Stylometry: Federalist Papers and Word and Sentence Length

After completing this module, students should be able to:

- Give examples of ways you might distinguish authorship in a quantitative way.
- Describe a method for comparing how close together two distributions or bar charts are.
- Adapt existing R code to gather evidence about authorship based on distributions of word length and sentence length.
Before class:

- Read the first few paragraphs of the four short story excerpts posted at https://tinyurl.com/uncMath115StudentSpreadsheets. You DO NOT have to read the entire excerpts.

- Two of these stories are written by one author and two by a different author. Try to decide which stories are by the same authors. Please do it without googling!
Introduction to Stylometry

**Question.** What is stylometry?

**Question.** In the before class assignment, how did you decide which stories were written by the same author and which by different authors? What quantitative measures could you use to decide?
Word length

What can word length tell us? What should we be looking for in these histograms?

Story 1

Word Length

Story 2

Word Length
Word length

STYLOMETRY: FEDERALIST PAPERS AND WORD AND SENTENCE LENGTH

![Histogram of word length for Story 3 and Story 4]
Sentence length

What can sentence length tell us?

**Story 1**

![Sentence Length in Words](image1)

**Story 2**

![Sentence Length in Words](image2)
The Federalist Papers


Historical Context

The Federalist Papers are a collection of 85 political theory articles published between October 1787 and May 1788. These papers presented the case for the system of government that the U.S. ultimately adopted.

Three people wrote the papers:

- Alexander Hamilton, first Secretary of the Treasury of the United States
- James Madison, fourth President of the United States
- John Jay, first Chief Justice of the United States
However, who wrote which of the papers was a matter of open debate for years, because:

- The Federalist Papers were published anonymously under the pseudonym “Publius”.
- Madison and Hamilton left conflicting testimonies regarding their roles.
  - Two days before he was killed in a duel, Hamilton wrote a note claiming 63 of the 85 Papers as his own work.
  - Ten years later, Madison refuted some of Hamilton’s claims, stating that he was the author of 12 of the papers on Hamilton’s list and that he had done most of the work on three more for which Hamilton claimed equal credit.
- Madison and Hamilton had unusually similar writing style.
Data Analysis

- Go to the stylometry assignment on RStudio.
- Click on the tab "federalistStylometry.Rmd" and scroll to the top of the file.
- Use the arrows to the left of the blocks of code to run them, one at a time, in order.
- Look at the histograms of word lengths. Does this help you determine authorship?
- Fill in the code at the places marked with ############### to find sentence length.
- Fill in the code at the last spot marked with ############### to draw histograms of sentence length. Does this help you determine authorship?
Homework

1. Add the code in three places marked by the following:

    # ADD CODE ABOVE TO DO THE SAME THING FOR MADISON, JAY, and DISPUTED

Also add code to draw histograms of sentence length in the code file, replacing the comment marked by ######, and run the code to get the histograms. When you hand this in,

(a) include a screenshot of the code

    # now repeat for sentences (part 1)
    # go back to the original corpuses and split them into sentence tokens instead
    sentencesHamilton = tokens(hamiltonCorpus, remove_numbers = TRUE, remove_punct = TRUE, what = "sentence")
    sentencesMadison = tokens(madisonCorpus, remove_numbers = TRUE, remove_punct = TRUE, what = "sentence")
    sentencesJay = tokens(jayCorpus, remove_numbers = TRUE, remove_punct = TRUE, what = "sentence")
    sentencesDisputed = tokens(disputedCorpus, remove_numbers = TRUE, remove_punct = TRUE, what = "sentence")

    # turn each lists of sentences into a spreadsheet, or "data frame" - sentences
dfSentHam = data.frame(text = sapply(unlist(sentencesHamilton, recursive = FALSE), as.character), stringsAsFactors = FALSE)
dfSentMad = data.frame(text = sapply(unlist(sentencesMadison, recursive = FALSE), as.character), stringsAsFactors = FALSE)
dfSentJay = data.frame(text = sapply(unlist(sentencesJay, recursive = FALSE), as.character), stringsAsFactors = FALSE)
dfSentDis = data.frame(text = sapply(unlist(sentencesDisputed, recursive = FALSE), as.character), stringsAsFactors = FALSE)

# find sentence length as number of words (part 3)
dfSentHam$SentenceLengthInWords = lengths(strsplit(dfSentHam$text, " "))
dfSentMad$SentenceLengthInWords = lengths(strsplit(dfSentMad$text, " "))
dfSentJay$SentenceLengthInWords = lengths(strsplit(dfSentJay$text, " "))
dfSentDis$SentenceLengthInWords = lengths(strsplit(dfSentDis$text, " "))

# draw histograms: adapt the code above that drew histograms of word length to draw histograms of sentence length
# HINT: you want to draw a histogram of dfSentHam$SentenceLengthInWords , and similar for Madison, etc.
# HINT: to set the bin sizes, use seq(0, 210, by = 10) to have bins from 0 to 210 that are 10 units wide
# HINT: use main = "title" to give your chart a title (fill in what you want instead of title)
#HINT: use xlab = "label" to give the label for the x-axis

hist(dfSentHam$SentenceLengthInWords, breaks = seq(0, 210, by = 10), main = "Hamilton", xlab = "Sentence Length in Words")
hist(dfSentMad$SentenceLengthInWords, breaks = seq(0, 210, by = 10), main = "Madison", xlab = "Sentence Length in Words")
hist(dfSentDis$SentenceLengthInWords, breaks = seq(0, 210, by = 10), main = "Disputed", xlab = "Sentence Length in Words")
hist(dfSentJay$SentenceLengthInWords, breaks = seq(0, 210, by = 10), main = "Jay", xlab = "Sentence Length in Words")
(b) include screenshots of each of the four histogram.

(c) In your opinion, does this give you enough info to decide who authored the disputed papers? If so, who does it look like the author is for the disputed papers and how can you tell? If not, why not?

(d) Would these histograms give you good evidence to rule out Jay as the author of the disputed papers? (not that anyone thought he wrote the disputed papers) Explain.