

[← BACK TO GRADES](#)

# Homework Assignment 4 (week 4)

## Faculty Feedback

**Score**

Last published: 8/5/2019 6:07 PM PDT

**98 / 100 (98.00%)****Comments****Comments****Overall:**

GREAT!! analysis here. Your effort has really paid off. See a few notes below.

**Introduction:**

Good introduction; however, some more detail framing the problem, issues, concerns, and various implications may help to motivate the proposed methodology.

**Distance Metrics and Experimental Design:**

GREAT INITIAL EDA!!! The paper would benefit from some more exploratory analysis wrt the character of the data, eg distribution of words, effect of stop words, ... . And some more supporting visualizations of the nature of the data.

EXCELLENT INVESTIGATION of the cluster analysis! WELL DONE!

**Conclusions:**

The discussion as to the disputed authorship could have been more deeply explored and discussed in the concluding remarks. But good high level analysis!

**Word Cloud:**

The wordcloud package can be difficult to install and can be a bit "finicky", so do not worry. If you have interest, feel free to read up on this specific package:

<https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf>

**General Notes Below.****Required Elements:**

In this homework you are provided with the Federalist Paper data set.

You can also use either the already processed .csv or you can create your own corpus. I recommend that you create the corpus and read in the corpus. (I posted exemplar code for either option!)

NOTE: The features are a set of "function words", for example, "upon". The feature value is the percentage of the word occurrence in an essay. For example, for the essay "Hamilton\_fed\_31.txt", if the function word "upon" appeared 3 times, and the total number of words in this essay is 1000, the feature value is  $3/1000=0.3\%$

- Now you are going to try solving this mystery using clustering algorithms k-Means, EM, and HAC.
- Document your analysis process and draw your conclusion on who wrote the disputed essays.
- Provide evidence for each method to demonstrate what patterns had been learned to predict the disputed papers, for example, visualize the clustering results and show where the disputed papers are located in relation to Hamilton and Madison's papers.

- For k-Means and EM, **analyze the centroids to explain which attributes are most useful** for clustering.
- Hint: the centroid values on these dimensions should be far apart from each other to be able to distinguish the clusters

**Headings Required:****Basis For Grades:**

100: This means that your Assignment was amazing. It covered everything – cleaning, prep – analysis that makes sense – visualizations – results (that are true) – etc. There is nothing really left to improve.

95: This means that your Assignment is really good! You covered most of the items noted above and perhaps a few others not noted. You can make some improvement on pre-processing and results analysis, as well as perhaps other visualizations. Overall – you have the idea and you did well.

90: This means that your Assignment is good, but could be a little better. Perhaps add items such as further data cleaning and pre-processing, data normalization, better or more visualizations, and/or more robust conclusions.

85: This means that your Assignment is a good start and largely meets the more general and overall requirements



Below 85 means that the level of 85 above was not quite met and many elements were missing.

**Student Submission** | [Homework Assignment 4 \(week 4\)](#)**Response**

Last submitted: 8/1/2019 12:41 AM PDT

*No response***Files** | [Download all \(3\)](#)

File Name	Uploaded	Feedback
-----------	----------	----------

File Name	Uploaded	Feedback
<a href="#">Ryan_Timbrook_HW4.docx</a>	8/1/2019 12:40 AM PDT	
<a href="#">ist707_hw4.Rmd</a>	8/1/2019 12:40 AM PDT	
<a href="#">rtimbroo_util.R</a>	8/1/2019 12:40 AM PDT	