INF2209

Fall 2022

Final Project Submission

Topic Modelling for the Herman Melville Corpus

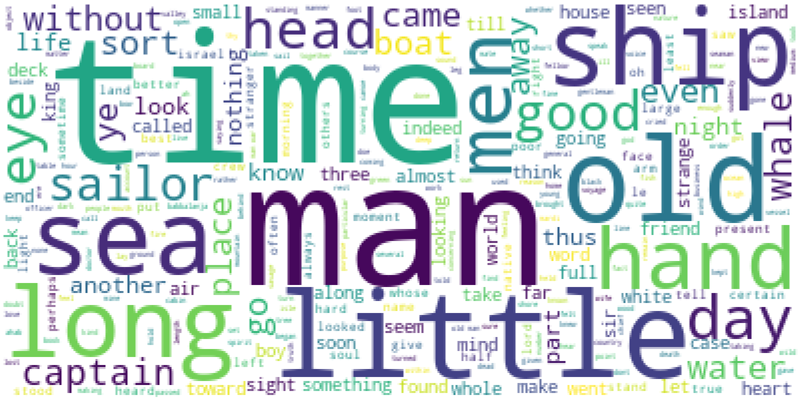
Paul King

996796506

Introduction

Herman Melville was a writer of novels, short fiction and poetry from the 19th century. He is widely considered one of the greatest American writers, and discussion of Melville's writing and biography is an ongoing scholarly pursuit, most notably by the Melville Society in their journal, *Leviathan*. As Melville is long-deceased, all of his published fiction exists in the public domain. Melville is known for his large vocabulary and influence on later writers. While a surface-level reading of Melville shows a writer who wrote about subjects he was familiar with such as life in New England or as a sailor, his books are rich with allusion and subtext, his experiences coloured through the lens of spirituality and symbolism (Maxwell, 2018). As such, his corpus is ideal for topic modelling, which can potentially reveal latent themes. Researchers have previously applied topic modelling to Melville; Du et al. compared the use of Latent Dirichlet Allocation with a novel topic model algorithm for *Moby-Dick* (2012). Using Correlation Explanation (Gallagher et al, 2021), I extracted 11 topics, broadly dividing documents into 4 categories of writing; natural, nautical, abstract and "life" (representing descriptions of day-to-day life of groups of people).

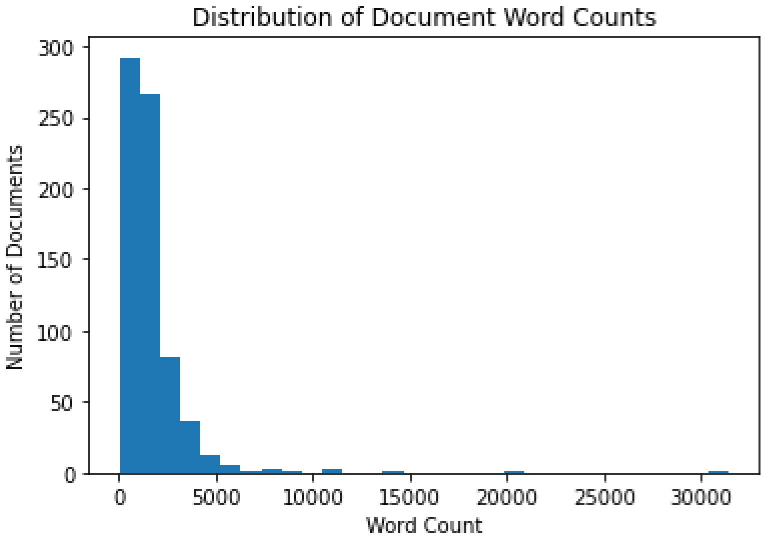
For analysis, I used a corpus consisting of 12 of his texts retrieved from Project Gutenberg, consisting of 10 novels and 2 collections of short stories. Originally, these texts were further split into chapters (for novels) or individual stories, which became the 704 documents for this Melville corpus. Later, I moved to using paragraphs as a unit of document analysis, which produced 8776 documents when controlling for word count. Melville's collection of poetry is considerably smaller than his fiction work, so it was omitted from the corpus. The source texts from Project Gutenberg were idiosyncratically formatted and required some text editing before preprocessing with Python; this included manually removing header and footer text and marking document boundaries for short story collections.



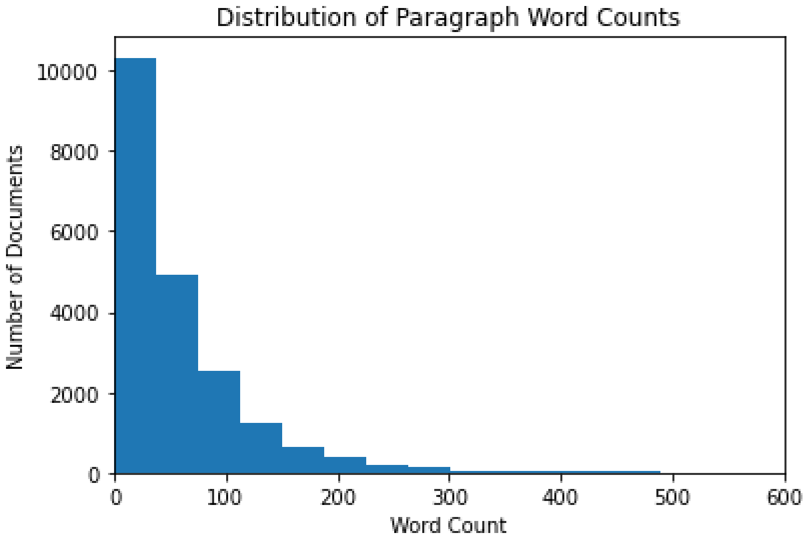
This is a wordcloud of the most common words in the Melville corpus. Many terms are nautical in nature, including "sea", "sailor", and "ship"; this makes sense, given that many of Melville's stories draw on his time at sea, in particular *Moby-Dick*, the largest book in the Melville Corpus. Other prominent words are harder to explain, such as “time”.

Book	**Word Count**
Moby-Dick	208458
Pierre	151615
White Jacket	138610
Redburn	117915
Omoo	101139
Mardi vol. 2	100192
Typee	98019
Mardi vol. 1	96020
Confidence-Man	92840
Piazza Tales	79225
Israel Potter	64481
Apple-Tree Table	60004

Analyzing our Document Word Counts gives us a mean document length of 1642 words and a median of 1223 words. We notice there are several unusually long documents, which are short stories (they are *“I and my Chimney”, “Bartleby the Scrivener”, “Benito Cereno”* and *“The Encantadas”*).



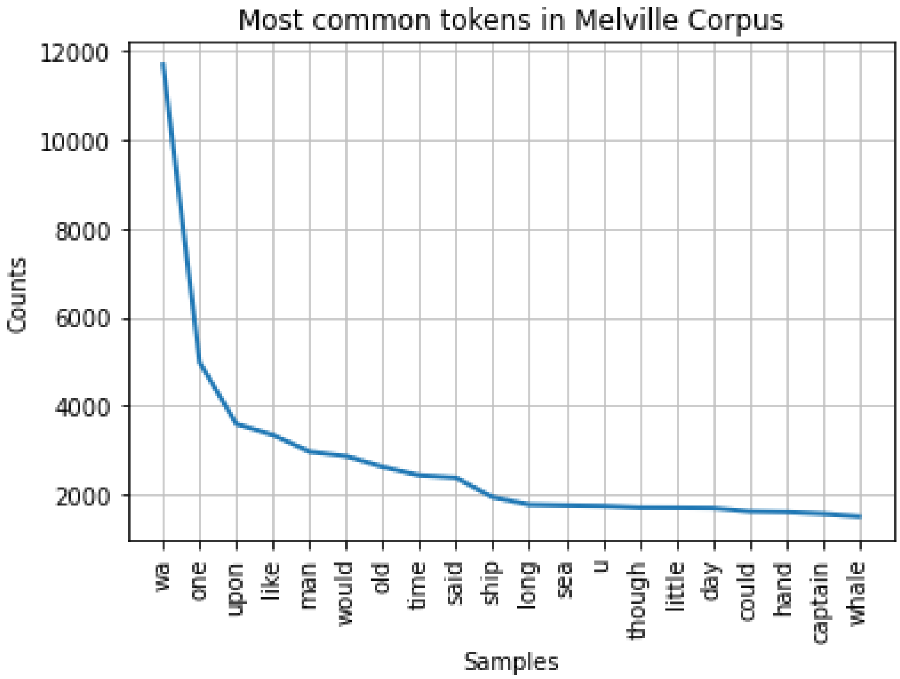
Breaking down into paragraph units, we find a mean wordcount of 56.3 and a standard deviation of 64.6, with a max document length of 1883 words. Given the wide spread of document lengths, there is a need to control for document length when modelling topics.



Methods

Model Algorithm and Topic Number Selection

The text of the documents of the Melville corpus were preprocessed into uniform word tokens that were understandable by topic modelling libraries. Preprocessing included the removal of punctuation and numbers, changing text to lower-case, and tokenization and lemmatization. Any words that were entirely upper-case were removed, as these generally contained chapter numbers, titles and other content that did not contain semantically interesting information. A list of English stopwords was available in NLTK; I extended this by extracting the 50 most common words in the Melville Corpus that were not in the existing stopword list, and manually examining the list for words without semantic content. This list was further refined with additional terms that appeared frequently in topics but did not appear semantically important.



3 topic modelling Python libraries were used, Gensim, Bitermplus and Corextopic, representing 4 topic modelling algorithms: Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), Biterm Topic Modelling (BMT) and Correlation Explanation (CorEx). In order to determine the optimal number of topics, I used a variety of metrics. All models were evaluated by coherence according to Gensim's CoherenceModel. LDA was also evaluated by perplexity. Bitermplus supports both perplexity and coherence; however, due to technical issues, I only evaluated BTM based on Gensim's CoherenceModel. CorEx was evaluated by total correlation based on an unsupervised model. While NMF and LDA were performed on the entire vocabulary, CorEx and BTM were limited to 10000 vectors (word tokens), due to time constraints, as well as higher vector lengths resulting in poorer coherence scores. For qualitative analysis of topics, I examined 5 topics produced by each algorithm. This number was selected as topic coherence and total correlation both dropped off rapidly with all algorithms. For comparative qualitative I also included a semi-supervised CorEx topic using thematically interesting anchor words.

Qualitative Data Analysis

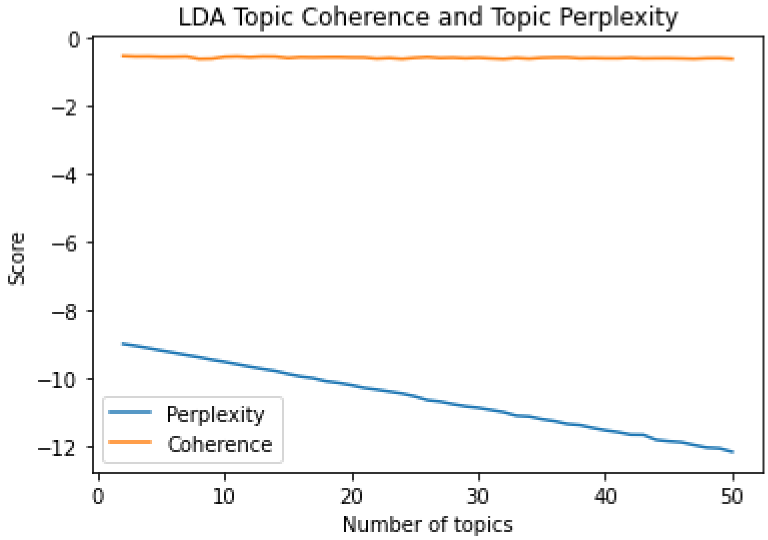
Based on the results of model and topic number selection, I proceeded with using CorEx as the most qualitatively interesting model. As well, I moved from using documents based on chapters to documents representing paragraphs of between 37 and 237 words inclusive. This decision was made based on the qualitative results of the previous stage, which showed that topics based on full-length chapters resulted in topics that correlated exclusively with specific stories. Full-length chapters were also time-consuming to analyze and did not present with clearly defined themes, which ran counter to my goals with topic extraction. Repeating topic number selection with this new constraint, the number of topics was expanded to 11 as coherence dropped down quicker and total correlation plateaued at this value (in line with the recommendations of Gallagher et al. in their original paper, who suggest "choos[ing] the number of topics by observing diminishing returns to the objective" and restarting the model to find the best total correlation score (2017)). The top 20 words of each topic were extracted, as well as the 5 top documents predicted by each topic. I then read and analyzed these 55 documents in a multi-step process, first identifying where high-probability topic words were used in documents, then reading each document and taking preliminary notes, then analyzing each topic as a whole.

Results

Model and Topic Number Selection

Note that this section of the results remains largely unchanged from the original midterm submission. However, the corresponding section on discussion has been updated.

LDA



LDA showed both negative perplexity and coherence. Coherence remained constant at around 0.5 while perplexity dropped in a linear fashion with increased number of topics. While perplexity (in the form of per-word likelihood bound) was only available for LDA, Chang et al. showed that perplexity scores do not correlate with interpretability (2009). Because of this I argue it is not a sufficient metric for evaluating topic models on literature.

LDA topic words for k = 5

Topic	Words
1	[old, time, ship, day, sea, captain, whale, hand, head, good]
2	[old, time, ship, hand, day, captain, good, men, sea, whale]
3	[old, ship, time, sea, captain, day, good, hand, men, whale]
4	[time, old, ship, sea, whale, good, day, water, hand, eye]
5	[old, time, day, men, hand, sea, ship, whale, captain, good]

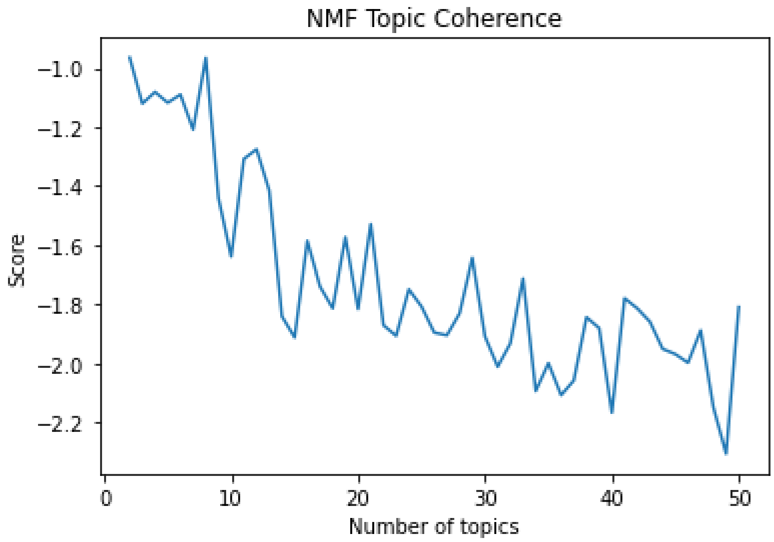
The topics for LDA showed considerable overlap, with the word “old” appearing in all topics. Increasing the number of passes to 10 marginally improved results but was considerably more expensive time-wise.

LDA topic words for k = 5, passes = 10

Topic	Words
1	[ship, captain, old, time, men, sea, deck, sailor, day, hand]
2	[time, old, hand, native, valley, day, place, kory, house, head]
3	[old, time, good, sir, friend, day, think, without, know, sort]
4	[medium, babbalanja, king, old, yoomy, lord, mohi, mardi, island, sea]
5	[whale, sea, boat, ship, ahab, time, head, old, ye, white]

The topics are largely common words, however they appear to refer to individual stories (topic 4 refers to *Mardi,* topic 5 refers to *Moby-Dick*).

NMF



NMF also had negative coherence, which decreased with increased number of topics. The results did take considerably less time to produce.

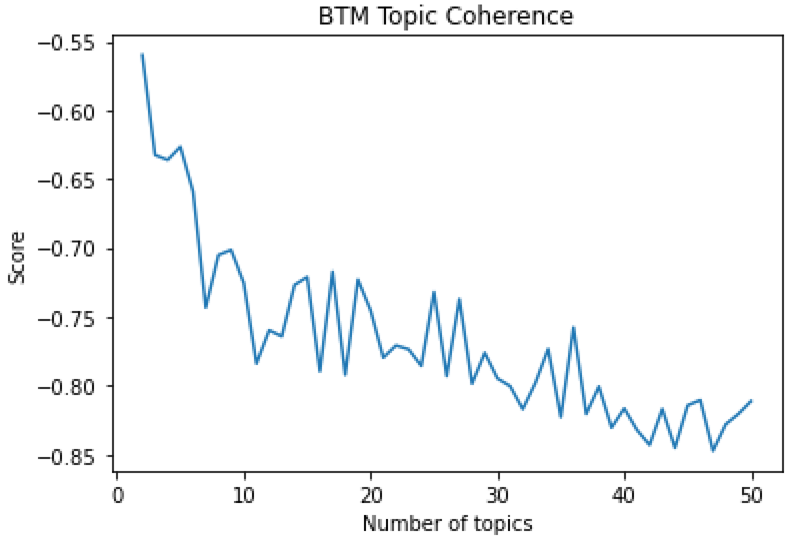
NMF topic words for k = 5

Topic	Words
1	[captain, ship, negro, delano, sea, boat, benito, time, hand, whale]
2	[old, isle, time, sea, good, men, king, boat, day, ever]
3	[old, sir, friend, day, chimney, time, house, wife, poor, sort]
4	[captain, delano, negro, benito, black, ship, white, whale, time, spaniard]
5	[good, captain, hand, time, boy, even, ship, know, lord, place]

The topics for NMF contained some common words but were more varied and interesting than LDA. I note that many of these topics correspond to individual stories; topics 1 and 4 relate to Benito Cereno, with references to race, sailing and captain Amasa Delano. Topic 3 relates to *I and my Chimney*. While these results are improved from LDA, they are largely still related to surface content and not underlying themes.

BTM

There were numerous technical issues with the BTM libraries. Bitermplus would unpredictably cause the Python kernel to crash, and calculating perplexity inside and outside of a loop resulted in different numbers, possibly due to memory issues while calculating the document-topic matrix. As a result, the perplexity results had to be discarded.



BTM topic coherence scores decreased with increasing number of topics but overall, the model performed best with regard to this measure. Calculating the perplexity for k = 5 returned a value of 0.00059, a positive result. However, it’s not certain if Gensim and Bitermplus use the same algorithm for calculating perplexity; Gensim’s log_perplexity function returns per-word likelihood bound, while Bitermplus’s documentation is unclear on how perplexity is calculated. As a result, these values are not comparable.

BTM topic words for k = 5

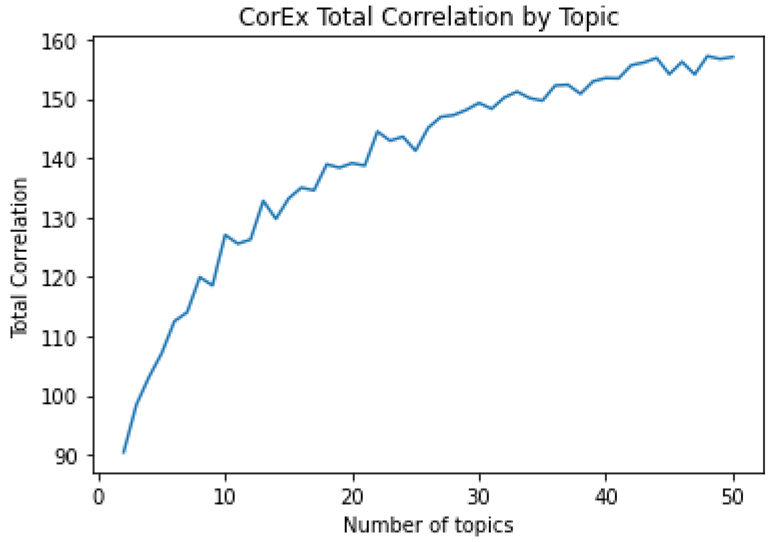
Topic	Words
1	[whale, sea, ship, boat, hand, water, head, white, sail, time]
2	[captain, ship, time, men, sailor, war, day, sea, good, officer]
3	[old, good, know, sir, time, go, ye, let, think, friend]
4	[old, time, king, eye, tree, sea, day, medium, lord, head]
5	[time, old, day, ship, hand, deck, house, captain, place, israel]

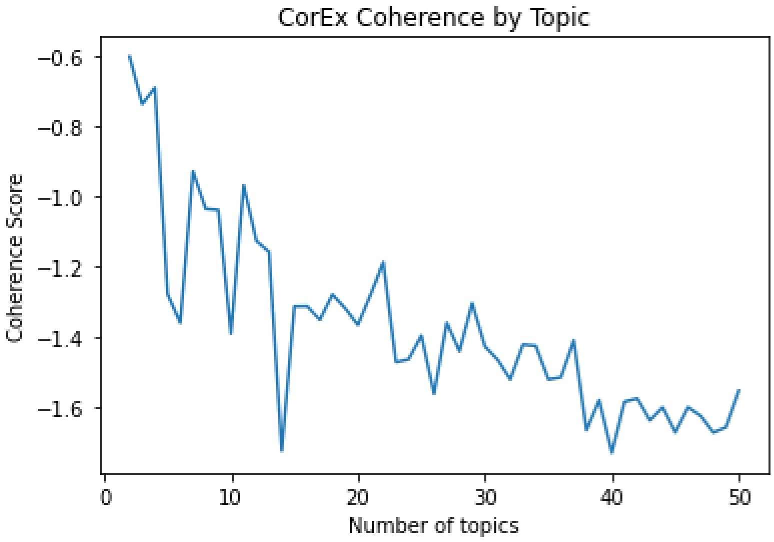
The topics for BTM were extremely similar to LDA, albeit slightly more varied.

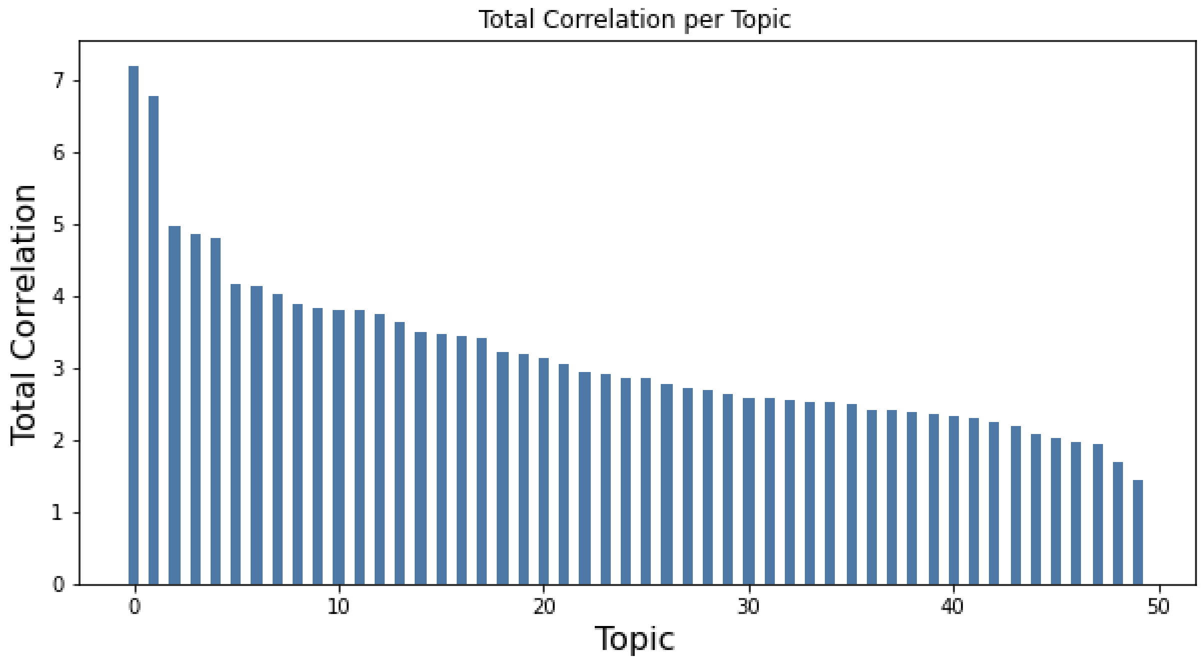
CorEx

Total correlation for CorEx improved with more topics, leveling out somewhat at k = 28.

Coherence dropped as number of topics increased.







For a CorEx model with 50 topics, the first 5 topics had the highest total correlation.

Unsupervised CorEx topic words for k = 5

1	[give, having, such, should, once, put, little, least, being, did]
2	[several, quite, during, two, however, case, indeed, fellow, very, most]
3	[towards, around, suddenly, again, those, whose, air, voice, hand, wild]
4	[sight, after, only, first, while, away, half, every, foot, over]
5	[deck, ship, crew, sailor, captain, seaman, mate, mast, forecastle, cabin]

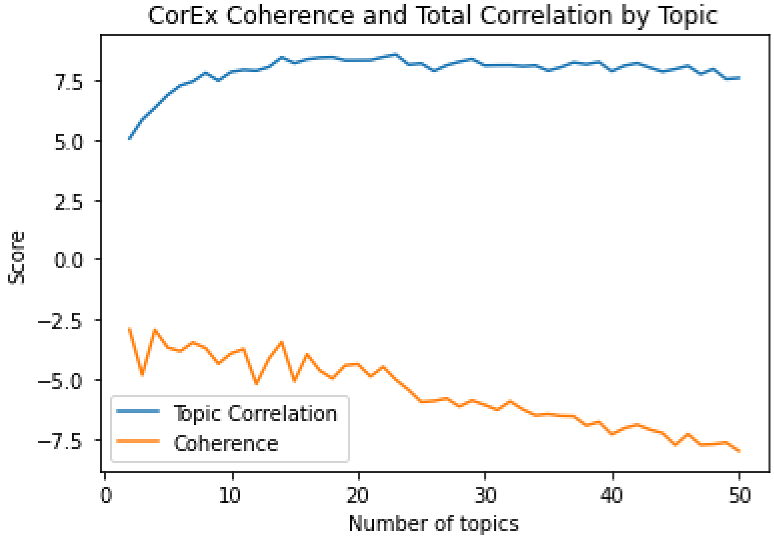
CorEx produced the most interesting topics. While Typee and Babbalanja clearly refer to particular stories, there is a lot more semantic content here than in other models.

Anchored CorEx topic words for k = 5; “prefer”, “sailor”, “time”

Topic	Words
1	[prefer, appeared, manner, truth, fact, moral, nature, understand, subject, trust]
2	[sailor, deck, ship, captain, crew, officer, mate, board, medium, main]
3	[time, put, well, under, about, such, very, being, going, after]
4	[shall, can, come, may, your, let, again, why, how, mean]
5	[towards, suddenly, whose, immediately, seemed, sight, moment, only, saw, effect]

The semi-supervised CorEx model had a total correlation of 111.43 for k = 5, comparable with unsupervised CorEx. Coherence was also similar, with a score of -1.31. The anchor words were selected from two common words (“sailor” and “time”), as well as “prefer” (from *Bartleby the Scrivener’s* response to all requests, “I would prefer not to”). Topic 1 is most interesting, with words like “truth”, “moral” and “nature”. Topics 4 and 5 appear to be largely functional in nature. Topic 3 returned largely functional words, suggesting that "time" may not be as semantically important as we might have expected.

Given the qualitative results of these topic modelling algorithms, which is covered in the discussion, I selected unsupervised CorEx as the best candidate for topic modelling. As well, due to challenges presented in the methods and discussion sections of this paper, I again calculated coherence and total correlation for topic numbers k = 2 to 50 with paragraph-level documents, rather than chapters.



For paragraph-length documents, total correlation gave diminishing returns above k = 11. As with all other models, coherence was negative and decreased in a linear fashion with increased number of topics. Given these results, k = 11 was selected as the number of topics.

Topic Words and Top Document Extraction

All 11 topics had readily-identifiable semantic themes. Topics could broadly be described in four categories, with some overlap: abstract (1, 5, 11), nautical (2, 3, 8, 9), natural (4, 9), and "life" topics (6, 7, 8, 10) that focused on a specific category of person. Topics could cover similar subject matter from different perspectives; for example, topics 3 and 9 both concerned whales, but topic 3 related to the action of whaling, with top words relating to the parts of a ship, and topic 9 related to the nature and biology of whales, with words relating to physical description of whales.

Summary of Topics

Topic number	Top 20 words sorted by probability	Source for Top 5 Documents	Themes of Top Documents
1	[however, part, indeed, time, degree, without, particular, least, general, perhaps, certain, matter, present, almost, even, instance, sometimes, generally, always, le]	Omoo, Typee, Mardi vol. 1	Positivity, Character
2	[boat, water, moment, foot, along, head, distance, towards, air, slowly, bow, shot, beneath, fell, instant, oar, stood, lay, came, sight]	Moby-Dick, Mardi vol. 1, Omoo	Nautical, Movement, Action
3	[deck, sail, sea, rope, main, mast, top, fore, gun, watch, wind, yard, hand, rigging, hole, aloft, night, iron, line, gale]	Moby-Dick	Nautical, Whaling (the action of)
4	[tree, nut, green, cocoa, grove, bough, leaf, wood, stone, fruit, palm, low, shell, red, mat, trunk, blue, eye, black, wall]	Omoo, Typee	Nature, Landscape
5	[mardi, king, sun, land, mountain, world, hill, isle, lord, sky, far, vast, ocean, star, golden, moon, vine, thousand, soul, sweet]	Mardi vol. 2, Moby-Dick	Size, Vastness
6	[native, valley, island, house, several, day, kory, islander, typee, inhabitant, toby, chief, appeared, nukuheva, mehevi, place, soon, three, ground, bread]	Typee	Polynesian life
7	[old, book, coat, room, dock, street, new, used, gentleman, york, dinner, fine, young, looking, town, year, liverpool, lady, church, hat]	Redburn, Israel Potter, The Apple-Tree Table and Other Sketches (Jimmy Rose)	Urban life (both refined and grimy)
8	[ship, captain, officer, sailor, board, seaman, mate, crew, vessel, frigate, cabin, commodore, navy, men, quarter, ashore, lieutenant, english, forecastle, port]	White Jacket	Nautical, Ship crew
9	[whale, sperm, ahab, fish, shark, leviathan, white, deep, tail, spout, jaw, lance, surface, moby, fin, body, dick, monster, fishery, bone]	Moby-Dick	Whaling (the physicality of the animal)
10	[ye, dont, let, thou, oh, sir, think, know, good, thy, shall, god, take, friend, dear, thee, go, cried, heart, tell]	Moby-Dick, Mardi vol. 2, The Confidence-Man	Dialogue (using Quaker language)
11	[war, american, fact, case, law, concerning, account, reason, known, circumstance, period, nature, story, deemed, therefore, service, person, event, usage, since]	Redburn, Typee, White Jacket, The Confidence-Man	History

Discussion

Model and Topic Number Selection

All coherence values calculated for all topic models were negative, indicating poor overlap in topics. I believe that the poor metrics may be explained by the rich, sparse vocabulary of Melville. Consider, as only a limited example, *Moby-Dick,* where passages can be technical and descriptive, or poetic and evocative. Compare these two passages from the chapter *Cetology,* regarding the "Narwhale" (both taken from the same paragraph-document)*:*

> Another instance of a curiously named whale, so named I suppose
> from his peculiar horn being originally mistaken for a peaked nose. The
> creature is some sixteen feet in length, while its horn averages five
> feet, though some exceed ten, and even attain to fifteen feet. Strictly
> speaking, this horn is but a lengthened tusk, growing out from the jaw
> in a line a little depressed from the horizontal.

And

> He is certainly a curious example of the
> Unicornism to be found in almost every kingdom of animated nature. From
> certain cloistered old authors I have gathered that this same
> sea-unicorn's horn was in ancient days regarded as the great antidote
> against poison, and as such, preparations of it brought immense prices.
> It was also distilled to a volatile salts for fainting ladies, the same
> way that the horns of the male deer are manufactured into hartshorn.

In the same chapter, regarding the same animal, Melville moves from discussing the animal descriptively, to an evocative historical anecdote, using words like "cloistered" and the neologism "Unicornism". (Naturally, this is the only appearance of this word in the entire corpus, and thus does not appear in the topic model.) The shifting of themes and vocabulary through passages may pose a challenge for topic modelling. As well, the Gensim CoherenceModel module is based on a metric that was evaluated using benchmark datasets that did not include literature (Röder et al, 2015), meaning it may not be an appropriate evaluation metric.

When using chapter-length documents, individual topics clearly related to individual stories, which doesn't give us any new information about the corpus as a whole; if we are interested in underlying themes rather than surface-level information, then we need to find topics that are novel and reveal connections in stories. For this reason I moved to using paragraph-length documents, which resulted in a better mixture of top documents in topics, at the cost of lower total correlation and coherence scores.

Comparing the evaluation metrics of these topic models in a quantitative manner is challenging due to the use of different metrics, as well as different algorithms used to calculate this metric. We do see that LDA performed best with regards to coherence, followed by LDA, CorEx and then NMF. That being said, no model produced positive coherence values, and improved coherence seemed to correlate with overfitting and less interesting topics.

While these algorithms cannot easily be compared in a quantitative manner, we can compare their topics in a qualitative manner. Both LDA and BTM produced uninteresting topics with too much overlap. NMF produced topics that related to individual stories, a marked improvement over the other two algorithms. CorEx produced topics that were comparatively rich in semantic content. The test of the semi-supervised topic model also produced interesting results. While it was not selected for qualitative analysis, there is plenty of future potential for well-selected anchor words by refining the resulting topics in this exploratory paper.

Extracted Topics and Top Documents

Topics ranged from abstract, such as topic 1, which expresses uncertainty and descriptions of character; to much more concrete topics, such as topic 10, which comprised dialogue from Quaker characters (members of the Religious Society of Friends), readily discernable from their use of the second-person pronouns thee, thy, and thou. Melville wrote many Quaker characters and “had a detailed knowledge of Quaker thought and practice” (Goering, 1981), and the use of Quaker dialogue identifies characters such as *Moby-Dick*’s Captain Ahab as a Quaker, despite the lack of an explicit religious denomination. The four categories of topic I identified relate well with Melville’s themes. Melville’s writing drew from life experiences as a sailor and whaler, living in America as well as on Polynesian islands, which relates to categories of Life and Nautical topics. Melville’s focus on natural topics also tracks with a period of American Romanticism in the mid-19th century, which often drew on natural themes and of which Melville is considered a part of (Yoder, 1973).

Category	Topic numbers	Document themes
Abstract	1,5,11	Character, Vastness, History
Nautical	2,3,8,9	Movement, Crew, Physicality
Natural	4,9	Biology (of animals, plants)
Life	6,7,8,10	Dependent on the experiences being described

The Abstract category is more challenging to explain, absent the context of Melville’s career as a Romanticist poet. Consider this passage from *Mardi, Volume 2,* appearing in topic 5:

> Not greener that midmost terrace of the Andes, which under a torrid
> meridian steeps fair Quito in the dews of a perpetual spring;—not

> greener the nine thousand feet of Pirohitee's tall peak, which, rising
> from out the warm bosom of Tahiti, carries all summer with it into the
> clouds;—nay, not greener the famed gardens of Cyrus,—than the vernal
> lawn, the knoll, the dale of beautiful Verdanna.

In this passage Melville draws comparison of the fictional island of Verdanna*,* in the South Pacific, to the mountains of the Andes and Tahiti, and the historical gardens of Persia (present-day Iran). Melville's writing is rich with these types of poetic allusion in both his prose and poetry, a consequence of his classical education and rich knowledge of literature (Sealts, 1988). These Abstract topics draw from his knowledge of history and literature and skill in poetic allusion. Given these results, I believe that Melville's passages may broadly be divided into those 4 categories, though this is not exhaustive, nor is it the only way to categorize his writing, considering the example from *Cetology,* which could be divided into poetic and technical.

Some of the top documents I extracted contained none of the top 20 probability words of their topic, such as this document from Topic 3, from *Moby-Dick*:

> "Pip? whom call ye Pip? Pip jumped from the whale-boat. Pip's missing.
> Let's see now if ye haven't fished him up here, fisherman. It drags
> hard; I guess he's holding on. Jerk him, Tahiti! Jerk him off; we haul
> in no cowards here. Ho! there's his arm just breaking water. A hatchet!
> a hatchet! cut it off—we haul in no cowards here. Captain Ahab! sir,
> sir! here's Pip, trying to get on board again."

This document is clearly on-topic as it relates to whaling and whaling vessels. The likely reason why a document such as this appears as a top document is that it contains other, less-likely topic words that did not appear in the top 20 words. This does call into question how predictive these topics are of individual documents, though, as has already been discussed, measures like perplexity do not necessarily correlate with human understanding of topics.

While some topics (such as 1 and 2) displayed a healthy mixture of document sources, other topics referred to only a single story, such as *Moby-Dick* in topics 3 and 9. Certain stories were over-represented in top documents, such as *Moby-Dick* with 15 documents and *Typee* with 13 documents, compared to only 2 documents from *Israel Potter*. Some stories were entirely unrepresented in top documents, such as a large number of short stories, and the entirety of *Pierre*. This is somewhat intuitive, as *Moby-Dick* is the longest novel in the corpus, whereas short stories likely had poor total correlation with the novel-length stories, and we saw that all attempts at topic modelling produced negative coherence scores. Only one document from a short story appears, which is the introductory paragraph to *Jimmy Rose*. Interestingly enough, this is the very first document that appears in my dataset; I believe that there is a technical issue that caused earlier document numbers to be over-represented in probabilities, as *Jimmy*

Rose was also over-represented in CorEx topic models using chapter-level documents, appearing in every single topic as the most probable document.

Conclusion and Future Steps

My attempts to improve coherence scores were inconclusive. However, higher coherence scores were still negative and appeared to be the results of overfitting, leading to repetitive topics with nothing but the most common words. My best results came from following intuitions about the Melville corpus and reviewing the resulting topics and documents for themes. I argue that improved metrics such as coherence and total correlation did not necessarily correlate with more informative results, and may not be appropriate for a literature corpus such as the Melville corpus. Quantitatively finding an optimum number of topics in a literary corpus such as this one is challenging and may require novel techniques.

At the same time, some topics contained much surface-level information but did little to expose latent themes. For example, topic 10 revealed the use of language and dialogue characteristic of Quakers (members of the Religious Society of Friends), and while this topic confirms intuitions we may find from a cursory reading of Melville, this does not "allow new meanings to emerge" (Ramsay, 2003) as we might hope from computer-assisted text analysis. Since the purpose of this paper is more exploratory than hypothesis-confirming, we might wonder what new information this paper has added to Melville scholarship and what revisions we might make in future work to elicit deeper, more latent information.

The number of topics was limited based largely on the fact that larger topic numbers led to lower coherence scores. However, given the size and breadth of the corpus, I do not believe that the topics I extracted from this corpus were exhaustive. For instance, the resultant topics say little about themes such as religion and spirituality, whose importance authors like Goering (1981) point to; despite this, these themes are only alluded to in Topics 5 (with terms like "lord" and "sky") and 10 (with identifiably Quaker features). Future analyses would likely benefit from a larger number of topics, with the expectation that topics may overlap considerably. Ultimately, however, the time to analyze the resulting topics thoroughly is a limiting factor.

For future directions, one could use the extracted document-topic matrix as independent variables for regression. This could be used to find relationships between the year of publication and topics, or between chapter number and topics. As an example, extracted topics from *Moby-Dick* could be used to show how themes change in the novel as the story progresses. More work will be required to refine the metrics used in order to give us confidence in extracted topics. We might require novel ways to handle the sparse vocabulary and capture its importance in the topic model. For instance, we may want to group words that are Biblical or Shakespearean or neologistic in origin, given that their appearance in the vocabulary is informative but their individual counts are too small to be considered in the topic model. Topic models are unable to represent the allusions to history, literature and religion that are present in Melville's writing but are lost due to their ignorance of syntax. Another limitation to be addressed in future work is the number of documents from the paragraph-level analysis that were removed due to being too long or short, comprising over 50% of documents the entire corpus (though only about 25% of word tokens). A more sophisticated method of defining documents to be consistent in length, but also reflect intentional passages, is required.

Appendix: Computation Times for Topic Models

Time and energy considerations are important in machine learning; authors such as Bender et al. (2021) and Strubell et al. (2019) have emphasized the importance of evaluating machine learning algorithms based on their energy consumption and environmental impact. With respect to these considerations, I recorded the time taken to compute topic models with each algorithm with all numbers of topics between 2 and 50 inclusive.

Time to compute topic models

Model Algorithm	Time to compute models k = 2 to 50
NMF	3:03.4
LDA	9:40.3
BTM	21:11.1
CorEx	37:39.4

NMF was by far the fastest algorithm, and produced results with less overlap than LDA, the close second algorithm. While CorEx produced the most interesting results from a qualitative perspective, we should bear in mind its time and energy cost when considering implementations at scale. I hope to see more efficient topic modelling algorithms for correlation and semi-supervised learning in the future.

References

About the society. The Melville Society. (2021). Retrieved November 25, 2022, from https://www.melvillesociety.org/aboutthesociety

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? &#x1f99c; *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, *22*.

Du, L., Buntine, W. & Jin, H. (2012). Modelling sequential text with an adaptive topic model. 535-545. https://doi.org/10.13140/2.1.4701.0249

Gallagher, R., Reing, K., Kale, D., & Ver Steeg, G. (2017). Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge. *Transactions of the Association for Computational Linguistics, 5*, 529-542. Retrieved from https://transacl.org/ojs/index.php/tacl/article/view/1244

Goering, W. M. (1981). "To Obey, Rebelling": The Quaker Dilemma in Moby-Dick. *The New England Quarterly*, *54*(4), 519–538. https://doi.org/10.2307/365151

Maxwell, D. E. S. (2018). *Herman Melville*. Routledge. https://doi.org/10.4324/9781315101804

Ramsay, S. (2003). Special Section: Reconceiving Text Analysis: Toward an Algorithmic Criticism. Literary and Linguistic Computing, 18(2), 167–174. https://doi.org/10.1093/llc/18.2.167

Röder, M., Both, A. & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining. 399-408. https://doi.org/10.1145/2684822.2685324

Sealts, M. M. (1988). Melville's reading / Merton M. Sealts, Jr. (Rev. and enl. ed. --). University of South Carolina Press.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. *ArXiv:1906.02243 [Cs]*. http://arxiv.org/abs/1906.02243

Yoder, R. A. (1973). The Equilibrist Perspective: Toward a Theory of American Romanticism. *Studies in Romanticism*, *12*(4), 705–740. https://doi.org/10.2307/25599899