Project Report on Frequency Distribution

CIS 6397 - 24353 – Text Mining

Fall 2023 Professor: Peggy Lindner

Nikhita Peddi 2215340

Yogitha Ishwarya Moganti 2216966

Sai Vikas Reddy Talasani 2217332

GitHub link: Group 12

Abstract

This document provides guidelines for preparing text data that will undergo preprocessing and for analyzing word distributions within two distinct corpuses, all accomplished using Python and NLTK, paying particular attention to identifying and documenting stop words, which played a crucial role in our research.

1. Introduction

Natural Language Processing is a field of artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language. It involves the development of algorithms and models that allow machines to process and work with text and speech data. NLP has a wide range of applications, including machine translation, sentiment analysis, chatbots, speech recognition, and more, making it a fundamental technology for enhancing human-computer interaction and automating language-related tasks.

In Natural Language Processing (NLP), the initial steps typically involve preprocessing text data, which includes tokenization (splitting text into words or phrases), removing stop words (common words like "the," "and," "in" that carry little meaning), and performing tasks like stemming or lemmatization to reduce words to their base form. After preprocessing, NLP tasks like sentiment analysis, named entity recognition, or machine translation can be performed on the cleaned and structured text data, enabling machines to understand, interpret, and generate human language effectively.

2. Methodology

The Python source code has been used for this project, making use of built-in modules from the NLTK library, including functions like word tokenize, stop words, Porter Stemmer, FreqDist, and ngrams to conduct various data preprocessing steps and determine the underlying topic of the corpus. The project pipeline consists of the following stages:

A. Staging involves navigating to the directory path and scan through all the existing text files to extract the raw corpus. The Python script has been optimized to enhance reading speed.

B. Pre-processing includes several data cleaning and formatting steps as:

- Removing the punctuation and special characters from the corpus.
- Converting the text into lower case.
- Tokenization of corpus.
- Collecting stop words.
- Stemming the respective tokens and removal of stop words

C. Experimentation entails computing frequency distributions for both unigrams and bi-grams in the corpus, varying the 'K' value to assess frequent occurrences. The study focuses primarily on eliminating irrelevant 'English' stop words. However, it is conceivable that the corpora may contain tokens from various languages. To address this issue, we have conducted our analysis by eliminating common global stop words to assess their influence.

3. Experimental Results

Case 1: Unigram

Corpus 1 - Stemming without the removal of stop words

Below output displays the words that appear most frequently in Corpus 1, but it includes stop words that do not assist in identifying the topic of Corpus 1. These individual tokens will not aid in achieving the objective of labeling the topic to which corpus1 pertains. Consequently, it is not feasible to derive the descriptor based on the tokens acquired.

Top 30 words:

```
[('the', 99996), ('of', 64773), ('and', 38888), ('to', 35795), ('in',
32335), ('a', 27129), ('it', 17147), ('that', 16991), ('is', 16736), ('for',
12513), ('be', 11655), ('as', 11396), ('wa', 9556), ('by', 9094), ('thi',
8815), ('on', 8692), ('or', 8154), ('with', 8104), ('not', 7811),
('which', 7709), ('are', 7393), ('at', 7275), ('bank', 7032), ('he',
6846), ('have', 6768), ('i', 5655), ('from', 5599), ('money', 5520),
('hi', 5436), ('but', 5305)]
```

Top 50 words:

```
[('the', 99996), ('of', 64773), ('and', 38888), ('to', 35795), ('in',
32335), ('a', 27129), ('it', 17147), ('that', 16991), ('is', 16736), ('for',
12513), ('be', 11655), ('as', 11396), ('wa', 9556), ('by', 9094), ('thi',
8815), ('on', 8692), ('or', 8154), ('with', 8104), ('not', 7811),
('which', 7709), ('are', 7393), ('at', 7275), ('bank', 7032), ('he',
6846), ('have', 6768), ('i', 5655), ('from', 5599), ('money', 5520),
('hi', 5436), ('but', 5305), ('all', 5244), ('they', 4806), ('you', 4519),
('their', 4272), ('had', 4190), ('would', 4190), ('ha', 4168), ('an',
4148), ('we', 4139), ('will', 4016), ('other', 3962), ('one', 3896), ('if',
3778), ('valu', 3774), ('state', 3766), ('ani', 3712), ('were', 3699),
('stock', 3674), ('been', 3648), ('no', 3322)]
```

Top 70 words:

```
[('the', 99996), ('of', 64773), ('and', 38888), ('to', 35795), ('in',
32335), ('a', 27129), ('it', 17147), ('that', 16991), ('is', 16736), ('for',
12513), ('be', 11655), ('as', 11396), ('wa', 9556), ('by', 9094), ('thi',
8815), ('on', 8692), ('or', 8154), ('with', 8104), ('not', 7811),
('which', 7709), ('are', 7393), ('at', 7275), ('bank', 7032), ('he',
6846), ('have', 6768), ('i', 5655), ('from', 5599), ('money', 5520),
('hi', 5436), ('but', 5305), ('all', 5244), ('they', 4806), ('you', 4519),
('their', 4272), ('had', 4190), ('would', 4190), ('ha', 4168), ('an',
4148), ('we', 4139), ('will', 4016), ('other', 3962), ('one', 3896), ('if',
3778), ('valu', 3774), ('state', 3766), ('ani', 3712), ('were', 3699),
('stock', 3674), ('been', 3648), ('no', 3322), ('there', 3314), ('so',
3186), ('new', 3071), ('time', 2988), ('may', 2944), ('work', 2939),
('year', 2890), ('est', 2881), ('our', 2820), ('can', 2814), ('than',
2797), ('who', 2784), ('these', 2784), ('when', 2747), ('more', 2717),
('gold', 2711), ('exchang', 2598), ('price', 2498), ('do', 2488),
('them', 2483)]
```

Corpus 1 - Stemming with the removal of stop words

Below output demonstrates the most commonly appearing words in Corpus 1 after the removal of stop words. It's clear that eliminating stop words has significantly reduced ambiguity in the data.

Top 30 words:

```
[('bank', 7032), ('money', 5520), ('would', 4190), ('one', 3896),
('valu', 3774), ('state', 3766), ('stock', 3674), ('new', 3071), ('time',
2988), ('may', 2944), ('work', 2939), ('year', 2890), ('est', 2881),
('gold', 2711), ('exchang', 2598), ('price', 2498), ('busi', 2368),
('great', 2365), ('countri', 2309), ('project', 2202), ('make', 2170),
('market', 2159), ('use', 2081), ('made', 2061), ('good', 1988),
('upon', 1931), ('interest', 1865), ('credit', 1860), ('gener', 1842),
('compani', 1819)]
```

Top 50 words:

```
[('bank', 7032), ('money', 5520), ('would', 4190), ('one', 3896),
('valu', 3774), ('state', 3766), ('stock', 3674), ('new', 3071), ('time',
2988), ('may', 2944), ('work', 2939), ('year', 2890), ('est', 2881),
('gold', 2711), ('exchang', 2598), ('price', 2498), ('busi', 2368),
('great', 2365), ('countri', 2309), ('project', 2202), ('make', 2170),
('market', 2159), ('use', 2081), ('made', 2061), ('good', 1988),
('upon', 1931), ('interest', 1865), ('credit', 1860), ('gener', 1842),
('compani', 1819), ('unit', 1799), ('york', 1768), ('per', 1729),
('much', 1697), ('cent', 1668), ('reserv', 1659), ('amount', 1641),
('larg', 1598), ('day', 1560), ('note', 1535), ('public', 1526),
('deposit', 1490), ('must', 1480), ('secur', 1477), ('issu', 1463),
('capit', 1448), ('banker', 1421), ('mani', 1400), ('nation', 1392),
('govern', 1379)]
```

Top 70 words:

```
[('bank', 7032), ('money', 5520), ('would', 4190), ('one', 3896),
('valu', 3774), ('state', 3766), ('stock', 3674), ('new', 3071), ('time',
2988), ('may', 2944), ('work', 2939), ('year', 2890), ('est', 2881),
('gold', 2711), ('exchang', 2598), ('price', 2498), ('busi', 2368),
('great', 2365), ('countri', 2309), ('project', 2202), ('make', 2170),
('market', 2159), ('use', 2081), ('made', 2061), ('good', 1988),
('upon', 1931), ('interest', 1865), ('credit', 1860), ('gener', 1842),
('compani', 1819), ('unit', 1799), ('york', 1768), ('per', 1729),
('much', 1697), ('cent', 1668), ('reserv', 1659), ('amount', 1641),
('larg', 1598), ('day', 1560), ('note', 1535), ('public', 1526),
('deposit', 1490), ('must', 1480), ('secur', 1477), ('issu', 1463),
('capit', 1448), ('banker', 1421), ('mani', 1400), ('nation', 1392),
('govern', 1379), ('theori', 1366), ('law', 1349), ('everi', 1345),
('loan', 1330), ('part', 1317), ('could', 1311), ('two', 1295), ('bond',
1283), ('trade', 1272), ('first', 1271), ('increas', 1254), ('case', 1232),
('silver', 1223), ('specul', 1215), ('fact', 1202), ('street', 1198),
('coin', 1187), ('way', 1172), ('take', 1171), ('share', 1169)]
```

Observation: When executing the script with various combinations of stemming and stop word removal, it was observed that stemming had a minor influence on the resulting tokens, whereas the removal of stop words had a substantial impact on shaping the corpus description.

Corpus 2 - Stemming without the removal of stop words

Below output reveals the words that occur most frequently in Corpus 2, including stop words. However, this information makes it somewhat unclear regarding the topic to which Corpus 2 pertains. Consequently, deriving the descriptor from the obtained tokens is not feasible.

Top 30 words:

```
[('the', 126172), ('and', 84474), ('to', 65979), ('of', 64662), ('a', 47545), ('i', 44453), ('in', 37032), ('he', 35860), ('it', 31314), ('that', 30940), ('wa', 29563), ('you', 25082), ('hi', 24546), ('her', 22776), ('with', 21929), ('not', 19767), ('she', 19523), ('had', 19369), ('for', 18796), ('as', 17429), ('but', 16565), ('at', 16094), ('be', 15134), ('on', 15108), ('him', 14903), ('is', 14599), ('s', 12931), ('my', 12326), ('me', 12306), ('all', 12291)]
```

Top 50 words:

```
[('the', 126172), ('and', 84474), ('to', 65979), ('of', 64662), ('a', 47545), ('i', 44453), ('in', 37032), ('he', 35860), ('it', 31314), ('that', 30940), ('wa', 29563), ('you', 25082), ('hi', 24546), ('her', 22776), ('with', 21929), ('not', 19767), ('she', 19523), ('had', 19369), ('for', 18796), ('as', 17429), ('but', 16565), ('at', 16094), ('be', 15134), ('on', 15108), ('him', 14903), ('is', 14599), ('s', 12931), ('my', 12326), ('me', 12306), ('all', 12291), ('have', 12268), ('thi', 12233), ('said', 11859), ('so', 9907), ('they', 9290), ('by', 9153), ('from', 9115), ('what', 8587), ('or', 8052), ('which', 7989), ('there', 7974), ('we', 7907), ('no', 7727), ('one', 7555), ('would', 7497), ('were', 7410), ('if', 7127), ('when', 6949), ('t', 6857), ('up', 6719)]
```

Top 70 words:

```
[('the', 126172), ('and', 84474), ('to', 65979), ('of', 64662), ('a', 47545), ('i', 44453), ('in', 37032), ('he', 35860), ('it', 31314), ('that', 30940), ('wa', 29563), ('you', 25082), ('hi', 24546), ('her', 22776), ('with', 21929), ('not', 19767), ('she', 19523), ('had', 19369), ('for', 18796), ('as', 17429), ('but', 16565), ('at', 16094), ('be', 15134), ('on', 15108), ('him', 14903), ('is', 14599), ('s', 12931), ('my', 12326), ('me', 12306), ('all', 12291), ('have', 12268), ('thi', 12233), ('said', 11859), ('so', 9907), ('they', 9290), ('by', 9153), ('from', 9115), ('what', 8587), ('or', 8052), ('which', 7989), ('there', 7974), ('we', 7907), ('no', 7727), ('one', 7555), ('would', 7497), ('were', 7410), ('if', 7127), ('when', 6949), ('t', 6857), ('up', 6719), ('out', 6703), ('do', 6453), ('them', 6341), ('are', 6275), ('an', 6166), ('then', 5868), ('could', 5846), ('been', 5801), ('will', 5705), ('who', 5304), ('look', 5251), ('go', 5161), ('your', 5075), ('come', 5071), ('like', 4982), ('see', 4956), ('more', 4941), ('now', 4924), ('know', 4866), ('their', 4790)]
```

Corpus 2 - Stemming with the removal of stop words

Below output illustrates the most encountered terms in Corpus 2 after eliminating stop words. The output tokens demonstrate that employing stemming and removing stop words significantly diminished ambiguity.

Top 30 words:

```
[('said', 11859), ('one', 7555), ('would', 7497), ('could', 5846), ('look', 5251), ('go', 5161), ('come', 5071), ('like', 4982), ('see', 4956), ('know', 4866), ('say', 4595), ('time', 4504), ('man', 4002), ('littl', 3787), ('must', 3504), ('hand', 3478), ('well', 3475), ('work', 3376), ('day', 3373), ('even', 3131), ('never', 3131), ('think', 3126), ('went', 3125), ('make', 3075), ('good', 3064), ('thought', 3020), ('thing', 2976), ('came', 2875), ('us', 2846), ('eye', 2823)]
```

Top 50 words:

```
[('said', 11859), ('one', 7555), ('would', 7497), ('could', 5846), ('look', 5251), ('go', 5161), ('come', 5071), ('like', 4982), ('see', 4956), ('know', 4866), ('say', 4595), ('time', 4504), ('man', 4002), ('littl', 3787), ('must', 3504), ('hand', 3478), ('well', 3475), ('work', 3376), ('day', 3373), ('even', 3131), ('never', 3131), ('think', 3126), ('went', 3125), ('make', 3075), ('good', 3064), ('thought', 3020), ('thing', 2976), ('came', 2875), ('us', 2846), ('eye', 2823), ('old', 2818), ('made', 2813), ('seem', 2720), ('much', 2712), ('back', 2648), ('long', 2643), ('take', 2627), ('love', 2606), ('shall', 2544), ('away', 2511), ('may', 2507), ('get', 2502), ('noth', 2468), ('way', 2462), ('face', 2408), ('without', 2404), ('two', 2365), ('ask', 2314), ('might', 2304), ('turn', 2295)]
```

Top 70 words:

```
[('said', 11859), ('one', 7555), ('would', 7497), ('could', 5846), ('look', 5251), ('go', 5161), ('come', 5071), ('like', 4982), ('see', 4956), ('know', 4866), ('say', 4595), ('time', 4504), ('man', 4002), ('littl', 3787), ('must', 3504), ('hand', 3478), ('well', 3475), ('work', 3376), ('day', 3373), ('even', 3131), ('never', 3131), ('think', 3126), ('went', 3125), ('make', 3075), ('good', 3064), ('thought', 3020), ('thing', 2976), ('came', 2875), ('us', 2846), ('eye', 2823), ('old', 2818), ('made', 2813), ('seem', 2720), ('much', 2712), ('back', 2648), ('long', 2643), ('take', 2627), ('love', 2606), ('shall', 2544), ('away', 2511), ('may', 2507), ('get', 2502), ('noth', 2468), ('way', 2462), ('face', 2408), ('without', 2404), ('two', 2365), ('ask', 2314), ('might', 2304), ('turn', 2295), ('tell', 2242), ('life', 2231), ('project', 2203), ('want', 2169), ('princ', 2150), ('first', 2147), ('though', 2144), ('saw', 2125), ('word', 2080), ('last', 2054), ('upon', 2037), ('great', 2014), ('let', 2003), ('still', 2002), ('give', 1997), ('feel', 1987), ('put', 1966), ('head', 1941), ('thou', 1897), ('took', 1882)]
```

Observation: As we executed the script with various combinations of stemming and stop word elimination, we observed that stemming had a minor effect on the resulting tokens, whereas the removal of stop words had a substantial impact on defining the corpus description.

Case 2: Bigrams

Corpus 1 - Stemming without the removal of stop words

Below output shows individual tokens do not aid in achieving the objective of annotating the topic to which Corpus 1 pertains. Consequently, it is unattainable to establish the descriptor using the tokens acquired.

Top 30 2-grams:

```
[(('of', 'the'), 17307), (('in', 'the'), 8832), (('to', 'the'), 5589), (('and', 'the'), 3709), (('on', 'the'), 3474), (('for', 'the'), 3034), (('it', 'is'), 2975), (('that', 'the'), 2877), (('est', 'est'), 2740), (('to', 'be'), 2633), (('by', 'the'), 2627), (('with', 'the'), 2287), (('the', 'bank'), 2222), (('of', 'a'), 2143), (('at', 'the'), 1937), (('new', 'york'), 1752), (('from', 'the'), 1651), (('of', 'thi'), 1610), (('unit', 'state'), 1571), (('in', 'a'), 1507), (('of', 'money'), 1402), (('is', 'a'), 1396), (('as', 'a'), 1392), (('the', 'same'), 1383), (('it', 'wa'), 1309), (('the', 'unit'), 1303), (('valu', 'of'), 1303), (('as', 'the'), 1283), (('is', 'the'), 1231), (('have', 'been'), 1229)]
```

Top 50 2-grams:

```
[(('of', 'the'), 17307), (('in', 'the'), 8832), (('to', 'the'), 5589), (('and', 'the'), 3709), (('on', 'the'), 3474), (('for', 'the'), 3034), (('it', 'is'),
2975), (('that', 'the'), 2877), (('est', 'est'), 2740), (('to', 'be'), 2633),
(('by', 'the'), 2627), (('with', 'the'), 2287), (('the', 'bank'), 2222),
(('of', 'a'), 2143), (('at', 'the'), 1937), (('new', 'york'), 1752), (('from',
'the'), 1651), (('of', 'thi'), 1610), (('unit', 'state'), 1571), (('in', 'a'),
1507), (('of', 'money'), 1402), (('is', 'a'), 1396), (('as', 'a'), 1392),
(('the', 'same'), 1383), (('it', 'wa'), 1309), (('the', 'unit'), 1303),
(('valu', 'of'), 1303), (('as', 'the'), 1283), (('is', 'the'), 1231), (('have',
'been'), 1229), (('all', 'the'), 1209), (('the', 'stock'), 1184), (('per',
'cent'), 1085), (('would', 'be'), 1077), (('of', 'it'), 1066), (('there', 'is'),
1064), (('in', 'thi'), 1047), (('ha', 'been'), 1046), (('may', 'be'), 1025),
(('will', 'be'), 995), (('the', 'valu'), 929), (('to', 'a'), 926), (('bank', 'of'),
921), (('part', 'of'), 897), (('the', 'countri'), 892), (('is', 'not'), 875),
(('one', 'of'), 868), (('stock', 'exchang'), 859), (('of', 'hi'), 856), (('for',
'a'), 855)]
```

```
Top 70 2-grams: [(('of', 'the'), 17307), (('in', 'the'), 8832), (('to',
'the'), 5589), (('and', 'the'), 3709), (('on', 'the'), 3474), (('for', 'the'),
3034), (('it', 'is'), 2975), (('that', 'the'), 2877), (('est', 'est'), 2740),
(('to', 'be'), 2633), (('by', 'the'), 2627), (('with', 'the'), 2287), (('the',
'bank'), 2222), (('of', 'a'), 2143), (('at', 'the'), 1937), (('new', 'york'),
1752), (('from', 'the'), 1651), (('of', 'thi'), 1610), (('unit', 'state'),
1571), (('in', 'a'), 1507), (('of', 'money'), 1402), (('is', 'a'), 1396),
(('as', 'a'), 1392), (('the', 'same'), 1383), (('it', 'wa'), 1309), (('the',
'unit'), 1303), (('valu', 'of'), 1303), (('as', 'the'), 1283), (('is', 'the'),
1231), (('have', 'been'), 1229), (('all', 'the'), 1209), (('the', 'stock'),
1184), (('per', 'cent'), 1085), (('would', 'be'), 1077), (('of', 'it'), 1066),
(('there', 'is'), 1064), (('in', 'thi'), 1047), (('ha', 'been'), 1046), (('may',
'be'), 1025), (('will', 'be'), 995), (('the', 'valu'), 929), (('to', 'a'), 926),
(('bank', 'of'), 921), (('part', 'of'), 897), (('the', 'countri'), 892), (('is',
'not'), 875), (('one', 'of'), 868), (('stock', 'exchang'), 859), (('of', 'hi'),
856), (('for', 'a'), 855), (('and', 'that'), 829), (('the', 'money'), 823),
(('the', 'state'), 820), (('that', 'it'), 805), (('amount', 'of'), 797), (('he',
'wa'), 794), (('the', 'market'), 788), (('the', 'project'), 784), (('and',
'in'), 784), (('wall', 'street'), 762), (('had', 'been'), 761), (('the', 'new'),
760), (('they', 'are'), 759), (('do', 'not'), 755), (('project', 'gutenberg'),
754), (('out', 'of'), 752), (('and', 'it'), 732), (('of', 'their'), 721), (('we',
'have'), 710), (('which', 'the'), 709)]
```

Corpus 1 - Stemming with the removal of stop words

Below output illustrates the words that appear most often in Corpus 1 after eliminating stop words. The use of stemming and the removal of stop words have significantly reduced ambiguity in the results.

```
Top 30 2-grams:
[(('est', 'est'), 2740), (('new', 'york'), 1752), (('unit', 'state'), 1571),
(('per', 'cent'), 1085), (('stock', 'exchang'), 867), (('wall', 'street'),
762), (('project', 'gutenberg'), 754), (('electron', 'work'), 648),
(('bank', 'england'), 485), (('valu', 'money'), 472), (('nation', 'bank'),
451), (('clear', 'hous'), 436), (('project', 'electron'), 432), (('trust',
'compani'), 377), (('quantiti', 'theori'), 348), (('feder', 'reserv'), 326),
(('gutenberg', 'literari'), 312), (('literari', 'archiv'), 312), (('bank',
'note'), 311), (('archiv', 'foundat'), 300), (('gold', 'silver'), 260),
(('reserv', 'bank'), 255), (('set', 'forth'), 253), (('million', 'dollar'),
241), (('project', 'work'), 240), (('money', 'market'), 217),
(('distribut', 'project'), 217), (('term', 'agreement'), 217), (('state',
'bank'), 215), (('san', 'francisco'), 207)]
```

```
Top 50 2-grams:
[(('est', 'est'), 2740), (('new', 'york'), 1752), (('unit', 'state'), 1571),
(('per', 'cent'), 1085), (('stock', 'exchang'), 867), (('wall', 'street'),
762), (('project', 'gutenberg'), 754), (('electron', 'work'), 648),
(('bank', 'england'), 485), (('valu', 'money'), 472), (('nation', 'bank'),
451), (('clear', 'hous'), 436), (('project', 'electron'), 432), (('trust',
'compani'), 377), (('quantiti', 'theori'), 348), (('feder', 'reserv'), 326),
(('gutenberg', 'literari'), 312), (('literari', 'archiv'), 312), (('bank',
'note'), 311), (('archiv', 'foundat'), 300), (('gold', 'silver'), 260),
(('reserv', 'bank'), 255), (('set', 'forth'), 253), (('million', 'dollar'),
241), (('project', 'work'), 240), (('money', 'market'), 217),
(('distribut', 'project'), 217), (('term', 'agreement'), 217), (('state',
'bank'), 215), (('san', 'francisco'), 207), (('rate', 'interest'), 205),
(('legal', 'tender'), 204), (('paper', 'money'), 203), (('project',
'licens'), 192), (('cent', 'per'), 173), (('full', 'project'), 168), (('bill',
'exchang'), 162), (('save', 'bank'), 160), (('york', 'stock'), 159),
(('suppli', 'demand'), 157), (('stock', 'market'), 155), (('bank',
'reserv'), 152), (('foreign', 'exchang'), 150), (('state', 'note'), 148),
(('year', 'ago'), 147), (('margin', 'util'), 147), (('uncl', 'sam'), 147),
(('copi', 'distribut'), 146), (('york', 'citi'), 145), (('let', 'us'), 143)]
```

```
Top 70 2-grams:
[(('est', 'est'), 2740), (('new', 'york'), 1752), (('unit', 'state'), 1571),
(('per', 'cent'), 1085), (('stock', 'exchang'), 867), (('wall', 'street'),
762), (('project', 'gutenberg'), 754), (('electron', 'work'), 648),
(('bank', 'england'), 485), (('valu', 'money'), 472), (('nation', 'bank'),
451), (('clear', 'hous'), 436), (('project', 'electron'), 432), (('trust',
'compani'), 377), (('quantiti', 'theori'), 348), (('feder', 'reserv'), 326),
(('gutenberg', 'literari'), 312), (('literari', 'archiv'), 312), (('bank',
'note'), 311), (('archiv', 'foundat'), 300), (('gold', 'silver'), 260),
(('reserv', 'bank'), 255), (('set', 'forth'), 253), (('million', 'dollar'),
241), (('project', 'work'), 240), (('money', 'market'), 217),
(('distribut', 'project'), 217), (('term', 'agreement'), 217), (('state',
'bank'), 215), (('san', 'francisco'), 207), (('rate', 'interest'), 205),
(('legal', 'tender'), 204), (('paper', 'money'), 203), (('project',
'licens'), 192), (('cent', 'per'), 173), (('full', 'project'), 168), (('bill',
'exchang'), 162), (('save', 'bank'), 160), (('york', 'stock'), 159),
(('suppli', 'demand'), 157), (('stock', 'market'), 155), (('bank',
'reserv'), 152), (('foreign', 'exchang'), 150), (('state', 'note'), 148),
(('year', 'ago'), 147), (('margin', 'util'), 147), (('uncl', 'sam'), 147),
(('copi', 'distribut'), 146), (('york', 'citi'), 145), (('let', 'us'), 143),
(('real', 'estat'), 139), (('note', 'issu'), 137), (('gold', 'coin'), 137),
(('year', 'est'), 137), (('est', 'year'), 135), (('public', 'domain'), 131),
(('one', 'hundr'), 131), (('valu', 'gold'), 129), (('professor', 'fisher'),
128), (('scheftel', 'compani'), 125), (('bank', 'bank'), 123), (('market',
'price'), 122), (('standard', 'valu'), 122), (('deposit', 'bank'), 122),
(('copi', 'project'), 120), (('bank', 'system'), 120), (('volum', 'trade'),
119), (('credit', 'bank'), 118), (('stock', 'bond'), 115), (('econom',
'valu'), 113)]
```

Observation: While executing the script using various combinations of stemming and stop word removal, we observed that stemming had minimal influence on the resulting tokens, whereas the removal of stop words had a substantial impact on shaping the description of the corpus.

Corpus 2 - Stemming without the removal of stop words

Below output illustrates the bigram words that appear most often in Corpus 2, without eliminating stop words. These individual tokens do not aid in achieving the objective of labelling the topic to which Corpus 2 pertains. Consequently, determining the descript based on these extracted tokens is not feasible.

```
Top 30 2-grams:
[(('of', 'the'), 14413), (('in', 'the'), 9998), (('to', 'the'), 6803), (('and',
'the'), 4998), (('on', 'the'), 4725), (('it', 'wa'), 4239), (('to', 'be'),
4102), (('at', 'the'), 3926), (('he', 'had'), 3820), (('he', 'wa'), 3615),
(('with', 'the'), 3101), (('and', 'i'), 2987), (('in', 'a'), 2984), (('of', 'hi'),
2857), (('of', 'a'), 2780), (('that', 'he'), 2760), (('for', 'the'), 2742),
(('it', 'is'), 2657), (('with', 'a'), 2626), (('i', 'am'), 2585), (('from', 'the'),
2571), (('had', 'been'), 2457), (('did', 'not'), 2346), (('i', 'have'),
2321), (('by', 'the'), 2243), (('all', 'the'), 2177), (('wa', 'a'), 2164),
(('don', 't'), 2157), (('and', 'he'), 2122), (('she', 'wa'), 2108)]
```

```
Top 50 2-grams: [(('of', 'the'), 14413), (('in', 'the'), 9998), (('to',
'the'), 6803), (('and', 'the'), 4998), (('on', 'the'), 4725), (('it', 'wa'),
4239), (('to', 'be'), 4102), (('at', 'the'), 3926), (('he', 'had'), 3820),
(('he', 'wa'), 3615), (('with', 'the'), 3101), (('and', 'i'), 2987), (('in', 'a'),
2984), (('of', 'hi'), 2857), (('of', 'a'), 2780), (('that', 'he'), 2760),
(('for', 'the'), 2742), (('it', 'is'), 2657), (('with', 'a'), 2626), (('i', 'am'),
2585), (('from', 'the'), 2571), (('had', 'been'), 2457), (('did', 'not'),
2346), (('i', 'have'), 2321), (('by', 'the'), 2243), (('all', 'the'), 2177),
(('wa', 'a'), 2164), (('don', 't'), 2157), (('and', 'he'), 2122), (('she',
'wa'), 2108), (('she', 'had'), 2099), (('in', 'hi'), 2085), (('of', 'her'),
2014), (('that', 'i'), 2008), (('to', 'her'), 1966), (('out', 'of'), 1913),
(('there', 'wa'), 1888), (('i', 'wa'), 1887), (('that', 'the'), 1805), (('he',
'said'), 1776), (('to', 'him'), 1768), (('into', 'the'), 1718), (('could',
'not'), 1707), (('to', 'me'), 1665), (('but', 'i'), 1610), (('for', 'a'), 1605),
(('him', 'and'), 1574), (('you', 'are'), 1545), (('the', 'princ'), 1540),
(('the', 'same'), 1507)]
```

```
Top 70 2-grams:
[(('of', 'the'), 14413), (('in', 'the'), 9998), (('to', 'the'), 6803), (('and',
'the'), 4998), (('on', 'the'), 4725), (('it', 'wa'), 4239), (('to', 'be'),
4102), (('at', 'the'), 3926), (('he', 'had'), 3820), (('he', 'wa'), 3615),
(('with', 'the'), 3101), (('and', 'i'), 2987), (('in', 'a'), 2984), (('of', 'hi'),
2857), (('of', 'a'), 2780), (('that', 'he'), 2760), (('for', 'the'), 2742),
(('it', 'is'), 2657), (('with', 'a'), 2626), (('i', 'am'), 2585), (('from', 'the'),
2571), (('had', 'been'), 2457), (('did', 'not'), 2346), (('i', 'have'),
2321), (('by', 'the'), 2243), (('all', 'the'), 2177), (('wa', 'a'), 2164),
(('don', 't'), 2157), (('and', 'he'), 2122), (('she', 'wa'), 2108), (('she',
'had'), 2099), (('in', 'hi'), 2085), (('of', 'her'), 2014), (('that', 'i'),
2008), (('to', 'her'), 1966), (('out', 'of'), 1913), (('there', 'wa'), 1888),
(('i', 'wa'), 1887), (('that', 'the'), 1805), (('he', 'said'), 1776), (('to',
'him'), 1768), (('into', 'the'), 1718), (('could', 'not'), 1707), (('to',
'me'), 1665), (('but', 'i'), 1610), (('for', 'a'), 1605), (('him', 'and'),
1574), (('you', 'are'), 1545), (('the', 'princ'), 1540), (('the', 'same'),
1507), (('a', 'littl'), 1501), (('that', 'she'), 1475), (('as', 'he'), 1451),
(('if', 'you'), 1431), (('wa', 'not'), 1426), (('seem', 'to'), 1426), (('look',
'at'), 1407), (('have', 'been'), 1403), (('of', 'thi'), 1364), (('it', 's'),
1361), (('to', 'hi'), 1345), (('go', 'to'), 1344), (('to', 'see'), 1342),
(('and', 'then'), 1339), (('of', 'it'), 1332), (('i', 'can'), 1327), (('one',
'of'), 1300), (('and', 'that'), 1290), (('wa', 'the'), 1289), (('i', 'had'),
1264)]
```

Corpus 2 - Stemming with the removal of stop words.

Below output demonstrates the words that occur most frequently in Corpus 2 after eliminating stop words. It's clear that applying stemming and removing stop words significantly reduced ambiguity.

```
Top 30 2-grams:
[(('project', 'gutenberg'), 743), (('electron', 'work'), 648), (('alexey',
'alexandrovitch'), 570), (('stepan', 'arkadyevitch'), 547), (('project',
'electron'), 432), (('captain', 'nemo'), 383), (('old', 'man'), 373),
(('could', 'see'), 360), (('unit', 'state'), 359), (('van', 'hels'), 315),
(('gutenberg', 'literari'), 312), (('literari', 'archiv'), 312), (('archiv',
'foundat'), 302), (('sergey', 'ivanovitch'), 290), (('young', 'man'),
275), (('let', 'us'), 261), (('come', 'back'), 251), (('one', 'day'), 241),
(('project', 'work'), 240), (('nastasia', 'philipovna'), 240), (('sir',
'jame'), 239), (('go', 'away'), 227), (('said', 'dorothea'), 227), (('next',
'day'), 222), (('first', 'time'), 221), (('distribut', 'project'), 216),
(('term', 'agreement'), 216), (('would', 'like'), 207), (('darya',
'alexandrovna'), 204), (('set', 'forth'), 203)]
```

```
Top 50 2-grams: [(('project', 'gutenberg'), 743), (('electron', 'work'),
648), (('alexey', 'alexandrovitch'), 570), (('stepan', 'arkadyevitch'),
547), (('project', 'electron'), 432), (('captain', 'nemo'), 383), (('old',
'man'), 373), (('could', 'see'), 360), (('unit', 'state'), 359), (('van',
'hels'), 315), (('gutenberg', 'literari'), 312), (('literari', 'archiv'), 312),
(('archiv', 'foundat'), 302), (('sergey', 'ivanovitch'), 290), (('young',
'man'), 275), (('let', 'us'), 261), (('come', 'back'), 251), (('one', 'day'),
241), (('project', 'work'), 240), (('nastasia', 'philipovna'), 240),
(('sir', 'jame'), 239), (('go', 'away'), 227), (('said', 'dorothea'), 227),
(('next', 'day'), 222), (('first', 'time'), 221), (('distribut', 'project'),
216), (('term', 'agreement'), 216), (('would', 'like'), 207), (('darya',
'alexandrovna'), 204), (('set', 'forth'), 203), (('young', 'ladi'), 202),
(('one', 'anoth'), 199), (('said', 'princ'), 196), (('ned', 'land'), 196),
(('project', 'licens'), 192), (('let', 'go'), 189), (('great', 'deal'), 182),
(('said', 'levin'), 176), (('open', 'door'), 173), (('look', 'like'), 172),
(('full', 'project'), 168), (('lizabetha', 'prokofievna'), 167), (('thou',
'art'), 162), (('one', 'thing'), 162), (('look', 'round'), 161), (('went',
'away'), 160), (('hester', 'prynn'), 157), (('must', 'go'), 156),
(('evgeni', 'pavlovitch'), 155), (('copyright', 'law'), 154)]
```

```
Top 70 2-grams: [(('project', 'gutenberg'), 743), (('electron', 'work'),
648), (('alexey', 'alexandrovitch'), 570), (('stepan', 'arkadyevitch'),
547), (('project', 'electron'), 432), (('captain', 'nemo'), 383), (('old',
'man'), 373), (('could', 'see'), 360), (('unit', 'state'), 359), (('van',
'hels'), 315), (('gutenberg', 'literari'), 312), (('literari', 'archiv'), 312),
(('archiv', 'foundat'), 302), (('sergey', 'ivanovitch'), 290), (('young',
'man'), 275), (('let', 'us'), 261), (('come', 'back'), 251), (('one', 'day'),
241), (('project', 'work'), 240), (('nastasia', 'philipovna'), 240),
(('sir', 'jame'), 239), (('go', 'away'), 227), (('said', 'dorothea'), 227),
(('next', 'day'), 222), (('first', 'time'), 221), (('distribut', 'project'),
216), (('term', 'agreement'), 216), (('would', 'like'), 207), (('darya',
'alexandrovna'), 204), (('set', 'forth'), 203), (('young', 'ladi'), 202),
(('one', 'anoth'), 199), (('said', 'princ'), 196), (('ned', 'land'), 196),
(('project', 'licens'), 192), (('let', 'go'), 189), (('great', 'deal'), 182),
(('said', 'levin'), 176), (('open', 'door'), 173), (('look', 'like'), 172),
(('full', 'project'), 168), (('lizabetha', 'prokofievna'), 167), (('thou',
'art'), 162), (('one', 'thing'), 162), (('look', 'round'), 161), (('went',
'away'), 160), (('hester', 'prynn'), 157), (('must', 'go'), 156),
(('evgeni', 'pavlovitch'), 155), (('copyright', 'law'), 154), (('came',
'back'), 151), (('could', 'help'), 146), (('one', 'side'), 145), (('everi',
'day'), 145), (('would', 'come'), 145), (('thou', 'hast'), 144), (('copi',
'distribut'), 144), (('would', 'never'), 144), (('last', 'night'), 142),
(('go', 'back'), 141), (('two', 'three'), 140), (('would', 'go'), 139),
(('said', 'lydgat'), 138), (('put', 'hand'), 137), (('said', 'valanc'), 132),
(('long', 'time'), 131), (('one', 'could'), 131), (('said', 'noth'), 129),
(('said', 'stepan'), 127), (('everi', 'one'), 126)]
```

Observation: As we executed the script with various combinations of stemming and stop word elimination, we observed that stemming had a minor effect on the resulting tokens, whereas the removal of stop words had a substantial impact on defining the corpus description.

Elimination of stop words from various languages:

In all of the previously mentioned scenarios, we have operated under the assumption that the raw corpus is monolingual, exclusively in the 'English' language. However, there is a potential scenario where the text may comprise tokens from multiple languages, essentially rendering the raw corpus multilingual. In such instances, it becomes imperative to eliminate non-English words or unproductive terms that have origins in languages other than '***English***.' To address this challenge, we have compiled a comprehensive list of nearly 10,000 stop words from various languages. The subsequent results derived from corpus1 and corpus2 underscore the most frequently occurring tokens after the removal of these global stop words, as mentioned earlier.

Corpus 1 - Unigrams with stemming and removal of global stop words.

```
Top 30 words:

[('bank', 7032), ('money', 5520), ('valu', 3774), ('state', 3766),
('stock', 3674), ('time', 2988), ('work', 2939), ('year', 2890), ('gold',
2711), ('exchang', 2598), ('price', 2498), ('busi', 2368), ('great',
2365), ('countri', 2309), ('project', 2202), ('make', 2170), ('market',
2159), ('made', 2061), ('interest', 1865), ('credit', 1860), ('gener',
1842), ('compani', 1819), ('unit', 1799), ('york', 1768), ('cent',
1668), ('reserv', 1659), ('amount', 1641), ('larg', 1598), ('day',
1560), ('note', 1535)]
```

Corpus 1 - Bigrams with stemming and removal of global stop words.

```
Top 30 2-grams:

[(('unit', 'state'), 1572), (('stock', 'exchang'), 868), (('wall', 'street'),
764), (('project', 'gutenberg'), 754), (('electron', 'work'), 648),
(('bank', 'england'), 494), (('valu', 'money'), 479), (('nation', 'bank'),
452), (('clear', 'hous'), 436), (('project', 'electron'), 432), (('trust',
'compani'), 377), (('quantiti', 'theori'), 348), (('feder', 'reserv'), 326),
(('bank', 'note'), 314), (('gutenberg', 'literari'), 312), (('literari',
'archiv'), 312), (('archiv', 'foundat'), 300), (('reserv', 'bank'), 269),
(('gold', 'silver'), 263), (('million', 'dollar'), 241), (('project', 'work'),
240), (('year', 'year'), 220), (('state', 'bank'), 218), (('money',
'market'), 218), (('distribut', 'project'), 217), (('term', 'agreement'),
217), (('rate', 'interest'), 209), (('san', 'francisco'), 207), (('paper',
'money'), 206), (('legal', 'tender'), 204)]
```

Corpus 2 – Unigrams with stemming and removal of global stop words.

```
Top 30 words:

[('time', 4504), ('hand', 3478), ('work', 3376), ('day', 3373), ('make',
3075), ('eye', 2823), ('made', 2813), ('back', 2648), ('long', 2643),
('love', 2606), ('away', 2511), ('turn', 2295), ('life', 2231), ('project',
2203), ('princ', 2150), ('word', 2080), ('great', 2014), ('feel', 1987),
('put', 1966), ('head', 1941), ('thou', 1897), ('look', 1893), ('give',
1877), ('felt', 1839), ('answer', 1836), ('hous', 1829), ('night', 1802),
('room', 1796), ('talk', 1685), ('open', 1657)]
```

Corpus 2 - Bigrams with stemming and removal of global stop words.

```
Top 30 2-grams:

[(('project', 'gutenberg'), 743), (('electron', 'work'), 648), (('alexey',
'alexandrovitch'), 570), (('stepan', 'arkadyevitch'), 547), (('project',
'electron'), 432), (('captain', 'nemo'), 383), (('unit', 'state'), 359),
(('gutenberg', 'literari'), 312), (('literari', 'archiv'), 312), (('archiv',
'foundat'), 302), (('sergey', 'ivanovitch'), 290), (('project', 'work'),
240), (('nastasia', 'philipovna'), 240), (('sir', 'jame'), 239),
(('distribut', 'project'), 216), (('term', 'agreement'), 216), (('darya',
'alexandrovna'), 204), (('young', 'ladi'), 202), (('project', 'licens'),
192), (('great', 'deal'), 182), (('full', 'project'), 168), (('lizabetha',
'prokofievna'), 167), (('thou', 'art'), 163), (('hester', 'prynn'), 157),
(('evgeni', 'pavlovitch'), 155), (('copyright', 'law'), 154), (('put',
'hand'), 151), (('love', 'love'), 148), (('copi', 'distribut'), 144),
(('thou', 'hast'), 144)]
```

Note: An important insight can be derived when examining these tokens. Numerous high-frequency words have undergone alterations compared to those found in both corpuses, where no English stop word removal were applied.

4. Conclusion

In summary, based on our experimental findings, stemming appears to have minimal impact on the final output tokens. However, when we removed stop words and conducted the analysis again, we observed significant changes in the most frequent unigrams and bigrams. This highlights the substantial influence that stop words can exert on frequency distributions, underscoring the importance of considering whether to remove them based on the specific task or objective at hand.

Notably, among the unigrams and bigrams, we found that words frequently occurring together provided more meaningful insights. This project has familiarized us with various preprocessing techniques, code modularization, and the analysis of word distributions in two distinct corpus using Python and the NLTK library.

LIST OF STOP WORDS - English

1) shan't
2) ain
3) ll
4) some
5) themselves
6) those
7) as
8) than
9) wouldn't
10) your
11) who
12) do
13) re
14) wouldn
15) but
16) with
17) the
18) of
19) it
20) don't
21) did
22) o
23) ourselves
24) and
25) hasn
26) before
27) about
28) no
29) s
30) yours
31) because
32) nor
33) she
34) any
35) up
36) there
37) theirs
38) not
39) didn't
40) all
41) him
42) we
43) you'd
44) has
45) through
46) in
47) don
48) weren't
49) having
50) were
51) had
52) am
53) he
54) below
55) ours
56) yourselves
57) hasn't
58) weren
59) until
60) by
61) it's
62) needn't
63) very
64) hers
65) been
66) mightn
67) after
68) once
69) mustn
70) are
71) out
72) few
73) was
74) a
75) that'll
76) how
77) my
78) at
79) his
80) same
81) other
82) their
83) didn
84) can
85) is
86) whom
87) ma
88) what
89) off
90) aren't
91) hadn't
92) during
93) so
94) i
95) which
96) mightn't
97) being
98) herself
99) should
100) doesn
101) over
102) shan
103) for
104) haven't
105) wasn
106) you've
107) wasn't
108) down
109) shouldn
110) should've
111) where
112) himself
113) couldn
114) that
115) won
116) m
117) now
118) haven
119) shouldn't
120) this
121) its
122) an
123) these
124) won't
125) each
126) me
127) just
128) you'll
129) from
130) itself
131) doing
132) why
133) needn
134) again
135) does
136) be
137) if
138) own
139) aren
140) hadn
141) under
142) myself
143) such
144) them
145) both
146) or
147) they
148) t
149) to
150) her
151) mustn't
152) while
153) here
154) too
155) you
156) d
157) have
158) on
159) between
160) doesn't
161) isn
162) when
163) y
164) she’s
165) most
166) couldn't
167) more
168) then
169) you're
170) further
171) isn't
172) yourself
173) ve
174) only
175) will
176) our
177) against
178) into
179) above