Homework 1 Comparing Corpora with Corpus Statistics
Syracuse University
IST 664 Natural Language Processing
Fall 2021

Task 1

In this assignment, I will be using text processing techniques to analyze two documents and characterize them with corpus statistics. Mainly, I will be looking at unigram, bigram, and trigram frequencies of the two documents and discussing the similarities and differences between them. The documents that I selected are “Moby-Dick” by Herman Melville and “Hamlet” by William Shakespeare.

“Moby-Dick” was a book written in 1851 about a voyage in the seas while “Hamlet” was a book written by Shakespeare in 1600 about a revenge seeking prince. These books were written more than 200 years apart from each other. Not only that, but the story plots are very different. Thus, because of the differences, these documents should be appropriate for the analysis.

Both documents were added to the collection of Gutenberg Etexts under the heavy literature portion around 1999-2001. I will be accessing the text from the Gutenberg collection via the Natural Language Toolkit (NLTK) package in Python. Since this is an open-source library, anybody is able to download them for free.

Task 2

a) Briefly state why you chose the processing options that you did

First, I changed all of the text to lowercase. I decided to do this to ensure that the same words are treated the same regardless of their capitalization. In some natural language processing tasks, such as part of speech tagging, this could be problematic. For example, proper nouns can be identified by starting with a capital letter when the word is not at the beginning of the sentence. But since this task is just looking at frequencies to characterize text, it makes sense to use lowercase.

There are many different ways of doing tokenization but I chose to use the word_tokenize() function from the NLTK package. One of the benefits of this tokenizer is that it splits on more than just white space but one of the drawbacks is that it does not process as fast as some of the other tokenizers. But since the text that I am processing in this task is not of significant proportions, it is an effective tokenizer.

I also removed stopwords from the text. I started by removing the NLTK default stopwords. But I found that there were still some words that I wanted to remove. For example, there were some leftovers from the tokenization such as “’d”, “’s”, “n’t”. I went back and added these to the list of stopwords and processed it again. The tokenizer also separated punctuation such as periods, semicolons, commas, etc. I used the string package to retrieve a list of punctuation marks and went back and removed those as well.

Lastly, I used the Porter stemmer to stem all of the tokens. I chose to use the Porter stemmer instead of a Snowball stemmer or Lancaster stemmer because it is the most lenient. I do not want to trim the words too aggressively because then I might be losing out on some insights. In the following chart, I provide a summary of the vocabulary size at the different stages of text processing that I just covered for the two documents.

Summary of vocabulary size at different stages of text processing

Text Processing Step	Moby Dick		Hamlet	
	Number of Tokens	Number of Unique Tokens	Number of Tokens	Number of Unique Tokens
Raw Text (No Processing)	255,028	20,742	36,372	5,535
Changed Tokens to All Lowercase	255,028	18,701	36,372	4,807
Removed Stopwords and Punctuation Marks	108,649	18,543	15,803	4,685
Stemmed using the Porter Stemmer	108,649	12,337	15,803	3,780

The first thing that stands out to me is that “Moby-Dick” has a much larger volume of text than “Hamlet” does. Changing the tokens to lowercase seemed to have a similar effect on both of the documents. The reduction in size of number of unique tokens was about -10% for “Moby-Dick” and about -13% for “Hamlet”. The same goes for removing the stopwords. 158 unique stopwords were removed from “Moby-Dick”, reducing the size of the number of tokens by about -57%. 122 unique stopwords were removed from “Hamlet”, reducing the size of the number of tokens by about -57%. A little over half of all of the tokens in both of these texts were stopwords. Interestingly, the stemmer did have a slightly greater impact on “Moby-Dick” than it did on “Hamlet”. The reduction in size of number of unique tokens in “Moby-Dick” was about -33% while in “Hamlet” it was about -19%.

b) Are there any problems with the word or bigram lists that you found? Could you get a better list of bigrams?

After I completed these text processing steps, I created lists of unigram, bigram, and trigram frequencies. For the bigrams, I also included a Mutual Information score with a minimum frequency of 5. I then narrowed down these lists to the top 50 in each list. Lastly, I put these lists into Pandas dataframes for easier analysis.

I did not find any problems with the word or bigram lists that I found. I was surprised to see that the two lists were very distinctive from each other. In fact, only 11 words out of the top 50 words in each of "Moby-Dick" and "Hamlet" appeared in both lists. For these 11 words, Shakespeare tended to have a slightly higher normalized frequency than did Melville. I do not think that I could have gotten a much better list of bigrams for this task.

c) How are the top 50 bigrams by frequency different from the top 50 bigrams scored by Mutual Information?

In "Hamlet", the bigrams by raw frequency and the bigrams by mutual information were similar to each other. In fact, only 2 bigrams out the top 50 bigrams scored by mutual information were not in the top 50 list for raw frequency. That means 48 / 50 bigrams were in both of the lists. This is interesting because it means that those words are paired together with each other throughout the book and were not frequently interchanged with other words.

In "Moby-Dick" on the other hand, 46 out of the 50 bigrams scored by Mutual information were not in the top 50 list for raw frequency. That means that 4 / 50 bigrams were in both of the lists. It appears that this is because the words that were part of those bigrams were frequently interchanged. For example, the bigram ('sperm', 'whale') and ('white', 'whale') appeared at the top of the list when scored by frequency but did not appear on the Mutual Information list. I suspect that this is because the word "whale" is interchanged with a lot of other words throughout the book.

Task 3

a) Clearly describe the problem or question you are trying to address throughout the comparison between the two selected documents

Based on the top 50 unigrams by normalized frequency, bigrams by normalized frequency, bigrams by Mutual Information score, and trigrams by normalized frequency, what were the main distinctions between Shakespeare's "Hamlet" and Melville's "Moby-Dick"?

b) Present and explain insights or conclusions based on the comparison to answer the question (do not just report numbers)

One of the major differences between the two documents is the content. In "Hamlet", the most frequent words include words such as "ham", "lord', and "king" while the most frequent words in "Moby-Dick" include words such as "whale", "ship", "sea". This is not surprising because these words are more specific to the story plots rather than being more general words.

Another difference is that Shakespeare had different vocabulary usage than did Melville in these books. For example, Shakespeare used the word "thou" (0.7%), "thee" (0.4%), "shall" (0.7%). This distinguishes "Hamlet" from "Moby-Dick" because these words were rarely used in "Moby-Dick" (less than 0.001%). A the time “Moby-Dick” was written, these words were still popular in literature, according to the Google Ngram Viewer, but then began to decline in usage thereafter.

There were several words that were used in “Moby-Dick” and were not used in “Hamlet” as well. The word “thing” was #23 on the list for “Moby-Dick” (0.3%) while it was only used a total of 24 times in “Hamlet” (0.0007%). The word “though” was # 15 on the list for “Moby-Dick” (0.4%) and it was used 22 in “Hamlet” (0.0006%).

I created a list of trigrams as well for my analysis. Unfortunately, I did not find that the trigrams were very insightful. The trigrams are even more specific to the text at hand, for example, there were only 9 trigrams that existed both in in “Moby-Dick” and “Hamlet”. Furthermore, I suspect that most of the trigrams are specific to a certain chapter, scene, or character within the books. The trigrams do provide some more context, however, given a group of words rather than just one or two words.

Top 50 unigram tokens for Hamlet and Moby-Dick

Top 50 Tokens Hamlet				Top 50 Tokens Moby-Dick		
Token	Count	PctTotal		Token	Count	PctTotal
ham	337	2.1%		whale	1468	1.4%
lord	216	1.4%		one	932	0.9%
king	180	1.1%		like	594	0.5%
haue	175	1.1%		upon	565	0.5%
come	128	0.8%		ship	558	0.5%
hamlet	107	0.7%		ye	510	0.5%
let	107	0.7%		ahab	509	0.5%
shall	107	0.7%		man	503	0.5%
thou	104	0.7%		sea	471	0.4%
good	98	0.6%		seem	462	0.4%
hor	95	0.6%		old	443	0.4%
thi	90	0.6%		time	441	0.4%
enter	86	0.5%		would	435	0.4%
like	82	0.5%		boat	431	0.4%
oh	81	0.5%		though	383	0.4%
make	74	0.5%		hand	349	0.3%
know	74	0.5%		captain	346	0.3%
loue	71	0.4%		yet	344	0.3%
well	70	0.4%		head	340	0.3%
father	70	0.4%		look	330	0.3%
would	68	0.4%		say	327	0.3%
self	66	0.4%		long	323	0.3%
giue	66	0.4%		thing	320	0.3%

may	65	0.4%		still	315	0.3%
sir	65	0.4%		see	309	0.3%
speak	64	0.4%		great	305	0.3%
qu	62	0.4%		said	304	0.3%
't	61	0.4%		two	291	0.3%
vs	61	0.4%		must	284	0.3%
laer	60	0.4%		come	284	0.3%
thee	58	0.4%		way	283	0.3%
say	58	0.4%		last	279	0.3%
ile	58	0.4%		white	278	0.3%
must	58	0.4%		go	277	0.3%
hath	57	0.4%		eye	271	0.2%
oph	56	0.4%		thou	268	0.2%
think	54	0.3%		round	258	0.2%
one	54	0.3%		stubb	256	0.2%
heauen	53	0.3%		queequeg	252	0.2%
time	51	0.3%		littl	249	0.2%
man	51	0.3%		harpoon	246	0.2%
doe	51	0.3%		three	242	0.2%
vpon	50	0.3%		day	240	0.2%
see	50	0.3%		sperm	240	0.2%
heer	50	0.3%		men	237	0.2%
go	50	0.3%		may	237	0.2%
pol	49	0.3%		water	236	0.2%
queen	47	0.3%		first	235	0.2%
mother	46	0.3%		everi	232	0.2%
look	45	0.3%		us	228	0.2%

Top 50 bigram tokens for Hamlet and Moby-Dick

Top 50 Bigrams Hamlet				Top 50 Bigrams Moby-Dick		
Bigram	Count	PctTotal		Bigram	Count	PctTotal
('lord', 'ham')	70	0.4%		('sperm', 'whale')	173	0.2%
('good', 'lord')	23	0.1%		('white', 'whale')	106	0.1%
('enter', 'king')	15	0.1%		('moby', 'dick')	81	0.1%
('hamlet', 'ham')	15	0.1%		('old', 'man')	75	0.1%
('wee', 'l')	13	0.1%		('captain', 'ahab')	64	0.1%
('haue', 'seene')	11	0.1%		('right', 'whale')	52	0.0%
('lord', 'hamlet')	11	0.1%		('cried', 'ahab')	33	0.0%
('enter', 'hamlet')	10	0.1%		('captain', 'peleg')	32	0.0%

('exeunt', 'enter')	10	0.1%		('aye', 'aye')	30	0.0%
('ham', 'oh')	10	0.1%		('mr.', 'starbuck')	29	0.0%
('ham', 'sir')	10	0.1%		('one', 'hand')	28	0.0%
('haue', 'heard')	9	0.1%		('let', 'us')	27	0.0%
('hor', 'lord')	9	0.1%		('look', 'ye')	26	0.0%
('king', 'queene')	9	0.1%		('every', 'one')	24	0.0%
('lord', 'haue')	9	0.1%		('cried', 'stubb')	23	0.0%
('ophe', 'lord')	9	0.1%		('one', 'side')	23	0.0%
('thou', 'hast')	9	0.1%		('never', 'mind')	22	0.0%
('enter', 'polonius')	8	0.1%		("'ye", 'see')	21	0.0%
('fathers', 'death')	8	0.1%		('thou', 'art')	21	0.0%
('ham', 'nay')	8	0.1%		('whale', 'head')	21	0.0%
('lord', 'polon')	8	0.1%		('ye', 'ye')	19	0.0%
('dost', 'thou')	7	0.0%		('new', 'bedford')	18	0.0%
('good', 'friends')	7	0.0%		('round', 'round')	18	0.0%
('let', 'see')	7	0.0%		('said', 'stubb')	18	0.0%
('let', 'vs')	7	0.0%		('sperm', 'whales')	18	0.0%
('rosincrance', 'guildensterne')	7	0.0%		('years', 'ago')	18	0.0%
('set', 'downe')	7	0.0%		('cried', 'starbuck')	17	0.0%
('thou', 'art')	7	0.0%		('old', 'ahab')	17	0.0%
('would', 'haue')	7	0.0%		('thee', 'thou')	17	0.0%
('ham', 'come')	6	0.0%		('boat', 'crew')	16	0.0%
('ile', 'haue')	6	0.0%		('cape', 'horn')	16	0.0%
('king', 'haue')	6	0.0%		('greenland', 'whale')	16	0.0%
('let', 'come')	6	0.0%		('lower', 'jaw')	16	0.0%
('mine', 'owne')	6	0.0%		('something', 'like')	16	0.0%
('well', 'lord')	6	0.0%		('whale', 'fishery')	16	0.0%
("'t", 'true')	5	0.0%		('would', 'seem')	16	0.0%
('enter', 'ghost')	5	0.0%		('young', 'man')	16	0.0%
('enter', 'horatio')	5	0.0%		('ivory', 'leg')	15	0.0%
('hamlet', 'hamlet')	5	0.0%		('stubb', 'flask')	15	0.0%
('heauen', 'earth')	5	0.0%		('thus', 'far')	15	0.0%
('horatio', 'marcellus')	5	0.0%		('well', 'known')	15	0.0%
('king', 'oh')	5	0.0%		('whaling', 'voyage')	15	0.0%
('let', 'know')	5	0.0%		('would', 'fain')	15	0.0%
('lord', 'exeunt')	5	0.0%		('ye', 'see')	15	0.0%
('qu', 'oh')	5	0.0%		('captain', 'bildad')	14	0.0%
('reynol', 'lord')	5	0.0%		('chief', 'mate')	14	0.0%
('rosin', 'lord')	5	0.0%		('ere', 'long')	14	0.0%
('shall', 'heare')	5	0.0%		('ever', 'since')	14	0.0%

('sit', 'downe')	5	0.0%		('three', 'four')	14	0.0%

Top 50 Bigram tokens for Hamlet and Moby-Dick (Mutual Information)

Top 50 Hamlet			Top 50 Moby-Dick	
Bigram	PMI		Bigram	PMI
('rosincrance', 'guildensterne')	9		('caw', 'caw')	14
('wee', 'l')	9		('samuel', 'enderby')	13
('sit', 'downe')	9		('warp', 'woof')	13
('horatio', 'marcellus')	8		('latitude', 'longitude')	13
('set', 'downe')	8		('straits', 'sunda')	13
('fathers', 'death')	7		('mrs.', 'hussey')	13
('dost', 'thou')	7		('um', 'um')	13
('wilt', 'thou')	7		('iii', 'duodecimo')	13
('heauen', 'earth')	6		('heidelburgh', 'tun')	13
('exeunt', 'enter')	6		('ii', 'octavo')	12
('enter', 'polonius')	6		('st.', 'george')	12
('mine', 'owne')	6		('father', 'mapple')	12
('thou', 'art')	6		('hither', 'thither')	12
("'t", 'true')	6		('huzza', 'porpoise')	12
('good', 'friends')	6		('fiery', 'pit')	12
('haue', 'heard')	6		('beef', 'bread')	11
('thou', 'hast')	6		('steering', 'oar')	11
('haue', 'seene')	6		('hundred', 'seventy-seventh')	11
('enter', 'ghost')	5		('wife', 'child')	11
('tell', 'vs')	5		('fore', 'aft')	11
('reynol', 'lord')	5		('centuries', 'ago')	11
('shall', 'heare')	5		('cape', 'horn')	11
('enter', 'horatio')	5		('seven', 'hundred')	11
('let', 'see')	5		('blows', 'blows')	10
('king', 'queene')	4		('moby', 'dick')	10
('enter', 'hamlet')	4		('new', 'york')	10
('good', 'lord')	4		('new', 'zealand')	10
('let', 'vs')	4		('new', 'bedford')	10
('lord', 'exeunt')	4		('ha', 'ha')	10
('enter', 'king')	4		('book', 'ii')	10
('lord', 'polon')	4		('harpoons', 'lances')	10

('qu', 'oh')	4		('gave', 'understand')	10
('lord', 'ham')	4		('book', 'folio')	10
('ham', 'nay')	4		('saturday', 'night')	10
('ophe', 'lord')	4		('drew', 'nigh')	9
('let', 'know')	3		('chief', 'mate')	9
('ile', 'haue')	3		('eight', 'ten')	9
('would', 'haue')	3		('years', 'ago')	9
('let', 'come')	3		('english', 'whalers')	9
('rosin', 'lord')	3		('forty', 'years')	9
('lord', 'hamlet')	3		('brought', 'alongside')	9
('hamlet', 'hamlet')	3		('lower', 'jaw')	9
('ham', 'sir')	3		('four', 'oceans')	9
('hor', 'lord')	3		('thousand', 'miles')	9
('hamlet', 'ham')	3		('drawing', 'nigh')	9
('well', 'lord')	3		('new', 'england')	9
('ham', 'oh')	3		('pagan', 'harpooneers')	9
('king', 'oh')	3		('closed', 'eyes')	9
('lord', 'haue')	2		('queer', 'queer')	9

Top 50 Trigram tokens for Hamlet and Moby-Dick

Top 50 Trigrams Hamlet				Top 50 Trigrams Moby Dick		
Trigram	Count	PctTotal		Trigram	Count	PctTotal
('enter', 'king', 'queene')	7	0.0%		('great', 'sperm', 'whale')	11	0.0%
('exeunt', 'enter', 'hamlet')	5	0.0%		('sperm', 'whale', 'head')	11	0.0%
('good', 'lord', 'ham')	5	0.0%		('every', 'one', 'knows')	9	0.0%
('ophe', 'lord', 'ham')	5	0.0%		('seen', 'white', 'whale')	8	0.0%
('enter', 'hamlet', 'ham')	4	0.0%		('book', 'ii', 'octavo')	7	0.0%
('enter', 'hamlet', 'horatio')	3	0.0%		('book', 'folio', 'chapter')	6	0.0%
('enter', 'polonius', 'pol')	3	0.0%		('cape', 'good', 'hope')	6	0.0%
('exeunt', 'manet', 'hamlet')	3	0.0%		('right', 'whale', 'head')	6	0.0%
('good', 'lord', 'polon')	3	0.0%		('seven', 'hundred', 'seventy-seventh')	6	0.0%
('hor', 'good', 'lord')	3	0.0%		('greenland', 'right', 'whale')	5	0.0%
('lord', 'exeunt', 'enter')	3	0.0%		('hast', 'seen', 'white')	5	0.0%
('lord', 'ham', 'sir')	3	0.0%		('ii', 'octavo', 'chapter')	5	0.0%
('reynol', 'good', 'lord')	3	0.0%		('see', 'cried', 'ahab')	5	0.0%
("'t", 'true', "'t")	2	0.0%		('sperm', 'whale', 'fishery')	5	0.0%
('buried', 'christian', 'buriall')	2	0.0%		('stubb', 'second', 'mate')	5	0.0%
('charge', 'thee', 'speake')	2	0.0%		('whale', 'white', 'whale')	5	0.0%
('christian', 'buriall', 'clo')	2	0.0%		('would', 'almost', 'thought')	5	0.0%

('clay', 'made', 'guest')	2	0.0%		('aye', 'aye', 'sir')	4	0.0%
('deere', 'brothers', 'death')	2	0.0%		('bildad', 'said', 'peleg')	4	0.0%
('deere', 'lord', 'ham')	2	0.0%		('book', 'iii', 'duodecimo')	4	0.0%
('doe', 'lord', 'ham')	2	0.0%		('captain', 'ahab', 'said')	4	0.0%
('dost', 'thou', 'heare')	2	0.0%		('captain', 'peleg', 'said')	4	0.0%
('dye', 'sleepe', 'sleepe')	2	0.0%		('caw', 'caw', 'caw')	4	0.0%
('enter', 'enter', 'queene')	2	0.0%		('chase', 'moby', 'dick')	4	0.0%
('enter', 'horatio', 'marcellus')	2	0.0%		('even', 'present', 'day')	4	0.0%
('enter', 'king', 'king')	2	0.0%		('god', 'bless', 'ye')	4	0.0%
('ere', 'go', 'bed')	2	0.0%		('old', 'manx', 'sailor')	4	0.0%
('exeunt', 'enter', 'horatio')	2	0.0%		('round', 'cape', 'horn')	4	0.0%
('exeunt', 'scena', 'secunda')	2	0.0%		('sleep', 'two', 'bed')	4	0.0%
('father', 'much', 'offended')	2	0.0%		('sperm', 'whale', 'right')	4	0.0%
('follow', 'exeunt', 'enter')	2	0.0%		('starbuck', 'stubb', 'flask')	4	0.0%
('gho', 'sweare', 'ham')	2	0.0%		('thee', 'old', 'man')	4	0.0%
('go', 'exeunt', 'enter')	2	0.0%		('thou', 'clear', 'spirit')	4	0.0%
('god', 'blesse', 'sir')	2	0.0%		('whale', 'right', 'whale')	4	0.0%
('goe', 'ile', 'follow')	2	0.0%		('whale', 'sperm', 'whale')	4	0.0%
('guest', 'meete', 'ham')	2	0.0%		('white', 'whale', 'white')	4	0.0%
('guild', 'good', 'lord')	2	0.0%		("'ll", 'give', 'ye')	3	0.0%
('guild', 'lord', 'ham')	2	0.0%		("'ye", 'see', 'cried')	3	0.0%
('ham', 'glad', 'see')	2	0.0%		('aye', 'sir', 'said')	3	0.0%
('ham', 'nay', 'know')	2	0.0%		('blows', 'blows', 'blows')	3	0.0%
('ham', 'oh', 'wonderfull')	2	0.0%		('brought', 'bear', 'upon')	3	0.0%
('ham', 'sir', 'guild')	2	0.0%		('captain', 'ahab', 'captain')	3	0.0%
('hamlet', 'ham', 'good')	2	0.0%		('cook', 'said', 'stubb')	3	0.0%
('hamlet', 'horatio', 'ham')	2	0.0%		('corpusants', 'mercy', 'us')	3	0.0%
('hamlet', 'thou', 'hast')	2	0.0%		('dick', 'moby', 'dick')	3	0.0%
('hath', 'made', 'mad')	2	0.0%		('ding', 'dong', 'ding')	3	0.0%
('haue', 'newes', 'tell')	2	0.0%		('fifty', 'years', 'ago')	3	0.0%
('haue', 'seene', 'night')	2	0.0%		('first', 'congregational', 'church')	3	0.0%
('heauen', 'earth', 'must')	2	0.0%		('first', 'night', 'watch')	3	0.0%