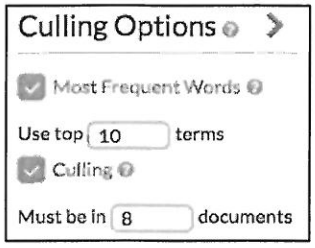


and **Apply Cuts** Apply Cuts

(5) From the Analyze menu, select **Clustering**, then **K-means Clustering**

(6) Change the *number of clusters* to four (**4**) and select **Get K-Means** 

(7) From the Analyze menu, select **Topwords**
   a) Turn off the **Culling** options (for now)
   b) **Tokenize** by 1-gram by Tokens (Words)
   c) Select **Get Topwords** 



(8) From the Visualize menu, select **Word Cloud** (one file at a time), **Multicloud** (side-by-side wordclouds for more than one), and/or **BubbleViz** for three visualizations of word frequencies.