Comparable to the Incomparable? Exploratory Text Analysis of Charles Dickens and Mark Twain

INTRODUCTION: Charles Dickens and Samuel Clemens, better known as Mark Twain, are two of the most formidable literary giants of the nineteenth century. While Dickens is considered by many as "the greatest writer of the Victorian era",[1] Twain is "one of America's best and most beloved authors."[2] Though Charles Dickens was born in England in 1812 and died in 1870, only shortly after the American Mark Twain (1835-1910) published one of his first works, there is evidence that Twain read and grudgingly (at the very least) admired Dickens,[3] and that, in fact, the two authors were quite similar. From their writing styles and content – the use of humor and satire to explore social justice and institutional failures, and the wide range of mediums (novels, short stories, speeches, essays, travelogs) they explored – to the surprising number of coincidences in their life stories,[4] these two authors appear intertwined in more ways than one. At the very least, they both enjoy acclaim even today, and even from those relatively unfamiliar with their work.

Therefore, an investigation of the commonalities between Dickens and Twain seems warranted, and indeed is not unprecedented.[5,6] For those uninitiated into the many works of these two prolific authors, this would be a daunting task. However, text analytics can help quickly and fairly adeptly pinpoint some of the parallels between Dickens and Twain. By creating three corpuses and vocabularies – one for each author and one for both authors – and employing foundational text analytics strategies, including language models; vector space models; similarity and clustering methods, Principal Component Analysis (PCA); topic modeling; word embeddings; and sentiment analysis, one can readily gain an understanding of and appreciation for the depth, breadth, and influence of Dickens and Twain, individually and collectively.

RESULTS: Relevant findings from 50 works of Dickens and 45 of Twain curated from Project Gutenberg[7,8] and parsed using an ordered hierarchy of content objects (OHCO) of book, chapter, paragraph, sentence, and token are summarized below.

Language Models: When considering trigrams that do not contain either beginning or end of sentences nor stop words, the several most frequent ones in Dickens's works tend to follow the general pattern "said mr *male surname*", and all such surnames appear in different novels of Dickens: "Pickwick" from *The Pickwick Papers*, "Pecksniff" from *Martin Chuzzlewit*, "Boffin" from *Our Mutual Friend*, and "Dombey" from *Dombey and Sons*.[9] In contrast, many of the most frequent trigrams in Twain's works relate to time, for example, "hundred years ago", "wilson's new calendar", "twenty four hours", and "fifty years ago".[10] However, even the top trigrams in the Twain corpus appear in far reduced numbers compared to Dickens, with about an eight-fold difference between the two authors. This disparity is reflected when examining the combined corpus, as the terms "said" and "mr" are two of the most commonly used terms, highlighting differences in the number of works of each type (novel, story, or nonfiction) between the authors, with Dickens writing more stories and novels while Twain wrote more nonfiction (based on the contents of the current corpus).[11]

Vector Space Models: Using the max method to compute term frequency (TF) for each term, i.e., $TF_{t,d} = \frac{(tf_{t,d})}{tf_{max}(d)}$,[12] and the standard method to compute Inverse Document Frequency (IDF), i.e., $IDF = log_2(\frac{N_{docs}}{Document\ Frequency})$[12] to calculate term frequency – inverse document frequency (TFIDF), the ten most nouns (excluding proper nouns) based on document frequency – inverse document frequency (DFIDF) for Dickens, Twain, and the combined corpus are as follows.[13,14,15]

Corpus	**Significant Common Nouns Based on DFIDF**									
Dickens	sleep	windows	top	thank	dress	confidence	breast	notice	duty	bless
Twain	indeed	door	deal	children	money	ones	miles	family	everybody	ground
Combined	sound	minutes	change	purpose	character	father	foot	attention	mother	arm

Table 1: Top 10 Most Significant Common Nouns Based on DFIDF for Each Corpus

These nouns may relate to topics common across works by both authors, such as society in the domestic and natural sense; learning; emotions and relationships, with a greater emphasis for Dickens on civility and propriety, versus familial and amical interactions and location for Twain.

Similarity and Clustering: Various clustering methods uncovered more points of divergence within and across corpuses. Multiple distance metrics and linkage methods were explored for hierarchical agglomerative clustering (HCA), but for the sake of brevity, only cityblock, Jaccard, and Jensen-Shannon with weighted linkage and cosine and Euclidean with ward linkage will be discussed here. When clustering Dickens' works separately, cityblock, cosine, and Euclidean differentiated among novels, speeches or works of nonfiction, and stories fairly well, as well as travelogs written in distinct locations (**Figure 1**).[16] For Twain, cityblock picked up on vernacular, since many of Twain's novels and short stories incorporate Southern dialects, whereas his more formal essays and speeches sometimes draw on languages such as German or French. Cosine and Euclidean also fell prey to these more superficial differences, although they did also seem to reflect differences in location presented in travel narratives (**Figure 2**).[17] All methods with HCA with the combined corpus resulted in clusters with a mix of works from both authors, yet were largely able to separate works based on type, indicating that defining characteristics of novels, short stories, and nonfiction were more prominent than differences between the authors themselves.[18]

K-means clustering offered control over the number of clusters to create when attempting to find works associated with one another based on similarities in significant term TFIDF values. The best combination of the hyperparameters number of clusters and TFIDF matrix normalization (raw values, binary values, or L1 or L2 normalized) based on the maximum silhouette score was found, yet none of these clustering methods proved to be particularly informative, especially for the corpuses separated by author. The combined corpus, though, produced some surprising results. Two clusters using the binarized feature vector maximized the silhouette score at 0.359987 (which is still rather low), yet the two clusters did not reflect genre (fiction versus nonfiction) or author. More works of every type appeared in cluster 0 than cluster 1, but almost all novels (including all of Dickens's) and works of nonfiction appeared in cluster 0, whereas there was a more even split in the distribution of short stories between clusters.[19]

Principal Component Analysis (PCA): Some of the findings from PCA mirrored those from K-means clustering. A manual approach to PCA (as opposed to the implementation by the package Prince) using the 1000 most significant terms based on DFIDF (excluding proper nouns) with ten principal components revealed some separation along Principal Component (PC) 0 between Dickens and Twain, but there was little separation along the axis of any other PC between the authors (**Figure 3**). Likewise, there was little separation between works of distinct types, although there was more variation along PC 0 and PC 1 for novels compared to short stories and nonfiction. Attempting to gloss PC 0, with the terms "river", "miles", "land", "war", and "city" associated associated with the positive end of the spectrum, where most Twain works cluster, and "sir", "replied", "dear", "gentleman", and "cried" associated with the negative end of the spectrum, where most Dickens works cluster, this PC seems to capture the disparate subject matter, settings, and styles of the two authors. Twain's novels and stories in particular adopt a less formal air than Dickens, and often focus on adventures that take place in the natural world, whereas Dickens writes more about societal and domestic affairs. PC 1 also reflects some of these disparities, although it emphasizes more changes in vernacular, education, and class among characters in novels and stories, since the terms "says", "ain't", "didn't", "couldn't", and "that's" oppose terms "city", "war", "church", "sea", and "love" (**Figure 4**).[20]

With PCA using six principal components and the Python package Prince for all of the terms in the combined vocabulary, at first outliers skewed the results. These outliers related to the frequent use of Southern or Cockney dialects and other languages in certain works, so all of the outliers at the upper end of the spectrum for both authors and proper nouns from the combined corpus and full vocabulary, reducing them by ~7% and ~19%, respectively. When running PCA again with the reduced corpus and vocabulary, the major point of variation among the works could not be attributed to author or work type, since neither categorization aligned with PC 0 (although works by the two authors did separate along PC 1) (**Figures 5, 6**).[21]

Topic Models: The Latent Dirichlet Allocation (LDA) topic model method, implemented with the Python package scikit-learn and using a predefined number of topics (40) and words (2000) helped elucidate some of the major themes favored both by individual authors and shared by them. The table below provides five

of the top terms associated with the four most frequent topics (in order of descending frequency) from each corpus, works associated with that topic, and a gloss for each topic.[22,23,24]

Corpus	Example Works	Top Terms	Gloss
Dickens	Master Humphrey's Clock, Speeches of Charles Dickens	institution, audience, association, education, class, social, distinguished, reader, highest	Propriety, education, class, status, hierarchy
	The Holly Tree, Great Expectations	chap, bundle, lot, leg, thankee, churchyard, pipe, fur, nod	Lower class daily rhythms (work, religion, speech)
	A Tale of Two Cities, Hard Times	sorrow, shadow, shadows, hurriedly, clasped, compassion, pain, trembled, crept	Somber and eerie setting, moral dilemmas
	Pictures from Italy, American Notes, Reprinted Pieces	buildings, houses, gardens, trees, built, grim, painted, ancient, season	Traveling and the associated architecture and weather
Twain	What is Man, Following the Equator, Mark Twain Speeches	science, reverence, civilization, government, nations, political, religious, british, instance	Society, structure, (geo)politics
	The Adventures of Tom Sawyer, A Connecticut Yankee in King Arthur's Court	awake, stillness, cheer, whispered, sprang, wound, crept, hurried, resolved	Adventure, action, secrecy
	The American Claimant, Alonzo Fitz and Other Stories	earl, sack, song, loving, glanced, confess, gratitude, foolish, sigh	Court intrigue, comedy
	A Tramp Abroad, Life on the Mississippi, Roughing It	glacier, steep, summit, rope, scenery, ice, gloom, mountain, huge	Travel, scenery, outdoors, ruggedness
Combined	What is Man, The Gilded Age, The American Claimant	detail, color, doesn't, details, recognized, honor, rule, nation, afterward	Royalty, society
	The Mystery of Edwin Drood, David Copperfield, Hard Times	guardian, cousin, assure, sister, confidence, pursued, dearest, madam, agreeable	Familial relationships and protection
	Mark Twain Speeches, The Letters of Mark Twain, The Poems and Verses of Charles Dickens	lecture, 3, wrote, literary, 2, author, letters, machine, print	Formal styles of writing and speaking that directly reference writing speeches, letters, etc.
	American Notes, Speeches of Charles Dickens, The Treaty with China and Its Provisions Explained	institution, science, political, class, national, social, association, education, legal	(Inter)national social institutions (e.g., education, justice)

Table 2: Works, Top Terms, and Topic Glosses by Topic and Corpus

Though the exact terms and possible glosses for each of the corpuses differ, they all focus on societal institutions and structures, such as education, the justice system, and classes; familial relationships; and travel and sights observed on adventures and journeys (architecture, nature, etc.).

Word Embeddings: Examining word embeddings visualized with t-distributed Stochastic Neighbor Embedding (tSNE) plots provided more evidence for some of the conclusions drawn with topic models. Clusters of words found in similar contexts in the combined corpus included one that could be briefly described as "travel and journeys" (terms: "luggage", "country", "town", "borough", "inn", "tent"), while another related to nature and the outdoors ("woods", "ice", "gravel", "dust", "deserts", "valleys", "plants").[25] In the Dickens corpus, certain clusters related to systemic social hierarchies (the proximity of "ease", "liberty", "credit", "comfort", "reign" potentially linking the ideas of comfort to financial stability and a higher station in society; "nurse", "servant", "housekeeper", "lad", "physician" to domestic occupations and potentially gender roles),[26] patterns echoed in the Twain corpus ("experts", "residents", "statesmen", "mormons", "historians", "establishments" relating to government and religion; "care", "excuse", "play", "blow", "cheer" relating to antics, mischievousness, adventure).[27]

Similarities between words based on shared contexts also illustrated some of these themes. Looking at the word "poor", for instance, returned words such as "miserable", "wretched", "wicked", "sick", "foolish", and

"friendless" for the combined corpus, yet also "brave" and "darling" for all three corpuses, and "gentle" and "innocent" in the Twain corpus.[28,29,30] While this finding points to the undesirable condition of the poor in society, it also underscores certain laudable characteristics born out of poverty and adversity that can help those in dire straits overcome class barriers and other obstacles. On the other hand, "rich" is associated with many positive words, such as "healthy", "clever", "picturesque", and "pure" for the combined and Twain corpuses,[28,30] whereas the most similar words in the Dickens corpus are "shabby", "rare", "hungry", "lazy", showing that Dickens placed a greater emphasis on the differences between the rich and poor, along with the harmful attributes of wealth and the ways in which the rich can use their station to disadvantage the less fortunate.[26] In the same vein, "money" is closely most closely associated with "trouble" across all corpuses, yet also "profit" and "stability", drawing a stark picture of the double-edged sword that money represents – both a means to an end and survival and the potential ruin of those with and without it.[28,29,30]

Sentiment Analysis: The majority of works by both authors had positive polarity, with only 17 of 95 works, or approximately 18%, or with neutral (i.e., 0) or negative polarity. Just over half of those works fall into the short stories category, and over two-thirds of them were written by Twain. When looking at overall sentiment values for eight core emotions – joy, trust, anticipation, sadness, fear, anger, surprise, and disgust – for each of the works, as one would expect, works with lower values for anger, sadness, disgust, and fear and higher values for anticipation and joy have higher polarity, but overall polarity values show less clear correlations with surprise and trust.[31]

Some novels by Dickens and Twain employ young male adolescents as their protagonists, specifically *Great Expectation*, *Oliver Twist*, *The Old Curiosity Shop*, *David Copperfield*, and *Dombey and Sons* by Dickens and *The Adventures of Tom Sawyer* and *The Adventures of Huckleberry Finn* by Twain,[5] and though the circumstances of these characters differ, an exploration of the sentiment shifts based on sentence-level analysis using the Python package VADER revealed that certain plots follow similar patterns. *Oliver Twist* and *The Adventures of Tom Sawyer* exhibit a characteristic two peaks in sentiment (indicating an upward trend) towards the latter halves of the novels (**Figure 7**), and the emotional progression of *David Copperfield* resembles that of both the novels analyzed by Twain (**Figure 8, 9**), although Twain's works as a whole generally report lower sentiment values than those of Dickens.[31]

In much the same way, both authors wrote satirical works that exposed stark realities in their own home countries, Britain for Dickens and America for Twain, and that of the other author.[32] Dickens wrote the novel *Martin Chuzzlewit* in part in reaction to his visit to the United States in the 1840s and his disgust with societal wrongs, such as slavery.[33,34] *The American Claimant* by Twain is a satirical novel that examines American and British governments and their idiosyncrasies and shortcomings, such as the roles of journalism, capitalism, and industrialization.[35] Though once again the overall sentiment values of Twain's work are lower than that of Dickens, these two novels follow similar sentiment trajectories: they begin on relatively good terms, then quite quickly endure a series of rises and falls, after which a final precipitous decline precedes a potential return to better times at the end (**Figure 10**). As such, these works that combine humor and moral truths may cause readers to both consider the ills in society, yet also have hope that such flaws can be remedied through reform.

CONCLUSION: Despite the differences in the backgrounds and lives of Dickens and Twain, from their nationalities to the time periods in which they lived, analysis of their works of fiction and nonfiction provide insights into their shared use of humor to address topics such as the shortcomings of societal institutions, prominently disparities between classes, and familial and amical relationships, with humor and cogent observations about their surroundings and political events. Their worldview that defects in society – most notably poverty and imperialism – must be recognized and fixed; the importance of companionship; and the benefits of travel and a desire to experience the world are notable even in this cursory exploration of their diverse sets of works. Future analysis of the narrative archetypes commonly employed by each author, as well as the influence of years of publication and their coinciding with the development of Britain, America, and the world at large, would augment the conclusions presented in this paper. However, for the person aware of the prowess of Dickens and Twain, yet largely naïve of their diverse works, the analysis in its current state offers a helpful starting point for understanding these literary giants.

FIGURES

Figure 1: Representative Hierarchical Agglomerative Clustering Plot for Dickens – Euclidean distance, L2 normalization, ward linkage

Figure 2: Representative Hierarchical Agglomerative Clustering Plot for Twain – cityblock distance, raw, weighted linkage

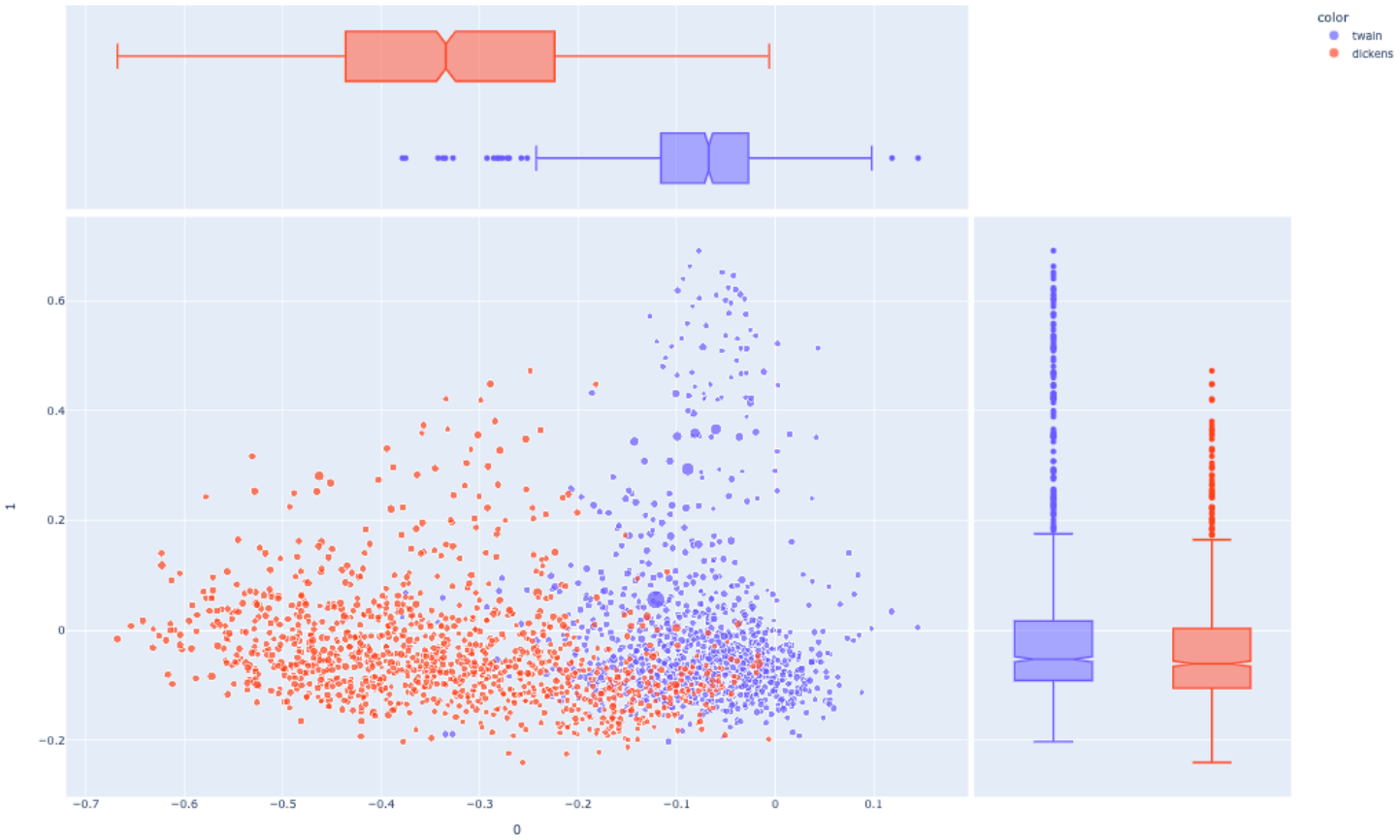

Figure 3: Manual PCA for Combined Corpus with 1000 Most Significant Terms Based on DFIDF (no proper nouns) by author – PC 0 vs. 1

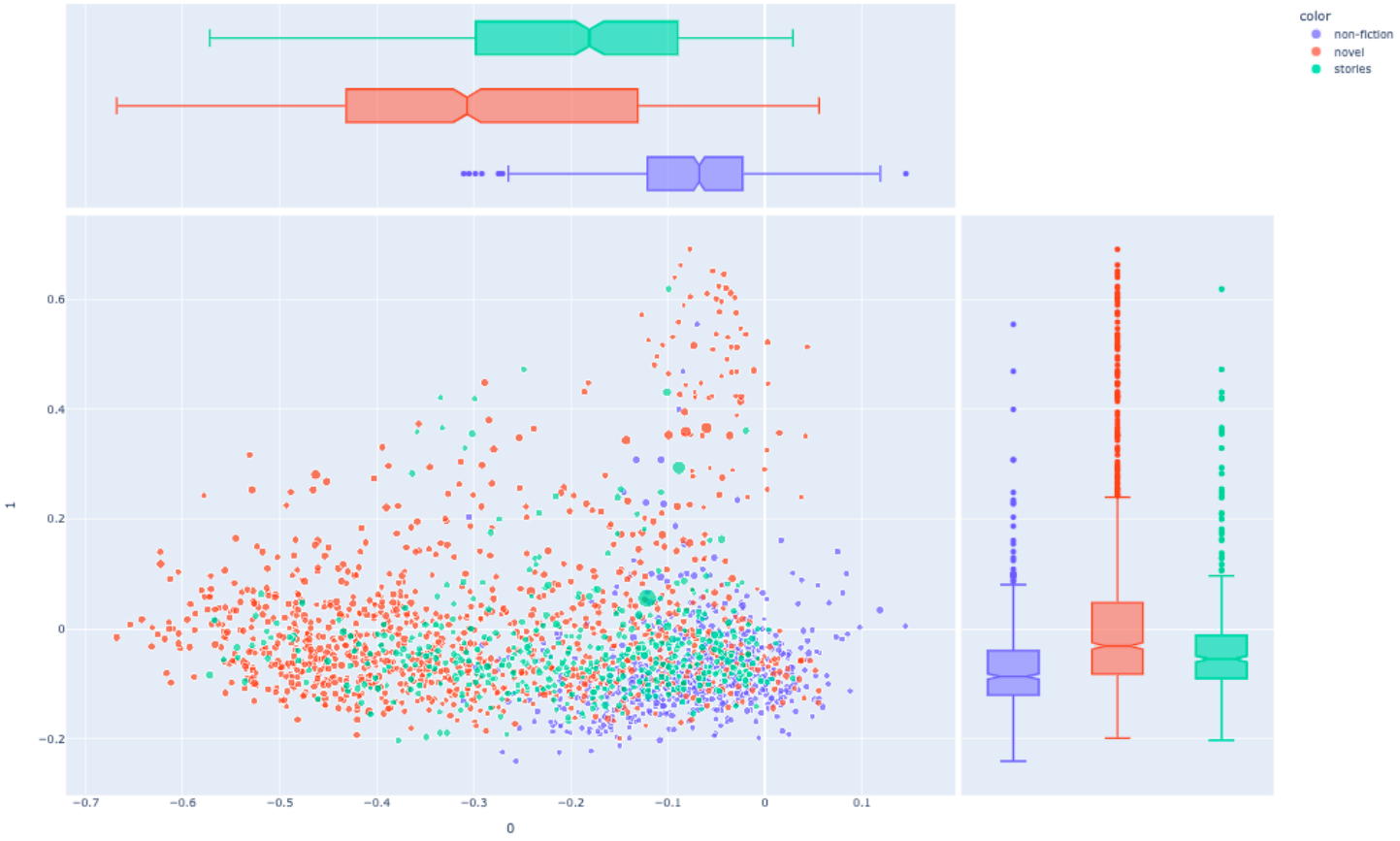

Figure 4: PCA for Combined Corpus with 1000 Most Significant Terms Based on DFIDF (no proper nouns) by work type – PC 0 vs. 1

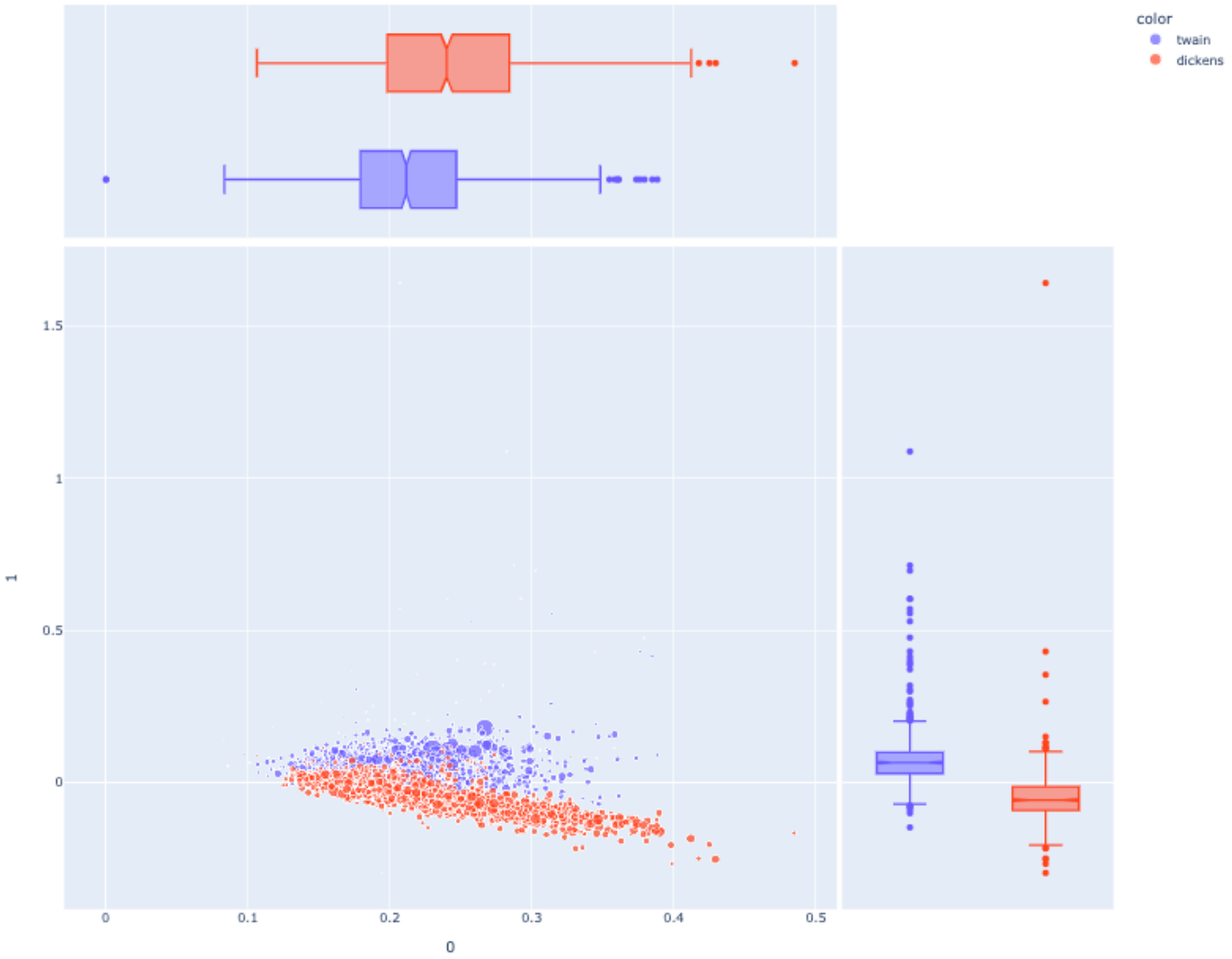

Figure 5: Prince PCA for Combined Corpus with Outliers Removed from Combined Corpus and Vocab by author – PC 0 vs. 1

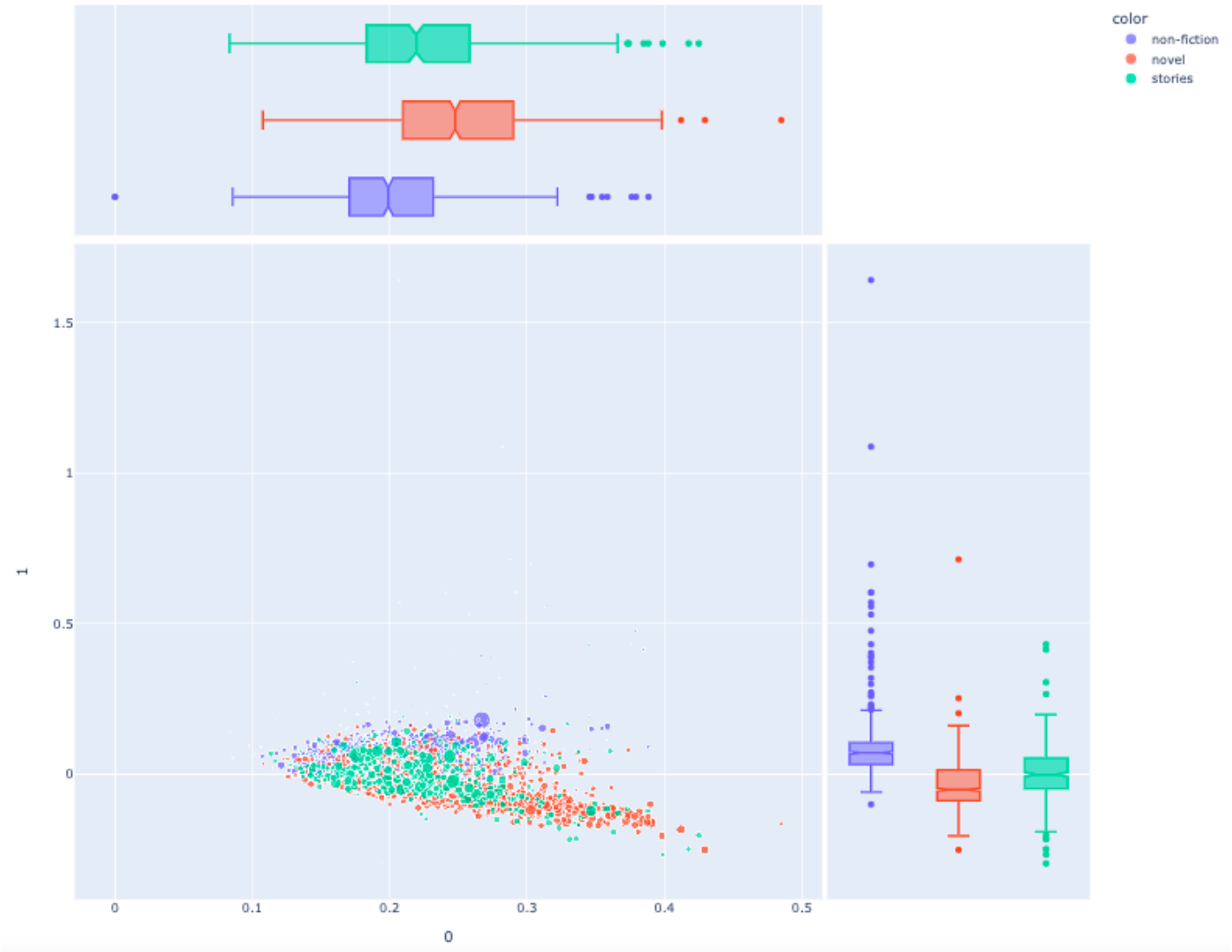

Figure 6: Prince PCA for Combined Corpus with Outliers Removed from Combined Corpus and Vocab by work type – PC 0 vs. 1

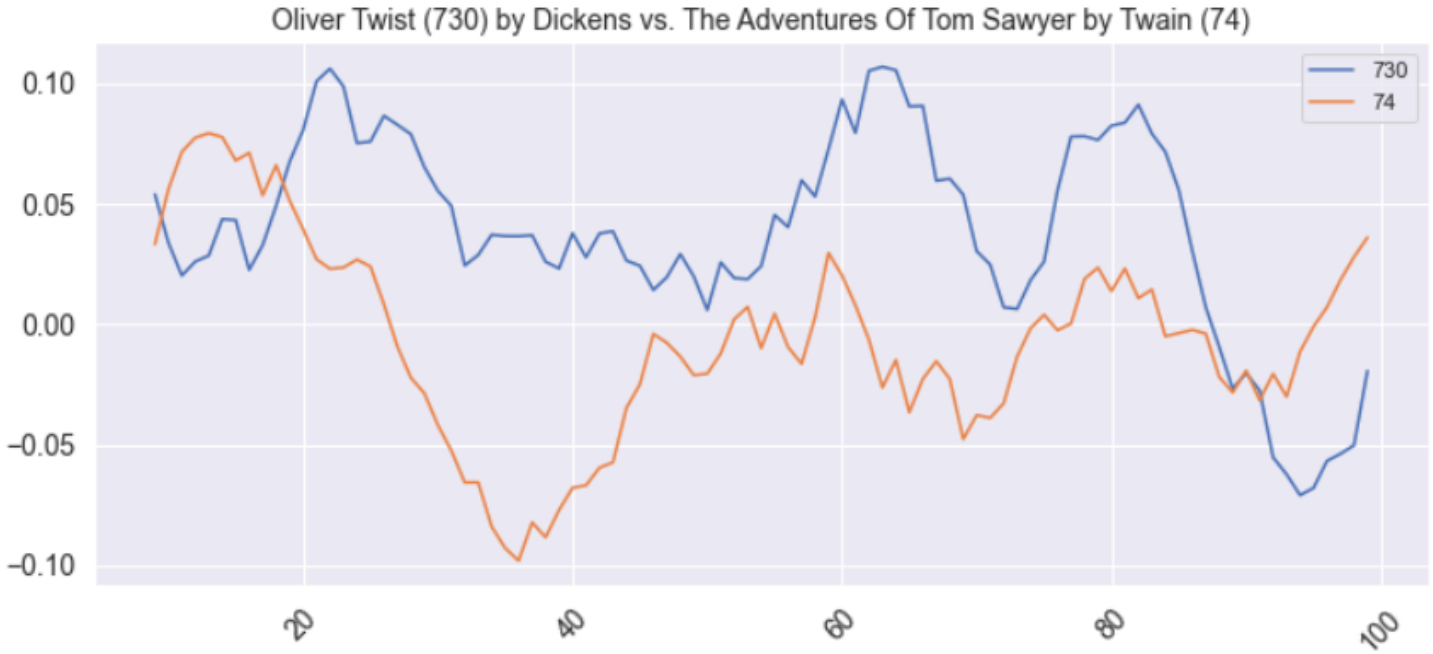

Figure 7: Sentiment Analysis at the Sentence Level with VADER – Polarity for Oliver Twist vs. The Adventures of Tom Sawyer

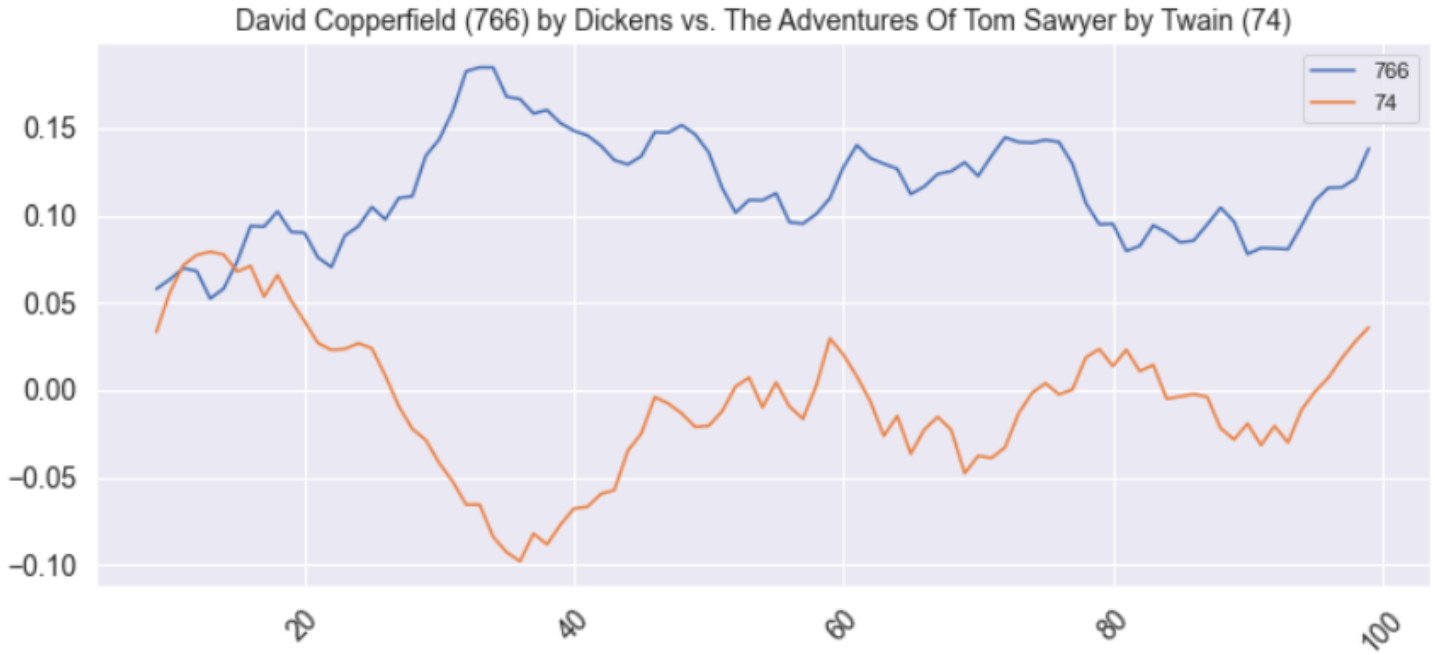

Figure 8: Sentiment Analysis at the Sentence Level with VADER – Polarity for David Copperfield vs. The Adventures of Huckleberry Finn

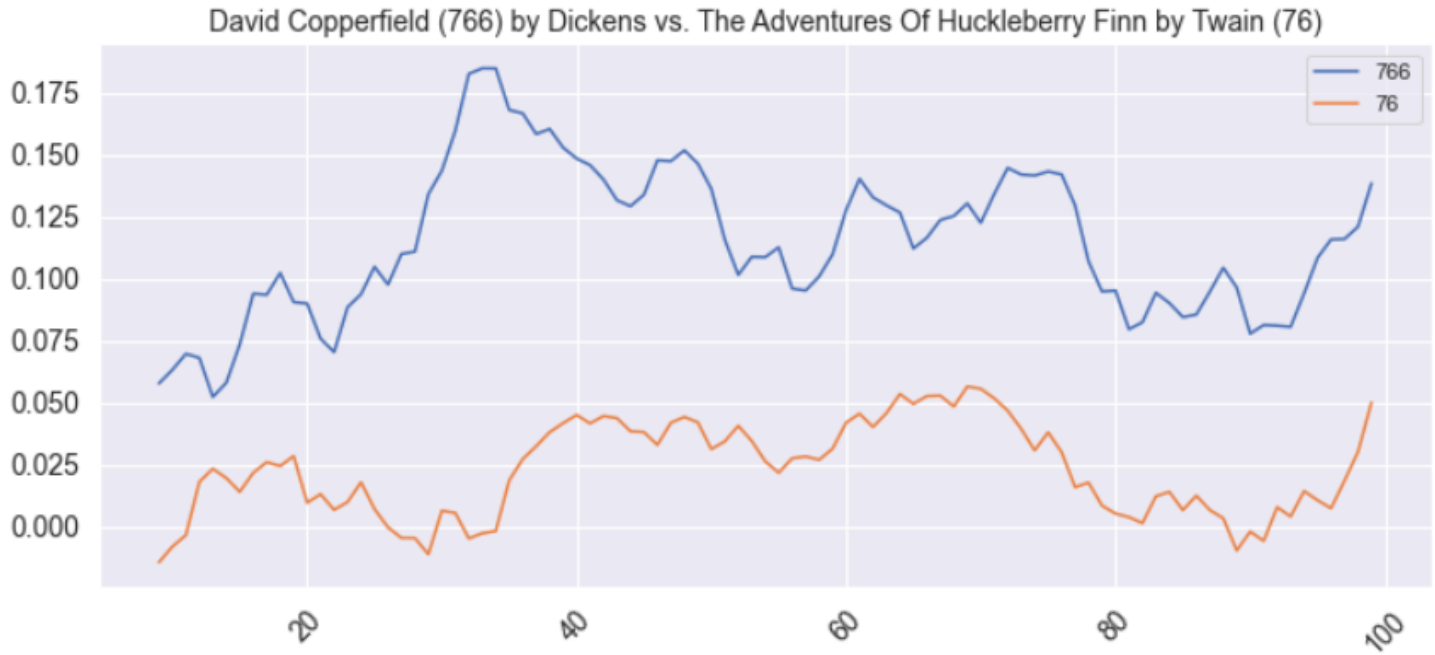

Figure 9: Sentiment Analysis at the Sentence Level with VADER – Polarity for David Copperfield vs. The Adventures of Tom Sawyer

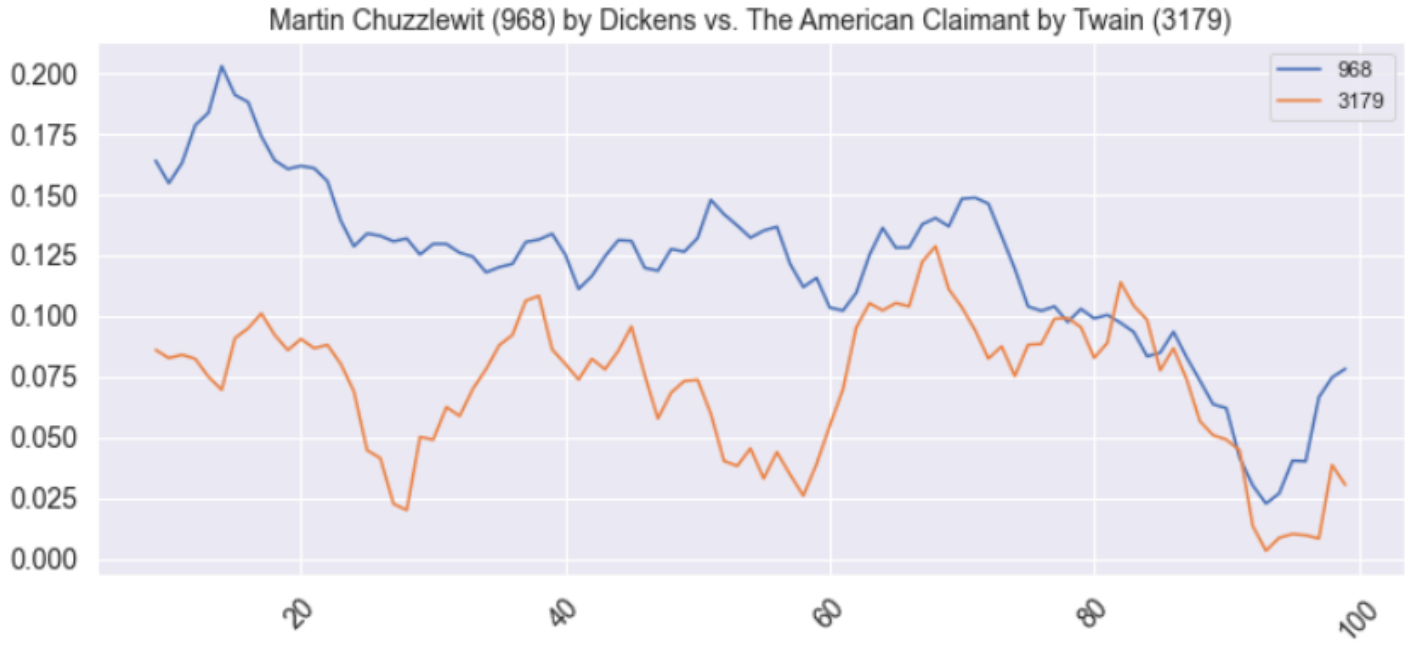

Figure 10: Sentiment Analysis at the Sentence Level with VADER – Polarity for Martin Chuzzlewit vs. The American Claimant

SOURCES

1. Collins, Philip. "Charles Dickens." *Encyclopædia Britannica*, Encyclopædia Britannica, Inc., 3 Feb. 2022, https://www.britannica.com/biography/Charles-Dickens-British-novelist. Accessed 28 April 2022.

2. Quirk, Thomas V. "Mark Twain" *Encyclopædia Britannica*, Encyclopædia Britannica, Inc., 17 Apr. 2022, https://www.britannica.com/biography/Mark-Twain. Accessed 28 April 2022.

3. Gardner, Joseph H. "Mark Twain and Dickens." *Publications of the Modern Language Association*, vol. 84, no. 1, pp. 90-101, 1969, https://www.jstor.org/stable/1261160?saml_data=eyJzYW1sVG9rZW4iOiIzYzFjMTc3Ny0xYjk4LTRhNTUtOGIyNC1mOGRlNmY4ZjI5MTkiLCJlbWFpbCI6ImNldzRwZkB2aXJnaW5pYS5lZHUiLCJpbnN0aXR1dGlvbklkcyI6WyJmOGFmN2FjOS01MWVjLTQ0YzgtODFhOS0wNzlhNzQzMjgzOTMiXX0&seq=1. Accessed 28 April 2022.

4. Dawidziak, Mark. "Charles Dickens and Mark Twain: Separated at Birth?" *WordPress.com*, On the road with Gerald Dickens, https://geralddickens.wordpress.com/2021/03/12/guest-blog-mark-twain-and-charles-dickens/. Accessed 28 April 2022.

5. Yuan, Siyu. "A Comparative Analysis of Charles Dickens and Mark Twain." *Academic Journal of Humanities and Social Sciences*, vol. 3, no. 7, 2016, https://francis-press.com/uploads/papers/JWqvK6RY0I3SK3KiHrC5SW6YIg0ns2yXWDDmHGqg.pdf. Accessed 28 April 2022.

6. Blair, Walter. "The French Revolution and 'Huckleberry Finn.'" *Modern Philology*, vol. 55, no. 1, pp. 21-35, 1957, https://www.jstor.org/stable/pdf/435269.pdf?casa_token=xS_hf6CBXxEAAAAA:qwt0tRAzfBH9nxq4qLWziN

ndltHojhtwYWt9pWwndtwUTDAMo0MwDEbT3G3EjG24rgnJd4n9RhvIRWNNN_VdScR8MXTNHkamg6cNkU_TeOR3oUx4n_k. Accessed 28 April 2022.

7. Widger, David. "Index of the Project Gutenberg Works by Charles Dickens." *Project Gutenberg*, Project Gutenberg, 2018. https://www.gutenberg.org/ebooks/58157. Accessed 28 April 2022.

8. Widger, David. "The Works of Mark Twain." *Project Gutenberg*, Project Gutenberg, 2019. https://www.gutenberg.org/files/28803/28803-h/28803-h.htm. Accessed 28 April 2022.

9. dickens_analysis_M3-7.ipynb → section M03: Language Models, subsection: Trigram table

10. twain_analysis_M3-7.ipynb → section M03: Language Models, subsection: Trigram table

11. full_analysis_M3-7.ipynb → section M03: Language Models, subsection: Trigram table

12. "Maximum Tf Normalization." *Maximum TF Normalization*, Cambridge University Press, 4 July 2009, https://nlp.stanford.edu/IR-book/html/htmledition/maximum-tf-normalization-1.html. Accessed 28 April 2022.

13. dickens_analysis_M3-7.ipynb → section: M06: Similarity and Clustering, subsection: Top 20 nouns by DFIDF, sorted in descending order (including plural nouns but not proper nouns)

14. twain_analysis_M3-7.ipynb → section: M06: Similarity and Clustering, subsection: Top 20 nouns by DFIDF, sorted in descending order (including plural nouns but not proper nouns)

15. full_analysis_M3-7.ipynb → section: M06: Similarity and Clustering, subsection: Top 20 nouns by DFIDF, sorted in descending order (including plural nouns but not proper nouns)

16. dickens_analysis_M3-7.ipynb → section: M06: Similarity and Clustering, subsection: Hierarchical agglomerative cluster diagrams for the distance measures

17. twain_analysis_M3-7.ipynb → section: M06: Similarity and Clustering, subsection: Hierarchical agglomerative cluster diagrams for the distance measures

18. full_analysis_M3-7.ipynb → section: M06: Similarity and Clustering, subsection: Hierarchical agglomerative cluster diagrams for the distance measures

19. full_ananlysis_M3-7.ipynb → section: M06: Similarity and Clustering, subsection: K-Means

20. full_ananlysis_M3-7.ipynb → section: M07: Principal Component Analysis, subsection: Manual PCA Methods with Only 10000 Most Significant Terms (excluding proper nouns)

21. full_ananlysis_M3-7.ipynb → section: M07: Principal Component Analysis, subsection: Prince PCA with Outliers Removed

22. dickens_tmodel_wordem.ipynb → section: M08: Topic Models, subsection: Works and Top Terms Associated with Each Topic

23. twain_tmodel_wordem.ipynb → section: M08: Topic Models, subsection: Works and Top Terms Associated with Each Topic

24. full_tmodel_wordem.ipynb → section: M08: Topic Models, subsection: Works and Top Terms Associated with Each Topic

25. full_tmodel_wordem.ipynb → section: M09: Word Embeddings, subsection: Noun tSNE plot

26. dickens_tmodel_wordem.ipynb → section: M09: Word Embeddings, subsection: Noun tSNE plot

27. twain_tmodel_wordem.ipynb → section: M09: Word Embeddings, subsection: Noun tSNE plot

28. full_tmodel_wordem.ipynb → section: M09: Word Embeddings, subsection: Similarities

29. dickens_tmodel_wordem.ipynb → section: M09: Word Embeddings, subsection: Similarities

30. twain_tmodel_wordem.ipynb → section: M09: Word Embeddings, subsection: Similarities

31. sentiment_analysis.ipynb → section: Sentiment by Book

32. Gardner, Joseph. *Dickens in America: Twain, Howells, James, and Norris*. E-book, Routledge, 1988, https://books.google.com/books?id=2UdnDwAAQBAJ&pg=PT77&lpg=PT77&dq=twain%27s+most+similar+book+to+dickens&source=bl&ots=iesmzfcUWo&sig=ACfU3U3OtqVOreabkL4HHt1BE_Lt2tEvWA&hl=en&sa=X&ved=2ahUKEwjqlIqrvab3AhX4knIEHeyEDgsQ6AF6BAg2EAM#v=onepage&q=twain's%20most%20similar%20book%20to%20dickens&f=false. Accessed 28 April 2022.

33. Chesterton, G.K. *Martin Chuzzlewit - Introduction by G.K. Chesterton*, American Literature, https://americanliterature.com/author/charles-dickens/book/martin-chuzzlewit/introduction-by-gk-chesterton. Accessed 28 April 2022.

34. Burke, Jackson. "History of Scams: Nothing New under the Sun." *CNBC*, CNBC, 17 Feb. 2015, https://www.cnbc.com/2015/02/17/scams-hacking-spanish-prisoner.html. Accessed 28 April 2022.

35. Messent, Peter. "The American Claimant Review." *Goodreads*, Goodreads, https://www.goodreads.com/book/show/2010710.The_American_Claimant. Accessed 28 April 2022.