Week 2 Assignment

The Data:

https://drive.google.com/drive/folders/1993Yf-s15-Ni_odwpWVga5AVvN1QgXQc?usp=sharing

RE: https://guides.loc.gov/federalist-papers/full-text

Directions: Complete this Assignment in R.

1) Using the **Federalist Papers Corpus**, use R to create a labeled data frame and a matrix (either or both is fine – see the code for examples) with the **words as the column names**, the **rows as each paper**, and the data values as word counts.

The following is an example of placing words as columns and rows as papers for authors. You will create a similar format for your Federalists Papers. So this is an example of this type of format.

```
                Terms
Docs             abbey abbots abdy abhor abhorred abide abilities
  Austen_Emma.txt    31      1    1     1        1     1         3
  Austen_Pride.txt    0      0    0     0        0     1         6
  Austen_Sense.txt    0      0    0     1        2     0         9
```

Be sure to remove useless words – like “and”, “but”, “it”, etc.

You are free to determine which words to remove and to try different options.

2) Also **print out the top 10 most frequent words**. You may also choose to print the top 10 least frequent words. Think about what this tells you.
3) **Print out the number of words in each document** (after removing any stopwords like “and”, “but”, etc.)
4) Create a wordcloud for Hamilton and one wordcloud for Madison and compare them. What do they show and suggest?

You are free to clean up and process the data as much as you wish. More is always better.

Optional: Use the code example to try to cluster your data.

The goals of this assignment are to:

1) Use R to work with text data

2) Get the text data into an analyzable format (dataframe and/or matrix as shown in the example above).
3) Think about what you have found and what it means.

Optional Additional Practice ad Steps:

If you complete the above and wish to practice more – think about how you might determine who wrote the disputed (disp) papers.

Notice that some papers are written by Madison, some by Hamilton, some by Jay, and some by HM (Hamilton and/or Madison).

However, some of the papers are unknown and are called “disputed”.

How might you determine who wrote each?

Was is Madison? Was is Hamilton?