# Report on Frequency Distributions of Unigrams and bigrams

**Sai Phani Ram Popuri - 2205577**
MS Data Science, University of Houston
Conceptualization, Formal Analysis, Review and Editing
spopuri2@cougarnet.uh.edu

**Prathima Mettu - 2160335**
MS Data Science, University of Houston
Methodology, Writing Original Draft, Supervision
mettu.prathima@uh.edu

### Abstract

This document contains the instructions for preparing a text data which undergoes the preprocessing and to analyze word distributions in two different corpora using Python and NLTK.

## 1 Introduction

The purpose of this project is to analyze word distributions in two different corpora using Python and Natural Language Toolkit (NLTK). The two corpora are stored in separate directories, and the word distributions for the respective corpora have been analyzed. We will then answer several questions regarding the top 30 most common words in each corpus and run the script with different values for the k most frequent words to document our findings.

## 2 Methodology

The source code has been written in Python programming language. Various inbuilt modules from the NLTK library such as word_tokenize, stopwords, PorterStemmer, FreqDist, and ngrams have been used to perform various preprocessing steps and infer the topic to which the corpus belongs to. The following stages are involved in the entire project pipeline:

A. **Staging** : Navigate to the directory path and walk through all the existing text files and fetch the raw corpus. The python script has been optimized to speed up reading.

B. **Pre-processing**: It involves various stages of cleaning and formatting of data such as:

   a. Removing the punctuation and special characters from the corpus.
   b. Converting the text into lower case.
   c. Tokenization of corpus.
   d. Collecting stop words.
   e. Stemming the respective tokens and removal of stop words

C. **Experimentation** :

a. Compute k frequency distribution for uni-gram of a corpus .
b. Compute k frequency distribution bi-grams of a corpus.

Once the processed corpora is obtained, the aforementioned steps have been performed by varying the value of ‘K’, which denotes the count of frequent occurrence.

The frequency distributions of the unigrams and the bigrams have been assessed by considering various combinations of stemming and stop words removal. The stop words have been further classified based on their language of origin. The major portion of the work aims at removing the irrelevant words from the ‘English’ language. But, there might be a possibility of having tokens from multiple languages in the corpora. To handle this, we have even performed our analysis by removing the global stop words to check the impact. The stemming and stop word removal combinations have been tabulated below.

| *Stemming* | *Stop word removal* |
|---|---|
| True | False |
| True | True |
| False | True |
| False | False |

## 3 Experimental Results

Once the processed tokens are obtained, the frequency distributions of the unigrams and the bigrams have been computed. The results are as follows:

### Corpus 1: Unigrams

#### a. *Stemming with no stop words removal.*

Top 30 words:

```
[('the', 633529), ('and', 357113), ('to', 295245), ('of', 269229), ('a', 231626), ('i', 166049), ('in',
153709), ('he', 149938), ('wa', 141377), ('it', 133930), ('that', 129274), ('you', 110554), ('hi',
101848), ('as', 95162), ('had', 89905), ('for', 86708), ('with', 84394), ('at', 70806), ('be', 70470),
('but', 69777), ('on', 68141), ('not', 64546), ('they', 60766), ('said', 60590), ('have', 57117), ('is',
54796), ('her', 52130), ('were', 49514), ('him', 49481), ('we', 47192)]
```

Top 50 words:

```
[('the', 633529), ('and', 357113), ('to', 295245), ('of', 269229), ('a', 231626), ('i', 166049), ('in',
153709), ('he', 149938), ('wa', 141377), ('it', 133930), ('that', 129274), ('you', 110554), ('hi',
101848), ('as', 95162), ('had', 89905), ('for', 86708), ('with', 84394), ('at', 70806), ('be', 70470),
('but', 69777), ('on', 68141), ('not', 64546), ('they', 60766), ('said', 60590), ('have', 57117), ('is',
54796), ('her', 52130), ('were', 49514), ('him', 49481), ('we', 47192), ('she', 45367), ('do', 43432),
('all', 43343), ('by', 43312), ('my', 40931), ('so', 39461), ('there', 38559), ('which', 37556), ('from',
37348), ('up', 37097), ('no', 36654), ('one', 36632), ('me', 35835), ('would', 35707), ('thi', 35321),
('their', 34996), ('if', 33635), ('them', 33177), ('out', 31459), ('when', 31440)]
```

Top 70 words:

```
 [('the', 633529), ('and', 357113), ('to', 295245), ('of', 269229), ('a', 231626), ('i', 166049), ('in',
153709), ('he', 149938), ('wa', 141377), ('it', 133930), ('that', 129274), ('you', 110554), ('hi',
101848), ('as', 95162), ('had', 89905), ('for', 86708), ('with', 84394), ('at', 70806), ('be', 70470),
('but', 69777), ('on', 68141), ('not', 64546), ('they', 60766), ('said', 60590), ('have', 57117), ('is',
54796), ('her', 52130), ('were', 49514), ('him', 49481), ('we', 47192), ('she', 45367), ('do', 43432),
('all', 43343), ('by', 43312), ('my', 40931), ('so', 39461), ('there', 38559), ('which', 37556), ('from',
37348), ('up', 37097), ('no', 36654), ('one', 36632), ('me', 35835), ('would', 35707), ('thi', 35321),
('their', 34996), ('if', 33635), ('them', 33177), ('out', 31459), ('when', 31440), ('been', 30561),
('will', 30522), ('what', 28826), ('or', 28789), ('an', 28539), ('are', 27500), ('who', 27099), ('go',
26227), ('then', 26139), ('could', 25451), ('now', 23880), ('time', 23002), ('down', 22834), ('come',
22063), ('littl', 21798), ('look', 21725), ('your', 21723), ('did', 21403), ('into', 20962), ('some',
20590)]
```

The above output show the most frequently occurring words in Corpus 1 without the removal of stop words. These individual tokens will not contribute to the goal of annotating the topic the corpus1 belongs to. Therefore, it is impossible to arrive at the descriptor based on the obtained tokens.

#### *b. Stemming with stop words removal.*

Top 30 words:

```
[('said', 60590), ('one', 36632), ('would', 35707), ('go', 26227), ('could', 25451), ('time', 23002),
('come', 22063), ('littl', 21798), ('look', 21725), ('see', 20202), ('like', 20109), ('say', 18207),
('know', 17738), ('well', 17264), ('man', 17044), ('two', 16864), ('boy', 15485), ('upon', 15412),
('get', 15193), ('day', 15165), ('hand', 14739), ('way', 14661), ('men', 14323), ('think', 14314),
('back', 14272), ('good', 14179), ('make', 14129), ('take', 14017), ('us', 13981), ('made', 13335)]
```

Top 50 words:

```
[('said', 60590), ('one', 36632), ('would', 35707), ('go', 26227), ('could', 25451), ('time',
23002), ('come', 22063), ('littl', 21798), ('look', 21725), ('see', 20202), ('like', 20109),
('say', 18207), ('know', 17738), ('well', 17264), ('man', 17044), ('two', 16864), ('boy',
15485), ('upon', 15412), ('get', 15193), ('day', 15165), ('hand', 14739), ('way', 14661),
('men', 14323), ('think', 14314), ('back', 14272), ('good', 14179), ('make', 14129), ('take',
14017), ('us', 13981), ('made', 13335), ('came', 13040), ('much', 12639), ('great', 12445),
('place', 12331), ('must', 12270), ('long', 12238), ('seem', 11966), ('old', 11943), ('away',
11719), ('went', 11624), ('work', 11429), ('ask', 10679), ('never', 10656), ('ye', 10563),
('even', 10443), ('cri', 10237), ('may', 10156), ('first', 10119), ('got', 10057), ('thought',
9994)]
```

Top 70 words:

```
[('said', 60590), ('one', 36632), ('would', 35707), ('go', 26227), ('could', 25451), ('time',
23002), ('come', 22063), ('littl', 21798), ('look', 21725), ('see', 20202), ('like', 20109),
('say', 18207), ('know', 17738), ('well', 17264), ('man', 17044), ('two', 16864), ('boy',
15485), ('upon', 15412), ('get', 15193), ('day', 15165), ('hand', 14739), ('way', 14661),
('men', 14323), ('think', 14314), ('back', 14272), ('good', 14179), ('make', 14129), ('take',
14017), ('us', 13981), ('made', 13335), ('came', 13040), ('much', 12639), ('great', 12445),
('place', 12331), ('must', 12270), ('long', 12238), ('seem', 11966), ('old', 11943), ('away',
11719), ('went', 11624), ('work', 11429), ('ask', 10679), ('never', 10656), ('ye', 10563),
('even', 10443), ('cri', 10237), ('may', 10156), ('first', 10119), ('got', 10057), ('thought',
9994), ('tell', 9893), ('turn', 9829), ('right', 9634), ('head', 9590), ('father', 9477), ('last',
9468), ('thing', 9450), ('might', 9319), ('shall', 9114), ('captain', 9075), ('want', 8992),
('give', 8803), ('let', 8698), ('sir', 8694), ('eye', 8614), ('put', 8453), ('young', 8421),
('face', 8354), ('water', 7896), ('de', 7864)]
```

The above snippet shows the most frequently occurring words in Corpus 1 by removing stop words. It is evident that stop word removal has reduced the ambiguity to a greater extent. A quick glance at the tokens indicates that the words ***“boy”, “water”, “father”, “captain”, “young”, “men”, “work”*** etc., might speak of an anecdote of a young boy and the relation with his father who is a navy captain.

**Note**: When running the script with different combinations of stemming and stop word removal, it has been that *stemming has a little impact on the generated tokens whereas the stop words played a significant role in arriving at the corpus description.*

### Corpus 2: Unigrams

#### *a. Stemming with no stop words removal.*

Top 30 words:

```
[('the', 553282), ('and', 346435), ('to', 318079), ('of', 308939), ('a', 241644), ('in',
170173), ('i', 169126), ('he', 160986), ('wa', 151919), ('it', 142051), ('that', 141118),
('her', 133580), ('you', 128770), ('hi', 124640), ('she', 108947), ('had', 105154), ('with',
98676), ('as', 93591), ('not', 92031), ('for', 91467), ('is', 78812), ('be', 77662), ('but',
76177), ('at', 69427), ('have', 68626), ('him', 58743), ('on', 54738), ('by', 53079),
('which', 49013), ('said', 47618)]
```

Top 50 words:

```
[('the', 553282), ('and', 346435), ('to', 318079), ('of', 308939), ('a', 241644), ('in',
170173), ('i', 169126), ('he', 160986), ('wa', 151919), ('it', 142051), ('that', 141118),
('her', 133580), ('you', 128770), ('hi', 124640), ('she', 108947), ('had', 105154), ('with',
98676), ('as', 93591), ('not', 92031), ('for', 91467), ('is', 78812), ('be', 77662), ('but',
76177), ('at', 69427), ('have', 68626), ('him', 58743), ('on', 54738), ('by', 53079),
('which', 49013), ('said', 47618), ('all', 46834), ('so', 46653), ('my', 44692), ('thi', 44074),
('me', 42817), ('do', 42735), ('would', 41429), ('from', 40148), ('no', 39642), ('one',
36776), ('if', 36499), ('were', 36042), ('they', 36019), ('what', 35999), ('been', 35639),
('there', 35616), ('who', 32948), ('an', 31473), ('will', 30714), ('or', 30599)]
```

Top 70 words:

```
[('the', 553282), ('and', 346435), ('to', 318079), ('of', 308939), ('a', 241644), ('in', 170173), ('i',
169126), ('he', 160986), ('wa', 151919), ('it', 142051), ('that', 141118), ('her', 133580), ('you',
128770), ('hi', 124640), ('she', 108947), ('had', 105154), ('with', 98676), ('as', 93591), ('not',
92031), ('for', 91467), ('is', 78812), ('be', 77662), ('but', 76177), ('at', 69427), ('have', 68626),
('him', 58743), ('on', 54738), ('by', 53079), ('which', 49013), ('said', 47618), ('all', 46834), ('so',
46653), ('my', 44692), ('thi', 44074), ('me', 42817), ('do', 42735), ('would', 41429), ('from',
40148), ('no', 39642), ('one', 36776), ('if', 36499), ('were', 36042), ('they', 36019), ('what',
35999), ('been', 35639), ('there', 35616), ('who', 32948), ('an', 31473), ('will', 30714), ('or',
30599), ('when', 30594), ('are', 29296), ('could', 28368), ('your', 27394), ('more', 24931), ('veri',
24277), ('out', 23608), ('their', 23477), ('we', 23330), ('like', 23292), ('look', 23020), ('know',
23003), ('up', 22695), ('them', 22605), ('then', 22372), ('did', 22365), ('now', 22139), ('littl',
21916), ('man', 20879), ('time', 20673)]
```

The above snippet show the most frequently occurring words in Corpus 2 without the removal of stop words. Based on these, it is a bit ambiguous as to what topic the corpus2 belongs to. Therefore, it is impossible to arrive at the descriptor based on the obtained tokens.

### ***b. Stemming with stop words removal.***

Top 30 words:

```
[('said', 47618), ('would', 41429), ('one', 36776), ('could', 28368), ('like', 23292), ('look', 23020),
('know', 23003), ('littl', 21916), ('man', 20879), ('time', 20673), ('come', 19843), ('go', 19313),
('say', 19145), ('see', 18537), ('upon', 18437), ('think', 18373), ('never', 17598), ('old', 16478),
('must', 16308), ('well', 16133), ('much', 15887), ('hand', 15820), ('ladi', 15562), ('thought',
15427), ('good', 15149), ('even', 14636), ('day', 14518), ('make', 14187), ('made', 13819), ('love',
13532)]
```

Top 50 words:

```
[('said', 47618), ('would', 41429), ('one', 36776), ('could', 28368), ('like', 23292), ('look', 23020),
('know', 23003), ('littl', 21916), ('man', 20879), ('time', 20673), ('come', 19843), ('go', 19313),
('say', 19145), ('see', 18537), ('upon', 18437), ('think', 18373), ('never', 17598), ('old', 16478),
('must', 16308), ('well', 16133), ('much', 15887), ('hand', 15820), ('ladi', 15562), ('thought',
15427), ('good', 15149), ('even', 14636), ('day', 14518), ('make', 14187), ('made', 13819), ('love',
13532), ('eye', 13336), ('way', 13132), ('seem', 12979), ('might', 12550), ('take', 12540), ('young',
12437), ('miss', 12119), ('life', 12080), ('thing', 11962), ('ask', 11623), ('face', 11413), ('great',
11181), ('tell', 11042), ('long', 10969), ('noth', 10924), ('sir', 10911), ('two', 10828), ('first',
10592), ('word', 10514), ('back', 10441)]
```

Top 70 words:

```
[('said', 47618), ('would', 41429), ('one', 36776), ('could', 28368), ('like', 23292), ('look', 23020),
('know', 23003), ('littl', 21916), ('man', 20879), ('time', 20673), ('come', 19843), ('go', 19313),
('say', 19145), ('see', 18537), ('upon', 18437), ('think', 18373), ('never', 17598), ('old', 16478),
('must', 16308), ('well', 16133), ('much', 15887), ('hand', 15820), ('ladi', 15562), ('thought',
15427), ('good', 15149), ('even', 14636), ('day', 14518), ('make', 14187), ('made', 13819), ('love',
13532), ('eye', 13336), ('way', 13132), ('seem', 12979), ('might', 12550), ('take', 12540), ('young',
12437), ('miss', 12119), ('life', 12080), ('thing', 11962), ('ask', 11623), ('face', 11413), ('great',
11181), ('tell', 11042), ('long', 10969), ('noth', 10924), ('sir', 10911), ('two', 10828), ('first',
10592), ('word', 10514), ('back', 10441), ('father', 10347), ('may', 10264), ('came', 10161),
('shall', 10151), ('though', 10061), ('turn', 10042), ('last', 9898), ('away', 9840), ('place', 9782),
('still', 9571), ('friend', 9409), ('hous', 9385), ('heart', 9288), ('us', 9198), ('without', 9158), ('yet',
9022), ('ye', 9012), ('dear', 9010), ('want', 9001), ('mother', 8967)]
```

The above snippet shows the most frequently occurring words in Corpus 2 by removing stop words. From the output tokens,  it can be seen that stemming and stop word removal reduced the ambiguity to a greater extent. A quick glance at the tokens indicates that the words ***“father”, “mother”, “love”, “heart”, “life”***, etc., might speak of an of family relationships and emotions.

**Note**: When running the script with different combinations of stemming and stop word removal, we witnessed that stemming has a little impact on the generated tokens whereas the stop words played a significant role in arriving at the corpus description.

## Corpus 1 - Bigrams

### ***a. Stemming with no stop words removal.***

Top 30 bi-grams:

```
[(('of', 'the'), 78767), (('in', 'the'), 43328), (('to', 'the'), 36874), (('and', 'the'), 25528),
(('on', 'the'), 22783), (('it', 'wa'), 22611), (('at', 'the'), 19210), (('to', 'be'), 19114), (('for',
'the'), 16099), (('he', 'wa'), 15348), (('he', 'had'), 15004), (('in', 'a'), 14776), (('with', 'a'),
14086), (('had', 'been'), 13878), (('with', 'the'), 13666), (('from', 'the'), 13431), (('by',
'the'), 13310), (('and', 'i'), 13029), (('that', 'the'), 12646), (('wa', 'a'), 11851), (('of', 'a'),
11683), (('it', 'is'), 11538), (('of', 'hi'), 11386), (('as', 'he'), 11257), (('that', 'he'), 10591),
(('there', 'wa'), 10008), (('and', 'then'), 9616), (('into', 'the'), 9397), (('i', 'do'), 9244),
(('for', 'a'), 9199)]
```

Top 50 bi-grams:

```
[(('of', 'the'), 78767), (('in', 'the'), 43328), (('to', 'the'), 36874), (('and', 'the'), 25528),
(('on', 'the'), 22783), (('it', 'wa'), 22611), (('at', 'the'), 19210), (('to', 'be'), 19114), (('for',
'the'), 16099), (('he', 'wa'), 15348), (('he', 'had'), 15004), (('in', 'a'), 14776), (('with', 'a'),
14086), (('had', 'been'), 13878), (('with', 'the'), 13666), (('from', 'the'), 13431), (('by',
'the'), 13310), (('and', 'i'), 13029), (('that', 'the'), 12646), (('wa', 'a'), 11851), (('of', 'a'),
11683), (('it', 'is'), 11538), (('of', 'hi'), 11386), (('as', 'he'), 11257), (('that', 'he'), 10591),
(('there', 'wa'), 10008), (('and', 'then'), 9616), (('into', 'the'), 9397), (('i', 'do'), 9244),
(('for', 'a'), 9199), (('i', 'have'), 9120), (('one', 'of'), 8936), (('he', 'said'), 8616), (('and',
'he'), 8553), (('i', 'am'), 8466), (('they', 'were'), 8179), (('have', 'been'), 8043), (('i', 'wa'),
7784), (('the', 'other'), 7710), (('did', 'not'), 7699), (('as', 'the'), 7694), (('and', 'a'), 7619),
(('said', 'the'), 7486), (('in', 'hi'), 7366), (('would', 'be'), 7354), (('out', 'of'), 7299), (('but',
'i'), 7195), (('all', 'the'), 7030), (('that', 'i'), 6756), (('seem', 'to'), 6738)]
```

Top 70 bi-grams:

```
[(('of', 'the'), 78767), (('in', 'the'), 43328), (('to', 'the'), 36874), (('and', 'the'), 25528),
(('on', 'the'), 22783), (('it', 'wa'), 22611), (('at', 'the'), 19210), (('to', 'be'), 19114), (('for',
'the'), 16099), (('he', 'wa'), 15348), (('he', 'had'), 15004), (('in', 'a'), 14776), (('with', 'a'),
14086), (('had', 'been'), 13878), (('with', 'the'), 13666), (('from', 'the'), 13431), (('by',
'the'), 13310), (('and', 'i'), 13029), (('that', 'the'), 12646), (('wa', 'a'), 11851), (('of', 'a'),
11683), (('it', 'is'), 11538), (('of', 'hi'), 11386), (('as', 'he'), 11257), (('that', 'he'), 10591),
(('there', 'wa'), 10008), (('and', 'then'), 9616), (('into', 'the'), 9397), (('i', 'do'), 9244),
(('for', 'a'), 9199), (('i', 'have'), 9120), (('one', 'of'), 8936), (('he', 'said'), 8616), (('and',
'he'), 8553), (('i', 'am'), 8466), (('they', 'were'), 8179), (('have', 'been'), 8043), (('i', 'wa'),
7784), (('the', 'other'), 7710), (('did', 'not'), 7699), (('as', 'the'), 7694), (('and', 'a'), 7619),
(('said', 'the'), 7486), (('in', 'hi'), 7366), (('would', 'be'), 7354), (('out', 'of'), 7299), (('but',
'i'), 7195), (('all', 'the'), 7030), (('that', 'i'), 6756), (('seem', 'to'), 6738), (('go', 'to'), 6483),
(('to', 'hi'), 6424), (('a', 'few'), 6386), (('a', 'littl'), 6296), (('they', 'had'), 6202), (('do',
'you'), 6165), (('you', 'are'), 6153), (('to', 'do'), 6049), (('if', 'you'), 6043), (('wa', 'the'),
6040), (('wa', 'not'), 5968), (('i', 'had'), 5924), (('the', 'boy'), 5888), (('and', 'that'), 5828),
(('would', 'have'), 5791), (('at', 'onc'), 5769), (('to', 'see'), 5673), (('but', 'the'), 5579), (('as',
'if'), 5454), (('will', 'be'), 5325)]
```

These individual tokens will not contribute to the goal of annotating the topic the corpus2 belongs to. Therefore, it is impossible to arrive at the descriptor based on the obtained tokens.

Case 2: Stemming with stop words removal.

Top 30 2-grams:

```
[(('brer', 'rabbit'), 1599), (('could', 'see'), 1595), (('old', 'man'), 1506), (('come', 'back'),
1210), (('two', 'three'), 1186), (('sir', 'said'), 1100), (('let', 'us'), 1079), (('said', 'mr'),
1046), (('good', 'deal'), 1041), (('ye', 'said'), 1027), (('project', 'gutenberg'), 1018),
(('brer', 'fox'), 980), (('next', 'day'), 977), (('said', 'captain'), 975), (('next', 'morn'), 958),
(('let', 'go'), 957), (('one', 'day'), 953), (('everi', 'one'), 934), (('electron', 'work'), 918),
(('young', 'man'), 905), (('said', 'doctor'), 884), (('littl', 'girl'), 864), (('go', 'back'), 854),
(('uncl', 'remu'), 846), (('said', 'uncl'), 829), (('could', 'get'), 816), (('half', 'hour'), 807),
(('great', 'deal'), 806), (('look', 'like'), 788), (('would', 'like'), 787)]
```

Top 50 2-grams:

```
[(('brer', 'rabbit'), 1599), (('could', 'see'), 1595), (('old', 'man'), 1506), (('come', 'back'), 1210),
(('two', 'three'), 1186), (('sir', 'said'), 1100), (('let', 'us'), 1079), (('said', 'mr'), 1046), (('good',
'deal'), 1041), (('ye', 'said'), 1027), (('project', 'gutenberg'), 1018), (('brer', 'fox'), 980), (('next',
'day'), 977), (('said', 'captain'), 975), (('next', 'morn'), 958), (('let', 'go'), 957), (('one', 'day'), 953),
(('everi', 'one'), 934), (('electron', 'work'), 918), (('young', 'man'), 905), (('said', 'doctor'), 884),
(('littl', 'girl'), 864), (('go', 'back'), 854), (('uncl', 'remu'), 846), (('said', 'uncl'), 829), (('could', 'get'),
816), (('half', 'hour'), 807), (('great', 'deal'), 806), (('look', 'like'), 788), (('would', 'like'), 787),
(('littl', 'boy'), 769), (('look', 'round'), 769), (('said', 'dick'), 757), (('one', 'side'), 738), (('long',
'time'), 732), (('well', 'said'), 728), (('said', 'would'), 720), (('one', 'two'), 705), (('could', 'help'),
698), (('come', 'along'), 685), (('took', 'place'), 669), (('must', 'go'), 666), (('ye', 'sir'), 657),
(('made', 'way'), 648), (('never', 'mind'), 644), (('oh', 'ye'), 637), (('boy', 'said'), 629), (('take',
'place'), 627), (('would', 'go'), 621), (('project', 'electron'), 612)]
```

Top 70 2-grams:

```
[(('brer', 'rabbit'), 1599), (('could', 'see'), 1595), (('old', 'man'), 1506), (('come', 'back'), 1210),
(('two', 'three'), 1186), (('sir', 'said'), 1100), (('let', 'us'), 1079), (('said', 'mr'), 1046), (('good',
'deal'), 1041), (('ye', 'said'), 1027), (('project', 'gutenberg'), 1018), (('brer', 'fox'), 980), (('next',
'day'), 977), (('said', 'captain'), 975), (('next', 'morn'), 958), (('let', 'go'), 957), (('one', 'day'), 953),
(('everi', 'one'), 934), (('electron', 'work'), 918), (('young', 'man'), 905), (('said', 'doctor'), 884),
(('littl', 'girl'), 864), (('go', 'back'), 854), (('uncl', 'remu'), 846), (('said', 'uncl'), 829), (('could', 'get'),
816), (('half', 'hour'), 807), (('great', 'deal'), 806), (('look', 'like'), 788), (('would', 'like'), 787),
(('littl', 'boy'), 769), (('look', 'round'), 769), (('said', 'dick'), 757), (('one', 'side'), 738), (('long',
'time'), 732), (('well', 'said'), 728), (('said', 'would'), 720), (('one', 'two'), 705), (('could', 'help'),
698), (('come', 'along'), 685), (('took', 'place'), 669), (('must', 'go'), 666), (('ye', 'sir'), 657),
(('made', 'way'), 648), (('never', 'mind'), 644), (('oh', 'ye'), 637), (('boy', 'said'), 629), (('take',
'place'), 627), (('would', 'go'), 621), (('project', 'electron'), 612), (('know', 'said'), 608), (('two',
'day'), 606), (('take', 'care'), 605), (('would', 'take'), 605), (('would', 'come'), 601), (('first', 'time'),
601), (('two', 'men'), 595), (('shook', 'head'), 588), (('hundr', 'yard'), 573), (('said', 'bob'), 572),
(('last', 'night'), 569), (('make', 'way'), 567), (('short', 'time'), 565), (('fer', 'ter'), 557), (('one',
'anoth'), 550), (('said', 'tom'), 546), (('could', 'make'), 542), (('open', 'door'), 535), (('wait', 'till'),
534), (('go', 'away'), 533)]
```

The above snippet shows the most frequently occurring words in Corpus 2 by removing stop words. It is evident that stemming and stop word removal reduced the ambiguity to a greater extent. A quick glance at the tokens indicates that the words ***"boy ", "man", "cloth", "deal", "old"***, etc., might speak of an anecdote of a Conversation that centers around a young boy trying to negotiate a cloth purchase deal with an old man.

**Note:** When running the script with different combinations of stemming and stop word removal, we witnessed that stemming has a little impact on the generated tokens whereas the stop words played a significant role in arriving at the corpus description.

## Corpus 2 - Bigrams

### *a. Stemming with No stop words removal.*

Top 30 2-grams:

```
[(('of', 'the'), 66040), (('in', 'the'), 42953), (('to', 'the'), 29798), (('it', 'wa'), 23761), (('to', 'be'),
22462), (('and', 'the'), 20606), (('he', 'had'), 20178), (('it', 'is'), 17756), (('he', 'wa'), 17368), (('on',
'the'), 17332), (('with', 'a'), 16695), (('at', 'the'), 16593), (('of', 'hi'), 16009), (('had', 'been'), 15232),
(('for', 'the'), 14823), (('of', 'a'), 14636), (('i', 'am'), 14313), (('in', 'a'), 14233), (('she', 'had'),
13438), (('i', 'have'), 12923), (('with', 'the'), 12586), (('that', 'he'), 12096), (('of', 'her'), 11989),
(('by', 'the'), 11821), (('wa', 'a'), 11632), (('she', 'wa'), 11544), (('from', 'the'), 11154), (('in', 'hi'),
10836), (('to', 'her'), 10725), (('and', 'i'), 10265)]
```

Top 50 2-grams:

```
[(('of', 'the'), 66040), (('in', 'the'), 42953), (('to', 'the'), 29798), (('it', 'wa'), 23761), (('to', 'be'),
22462), (('and', 'the'), 20606), (('he', 'had'), 20178), (('it', 'is'), 17756), (('he', 'wa'), 17368), (('on',
'the'), 17332), (('with', 'a'), 16695), (('at', 'the'), 16593), (('of', 'hi'), 16009), (('had', 'been'), 15232),
(('for', 'the'), 14823), (('of', 'a'), 14636), (('i', 'am'), 14313), (('in', 'a'), 14233), (('she', 'had'),
13438), (('i', 'have'), 12923), (('with', 'the'), 12586), (('that', 'he'), 12096), (('of', 'her'), 11989),
(('by', 'the'), 11821), (('wa', 'a'), 11632), (('she', 'wa'), 11544), (('from', 'the'), 11154), (('in', 'hi'),
10836), (('to', 'her'), 10725), (('and', 'i'), 10265), (('did', 'not'), 9991), (('have', 'been'),
9632), (('i', 'do'), 9438), (('there', 'wa'), 9358), (('that', 'the'), 9265), (('as', 'he'), 9002),
(('wa', 'not'), 8592), (('for', 'a'), 8294), (('you', 'are'), 8225), (('in', 'her'), 8149), (('a', 'littl'),
8141), (('that', 'she'), 7716), (('that', 'i'), 7667), (('all', 'the'), 7551), (('to', 'hi'), 7520),
(('and', 'he'), 7475), (('out', 'of'), 7458), (('he', 'said'), 7304), (('would', 'have'), 7269),
(('and', 'then'), 7259)]
```

Top 70 2-grams:

```
[(('of', 'the'), 66040), (('in', 'the'), 42953), (('to', 'the'), 29798), (('it', 'wa'), 23761), (('to',
'be'), 22462), (('and', 'the'), 20606), (('he', 'had'), 20178), (('it', 'is'), 17756), (('he', 'wa'),
17368), (('on', 'the'), 17332), (('with', 'a'), 16695), (('at', 'the'), 16593), (('of', 'hi'), 16009),
(('had', 'been'), 15232), (('for', 'the'), 14823), (('of', 'a'), 14636), (('i', 'am'), 14313), (('in',
'a'), 14233), (('she', 'had'), 13438), (('i', 'have'), 12923), (('with', 'the'), 12586), (('that',
'he'), 12096), (('of', 'her'), 11989), (('by', 'the'), 11821), (('wa', 'a'), 11632), (('she', 'wa'),
11544), (('from', 'the'), 11154), (('in', 'hi'), 10836), (('to', 'her'), 10725), (('and', 'i'),
10265), (('did', 'not'), 9991), (('have', 'been'), 9632), (('i', 'do'), 9438), (('there', 'wa'),
9358), (('that', 'the'), 9265), (('as', 'he'), 9002), (('wa', 'not'), 8592), (('for', 'a'), 8294),
(('you', 'are'), 8225), (('in', 'her'), 8149), (('a', 'littl'), 8141), (('that', 'she'), 7716), (('that',
'i'), 7667), (('all', 'the'), 7551), (('to', 'hi'), 7520), (('and', 'he'), 7475), (('out', 'of'), 7458),
(('he', 'said'), 7304), (('would', 'have'), 7269), (('and', 'then'), 7259), (('could', 'not'),
7059), (('one', 'of'), 6974), (('would', 'be'), 6952), (('into', 'the'), 6872), (('do', 'you'),
6854), (('wa', 'the'), 6781), (('and', 'a'), 6764), (('seem', 'to'), 6688), (('but', 'i'), 6529),
(('is', 'a'), 6508), (('to', 'him'), 6429), (('if', 'you'), 6422), (('of', 'it'), 6332), (('you', 'have'),
6271), (('to', 'have'), 6270), (('as', 'she'), 6214), (('as', 'a'), 6107), (('to', 'see'), 6075), (('as',
'the'), 5995), (('as', 'if'), 5984)]
```

The above snippet show the most frequently occurring bigram words in Corpus 2 without the removal of stop words. These individual tokens will not contribute to the goal of annotating the topic the corpus2 belongs to. Therefore, it is impossible to arrive at the descriptor based on the obtained tokens.

### *b. Stemming with stop words removal.*

Top 30 2-grams:

```
[(('young', 'man'), 2361), (('cloth', 'extra'), 2075), (('old', 'man'), 1821), (('young', 'ladi'),
1496), (('crown', 'svo'), 1355), (('vol', 'ii'), 1270), (('let', 'us'), 1243), (('great', 'deal'),
1106), (('svo', 'cloth'), 1024), (('look', 'upon'), 1001), (('would', 'like'), 992), (('come',
'back'), 957), (('crown', 'cloth'), 952), (('first', 'time'), 925), (('good', 'deal'), 914),
(('would', 'never'), 870), (('could', 'help'), 853), (('vol', 'iii'), 843), (('could', 'see'), 829),
(('one', 'day'), 811), (('go', 'away'), 782), (('must', 'go'), 780), (('one', 'anoth'), 762),
(('two', 'three'), 757), (('one', 'thing'), 733), (('old', 'friend'), 712), (('would', 'come'), 704),
(('oh', 'ye'), 698), (('sir', 'harri'), 683), (('young', 'men'), 669)]
```

Top 50 2-grams:

```
[(('young', 'man'), 2361), (('cloth', 'extra'), 2075), (('old', 'man'), 1821), (('young', 'ladi'),
1496), (('crown', 'svo'), 1355), (('vol', 'ii'), 1270), (('let', 'us'), 1243), (('great', 'deal'),
1106), (('svo', 'cloth'), 1024), (('look', 'upon'), 1001), (('would', 'like'), 992), (('come',
'back'), 957), (('crown', 'cloth'), 952), (('first', 'time'), 925), (('good', 'deal'), 914),
(('would', 'never'), 870), (('could', 'help'), 853), (('vol', 'iii'), 843), (('could', 'see'), 829),
(('one', 'day'), 811), (('go', 'away'), 782), (('must', 'go'), 780), (('one', 'anoth'), 762),
(('two', 'three'), 757), (('one', 'thing'), 733), (('old', 'friend'), 712), (('would', 'come'), 704),
(('oh', 'ye'), 698), (('sir', 'harri'), 683), (('young', 'men'), 669), (('everi', 'one'), 648),
(('take', 'care'), 647), (('year', 'ago'), 644), (('shook', 'head'), 642), (('ye', 'said'), 637),
(('last', 'night'), 629), (('go', 'back'), 626), (('would', 'say'), 618), (('thought', 'would'),
611), (('open', 'door'), 607), (('would', 'make'), 596), (('sir', 'franci'), 580), (('young',
'fellow'), 578), (('look', 'like'), 575), (('long', 'time'), 562), (('would', 'take'), 558), (('sir',
'john'), 556), (('would', 'go'), 553), (('well', 'said'), 551), (('old', 'ladi'), 548)]
```

Top 70 2-grams:

```
[(('young', 'man'), 2361), (('cloth', 'extra'), 2075), (('old', 'man'), 1821), (('young', 'ladi'),
1496), (('crown', 'svo'), 1355), (('vol', 'ii'), 1270), (('let', 'us'), 1243), (('great', 'deal'),
1106), (('svo', 'cloth'), 1024), (('look', 'upon'), 1001), (('would', 'like'), 992), (('come',
'back'), 957), (('crown', 'cloth'), 952), (('first', 'time'), 925), (('good', 'deal'), 914),
(('would', 'never'), 870), (('could', 'help'), 853), (('vol', 'iii'), 843), (('could', 'see'), 829),
(('one', 'day'), 811), (('go', 'away'), 782), (('must', 'go'), 780), (('one', 'anoth'), 762),
(('two', 'three'), 757), (('one', 'thing'), 733), (('old', 'friend'), 712), (('would', 'come'), 704),
(('oh', 'ye'), 698), (('sir', 'harri'), 683), (('young', 'men'), 669), (('everi', 'one'), 648),
(('take', 'care'), 647), (('year', 'ago'), 644), (('shook', 'head'), 642), (('ye', 'said'), 637),
(('last', 'night'), 629), (('go', 'back'), 626), (('would', 'say'), 618), (('thought', 'would'),
611), (('open', 'door'), 607), (('would', 'make'), 596), (('sir', 'franci'), 580), (('young',
'fellow'), 578), (('look', 'like'), 575), (('long', 'time'), 562), (('would', 'take'), 558), (('sir',
'john'), 556), (('would', 'go'), 553), (('well', 'said'), 551), (('old', 'ladi'), 548), (('one',
'would'), 548), (('next', 'day'), 542), (('never', 'mind'), 537), (('made', 'mind'), 530),
(('could', 'tell'), 530), (('turn', 'away'), 530), (('one', 'els'), 530), (('let', 'go'), 528), (('post',
'svo'), 528), (('sir', 'said'), 524), (('one', 'could'), 522), (('look', 'round'), 516), (('one',
'two'), 515), (('door', 'open'), 513), (('illustr', 'board'), 511), (('know', 'said'), 509), (('said',
'old'), 509), (('would', 'give'), 506), (('could', 'make'), 506), (('think', 'would'), 505)]
```

The above snippet shows the most frequently occurring words in Corpus 2 by removing stop words. It is evident that stemming and stop word removal reduced the ambiguity to a greater extent. A quick glance at the tokens indicates that the words ***"young", "man", "cloth", "crown", "deal"***, etc., might speak of an anecdote of a Conversation that centers around a young boy trying to negotiate a cloth purchase deal with an old man.

### Removal of stop words belonging to multiple languages:

In all the above cases, it has been assumed that the raw corpus is unilingual. That is, the entire text has been presented in '***English'*** language. But, there might be a possibility that the text contains tokens from various languages or in other words, the raw corpus is multi-lingual.

In this case, it is essential to remove the unproductive words that find their etimology apart from 'English'. To handle this situation, we have gathered close to ***10,000*** stop words from different languages. The following outputs from the corpus1 and corpus2 highlights the most frequently occurring tokens after removing the global stop words as mentioned above.

**Corpus 1:** Unigrams without stemming and global stop words removal.

```
Top 30 words:

[('time', 20174), ('back', 13809), ('made', 13335), ('great', 12415), ('away', 11716), ('long',
11656), ('day', 10752), ('make', 10231), ('boy', 10147), ('father', 9290), ('captain', 8952), ('place',
8748), ('sir', 8674), ('hand', 8458), ('young', 8421), ('work', 8236), ('head', 8148), ('cried', 7770),
('found', 7754), ('asked', 7501), ('round', 7458), ('put', 7390), ('water', 7305), ('night', 7268),
('looked', 7078), ('left', 7012), ('give', 6971), ('eyes', 6965), ('heard', 6860), ('side', 6778)]
```

**Corpus 1:** Bigrams without stemming and global stop words removal.

```
Top 30 2-grams:

[(('brer', 'rabbit'), 1599), (('project', 'gutenberg'), 1018), (('brer', 'fox'), 980), (('uncle', 'remus'),
846), (('half', 'hour'), 806), (('great', 'deal'), 806), (('long', 'time'), 740), (('project', 'electronic'),
612), (('short', 'time'), 566), (('hundred', 'yards'), 566), (('shook', 'head'), 553), (('electronic',
'works'), 544), (('made', 'mind'), 457), (('brer', 'wolf'), 456), (('short', 'distance'), 455), (('cried',
'dick'), 449), (('set', 'work'), 446), (('r', 'harry'), 445), (('gutenberg', 'literary'), 442), (('literary',
'archive'), 442), (('archive', 'foundation'), 442), (('united', 'states'), 423), (('uncle', 'jack'), 421),
(('poor', 'fellow'), 411), (('time', 'time'), 391), (('run', 'away'), 386), (('fell', 'back'), 385),
(('electronic', 'work'), 374), (('caught', 'sight'), 366), (('uncle', 'dick'), 366)]
```

A key observation can be made by looking at these tokens. Many of the most frequently occurring words have been modified from those of the ones in the corpus 1 (without stemming and with stop word removal in English).

Similarly, the corpus 2 tokens are as follows:

**Corpus 2:** Unigrams without stemming and global stop words removal.

```
Top 30 words:

[('time', 17710), ('made', 13819), ('lady', 13266), ('young', 12437), ('life', 12080), ('miss',
11479), ('eyes', 11333), ('great', 10991), ('sir', 10847), ('long', 10113), ('father', 10087),
('back', 10052), ('make', 10019), ('day', 9999), ('away', 9840), ('hand', 9316), ('love',
9171), ('dear', 8930), ('mother', 8713), ('poor', 8609), ('house', 8594), ('looked', 8532),
('heart', 8440), ('mind', 8169), ('knew', 7703), ('moment', 7626), ('woman', 7569),
('place', 7417), ('felt', 7216), ('head', 7089)]
```

**Corpus 2:** Bigrams without stemming and global stop words removal.

```
Top 30 2-grams:

[(('cloth', 'extra'), 2072), (('crown', 'svo'), 1355), (('young', 'lady'), 1142), (('great', 'deal'),
1108), (('svo', 'cloth'), 1021), (('crown', 'cloth'), 953), (('vol', 'iii'), 841), (('sir', 'harry'),
677), (('shook', 'head'), 610), (('sir', 'francis'), 580), (('years', 'ago'), 576), (('long', 'time'),
569), (('sir', 'john'), 560), (('post', 'svo'), 528), (('young', 'fellow'), 510), (('illustrated',
'boards'), 508), (('lady', 'muriel'), 504), (('made', 'mind'), 499), (('sir', 'tom'), 494),
(('miss', 'heath'), 492), (('cloth', 'limp'), 472), (('long', 'ago'), 466), (('miss', 'dart'), 461),
(('thousand', 'pounds'), 450), (('sir', 'geoffrey'), 433), (('turned', 'away'), 419), (('lord',
'erradeen'), 393), (('lady', 'ridgeway'), 380), (('short', 'time'), 378), (('beg', 'pardon'),
373)]
```

### Conclusion:

In conclusion, based on the experimental results stemming seems to have almost no impact on the final output tokens. Furthermore, removal of stop words and repeating the analysis, we noticed that the most common unigrams and bigrams change significantly. This suggests that stop words can heavily influence the frequency distribution, and it is important  consider whether to remove stop words or not depending on the specific task or goal.

An interesting observation is that when the stop words from multiple languages have been considered for removal, the output tokens seem to have lesser noise. This helps us in further refining of tokens enabling us to better describe the underlying topic.

Among the unigrams and bigrams, the words that commonly occur together provided a more meaningful insights. This project has familiarized us with various preprocessing techniques, modularization of code, and to analyze the word distributions in two different corpora using python and NLTK library.

## LIST OF STOP WORDS – English

haven
we
wasn
d
their
do
o
up
his
a
both
s
mightn't
theirs
should
you
while
isn
whom
no
hasn
some
then
yourselves
don
themselves
with
against
wasn't
doing
doesn
this
couldn
its
over
i
re
didn
are
again
having
you'd
he
ve
herself own
hadn't
won't
doesn't
which
were
am
on
is
won
hers
y
shan
haven't
as
through
that'll
between
she
had
not
same
didn't
off
further
these
once
aren
you've
that
just
there
should've
to
why
weren
here
any
ma
hadn
it's
needn't
if
or
an
below
would
wouldn’t
but
have
where
himself
aren't
t
because
those
shouldn
ourselves
until
mustn
they
when
from
at
isn't
him
all
what
shouldn't
nor
was
shan't
by
wouldn't
such
mightn
needn
few
the
it
only
don't
most
yours
each
can
does
been
has
out
will
other
hasn't
couldn’t
more
being
before
down
you'll
weren't
be
under
itself
into
who
your
so
now
my
her
myself
for
me
how
ain
in
too
yourself
about
ours
than
during
our
m
very
wouldn
of
ll
she's
and
above
them
you're
after
did
mustn't