|  |
| --- |
|  |
| A picture containing logo  Description automatically generated |

|  |  |
| --- | --- |
| Applied Data Science Portfolio | Beverlyn Tucker |
| SUID | 255962510 |
| Professor | Stephen Wallace |
| Date Submitted | August 30, 2021 |
| Website | [Portfolio](https://bevstucker1976.wixsite.com/bevtucker) [GitHub](https://github.com/BevTucker?tab=repositories) |

Portfolio Beverlyn Tucker

# **Introduction**

## I got interested in data when I was working as a quality technician. Data is the output of the daily activities. Then later, I became an Internal auditor at Boeing. I utilized data in my audit; these give me a fascination and how data is interconnected to each process. This fascination leads me on the journey as a Data Scientist.

# **Journey to Data Science**

During this journey, I came across a Deloitte company YouTube interview. The interview was about how the company utilized data Analytics for their internal audits, how data analytics can be applied in internal audits (Janet L and Terry H,2012). These ignite me to go back to school to understand better how to use data for the audit. Then a few months later, I attend a conference at Washington University about data analytics one that captures my attention is the ethical dilemmas on the data. After the conference, I decided to take some professional programs/studies specific to data science. So, in April 2019, I started doing the Applied Data Science program at Syracuse.

# **Data Science Process**

# 

# The Data Science Process is a problem-solving process. Before getting into the process, one has to have an apparent problem to solve. Sometimes the problem is given and trivial. Other times, the problem may be buried deep in the business requirements and processes for problem formulation.

**IST652\_Scripting Project Ecommerce Women’s Clothing**

**Goal:** This project aims to help eCommerce clothing Investigate ratings to see if businesses can utilize this information to formulate actionable insights for improvements. Rating is fundamental in the online platform.

First, we will concentrate on the ratings and compare ratings one and two with ratings three through five. Second, based on this group, is the demographic change? Third, apply polarity analysis to the text reviews to see if the results are positive or negative? Fourth, what products to recommend to the customer, then focus on target products? Fifth, leveraged three models, Naïve Bayes(BNB), Bernoulli, and Support Vector Machine(SVM), to predict which reviews would get various rating levels, whether costumers will recommend, and whether they would receive positive feedback from other customers.

**Technologies:** Pandas WordCloud Word tokenize, WordVectorizer ,TextBlob and Steaming

**Techniques:** Data mining, descriptive analytics, grouping and aggregation, plotting, data transformation, illustration

**Models:** Naïve Bayes(BNB), Bernoulli, and Support Vector Machine(SVM)

As the world shrinks, due to rapid growth in technology, every conceivable industry's landscape has changed over a relatively short amount of time to keep pace with technological advancement. On the consumers' products and services side, people can purchase more various goods than ever before on platforms like Amazon to Google's services.

With so much access and the ability to tap into a vast consumer base, businesses can see what drives consumer purchasing decisions to leverage customers' experiences before making a purchasing decision or investing in a marketing strategy. Without spending an enormous budget on marketing, online reviews will provide direct feedback on how the business can improve. Moreover, with the online economy, companies can incentivize feedback with better shipping rates or a discount. The establishment was always looking at the demographics.

The companies may ask the data scientists to help have an insight with their products and help with their business to keep up with competitors. Rating is fundamental in the online platform. A client may ask to Investigate the ratings. We will concentrate on the ratings and compare ratings one and two with ratings three through five. Base on this group, is the demographic change? When applying polarity analysis to the text reviews, Is the results more positive or negative? What products do you recommend to the customer to focus on or to target products? What model is appropriate for this classification?

One essential benefit a company may exploit from online reviews is that the analytical insights may empower businesses to improve without spending a huge amount of money on traditional marketing research. This also creates opportunities for direct interactions between the business and its customers. For example, companies can incentivize customers to provide direct feedback by offering perks and discounts. Of course, the drawback is that negative sentiment may quickly spread from one unhappy customer to a larger crowd. However, positive feedback from former customers may very likely resonate with new customers, which resonates and realizes a purchase.

Not surprisingly, some of the most detailed feedback comes from personal experiences, and they are usually unique to individuals. Prospective customers may find this very helpful as there are always minute aspects learned from someone else's perspective. These unique aspects echo among customers, and the prospective customers may leverage reviews to make purchasing decisions and generate more feedback extolling their joy or grief. This mitigates one downside of the online clothing e-commerce -- one cannot try on the clothing before purchasing. With so much information in these online reviews, how can companies (and even some customers) churn through a sea of reviews to understand how people feel relatively quickly? Thankfully, machine learning algorithms were implemented to determine whether reviews in message are related to positive or negative sentiment. Because algorithms are so fast and can run in perpetuity, companies and customers alike can look at the produced aggregated information to divine meaning. The only questions remaining are "which algorithm do you choose?", "How difficult is it to produce?" and "how accurate is it?"

# **ANALYSIS**

This project aims to ingest data from Kaggle.com regarding women's clothing reviews, clean it, use unsupervised and supervised methodologies to understand the customers' opinions towards various clothing pieces that have been purchased. The variables to be predicted Rating, it is possible to look deeper into the reviews and see which terms are strongly associated with recommendations and high ratings. From there, Support Vector Machines and Naïve Bayes Algorithms were leveraged to predict which reviews would get various rating levels, whether costumers will recommend, and whether they would receive positive feedback from other customers. Ultimately, the goal is to see if this information can be utilized by businesses to formulate actionable insights for improvements.

All Departments All rating

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

# Chart, bar chart Description automatically generatedChart, bar chart Description automatically generated

# Chart, bar chart Description automatically generatedChart, bar chart Description automatically generated

# Chart, bar chart Description automatically generatedChart, bar chart Description automatically generated

# Rating 1 and 2 Rating 3 to 5

# Text Description automatically generatedText Description automatically generated

# Chart Description automatically generatedChart, histogram Description automatically generated

# Chart, bar chart Description automatically generatedChart, histogram Description automatically generated

# **Chart, bar chart Description automatically generated Chart, bar chart Description automatically generated**

# **Clooking for collinearity**

# Rating 1 and 2 Rating 3 to 5

# Diagram Description automatically generatedChart, diagram Description automatically generated

# **Model**

SVM Accuracy: 0.8090909090909091

array([[ 19, 55],

[ 8, 248]], dtype=int64)

len\_score= cross\_val\_score(modelDT,x\_train\_ld, y\_train, cv=5)

len\_score.mean()

0.7895522388059703

MNB Accuracy: 0.7757575757575758

array([[ 0, 74],

[ 0, 256]], dtype=int64)

MNB\_score= cross\_val\_score(Mul\_modelNB,x\_train\_ld, y\_train, cv=5)

MNB\_score.mean()

0.7805970149253731

Bern: 0.7727272727272727

array([[ 10, 64],

[ 11, 245]], dtype=int64)

Bern\_score= cross\_val\_score(Bern\_clf\_brn,x\_train\_ld, y\_train, cv=5)

Bern\_score.mean()

0.755223880597015

# **RESULTS SUMMARY**

Fashion is customary to its origin but also loved in different parts of the world. Who does not love style when clothing is worn every day? Statistics from Inc website states, "91% of people read online reviews regularly or occasionally." "84% of people trust online reviews as much as friends." "68% of people form an opinion after reading one to six online reviews." Businesses could turn reviews into revenues. Therefore, how would customers wade through all these products and services to understand which platforms, products, and services are good or bad choices based on reviews from Kaggle's E-commerce Women's Clothing?

After comparing ratings one and two to ratings three through five, the data showed the majority demographics did not change and was still the same age group from ages 25 – 55.

The departments to consider for improvements based on reviews are Dresses and Tops, on tops specifically to knits. These two departments have approximately 1,000 counts of individuals not in ratings three through five.

After performing sentiment analysis to the target products/focus products, specifically polarity analysis, the data shows promising results, about 400 reviews that show positive and about 100 negative results. These have a higher probability of turning this product rating 3-5. By improving the product from fabric quality, show true color, not photoshop/photo enhancer. By utilizing tools like a camera to reflect true color in the pictures and easy access to the size charts.

This project is one of the classification problems. Out of three models Naïve Bayes(BNB), Bernoulli, and Support Vector Machine(SVM). SVM linear shows is a good algorithm to use, although it is not perfect. I suggest training all three models with large datasets to see the best results.

# **LESSON LEARN**

# Lesson learn When I was experimenting and analyzing the word cloud, something missing. This time I made a different approach. All processes are the same; besides the following, I did not use a stemmer I CountVectorizer and TFIDF, then shape the data, then pass through to the World Cloud this time works the way I expected. As data scientist essential to have an open mind to try different methodology to test against each other to see if the results are comparable.

# **CONCLUSIONS**

# This analysis has shown two-fold: first, no single approach to a problem will be a panacea. If time permits, trying as many strategies and permutations of those strategies as you can is the best way to achieve good results. Assuming one strategy will perform poorly and not testing it out could lead to lost opportunities to produce better results. Second, looking at accuracy as a measurement is not good enough. Data scientists must use alternate, more nuanced analysis to gain better insights into what their methods achieve.

From a customers' perspective, reviews can help people make more informed decisions, and understanding how machine learning works, even at a basic level, may help customers write better reviews. How? It isn't easy to be heard when there are so many voices online. Suppose word choice is essential data aggregation and understanding sentiment more quickly, perhaps by modifying one's reviews to be more easily understood by machines. In that case, executives will become more informed about customers' opinions. This could lead to better products and services since, ultimately, businesses want to sell more, and happy customers are more likely to be returning customers.

From a business standpoint, as technology continues its march forward, human society will move along with it and attempt to utilize it in a myriad of ways to better our lives and make business easier. Data analytics techniques are rapidly being developed and deployed around the internet, from condensing information to identifying bath-faith actors. Even in phones and smartwatches, people carry around every day. With any luck, mathematicians and data scientists will keep up with the internet and data to better understand our fellow citizens and truly know if that clothing store is worthy of its five-star Rating. [More info](https://github.com/BevTucker/Data_Science_Python/blob/master/Women's_Clothing_Review.ipynb)

# **IST718 Big Data Analytics Big Data PROBLEM MNIST dataset**

# **Goal:** To illustrates which machine learning performed the best between GaussianNB, Decision Tree Classifier, and Neural Network compared to the MNIST techniques that have been found upper 98-99%. On MNIST dataset is a collection of images created by the National Institute of Standards and Technology (NIST). Each image in the dataset consists of 28 x 28 pixels with 60,000 images for training and 10,000 for testing.

**Technologies:**

tensorflow.keras import datasets # for keras MNIST datasets

sklearn.metrics import confusion\_matrix # for prediction verification

matplotlib.pyplot # for plotting

numpy # for np functions

seaborn as sns; sns.set() # for heatmap

sklearn.metrics import classification\_report # to print classification accuracy

tensorflow.keras.utils import to\_categorical # to convert data to categorical

tensorflow.keras import layers # for Neural Network layers

tensorflow.keras import Sequential # for Neural Network

keras.callbacks import ModelCheckpoint #for Neural Network callback

sklearn.naive\_bayes import GaussianNB # for GNB

sklearn.tree import DecisionTreeClassifier # for decision Tree classfier

**Techniques:** Data gathering, descriptive analytics plotting, data transformation, illustration, predictive analytics, machine learning, Confusion matrix

**Models:** GaussianNB, Decision Tree Classifier, and Neural Network

# In this generation, as technology improves, the amount of digital data generated every day in stores increases rapidly—the company hires data scientists on how to maximize the business and gain insight from it.

# The MNIST dataset is a collection of images created by the National Institute of Standards and Technology (NIST). Each image in the dataset consists of 28 x 28 pixels. With the provided 60,000 images for training and 10,000 for testing, users have typically found between 93 to near 100% accuracy. The successful implementation of varying models has been both a triumph and a problem for computer vision. Specifically, since the optimal solution has been attained, predicting handwritten images is no longer one of the defacto challenges.

# The dataset used for this study was obtained directly from the fashion mnist.load\_data straight from the cloud. Since this study has been a simplified computer vision problem, no additional data scrubbing was performed. However, future studies could possibly improve the analysis by including additional images that are not the target fashion groups. Utilized tenssorflow.keras to import the data sets, confusion matric for prediction verification, and classification report for classification accuracy.

# Application Description automatically generated with medium confidence

# A picture containing chart Description automatically generated Graphical user interface Description automatically generated

# Table Description automatically generated

# The study illustrates that the neural network implantation significantly better than the GaussianNB with weighted average of 64% and f1 score of 56%. And Decision Tree Classifier with weighted average of 79% and with f1score of 79%. It would be interesting to see if adjustments in the neural network weights, number of layers, and different backward propagation techniques could change the difference between the two implementations. Similarly, changing the parameters for GaussianNB and Decision Tree Classifier. That would be interesting to see if the outcome changes. Compared to the traditional MNIST number classification, the fashion variant is not as accurate. Specifically, the MNIST techniques have been found in the upper 98-99%. However, as stated earlier, more clever techniques could be imposed to draw higher results; this could involve a series of ensemble learners though a simple additional approach could involve appending to the provided dataset. In addition, data could be as simple as a set of images not belonging to any of the target classes. For example, additional images of non fashion images could be used during the train. This would be analogous to extending the one versus rest, taken advantage of within the sklearn framework.

# **Conclusions**

# Overall this study has succeeded in demonstrating how a multilayer neural network can outperform the GaussianNB and Decision Tree Classifier. However, as stated in the results, it would be interesting to explore the adjustment of parameters for both techniques. Specifically, adjusting the penalty and changing the kernel. It may or may not changes the result. At a higher level, the fashion MNIST dataset is more interesting than its predecessor. More improvements and additional techniques will be needed to achieve the same or close level of accuracy as the numbers example. [More info](https://github.com/BevTucker/Data_Science_Python/blob/master/IST718_Lab3.ipynb)

##### IST664\_NLP Corpus Statistics and Mutual Information

**Goal:** To instigate two articles on whether the writer has changed in his writing approached. Choosing two articles written by Russell Carlton for the year 2009 and the year 2018 in the website Baseball Prospectos, he relied heavily on statistics and the psychological side of the report. The main theme in NLP is text representation, which is fundamental and indispensable for text-based intelligent information processing. Generally, text representation word frequencies examining the relevance of keywords to documents in the corpus.

**Technologies:**

import nltk, from nltk.corpus import stopwords, from nltk import bigrams

from nltk.collocations import \*, import re for regular expression

# **Techniques:** data mining, descriptive analytics, grouping and aggregation

# 

# **Model:** TFIDF, Pointwise Mutual Information (PMI), bigram and ngram

Mainstream media is a traditional reference for news entities capable of shaping and molding opinions of the public. The writers capable of reaching out to a different level of audiences. The dissemination of knowledge covers a wide variety of platforms, including television, online, articles, social media, and more.

The Limitless capabilities of communication allow news agencies, writers to maximize information delivery. Moreover, benefits potentially include a better-informed and engaged general public. However, the writer can express their feelings or utilizing statistics to avoid biases. The writer can analyze and write an article based on experience, observation, psychology, statistical reports, and feelings, and ideas.

This article that is going to instigate whether the writer has change in his writing approached. Choosing two articles written by Russell Carlton for the year 2009 and the year 2018 in the website Baseball Prospectos was he rely heavily on statistics and the psychological side of the report.

One of the main themes in NLP is text representation, which is fundamental and indispensable for text-based intelligent information processing. Generally, text representation word frequencies examining the relevance of keywords to documents in the corpus.

We will begin loading the nltk and re packages that will allow us to process both texts and do some initial cleaning, such as lowercasing words and stemming.

We will first lowercase every word within each corpus. Then, we will remove stopwords from both corpora; however, we will not add any initial stopwords to nltk's stopwords list. If any additional words show up within the corpora that are deemed unnecessary, these words will then be added and will be reprocessed. Finally, we will stem every token within the corpora.

# **After removing stop words**[**¶**](https://render.githubusercontent.com/view/ipynb?color_mode=auto&commit=28221831217223a27362939726440453edc3b294&enc_url=68747470733a2f2f7261772e67697468756275736572636f6e74656e742e636f6d2f4265765475636b65722f446174615f536369656e63655f507974686f6e2f323832323138333132313732323361323733363239333937323634343034353365646333623239342f436f7270757325323053746174697374696373253230616e642532304d757475616c253230496e666f726d6174696f6e2e6970796e62&nwo=BevTucker%2FData_Science_Python&path=Corpus+Statistics+and+Mutual+Information.ipynb&repository_id=245831982&repository_type=Repository#After-removing-stop-words)

By removing stop words drop down dramatically 2009 articles drop 564 from 1369 to 805 tokens and 731 from 1650 to 919 tokens

**Most common words analysis**[**¶**](https://render.githubusercontent.com/view/ipynb?color_mode=auto&commit=28221831217223a27362939726440453edc3b294&enc_url=68747470733a2f2f7261772e67697468756275736572636f6e74656e742e636f6d2f4265765475636b65722f446174615f536369656e63655f507974686f6e2f323832323138333132313732323361323733363239333937323634343034353365646333623239342f436f7270757325323053746174697374696373253230616e642532304d757475616c253230496e666f726d6174696f6e2e6970796e62&nwo=BevTucker%2FData_Science_Python&path=Corpus+Statistics+and+Mutual+Information.ipynb&repository_id=245831982&repository_type=Repository#Most-common-words-analysis)

The 50 words frequency in the year 2009 is all about the team is seems blue jay appears (208) times and Halladay(189) is the main topic about prospect, proposal, and offer. The offer was not fear and someone walkaway.

On the other hand, in 2018, it is a different flavor about the skills; the ball appears 349 times left fielder; these always go together in baseball left fielder scores better than the centerfielder. Talking about the batter, the hit, worth, percentage, and also talks about the defense and all components in the baseball game.

**Bigram Analysis**[**¶**](https://render.githubusercontent.com/view/ipynb?color_mode=auto&commit=28221831217223a27362939726440453edc3b294&enc_url=68747470733a2f2f7261772e67697468756275736572636f6e74656e742e636f6d2f4265765475636b65722f446174615f536369656e63655f507974686f6e2f323832323138333132313732323361323733363239333937323634343034353365646333623239342f436f7270757325323053746174697374696373253230616e642532304d757475616c253230496e666f726d6174696f6e2e6970796e62&nwo=BevTucker%2FData_Science_Python&path=Corpus+Statistics+and+Mutual+Information.ipynb&repository_id=245831982&repository_type=Repository#Bigram-Analysis)

The bigrams are very interesting how words connect each other, although blue and jay appeared many times, to begin with still in the top one with the word frequency of 0.017% of likelihood ratio and raw frequency. Next is Roy, and the Hallway, machine knew that these words are pair together. Upon looking with the pair of words, is confirmed the topic between Blue Jays and Roy Hallaway, the writer(Russell) talking if Hallaway will pick two or three persons and seems blue jays walk away from the offer. The writer continued to talk about the blue jay. One that is very noticeable the word better describes in so many ways but has the lowest word frequency.

The year 2018 talking about a baseball game norm was the left fielder with a bigram score of 0.023%. And who make to the about the playoff. Upon looking at 1520, it puzzled me, but later I figure it out it is 15-20 percent, during the process it got deleted the dash. Base on this data, the writer uses statistics to perform the analysis. One that is very noticeable the word ball describes in so many ways. Besides that, the writer is very connected to the event.

Scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Bi gram forRussel\_Year\_09 Bi gram forRussel\_Year\_2018

Text

Description automatically generated Text

Description automatically generated

(('roy', 'halladay'), 5.250095350126295)

**Pointwise Mutual Information (PMI)**

**PMI year\_2009**

(('blue', 'jay'), 5.138370267567467)

**PMI year\_2018**

(('left', 'fielder'), 4.923406437897166)

**PMI Analysis**[**¶**](https://render.githubusercontent.com/view/ipynb?color_mode=auto&commit=28221831217223a27362939726440453edc3b294&enc_url=68747470733a2f2f7261772e67697468756275736572636f6e74656e742e636f6d2f4265765475636b65722f446174615f536369656e63655f507974686f6e2f323832323138333132313732323361323733363239333937323634343034353365646333623239342f436f7270757325323053746174697374696373253230616e642532304d757475616c253230496e666f726d6174696f6e2e6970796e62&nwo=BevTucker%2FData_Science_Python&path=Corpus+Statistics+and+Mutual+Information.ipynb&repository_id=245831982&repository_type=Repository#PMI-Analysis)

Surprisingly the PMI is very good. Only one word that don't occur together the world 'littil' for the word little in the 2009 article. All are good collocation pairs have high PMI because the probability of co-occurrence is only slightly lower than the probabilities of occurrence of each word. Conversely, a pair of words whose probabilities of occurrence are considerably higher than their likelihood of co-occurrence gets a small PMI score.

Next, using PMI setting a five-word frequency limit. Both articles have 50 qualifying bigrams. These all refer to characters and actions associated with the characters.

In the 2009 article, Russell uses a character-centered approach, focusing on the situation itself and how characters react to the same events.

On the other hand, the 2018 article. Action-based writing. The author keeps the action flowing. There are many descriptive bigrams that either involve adjectives, nouns, or adverbs.

**The Two Articles**

The two articles have a different flavor of the year, month, or day. The author is very focused on both topics.

The top 50 words very much are the summary of the story. Next is the value of each word frequency very much the same.

Base on two articles, the author changes his writing style base on the event and maintains the efficacy. Upon reading the top 50 words, the writer is has a few emotional peaks.

Overall, there aren't many differences between both articles' bigrams. However, there is a difference between corpora, denoting how Russell's writing style evolves from the one story to the next. In 2009 about the controversy on Blue Jay and Halladay. That Blue Jay walks away from the offer, and in the second article is the NBA norm game.

# **Conclusion**

Conducting analysis frequently requires skills and the wisdom to select effective methodologies. Without applying foundational principles for data preparation and associated analysis, could risk time management.

NLP is an exciting field for data scientists and engineers. Non-commercial packages and utilities allow coding challenges to be abstracted. Moreover, numerous off-shelf solutions require knowledge to select the right combination of frameworks. Often, the acumen in engineering a problem (or lack of), is noticeable within the associated results. Furthermore, in the area of text mining and NLP, the combination of tokenization and vectorization provide data scientists room for creativity. [More info](https://github.com/BevTucker/Data_Science_Python/blob/master/Corpus%20Statistics%20and%20Mutual%20Information.ipynb)

# **IST707** [**707 Data Analytics**](https://2su.datascience.syr.edu/ap/courses/985/sections/43ab3046-3e47-4db9-aa8e-427f120f3b1a/coursework) **Mystery of Federalist Clustering and Rediction**

# **Goal:** Instigate 11 disputed papers from our founding father. This study will use clustering techniques to determine distance measures, a sum of square errors, and corresponding visualizations. An attempt will be made to justify whether the disputed papers are more likely to be written by Alexander Hamilton or James Madison. Then use machine learning to predict 11 disputed papers from our founding father.

**Technologies:**

Caret., library(dplyr), library(tidyr),library(tidyverse),library(reshape2), library(reshape)

library(ggplot2),library(stats) ,library(flexclust),library(mclust),library(cluster)

library(stats), library(factoextra) # clustering algorithms & visualization

library(grDevices), library(scales), library(ggfortify),library(dendextend), library(knitr)

library(rpart), library(rpart.plot), library(caret)

# **Techniques:** data mining, descriptive analytics, grouping and aggregation, Affinity propagation,

**Model:** Clustering K-means, hclust, Euclidean, Dendogram, Train and predict

Introduction The ratification of the Constitution of the United States of America was led by three founding fathers, Alexander Hamilton, James Madison, and John Jay. Collectively, they contributed to the 85 papers, known to this day as The Federalist Papers. These papers provided the groundwork and the intent of the Constitution. However, each of the authors would often sign their work as the Latin ‘Publius’, rather than their names. It is known today that of the 85 articles, 74 have had identified with an author(s). The remaining 11 controversial papers have claimed to be written by either Hamilton or Madison. This controversy has led to various types of analysis. More recently, various types of text analysis approaches have exercised to help estimate the likelihood of one author over another.

The analysis provided in this study will attempt to use some clustering techniques to determine distance measures, a sum of square errors, and corresponding visualizations. An attempt will be made to justify whether the disputed papers are more likely to be written by Alexander Hamilton or James Madison.

Chart

Description automatically generated with medium confidence Chart, line chart

Description automatically generated

# Diagram Description automatically generated with medium confidence Chart, line chart Description automatically generated

# Chart, line chart Description automatically generated Chart, scatter chart Description automatically generated

**For this exploration there are collinear between writers**

Conclusions From this investigation, showed that the disputed authorship appears more closely related to Alexander Hamilton, then James Madison. When reviewing historical generalizations, Hamilton was most active between his time in the army, as well as congress. His experience, coupled with the perceived notion that states needed to be governed more closely, likely gave way to his stance on The Federalist Papers, and it needs to be ratified.

James Madison, on the other hand, was known to be very detail-oriented, and at times reclusive. It would be interesting to understand why Hamilton wrote so many more papers than Madison and understand the relationship between the two founding fathers. Specifically, why did Madison write more papers with Hamilton than himself writing his own papers independently? Similarly, why was there a dispute between authorship, rather than agreement on co-authorship?

Overall, this study has provided some good first steps for initial exploratory analysis. However, obtaining additional volumes, or papers authored by Madison could help further understand and differentiate from the Federalist Papers written by Hamilton. Similarly, since the study was to differentiate between the two authors, John Jay, could potentially be removed during the study and analysis. [More info](https://github.com/BevTucker/Rcode/blob/main/IST707%20707%20DATA%20ANALYTICS%20MYSTERY%20OF%20FEDERALIST%20CLUSTERING.Rmd)

**Prediction**

Diagram

Description automatically generated Table

Description automatically generated

Result base on the data presented, Alexander Hamilton was the author of the disputed papers. The interesting part is Jay showed that he wrote a disputed paper also. However this is an error since the disputed papers belong to James Madison and Alexander Hamilton.

Graphical user interface, application

Description automatically generated

Graphical user interface, diagram, text, application, chat or text message

Description automatically generated Text

Description automatically generated with low confidence

To tune the parameter to reduce tree levels to one allowing the outlier to be removed, this will allow the model to distinguish between James Madison and Alexander Hamilton.

Cross validation

Text, table

Description automatically generated with medium confidence

## Result on Disputed Papers

Graphical user interface, application, Word

Description automatically generated

**Conclusion**

From this analysis, it has been found that the disputed authorship appears more closely related to James Madison, then Alexander Hamilton. These results are quite contrary to the previous finding. However, as indicated in this study, implementing proper train and test sets would have provided additional error measures for the corresponding decision tree models. But, due to the relatively small dataset size, further splitting the provided dataset into a train and test set, would have reduced the already small dataset for training a model. This could likely increase the chances of overfitting.

As mentioned in a previous study, obtaining additional works produced by James Madison, to complement the initial dataset, would have produced a richer context to work. The same case for obtaining more writing samples for Alexander Hamilton would be valid. The best steps forward for both studies could likely be an integration of more data, as well as an adjustment to each preprocessing step.

Overall, it seems the results produced by this study are more likely to better predict than the clustering study. This is due to the fact of the incorrect assumption to retain the author’s attribute during clustering. Nevertheless, both approaches have concluded the same notion. Techniques implemented for prediction, are often limited by the size of a dataset. As shown in the result of this study, the decision trees were constructed using at most two words. Perhaps, only measuring word frequency, followed by a normalization technique, is not enough by itself. Measuring different arrangements of words, with corresponding distance measures relative to one another, could potentially be an improvement to these studies. Though it has been said many times, the simplest solution is often found best. [More info](https://github.com/BevTucker/Rcode/blob/main/Beverlyn-Tucker-IST707%20Mystery%20of%2011%20disputed%20Federalist.html)

**FIN654 Financial Analytics Portfolio CHEESECAKE and tesla**

**Goal:** To provide best portfolio to the clients

**Technologies:** quantmode, scatterD3, formattable

**Techniques:** descriptive analytics, grouping and aggregation, plotting and illustration

In today's generations, offering a variety of portfolio packages to investors is essential. Help them understand their risk tolerance.

Understand the Max Return/Risk Ratio Portfolio, Minimum Return/Risk Ratio Portfolio. All portfolios are excellent, but they will defend investor's risk tolerance.

​Chart, scatter chart

Description automatically generated![Graphical user interface, text, application

Description automatically generated]()

Above are all best portfolio clients will choose portfolio according to their risk tolerance.



![Text, application, chat or text message

Description automatically generated]()

* Overall, a well-diversified portfolio is your best bet for the consistent long-term growth of your investments.
* Determine the appropriate asset allocation for your investment goals and risk tolerance.
* Pick the right portfolio.
* Monitor the diversification of your portfolio, checking to see how weightings have changed.

Conservative vs. Aggressive Investors

The more risk you can bear, the more aggressive your portfolio will be, devoting a more significant portion to equities and less to bonds and other fixed-income securities. Conversely, the less risk you can assume, the more conservative your portfolio will be. Here are two examples, one for a conservative investor and one for a moderately aggressive investor. [More info](https://github.com/BevTucker/Rcode/blob/main/TSLA%20and%20CAKE%20Portfolio.R)

References

Tucker, B. (a). Beverlyn-tucker-IST707 mystery of 11 disputed federalist.html. Retrieved from <https://github.com/BevTucker/Rcode/blob/main/Beverlyn-Tucker-IST707%20Mystery%20of%2011%20disputed%20Federalist.html>

Tucker, B. (b). *Fin654 financial analytics portfolio cheesecake and tesla.* Retrieved from <https://github.com/BevTucker/Rcode/blob/main/TSLA%20and%20CAKE%20Portfolio.R>

Tucker, B. (c). *Ist652\_scripting project ecommerce women’s clothing.* Retrieved from [https://github.com/BevTucker/Data\_Science\_Python/blob/master/Women's\_Clothing\_Review.ipynb](https://github.com/BevTucker/Data_Science_Python/blob/master/Women)

Tucker, B. (d). Ist664\_nlp corpus statistics and mutual information. Retrieved from <https://github.com/BevTucker/Data_Science_Python/blob/master/Corpus%20Statistics%20and%20Mutual%20Information.ipynb>

Tucker, B. (e). Ist707 data analytics mystery of federalist clustering. Retrieved from <https://github.com/BevTucker/Rcode/blob/main/IST707%20707%20DATA%20ANALYTICS%20MYSTERY%20OF%20FEDERALIST%20CLUSTERING.Rmd>

Tucker, B. (f). Ist718 big data analytics big data problem mnist dataset. Retrieved from <https://github.com/BevTucker/Data_Science_Python/blob/master/IST718_Lab3.ipynb>