Course title: Digital Humanities Prof. J. Savoy

Serie 5. Probabilistic models and text preprocessing

A new folder was created in the folder "Exercises" entitled "Corpus" with different corpora useful for the next practical exercises. For this week, we will focus on the Federalist Papers (federailts-papersNew2.csv) which contains the 85 Federalist papers.

From this file, extract the papers written by 'Hamilton' (51 papers), 'Madison' (14 papers), and the test set (12 papers written by 'Hamilton OR Madison'). This test set is {49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 62, 63}.

The file contains a large data frame with row = paper and column = word-type. The field 'AUTHOR' indicates the author of the paper (thus the preprocessing is already done for this corpus).

1. Considering the words 'to', 'upon' and 'would', draw a graph representing the occurrences of those words in Hamilton and Madison's articles.
2. With these three words, model them as a binomial to reflect either occurrences in Hamilton or Madison's writing style.
3. Represent with a histogram the article length. Does it make sense to view this distribution as a Gaussian?

With the second corpus about SMSs (sms_spam.txt), you need to preprocessed the file to be able to extract only the good SMS ('ham') or the spam ones ('spam'). To preprocessed the text, you need to:

a. Transform the text to lowercase.
b. Normalize the tokens (replace the English contraction by their equivalent such 'can't' -> 'can not').
c. Be able to read a list of stopwords and to remove them when they appear in the text (two such lists are given in the folder 'Corpus').

4. Apply your preprocessing to both the spam and ham SMSs. Return the top 20 most frequent word-types for both categories.

(You will need your results for both the Federalist and SMSs corpora for the next exercises).