Teng Siong (T.S) Yeap
IST 664: Homework1

Homework 1: Corpus Statistics and Mutual Information

1. For this homework, I chose one poem document named “blake-poems.txt” from the Gutenberg collection and another story document from the Internet. From the Gutenberg collection, I chose blake-poems.txt. The text file contains 8239 words; for the second document, I found a text file named “The Bureau of Procuration (Story)” from the website http://textfiles.com/stories/. I saved the file to my local machine. To read this file to nltk, I used open() to create a holder for the file and call read() to read in the text in the file. It contains 30184 words.
   The code and output looked like this:

```
f=open('bureau.txt','rU')
raw=f.read()
tokens = nltk.word_tokenize(raw)
mywords = [w.lower( ) for w in tokens]
# show some of the words
print(len(mywords))
print(mywords[ :110])
```

```
C:\ProgramData\Anaconda3\lib\site-packages\ipykernel_launcher.py:1: DeprecationWarning: 'U' mode is deprecated
  """Entry point for launching an IPython kernel.
```

```
30184
['``', 'anyone', 'who', 'has', 'not', 'worked', 'for', 'them', 'simply', 'can', 'not', 'understand', 'them', '.', "''", '-', 'mil
le', 'vennamun', ',', 'introduction', 'to', ':', '``', 'the', 'use', 'of', 'ashes', ':', 'bureau', 'of', 'procuration', "manua
l'", 'half', 'past', 'eight', '.', 'the', 'bedside', 'alarm', 'woke', 'kelanie', 'up', 'with', 'the', 'sampled', 'victory-scree
ch', 'of', 'some', 'carnivorous', 'xenoform', '.', 'she', 'was', 'up', 'immediately', ',', 'eyes', 'wide', ',', 'fingers', 'cla
wing', 'the', 'pillow-pads', ',', 'gasping', 'with', 'shock', 'as', 'the', 'subconsciously-induced', 'adrenalin', 'shivered',
'through', 'her', 'system', '.', 'as', 'she', 'calmed', 'down', ',', 'her', 'pupils', 'dilated', 'out', 'from', 'crisis-', 'ind
uced', 'pinpricks', ',', 'her', 'breathing', 'and', 'pulse', 'rates', 'returned', 'to', 'normal', ',', 'and', 'she', 'wondere
d', ',', 'not', 'for', 'the', 'first', 'or', 'last', 'time']
```

2. To generate meaningful word frequencies distribution, I lowered case all the words. Then, I used the nltk.word_tokenize() as the tokenizer for both documents to keep them consistent. The other tokenizer nltk.corpus.gutenberg.words() only worked for Gutenberg related text file and it tokenized two symbols as one token. For example ':!" was considered as a token.

   a. list the top 50 words by frequency (normalized by the length of the document)

   I applied a filter named alpha filter what was shown in the class to remove all the non-alphabetic characters. This is because the comma and period will be the highest “word” in the word frequency distribution. The same rationale applied to the word “and”, “of” etc. I used the nltk.corpus.stopwords.words('english') to remove all the English stopwords. In addition to the standard English stopwords provided by the nltk, I modified the stopwords list to expand it a little by adding “us”. The word appeared often on both documents. After applying the stopwords filter, blake-peom.txt has 3644 words and The Bureau of Procuration (Story).txt has 12706 words. I then used these numbers to normalize my word frequency distribution. The result is shown below:

Teng Siong (T.S) Yeap
IST 664: Homework1

- blake-poems.txt
- The Bureau of Procuration (Story).txt

	Frequency	Normalized			Frequency	Normalized
little	45	0.012349		kelanie	189	0.014875
thee	42	0.011526		marek	133	0.010467
like	35	0.009605		nosan no os	100	0.00787
thou	35	0.009605		tsiry-feylen	95	0.007477
thy	31	0.008507		one	94	0.007398
love	29	0.007958		like	71	0.005588
sweet	28	0.007684		suit	69	0.005431
night	28	0.007684		back	62	0.00488
joy	25	0.006861		said	60	0.004722
away	24	0.006586		two	56	0.004407
weep	24	0.006586		ship	49	0.003856
father	22	0.006037		moridani	46	0.00362
sleep	21	0.005763		around	43	0.003384
happy	19	0.005214		something	40	0.003148
shall	19	0.005214		notepad	36	0.002833
day	19	0.005214		bythian	36	0.002833
mother	19	0.005214		turned	35	0.002755
child	18	0.00494		three	34	0.002676
every	17	0.004665		going	34	0.002676
never	17	0.004665		think	34	0.002676
thel	16	0.004391		earth	33	0.002597
hear	16	0.004391		eyes	31	0.00244
green	16	0.004391		bythians	31	0.00244
voice	16	0.004391		head	30	0.002361
infant	16	0.004391		parkry	29	0.002282
see	16	0.004391		end	29	0.002282
human	16	0.004391		xeno	28	0.002204
cloud	15	0.004116		found	27	0.002125
lamb	15	0.004116		human	27	0.002125
till	15	0.004116		asteroid	27	0.002125
bright	15	0.004116		still	26	0.002046
delight	14	0.003842		behind	25	0.001968
upon	14	0.003842		made	25	0.001968
head	14	0.003842		appeared	25	0.001968
weeping	14	0.003842		know	25	0.001968
holy	13	0.003568		away	25	0.001968

sit	12	0.003293		humans	25	0.001968
white	12	0.003293		wall	24	0.001889
care	12	0.003293		export	24	0.001889
oer	12	0.003293		translator	24	0.001889
face	12	0.003293		even	24	0.001889
children	12	0.003293		get	24	0.001889
tears	12	0.003293		front	24	0.001889
heard	12	0.003293		first	23	0.00181
sing	11	0.003019		time	23	0.00181
sun	11	0.003019		control	23	0.00181
birds	11	0.003019		side	23	0.00181
god	11	0.003019		way	23	0.00181
boy	11	0.003019		replied	22	0.001731
oh	11	0.003019		almost	22	0.001731

b. list the top 50 bigrams by frequencies

For bigram frequencies, I used the alpha filter to filter out all the non-alphabetical characters. Bigram frequency is the percentage of times that two words that occurs together (bigram) in all the bigram of the corpus.

blakepoem.txt			The Bureau of Procuration (Story).txt	
in', 'the'	0.005583		of', 'the'	0.006494
of', 'the'	0.003398		to', 'the'	0.003611
and', 'the'	0.003277		in', 'the'	0.003214
and', 'i'	0.002185		the', "nosan'no'os"	0.002584
on', 'the'	0.001699		on', 'the'	0.002418
the', 'little'	0.001699		from', 'the'	0.001756
to', 'the'	0.001699		at', 'the'	0.001657
in', 'a'	0.001578		that', 'the'	0.001491
i', 'am'	0.001456		into', 'the'	0.001425
like', 'a'	0.001456		the', 'ship'	0.001425
the', 'human'	0.001456		to', 'be'	0.001226
the', 'night'	0.001214		as', 'the'	0.001126
and', 'he'	0.001092		her', 'suit'	0.001126
when', 'the'	0.001092		we', 'have'	0.001126
from', 'the'	0.000971		with', 'a'	0.001126
no', 'more'	0.000971		it', 'was'	0.001093
the', 'sun'	0.000971		do', "n't"	0.00106
a', 'little'	0.00085		for', 'a'	0.00106
all', 'the'	0.00085		in', 'a'	0.00106
an', 'infant'	0.00085		of', 'a'	0.00106

hear', 'the'	0.00085		and', 'then'	0.001027
little', 'boy'	0.00085		kelanie', "'s"	0.000994
little', 'lamb'	0.00085		with', 'the'	0.000994
my', 'mother'	0.00085		one', 'of'	0.000961
the', 'vales'	0.00085		of', 'her'	0.000928
where', 'the'	0.00085		and', 'the'	0.000895
and', 'love'	0.000728		to', 'a'	0.000861
and', 'not'	0.000728		was', 'a'	0.000861
but', 'i'	0.000728		by', 'the'	0.000828
can', 'it'	0.000728		like', 'a'	0.000828
can', 'not'	0.000728		we', 'are'	0.000828
i', 'see'	0.000728		a', 'few'	0.000795
i', 'was'	0.000728		did', "n't"	0.000795
in', 'every'	0.000728		for', 'the'	0.000795
so', 'i'	0.000728		out', 'of'	0.000795
songs', 'of'	0.000728		the', 'moridani'	0.000795
voice', 'of'	0.000728		through', 'the'	0.000795
while', 'the'	0.000728		it', "'s"	0.000762
with', 'the'	0.000728		had', 'been'	0.000729
among', 'the'	0.000607		in', 'her'	0.000729
and', 'all'	0.000607		the', 'xeno'	0.000729
and', 'we'	0.000607		there', 'was'	0.000729
because', 'i'	0.000607		and', 'a'	0.000696
can', 'i'	0.000607		going', 'to'	0.000696
filled', 'with'	0.000607		her', 'notepad'	0.000696
human', 'form'	0.000607		she', 'was'	0.000663
i', 'can'	0.000607		have', 'been'	0.000629
it', 'be'	0.000607		she', 'had'	0.000629
of', 'my'	0.000607		that', 'they'	0.000629
of', 'thel'	0.000607		at', 'her'	0.000596

c. list the top 50 bigrams by their Mutual Information scores (using min frequency 5)
Unlike bigram frequency, Pointwise Mutual Information computes the probability of two words occurring in a corpus. It compares the likelihood of a pair of words to occur together with the likelihood of each word occur individually. The higher the PMI value means the stronger the two words will co-occur.

blakepoem.txt			The Bureau of Procuration (Story).txt	
no', 'more'	8.668404		carriage', 'return'	12.29653
human', 'form'	8.330182		artificial', 'intelligence'	11.81111
an', 'infant'	7.423291		miss', 'camden'	11.81111

Teng Siong (T.S) Yeap
IST 664: Homework1

little', 'boy'	6.864324		tickling', 'feeling'	11.44854
filled', 'with'	6.700825		return', 'line'	11.42206
little', 'lamb'	6.416865		line', 'feed'	11.19967
smiles', 'on'	6.411318		pthalklin', 'ervae'	10.97461
because', 'i'	5.985886		threat', 'termination'	10.43404
it', 'be'	5.828344		data', 'service'	10.33718
can', 'it'	5.735235		asteroid', 'belt'	9.641182
songs', 'of'	5.596037		bythian', 'scout'	9.033499
i', 'am'	5.570848		how', 'many'	8.431463
my', 'mother'	5.192642		pick', 'up'	8.25214
can', 'not'	5.166951		ca', "n't"	7.881496
i', 'see'	4.570848		wo', "n't"	7.881496
like', 'a'	4.475248		i', "'m"	7.826214
voice', 'of'	4.403391		staring', 'at'	7.762555
in', 'every'	4.366202		went', 'over'	7.552373
so', 'i'	4.178531		some', 'sort'	7.502985
of', 'thel'	4.140357		held', 'up'	7.444785
the', 'vales'	3.867606		at', 'least'	7.440627
the', 'human'	3.815139		filled', 'with'	7.405763
the', 'sun'	3.770745		cut', 'off'	7.39607
i', 'was'	3.616652		person', 'at'	7.384044
among', 'the'	3.552104		followed', 'by'	7.33075
a', 'little'	3.335071		their', 'way'	7.299041
but', 'i'	3.322921		did', "n't"	7.296534
can', 'i'	3.307814		stared', 'at'	7.277128
i', 'can'	3.307814		could', 'see'	7.213793
hear', 'the'	3.037531		looked', 'like'	7.146787
while', 'the'	3.007784		looking', 'for'	7.085309
and', 'we'	2.799775		should', 'be'	6.987545
the', 'weeping'	2.74475		has', 'been'	6.924975
the', 'night'	2.74475		glanced', 'at'	6.888086
till', 'the'	2.645214		more', 'than'	6.886012
in', 'the'	2.614187		something', 'like'	6.869253
when', 'the'	2.592746		fingers', 'into'	6.77716
in', 'a'	2.580457		"nosan'no'os", 'transport'	6.752213
on', 'the'	2.5781		one', 'side'	6.6107
the', 'green'	2.552104		do', "n't"	6.541646
the', 'voice'	2.552104		in', 'fact'	6.476355
the', 'little'	2.545678		we', "'re"	6.460133
where', 'the'	2.393675		"'d", 'like'	6.409821
and', 'love'	2.292292		it', 'seems'	6.274166

from', 'the'	2.230176		this', 'end'	6.268628
and', 'he'	1.92788		his', 'eyes'	6.204834
of', 'the'	1.847707		was', 'surprised'	6.1978
of', 'my'	1.765318		on', 'millimillenary'	6.175518
all', 'the'	1.752129		as', 'many'	6.158445
and', 'not'	1.724008		had', 'been'	6.155588

d. list the top 50 trigrams by frequencies

For trigram frequencies, I used the alpha filter to filter out all the non-alphabetical characters.

blakepoem.txt			The Bureau of Procuration (Story).txt	
can', 'it', 'be'	0.000607		of', 'the', "nosan'no'os"	0.00053
the', 'human', 'form'	0.000607		out', 'of', 'the'	0.000497
the', 'voice', 'of'	0.000607		there', 'was', 'a'	0.000497
never', 'can', 'it'	0.000485		for', 'a', 'moment'	0.000431
of', 'the', 'night'	0.000485		one', 'of', 'the'	0.000431
of', 'the', 'vales'	0.000485		appeared', 'to', 'be'	0.000364
the', 'little', 'boy'	0.000485		at', 'this', 'end'	0.000331
the', 'vales', 'of'	0.000485		person', 'at', 'this'	0.000331
vales', 'of', 'har'	0.000485		the', 'person', 'at'	0.000331
an', 'infant', 'small'	0.000364		in', 'front', 'of'	0.000298
and', 'not', 'sit'	0.000364		kelanie', "'s", 'notepad'	0.000298
book', 'of', 'thel'	0.000364		that', 'the', "nosan'no'os"	0.000298
can', 'i', 'see'	0.000364		back', 'to', 'the'	0.000265
day', 'and', 'night'	0.000364		bureau', 'of', 'procuration'	0.000265
garden', 'of', 'love'	0.000364		i', 'do', "n't"	0.000265
heard', 'on', 'the'	0.000364		of', 'the', 'ship'	0.000265
human', 'form', 'divine'	0.000364		over', 'to', 'the'	0.000265
i', 'went', 'to'	0.000364		the', 'pthalklin', 'ervae'	0.000265
in', 'the', 'year'	0.000364		the', 'ship', "'s"	0.000265
little', 'boy', 'lost'	0.000364		do', 'you', 'think'	0.000232
o', 'little', 'cloud'	0.000364		end', 'of', 'the'	0.000232
on', 'the', 'green'	0.000364		kelanie', 'and', 'marek'	0.000232
pretty', 'rose', 'tree'	0.000364		of', 'her', 'suit'	0.000232
seen', 'on', 'the'	0.000364		carriage', 'return', 'line'	0.000199
songs', 'of', 'innocence'	0.000364		front', 'of', 'the'	0.000199
the', 'book', 'of'	0.000364		of', 'the', 'moridani'	0.000199
the', 'echoing', 'green'	0.000364		on', 'the', 'screen'	0.000199
the', 'garden', 'of'	0.000364		return', 'line', 'feed'	0.000199
the', 'human', 'dress'	0.000364		side', 'of', 'the'	0.000199

the', 'little', 'ones'	0.000364		the', 'side', 'of'	0.000199
the', 'sun', 'does'	0.000364		to', 'be', 'a'	0.000199
thel', 'is', 'like'	0.000364		was', 'about', 'to'	0.000199
to', 'welcome', 'in'	0.000364		we', 'do', "n't"	0.000199
welcome', 'in', 'the'	0.000364		what', "'s", 'going'	0.000199
who', 'made', 'thee'	0.000364		"'s", 'going', 'to'	0.000166
"'d", 'her', 'pitying'	0.000243		by', 'the', "nosan'no'os"	0.000166
"'ll", 'tell', 'thee'	0.000243		edge', 'of', 'the'	0.000166
"'s", 'song', 'when'	0.000243		i', "'d", 'like'	0.000166
a', 'divine', 'image'	0.000243		in', 'her', 'suit'	0.000166
a', 'happy', 'blossom'	0.000243		in', 'the', "nosan'no'os"	0.000166
a', 'human', 'face'	0.000243		into', 'the', 'ship'	0.000166
a', 'human', 'heart'	0.000243		of', 'kelanie', "'s"	0.000166
a', 'land', 'of'	0.000243		of', 'the', 'hatch'	0.000166
a', 'shade', "o'er"	0.000243		some', 'sort', 'of'	0.000166
a', 'summer', 'morn'	0.000243		something', 'like', 'a'	0.000166
all', 'the', 'livelong'	0.000243		the', 'front', 'of'	0.000166
and', 'builds', 'a'	0.000243		to', 'one', 'side'	0.000166
and', 'gives', 'his'	0.000243		to', 'the', 'control'	0.000166
and', 'i', 'am'	0.000243		to', 'the', 'ground'	0.000166
and', 'i', 'can'	0.000243		to', 'think', 'that'	0.000166

3. Discussion:

From the word frequency distributions for both documents, we can illustrate couple examples to show the style of the author and the type of documents that we have. In blake-poem.txt, we can see “thee”, “thou” and “thy” these archaic words that reveal the age of the document. The author used a lot of adjectives like “little”, “sweet”, “happy”, “green” and “bright”. These archaic words and adjective word are found in the top 50 words frequency. Besides, this poem seems to express positive message as “love”, “joy”, “sweet”, “happy” and “delight” are greatly appeared in the poem. On the other hand, The Bureau of Procuration (Story).txt contains more names and verbs. In the top 50 words frequency list, we can see “kelanie”, “marek”, “nosan no os” and “tsiry-feylen” show up a lot in the text. They seem like a person’s name. Furthermore, the text has some verbs, for instance “turned”, “going”, “think”, “found”, are listed in the top 50 words frequency. The document is more narrative with actions and characters or subjects.

Secondly, for bigram frequency, the results for both are rather similar. Most of the time, there is a word wraps in front of behind the word “the”. I tried constructing bigram with different filters. I tried the stopwords filter but the bigram list was meaningless because it contained comma followed by a word or vise versa. An example is shown below:

Teng Siong (T.S) Yeap
IST 664: Homework1

```
(('.', "''"), 0.002427479062993082)
(('.', '``'), 0.0020633572035441195)
(('lamb', ','), 0.001213739531496541)
(('love', ','), 0.001213739531496541)
(('night', ','), 0.001213739531496541)
(('sleep', ','), 0.001213739531496541)
(('!', "''"), 0.0010923655783468867)
((',', 'like'), 0.0010923655783468867)
(('day', ','), 0.0010923655783468867)
```

If I apply the alpha filter and the stopwords filter, the bigram frequency will not be accurate. So, to have a meaningful analysis, we use Pointwise Mutual Information (PMI). From blake-poem.txt, the PMI proves that the unigram or words frequency discussed above. The poem has a great number of adjectives to describe a subject. For example, “little boy” and “little lamb”. In contrast, the narrative story, The Bureau of Procuration (Story).txt has a lot of verbs like “pick up”, “staring at” and “glanced at”.

Lastly, I experimented trigram to see the result. There were several tokenization errors after we performed the trigram analysis. In blake-poem, we can see ("'d", 'her', 'pitying'), ("'ll", 'tell', 'thee') and ("'s", 'song', 'when') are the errors of tokenization. The “d”, “ll” and “s” are tokenized by the apostrophe. The same thing goes to The Bureau of Procuration (Story).txt. We found (we', 'do', "n't"), (what', "'s", 'going') and (i', "'d", 'like'). Trigram also reveals more on the style of writing of the author. As mentioned before, poem has a lot of descriptive words. For instance, (a', 'summer', 'morn'), (a', 'divine', 'image'), (a', 'happy', 'blossom'), (a', 'human', 'face') and (a', 'human', 'heart'). On the other hand, the trigram analysis in The Bureau of Procuration (Story).txt showed some directional phrases. For example, (to', 'one', 'side'), (to', 'the', 'control'), (to', 'the', 'ground').

In conclusion, words frequency, bigram frequency, pointwise mutual information, trigram frequency or N-gram frequency are useful analysis to analyze a text. In this case, we see that the poem uses a lot of adjectives to describe a subject and positive words to convey a up-listing content. In contrast, we find a lot of verbs and names in the narrative story.