Task 1:

	Text1	Text3	Text3
10 frequently occuring Bigrams	[('[', 'Moby'), ('Moby', 'Dick'), ('Dick', 'by'), ('by', 'Herman'), ('Herman', 'Melville'), ('Melville', '1851'), ('1851', ']'), (']', 'ETYMOLOGY'), ('ETYMOLOGY', '.'), ('.', '(')]	[('[', 'Sense'), ('Sense', 'and'), ('and', 'Sensibility'), ('Sensibility', 'by'), ('by', 'Jane'), ('Jane', 'Austen'), ('Austen', '1811'), ('1811', ']'), (']', 'CHAPTER'), ('CHAPTER', '1')]	[('In', 'the'), ('the', 'beginning'), ('beginning', 'God'), ('God', 'created'), ('created', 'the'), ('the', 'heaven'), ('heaven', 'and'), ('and', 'the'), ('the', 'earth'), ('earth', '.')]
5 frequently occuring Trigrams	[('AFTER', 'EXCHANGING', 'HAILS'), ('Anacharsis', 'Clootz', 'deputation'), ('CAULKING', 'ITS', 'SEAMS'), ('ELIZABETH', 'OAKES', 'SMITH'), ('Et', 'tu', 'Brute')]	[('Austen', '1811', ']'), ('Jane', 'Austen', '1811'), ('200', 'L', 'per'), ('Drury', 'Lane', 'lobby'), ('L', 'per', 'annum')]	[('olive', 'leaf', 'pluckt'), ('sewed', 'fig', 'leaves'), ('yield', 'royal', 'dainties'), ('Fifteen', 'cubits', 'upward'), ('leaf', 'pluckt', 'o')]
Number of words with length > 16	11	3	0
Number of words with frequency > 500	58	54	54
Number of words ending in “ed”	2196	897	275

Task 2:

```
nltk.corpus.gutenberg.fileids()
```

```
['austen-emma.txt',
 'austen-persuasion.txt',
 'austen-sense.txt',
 'bible-kjv.txt',
 'blake-poems.txt',
 'bryant-stories.txt',
 'burgess-busterbrown.txt',
 'carroll-alice.txt',
 'chesterton-ball.txt',
 'chesterton-brown.txt',
 'chesterton-thursday.txt',
 'edgeworth-parents.txt',
 'melville-moby_dick.txt',
 'milton-paradise.txt',
 'shakespeare-caesar.txt',
 'shakespeare-hamlet.txt',
 'shakespeare-macbeth.txt',
 'whitman-leaves.txt']
```

```
gutenberg_sc = nltk.corpus.gutenberg.words('shakespeare-caesar.txt')
```

```
from nltk.corpus import brown
brown.words()
```

```
['The', 'Fulton', 'County', 'Grand', 'Jury', 'said', ...]
```

```
/home/pheonix/University/Semester-7/NLP Tasks/Tasks/nltk_data/corpora
```

```
fai_txt = nltk.data.load('/home/pheonix/University/Semester-7/NLP Tasks/Tasks/nltk_data/corpora/faizan.txt', format ='raw')
fai_txt = nltk.data.load('/home/pheonix/University/Semester-7/NLP Tasks/Tasks/nltk_data/corpora/faizan.txt', format ='text')
```

```
from nltk.util import ngrams
words = nltk.word_tokenize(fai_txt)
fai_bigrams = list(ngrams(words, 2))
fai_trigrams = list(ngrams(words, 3))
```

```
print(fai_bigrams[:50])
```

```
[('PM', 'denies'), ('denies', 'knowledge'), ('knowledge', 'of'), ('of', 'AWB'), ('AWB', 'kickbacks'), ('kickbacks', 'The'), ('The', 'Pri
', 'writing'), ('writing', 'to'), ('to', 'the'), ('the', 'wheat'), ('wheat', 'exporter'), ('exporter', 'asking'), ('asking', 'to'), ('to
and'), ('and', 'Deputy'), ('Deputy', 'Prime'), ('Prime', 'Minister'), ('Minister', 'Mark'), ('Mark', 'Vaile'), ('Vaile', 'to'), ('to', '
```

```
print(fai_trigrams[:50])
```

```
[('PM', 'denies', 'knowledge'), ('denies', 'knowledge', 'of'), ('knowledge', 'of', 'AWB'), ('of', 'AWB', 'kickbacks'), ('AWB', 'kickback
paying'), ('was', 'paying', 'kickbacks'), ('paying', 'kickbacks', 'to'), ('kickbacks', 'to', 'Iraq'), ('to', 'Iraq', 'despite'), ('Iraq'
', 'kept', 'fully'), ('kept', 'fully', 'informed'), ('fully', 'informed', 'on'), ('informed', 'on', 'Iraq'), ('on', 'Iraq', 'wheat'), ('
eputy', 'Prime', 'Minister'), ('Prime', 'Minister', 'Mark'), ('Minister', 'Mark', 'Vaile'), ('Mark', 'Vaile', 'to'), ('Vaile', 'to', 'AW
```

```
from nltk.corpus import PlaintextCorpusReader
import os
corpus_root = os.path.expanduser('/home/pheonix/University/Semester-7/NLP Tasks/Tasks/nltk_data/corpora/abc')
corpus = PlaintextCorpusReader(corpus_root, '.*', encoding='latin1')
wordlist = corpus.words()
bigramlist = list(ngrams(wordlist,2))
```

```
print(bigramlist[:50])
```

```
[('Australian', 'Broadcasting'), ('Broadcasting', 'Commission'), ('Commission', '2006'), ('2006', 'http'), ('http', '://'), ('://', 'www
'.'), ('.', 'abc'), ('abc', '.'), ('.', 'net'), ('net', '.'), ('.', 'au'), ('au', '/'), ('/', 'rural'), ('rural', '/'), ('/', 'news'), (
('/', 'news'), ('news', '/'), ('/', 'PM')]
```

Task 3:

<table>
<tr><th>Number of Words in Corpus</th><th>30 word generated sentence</th></tr>
<tr><td>766863</td><td>Broadcasting Commission (NFF) and the first time , the first time , the first time , the first time , the first time , the first time ,</td></tr>
<tr><td>13795</td><td>various files which make up the | | | | | | | | | | | | | | | | | | | | | | | |</td></tr>
</table>