



Name: Muhammad Ghazanfar

Roll#: 20P-0567

Sec: 7A

Course: NLP

Lab#02

# Task 1: Bigrams & Trigrams

## Task#01

```
import nltk
```

[1] ✓ 5.7s Python

```
from nltk.book import *
```

[2] ✓ 8.4s Python

```
... *** Introductory Examples for the NLTK Book ***
Loading text1, ..., text9 and sent1, ..., sent9
Type the name of the text or sentence to view it.
Type: 'texts()' or 'sents()' to list the materials.
text1: Moby Dick by Herman Melville 1851
text2: Sense and Sensibility by Jane Austen 1811
text3: The Book of Genesis
text4: Inaugural Address Corpus
text5: Chat Corpus
text6: Monty Python and the Holy Grail
text7: Wall Street Journal
text8: Personals Corpus
text9: The Man Who Was Thursday by G . K . Chesterton 1908
```

```
words = sorted(set(text1))[200:]
```

[3] ✓ 0.1s Python

```
longwords = [w for w in words if len(w) > 16]
print(longwords)
```

[4] ✓ 0.0s Python

```
... ['cannibalistically', 'characteristically', 'circumnavigations', 'comprehensiveness', 'indispensableness', 'preternaturalness', 'subterraneous']
```

```
fdist1 = FreqDist(text3)
high_freq = [w for w in words if fdist1[w] > 500]
print(high_freq)
```

[5] ✓ 0.0s Python

```
... ['And', 'and', 'he', 'his', 'in', 'of', 'that', 'the', 'to', 'unto']
```

```
# high_freq_idf = [w for w in words if IDF(text1,w) > 1]
# print(high_freq_idf)
```

[7] ✓ 0.0s Python

```
eign_words = [w for w in words if w.endswith('eign')]
print(eign_words)
```

[8] ✓ 0.0s Python

```
... ['Sovereign', 'foreign', 'reign', 'sovereign']
```

```
for w in words:
    if w.endswith('eign'):
        print(w)
```

[9] ✓ 0.0s Python

```
... Sovereign
foreign
reign
sovereign
```

```
list(bigrams(text1))[:10]
```

[10] ✓ 2.5s Python

```
... [('I', 'Moby'),
('Moby', 'Dick'),
('Dick', 'by'),
```

```

[11] ✓ 0.9s Python
... text1.collocations()
Sperm Whale; Moby Dick; White Whale; old man; Captain Ahab; sperm
whale; Right Whale; Captain Peleg; New Bedford; Cape Horn; cried Ahab;
years ago; lower jaw; never mind; Father Mapple; cried Stubb; chief
mate; white whale; ivory leg; one hand

[12] ✓ 0.2s Python
... from nltk.util import ngrams
list(ngrams(text1, 3))
[(' ', 'Moby', 'Dick'),
 ('Moby', 'Dick', 'by'),
 ('Dick', 'by', 'Herman'),
 ('by', 'Herman', 'Melville'),
 ('Herman', 'Melville', '1851'),
 ('Melville', '1851', ' '),
 ('1851', ' ', 'ETYMOLOGY'),

```

```

[13] ✓ 0.0s Python
... from nltk.collocations import *
from nltk.collocations import TrigramCollocationFinder, TrigramAssocMeasures

[14] ✓ 0.5s Python
... TrigramCollocationFinder.from_words(text3).nbest(TrigramAssocMeasures().pmi, 5)
[('olive', 'leaf', 'pluckt'),
 ('sewed', 'fig', 'leaves'),
 ('yield', 'royal', 'dainties'),
 ('Fifteen', 'cubits', 'upward'),
 ('leaf', 'pluckt', 'o')]

[15] ✓ 0.0s Python
... longwords = [w for w in words if len(w) > 16]

[16] ✓ 0.0s Python
... print(longwords)
['cannibalistically', 'characteristically', 'circumnavigations', 'comprehensiveness', 'indispensableness', 'preternaturalness', 'subterraneous

```

```

[17] ✓ 0.0s Python
... words = sorted(set(text3))

[18] ✓ 0.0s Python
... ed_words = [w for w in words if w.endswith('ed')]

[19] ✓ 0.0s Python
... len(ed_words)
281

```

|  | Text1   | Text2   | Text3   |
|--|---|---|---|
| <b>10 frequently occurring Bigrams</b>         | Sperm Whale; Moby Dick; White Whale; old man; Captain Ahab; sperm whale; Right Whale; Captain Peleg; New Bedford; Cape Horn;  | Colonel Brandon; Sir John; Lady Middleton; Miss Dashwood; every thing; thousand pounds; dare say; Miss Steeles; said Elinor; Miss Steele; | said unto; pray thee; thou shalt; thou hast; thy seed; years old; spake unto; thou art; LORD God; every living;                                     |
| <b>5 frequently occurring Trigrams</b>         | [('AFTER', 'EXCHANGING', 'HAILS'), ('Anacharsis', 'Cloutz', 'deputation'), ('CAULKING', 'ITS', 'SEAMS'), ('ELIZABETH', 'OAKES', 'SMITH'), ('Et', 'tu', 'Brute')]  | [('Austen', '1811', 'I'), ('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')] |
| <b>Number of words with length &gt; 16</b>     | ['cannibalistically', 'characteristically', 'circumnavigations', 'comprehensiveness', 'indispensableness', 'preternaturalness', 'subterraneousness', 'superstitiousness', 'uncomfortableness', 'uncompromisedness', 'uninterpenetratingly'] | ['companionableness', 'disinterestedness', 'disqualifications']   | [[] NO words greater than len 16  |
| <b>Number of words with frequency &gt; 500</b> | ['Ahab', 'But', 'I', 'The', 'a', 'all', 'an', 'and', 'are', 'as', 'at', 'be', 'but', 'by', 'for', 'from', 'had', 'have', 'he', 'him', 'his', 'in', 'into', 'is', 'it', 'like', 'man', 'me', 'more',   | ['"', 'I', 'a', 'all', 'and', 'as', 'at', 'be', 'but', 'by', 'could', 'for', 'from', 'had', 'have', 'he', 'her', 'him',                   | [',', 'and', 'he', 'his', 'in', 'of', 'that', 'the', 'to', 'unto']  |



```
[24] ✓ 3.7s Python
from nltk.util import ngrams
words = nltk.word_tokenize(psh_txt)
psh_bigrams = list(ngrams(words, 2))
psh_trigrams = list(ngrams(words, 3))

import os
from nltk import ngrams
from nltk.corpus import PlaintextCorpusReader

corpus_root = os.path.expanduser('/home/ghazanfar/nltk_data/corpora/abc')
corpus = PlaintextCorpusReader(corpus_root, '.*', encoding='latin-1')
wordlist = corpus.words()
bigramlist = list(ngrams(wordlist, 2))
print(bigramlist)

[26] ✓ 2.6s Python
... [['Australian
```

## Task 3: Generating Random Text with Bigrams

```
Task#03

[27] ✓ 16.5s Python
cfd = nltk.ConditionalFreqDist(bigramlist)

[28] ✓ 0.0s Python
print(cfd)
... <ConditionalFreqDist with 31893 conditions>

[32] ✓ 0.0s Python
def generate_model(cfdist, word, num):
    for i in range(num):
        print(word, end=' ')
        word = cfdist[word].max()

[37] ✓ 0.0s Python
generate_model(cfd, 'Australia', 30)
... Australia 's a new study , the first time , the first time , the first time , the first time , the first time , the first time

Ln 2, Col 64 Cell 32 of 34 Go Live Prettier
```

```
[34] ✓ 0.0s Python
len(wordlist)
... 766863
```

| Number of Words In Corpus | 30 word generated sentence   |
|---------------------------|--|
| 766863                    | Today , the first time , the first time , the first time , the first time , the first time , the first time , the first time , |
| 766863                    | <b>Australia ' s a new study , the first time , the first time , the first time , the first time , the first time</b>          |

