Computerized Text Analysis: Classwork 2
Basic Descriptive Text Statistics

Kenneth Benoit

The objective of this class exercise is to continue working with the French manifesto texts from Classwork 1, and to convert them into relative text frequencies and perform some basic descriptive analysis.

You will need the texts from Day 1 that you converted into plain text, UTF-8 files, as well as some additional utilities:

Instructions:

1. Inspect the converted French texts from Day 1. Inspection can be through a text editor, web browser, or using (UNIX) command line tools such as "less" and "file". The "file" command in particular can be used to verify that they are all in fact UTF-8 encoded, or seeing that they display properly when you view them in a viewer set to UTF-8. For example, you can open them in a web browser (e.g. Safari or Firefox) and override the "View–>Encoding" to `Unicode` or `UTF-8` and making sure it looks okay. (Some text editors can be set to a specific encoding display as well, although most are automatic.)

2. You may wish to *clean up* some of the converted texts in your text editor. This is especially an issue for text files that have been converted from pdf or Word and contain converted formatting characters such as page numbers, footers, and headers. Since none of this information is part of the text we are interested in analyzing, we should remove it.

3. Download the program Jfreq (**Windows exe** or **Mac dmg** version). This is a stand-alone Java executable that can be used to convert text files into word frequency matrixes.

4. Generate a word frequency matrix using JFreq and save it as `wordfreqm.csv`. ("csv" refers to Comma Separated values, a plain text format where each column is delimited by commas, which makes it easier for these sorts of files to be imported by spreadsheets and statistical packages.)

5. Import the word frequency matrix into your statistical package (or a spreadsheet program such as Excel).

   - In R:

     ```
     install.packages("foreign")
     library(foreign)
     setwd("your directory")
     wfm <- read.csv("wordfreqm.csv", header=TRUE)
     ```

   - In Stata:

     ```
     cd "your directory"
     insheet using wordfreqm.csv
     ```

   - You can also open the `.csv` file directly into a spreadsheet.

6. Inspect the matrix: does it look okay? (and: does the list of words, in whatever software you are using, look okay on the screen?)

7. Generate some basic descriptive statistics:

(a) What is the total number of types (unique words) per document?

(b) What are the total number of words (tokens) per document?

(c) What is the vocabulary diversity of each document?

(d) What is the median word frequency per document?

(e) What is the most frequently used word among all of the documents?