Exploring Word Distributions in Two Text Corpora: Unveiling Topics and Insight

CIS6397-Text mining

By Mukesh Reddy Mavurapu (Conceptualization, Coding, Software, Report editing), Sandeep Kukunuru (Formal analysis, Code validation, writing original draft), Tangella Sreekanth (Report review and editing)

GitHub Repo

Abstract

Using Python and the Natural Language Toolkit (NLTK), the document outlines instructions for preparing text data and word distribution analysis. It describes methods to examine word distributions in two different corpora after outlining a step-by-step methodology for preparing and cleaning text data. Natural language processing and text mining applications can both benefit from this process, which is crucial for understanding text data.

1 Introduction

In the era of information abundance, the analysis of text data has emerged as a pivotal avenue for driving meaning and extracting knowledge from vast textual collections. The purpose of this project is to analyze word distributions in two different corpora using Python and Natural Language Toolkit (NLTK). Our fundamental goal is to delve into the intricacies of these corpora, deciphering the patterns, frequencies, and semantic significance of words they contain. 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.

We tried to find the difference in the common words with stop words and without stop words. In doing so, we aim not only to quantify word occurrences but also to reveal the latent topics and narrative threads that underlie the language within these corpora, offering a deeper understanding of their textual essence.

2 Methodology

The source code has been written in python and used various import libraries from the NLTK library. These modules include essential functions such as word_tokenize, stopwords removal, Counter from Collections to know the frequency distribution analysis and ngram for extracting n-gram sequences. We meticulously orchestrated a multi-stage project pipeline to ensure robust preprocessing and insightful analysis of our text corpora.

The entire project unfolds across several distinct stages. First and foremost, we initiated the process by employing NLTK's word_tokenize to break down the text into individual tokens, ensuring that each word was adequately segmented. Subsequently, the stopwords module was used to eliminate common words, allowing us to focus on more meaningful terms. With a refined dataset, Counter was utilized to calculate word frequencies and identify the top 30 most common words in each corpus. To gain a deeper understanding, we conducted multiple iterations by varying the parameter 'k' to assess the influence of different word frequency thresholds. This iterative process allowed us to unveil nuanced linguistic patterns and uncover latent topics within the corpora. Finally, the ngrams module was employed to extract n-gram sequences, providing deeper insights into linguistic patterns. This systematic pipeline allowed us to gain a comprehensive understanding of the word distributions and unveil the underlying topics embedded within the corpora.

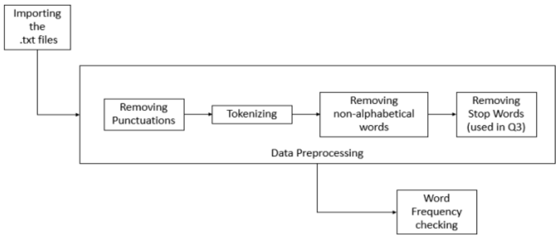

Figure 1: Flow chart of preprocessing.

3 Experimental Results

Once we got the tokens we check for the alphanumeric words and remove all the tokens which have punctuation. The frequency distributions for each of the unigram and bigram has been taken.

3.1 CORPUS1 and CORPUS2

Below are the results with no stop words for Unigrams. The most common words in corpus 1 are mostly stopwords which are highly frequent but carry very little meaning. These words are fundamental components of English language and are commonly used in various contexts.

Corpus1:

```
Top 30 words:
[('the', 99996), ('of', 64773), ('and', 38888), ('to', 35795), ('in', 32331), ('a', 27181), ('that',
16991), ('is', 16736), ('it', 13464), ('for', 12513), ('as', 11396), ('be', 10413), ('was', 9556),
('by', 9094), ('this', 8814), ('on', 8691), ('or', 8154), ('with', 8104), ('not', 7811), ('which',
7709), ('are', 7393), ('at', 7275), ('he', 6846), ('have', 6258), ('i', 5677), ('from', 5599),
('money', 5484), ('his', 5436), ('but', 5303), ('all', 5244)]

Top 10 words:
[('the', 99996), ('of', 64773), (‘and’ 38888), ('to', 35795), ('in', 32331), ('a', 27181), ('that', 1
6991), ('is', 16736), ('it', 13464), ('for', 12513)]

Top 100 words:
[('the', 99996), ('of', 64773), ('and', 38888), ('to', 35795), ('in', 32331), ('a', 27181), ('that', 1
6991), ('is', 16736), ('it', 13464), ('for', 12513), ('as', 11396), ('be', 10413), ('was', 9556), ('b
y', 9094), ('this', 8814), ('on', 8691), ('or', 8154), ('with', 8104), ('not', 7811), ('which', 7709
), ('are', 7393), ('at', 7275), ('he', 6846), ('have', 6258), ('i', 5677), ('from', 5599), ('money',
5484), ('his', 5436), ('but', 5303), ('all', 5244), ('they', 4806), ('you', 4519), ('their', 4256), ('
had', 4190), ('would', 4190), ('has', 4166), ('an', 4148), ('we', 4139), ('will', 3891), ('bank', 3
803), ('one', 3796), ('if', 3778), ('any', 3712), ('were', 3699), ('its', 3683), ('4', 3674), ('been',
 3647), ('other', 3481), ('there', 3314), ('no', 3310), ('so', 3186), ('new', 3071), ('value', 2975
), ('may', 2944), ('est', 2881), ('can', 2808), ('than', 2797), ('our', 2789), ('who', 2784), ('thes
e', 2773), ('when', 2747), ('more', 2716), ('gold', 2711), ('stock', 2643), ('them', 2483), ('onl
y', 2357), ('great', 2344), ('business', 2319), ('such', 2310), ('do', 2283), ('time', 2282), ('ban
ks', 2204), ('exchange', 2197), ('project', 2169), ('states', 2127), ('made', 2061), ('out', 2024)
, ('very', 1963), ('should', 1947), ('country', 1940), ('upon', 1931), ('market', 1916), ('some',
1915), ('what', 1903), ('york', 1768), ('work', 1753), ('per', 1729), ('much', 1697), ('up', 168
0), ('credit', 1672), ('into', 1663), ('about', 1654), ('united', 1650), ('same', 1612), ('years', 1
560), ('most', 1552), ('now', 1510), ('must', 1480), ('then', 1458), ('my', 1438)]
```

For the corpus 2 the most common words are also same as for the corpus 1 and Changing the K value significantly didn’t change anything for the corpus. The top words in both corpora are predominantly common English stop words, which do not provide specific information about the topics or content of the text. To understand the topics or descriptors, we need to analyze fewer common words and their context within the text.

Corpus2:

```
Top 30 words:
[('the', 126172), ('and', 84471), ('to', 65979), ('of', 64658), ('a', 47553), ('i', 44511), ('in',
37030), ('he', 35860), ('that', 30940), ('was', 29563), ('it', 28776), ('you', 25082), ('his',
24536), ('her', 22643), ('with', 21929), ('not', 19767), ('she', 19523), ('had', 19369), ('for',
18796), ('as', 17429), ('but', 16564), ('at', 16094), ('on', 15108), ('him', 14903), ('is', 14599),
('be', 13389), ('s', 12931), ('my', 12326), ('me', 12304), ('all', 12291)]

Top 10 words:
[('the', 126172), (‘and’,84471), ('to', 65979), ('of', 64658), ('a', 47553), ('i', 44511), ('in', 37
030), ('he', 35860), ('that', 30940), ('was', 29563)]

Top 100 words:

[('the', 126172), ('and', 84471), ('to', 65979), ('of', 64658), ('a', 47553), ('i', 44511), ('in', 37
030), ('he', 35860), ('that', 30940), ('was', 29563), ('it', 28776), ('you', 25082), ('his', 24536)
, ('her', 22643), ('with', 21929), ('not', 19767), ('she', 19523), ('had', 19369), ('for', 18796), (
'as', 17429), ('but', 16564), ('at', 16094), ('on', 15108), ('him', 14903), ('is', 14599), ('be', 13
389), ('s', 12931), ('my', 12326), ('me', 12304), ('all', 12291), ('said', 11859), ('have', 11313
), ('this', 11006), ('so', 9907), ('they', 9290), ('by', 9153), ('from', 9115), ('what', 8587), ('or',
 8052), ('which', 7989), ('there', 7974), ('we', 7907), ('no', 7724), ('would', 7497), ('were',
7410), ('one', 7392), ('if', 7126), ('when', 6949), ('t', 6857), ('up', 6712), ('out', 6701), ('them
', 6341), ('are', 6275), ('an', 6166), ('do', 5940), ('then', 5868), ('could', 5846), ('been', 5801)
, ('will', 5598), ('who', 5304), ('more', 4941), ('now', 4924), ('your', 4902), ('their', 4750), ('a
bout', 4729), ('can', 4681), ('did', 4483), ('into', 4386), ('see', 4312), ('some', 4261), ('any', 4
163), ('like', 4148), ('very', 4147), ('know', 4118), ('man', 3998), ('time', 3867), ('little', 378
6), ('come', 3727), ('how', 3725), ('than', 3718), ('down', 3542), ('before', 3526), ('only', 352
3), ('must', 3504), ('well', 3451), ('go', 3396), ('over', 3278), ('other', 3257), ('never', 3131),
('went', 3125), ('has', 3103), ('after', 3083), ('am', 2940), ('came', 2875), ('good', 2870), ('sh
ould', 2855), ('such', 2852), ('himself', 2852), ('us', 2846), ('old', 2818)]
```

Just from these frequencies its challenging to guess the specific topic or descriptor for each corpus. These words are generic and do not provide insights into the unique content of each corpus.

3.1.1 Bigrams

Below are the results for the most common bigrams without removing stop words. These individual tokens, even with two contiguous sequences, do not contribute to the goal of annotating the topic the both the corpus belong to. It is impossible to arrive at the descriptor based on the obtained tokens. Even with change in K value the most common words are with these stop words.

Corpus1:

```
Top 30 words:
[(('of', 'the'), 17240), (('in', 'the'), 8638), (('to', 'the'), 5524), (('and', 'the'), 3683), (('on',
'the'), 3462), (('for', 'the'), 3022), (('it', 'is'), 2975), (('that', 'the'), 2877), (('4', 'est'), 2877),
(('to', 'be'), 2632), (('by', 'the'), 2620), (('with', 'the'), 2285), (('of', 'a'), 2127), (('at', 'the'),
1924), (('new', 'york'), 1752), (('from', 'the'), 1648), (('of', 'this'), 1599), (('united', 'states'),
1571), (('in', 'a'), 1492), (('the', 'bank'), 1424), (('of', 'money'), 1394), (('is', 'a'), 1393), (('as',
'a'), 1391), (('the', 'same'), 1383), (('it', 'was'), 1309), (('as', 'the'), 1279), (('the', 'united'),
1275), (('is', 'the'), 1227), (('all', 'the'), 1209), (('have', 'been'), 1148)]
```

Corpus2:

```
Top 30 words:
[(('of', 'the'), 14411), (('in', 'the'), 9968), (('to', 'the'), 6803), (('and', 'the'), 4995), (('on',
'the'), 4725), (('it', 'was'), 4239), (('to', 'be'), 4083), (('at', 'the'), 3926), (('he', 'had'), 3820),
(('he', 'was'), 3615), (('with', 'the'), 3101), (('and', 'i'), 2990), (('in', 'a'), 2984), (('of', 'his'),
2857), (('of', 'a'), 2779), (('that', 'he'), 2760), (('for', 'the'), 2742), (('it', 'is'), 2655), (('with',
'a'), 2626), (('i', 'am'), 2585), (('from', 'the'), 2571), (('had', 'been'), 2457), (('did', 'not'),
2346), (('i', 'have'), 2318), (('by', 'the'), 2243), (('all', 'the'), 2177), (('was', 'a'), 2164),
(('don', 't'), 2157), (('and', 'he'), 2122), (('she', 'was'), 2108)]
```

:3.1.2 Stop word list:

Below are the results that show all the stopwords present in each of the corpus. After removing the stopwords the remaining common words provide some insights into the content of the corpus. However, it’s important to note that without context of the complete text it is challenging to make conclusions. Few of the common words in corpus 1 are “money”, “bank”, “gold”, “value”, “stock” and so on these words more relate to finance. There are few words which speak about the country, place so it might also relate to some geographical location as well. The presence of terms like “great” and “united” may indicate broader discussions related to the economy of the country. With the change in the K value the core themes and topics are more likely to remain constant. We have tried for the k value 100 and k value of 10 and with more K value we found more words on finance and economy of other country and regions.

Corpus1:

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

Top 30 stop words:
[('money', 5484), ('would', 4190), ('bank', 3803), ('one', 3796), ('4', 3674), ('new', 3071),
('value', 2975), ('may', 2944), ('est', 2881), ('gold', 2711), ('stock', 2643), ('great', 2344),
('business', 2319), ('time', 2282), ('banks', 2204), ('exchange', 2197), ('project', 2169),
('states', 2127), ('made', 2061), ('country', 1940), ('upon', 1931), ('market', 1916), ('york',
1768), ('work', 1753), ('per', 1729), ('much', 1697), ('credit', 1672), ('united', 1650), ('years',
1560), ('must', 1480)]
```

Below are all the stopwords that are present in the corpus2. And showing the top 30 most common words after removing these stopwords. Some of the common words are “said”, “would”, “could”, “see”, “like” etc. these words say more about stories or discussions and words like “could”, “like”, “know”, “must”, “may” talk more about uncertainty and might relate to the characters thoughts. Similar to corpus 1 we tried for different K values in corpus 2 and it tells the same that we found from k value of 30 but we can make stronger assumptions as we increase the k value.

Corpus2:

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

Top 30 stop words:
[('said', 11859), ('would', 7497), ('one', 7392), ('could', 5846), ('see', 4312), ('like', 4148),
('know', 4118), ('man', 3998), ('time', 3867), ('little', 3786), ('come', 3727), ('must', 3504),
('well', 3451), ('go', 3396), ('never', 3131), ('went', 3125), ('came', 2875), ('good', 2870),
('us', 2846), ('old', 2818), ('made', 2813), ('much', 2712), ('say', 2711), ('back', 2578),
('thought', 2575), ('shall', 2544), ('away', 2511), ('may', 2507), ('eyes', 2467), ('think',
2463)]
```

Removing stop words has significantly changed the descriptors for each corpus. While the initial analysis hinted at narrative or storytelling in both the corpus.

Corpus1 appears to finance, investments or economic discussions.

Corpus2 seems to be more focused on storytelling or narratives involving characters and their actions.

The use of stop words helps in highlighting the core contents and themes of each corpus. It’s essential to consider stopwords in text analysis to better understand the underlying topics and context.

3.1.3 Most common bigrams without stopwords

Below are the 30 most common bigrams for both the corpus documents after removing the stopwords. We made some assumptions from unigrams but bigrams which are pairs of consecutive words, can be more useful than unigrams in text analysis as they provide context and capture relationships between words.

For Corpus 1 the bigrams give further light on the prevalent themes and topics within the body of the text. The presence of bigrams like “new york” indicates a potential focus on locations or events related to New York. And all the bigrams are mostly related to money, stocks and finance or business related. Words like “project Gutenberg” and “electronic works” refer to literary or digital content.

Corpus1:

```
Top 30 stop words:
[(('4', 'est'), 2877), (('new', 'york'), 1752), (('united', 'states'), 1571), (('per', 'cent'), 1084),
(('stock', 'exchange'), 836), (('wall', 'street'), 762), (('project', 'gutenberg'), 754), (('bank',
'england'), 473), (('value', 'money'), 464), (('project', 'electronic'), 432), (('clearing',
'house'), 392), (('electronic', 'works'), 384), (('quantity', 'theory'), 347), (('federal', 'reserve'),
326), (('gutenberg', 'literary'), 312), (('literary', 'archive'), 312), (('archive', 'foundation'),
300), (('trust', 'company'), 272), (('electronic', 'work'), 264), (('gold', 'silver'), 256),
(('national', 'bank'), 242), (('set', 'forth'), 236), (('bank', 'notes'), 223), (('terms', 'agreement'),
217), (('san', 'francisco'), 207), (('money', 'market'), 201), (('legal', 'tender'), 200), (('paper',
'money'), 196), (('project', 'license'), 192), (('national', 'banks'), 186)]
```

For the corpus 2 use of bigrams has provided crucial context and specificity to our analysis. The words such as “project Gutenberg” focus on literary topics, digital publishing. There are few character names that are repeating like “captain Nemo” and “van Helsing” which are like works of fiction. In summary these bigrams talk more about the documents from what we saw from the unigrams. Unigrams didn’t introduce the nouns and are mostly about a narrative form, but we were not able to conclude. But the bigrams increase the confidence of the predictions.

Corpus 2:

```
Top 30 stop words:
(('project', 'gutenberg'), 743), (('alexey', 'alexandrovitch'), 570), (('stepan', 'arkadyevitch'),
547), (('project', 'electronic'), 432), (('electronic', 'works'), 384), (('captain', 'nemo'), 383),
(('old', 'man'), 373), (('could', 'see'), 360), (('united', 'states'), 359), (('van', 'helsing'), 315),
(('gutenberg', 'literary'), 312), (('literary', 'archive'), 312), (('archive', 'foundation'), 302),
(('sergey', 'ivanovitch'), 290), (('young', 'man'), 275), (('electronic', 'work'), 264), (('let',
'us'), 258), (('nastasia', 'philipovna'), 239), (('sir', 'james'), 239), (('one', 'day'), 230), (('said',
'dorothea'), 227), (('first', 'time'), 220), (('next', 'day'), 218), (('terms', 'agreement'), 216),
(('darya', 'alexandrovna'), 204), (('set', 'forth'), 203), (('one', 'another'), 198), (('said',
'prince'), 196), (('ned', 'land'), 196), (('project', 'license'), 192)
```

In summary, the utilization of bigrams has not only refined but also enriched the descriptor for both the corpus files. We have also tried for different k, it tells the same that we found from k value of 30 but we can make stronger assumptions as we increase the k value. The utilization of bigrams has thus yielded a more nuanced and context-rich understanding of both corpora, demonstrating the importance of considering higher-level linguistic units for text analysis.

Conclusion:

In conclusion, we observed that the most prevalent unigrams and bigrams vary dramatically once stop words are removed and the study is repeated. This implies that stop words can significantly affect the frequency distribution, so it's crucial to take the job or aim at hand into account while deciding whether to eliminate stop words or not.

An intriguing finding is that the output tokens appear to have less noise when stop words from several languages have been considered for elimination. This aids in the continued refinement of the tokens, allowing us to more accurately express the underlying subject.

The terms that frequently occur together among the unigrams and bigrams offered more significant information. With the help of the NLTK library and Python, we were able to examine the word distributions in two different corpora as well as learn about various preprocessing strategies.