Mini Project 1 Report

Mahesh R, Nalabolu: Conceptualization, Formal Analysis, Methodology, Writing original draft.
Minh V, Nguyen: Conceptualization, Formal Analysis, Validation, Supervision.
Shifafatima Khoja: Conceptualization, Formal Analysis, Writing – review & editing.
GitHub Link

Abstract

This document summarizes the findings after the completion of Mini Project 1. The task is to complete the word distributions of a set of corpora. It is to be accomplished using Python and NLTK libraries to ascertain the contents of each corpus using text mining techniques.

1. Introduction

The objective of this assignment is to analyze word distributions of two different corpora using Python and Natural Language Toolkit (NLTK). The two corpora are in separate directories, and the analysis involves computing the most common words and n-grams in each corpus. The main questions we aim to answer are: What are the 30 most common words in each corpus? What can we say about these top 30 words? Can we guess the topic or descriptor for each corpus just by looking at these word frequencies? We will also experiment with different values for k, the number of most frequent words to analyze.

2. Methodology

Before performing any Python tasks, we opened each corpus and took a look at how the data was presented to us. Just quickly reading a text file from each corpus, helped us get a better idea of how to proceed further with the task at hand. Also, it is crucial to know what kind of data you are working with to decide which preprocessing tasks will be necessary. The methodology for this assignment involved using Python and NLTK to process the text data. First, we needed to be able to access the corpora, so we started by defining paths to the two directories containing the corpora. Then, we iterated through all the files from the directory and saved the contents to a list. Additionally, with any text mining task, it is important to perform preprocessing, so we created a function to help us preprocess the text data. This function has many embedded functions, which helped in converting all the text to lowercase, removing punctuation, removing the common English stop words, removing all non-alphabetic tokens, and stemming the data. Preprocessing helps to ensure that the next tasks that we perform using this data are as accurate as possible.

After making sure that the data has been processed successfully, we computed the word distributions using the FreqDist method. For the word distribution function, we looped through all files in each corpus directory, preprocessed the text in each file, and then updated a Counter object with the preprocessed tokens. We then printed the 30 most common words in each corpus using the most_common() method of the Counter object. We also experimented with different values of k to see how the most common words change.

Lastly, we used methods from the NLTK Collocations class to retrieve the

word distributions of bi-grams. For these distributions, we followed a similar procedure as above, but instead of counting individual words, we counted pairs of consecutive words (bi-grams). We then printed the 30 most common bi-grams in each corpus.

3. Experimental Results

The results of our analysis showed that the top 30 words in each corpus are different, indicating that each corpora have different characteristics. For example, the top 30 words in Corpus 1 include :

```
[('said', 60602), ('one', 36983),
('would', 33896), ('go', 26525),
('could', 23917), ('time', 23335),
('look', 23039), ('come', 22228),
('littl', 21884), ('like', 20824),
('see', 20258), ('well', 19761),
('say', 18226), ('mr', 17882),
('know', 17659), ('man', 17427),
('two', 17114), ('day', 16481),
('good', 16148), ('boy', 15763),
('upon', 15418), ('get', 15227),
('hand', 15221), ('way', 15208),
('men', 14695), ('dont', 14685),
('back', 14420), ('think', 14352),
('make', 14217), ('take', 14054)]
```

It's difficult to determine a specific topic or descriptor without additional context. However, some of the most frequent words such as "said," "one," "would," "go," "time," and "come" are quite general and could apply to many different types of texts. The presence of words like "mr" and "man" suggest that the corpus may contain works of fiction or non-fiction that include male characters. In Corpus2, we have:

```
[('said', 48302), ('mr', 40725),
('would', 40337), ('one', 37016),
('could', 27343), ('look', 24405),
('like', 24098), ('know', 23019),
('littl', 21956), ('man', 21219),
('time', 21099), ('come', 19964),
('go', 19537), ('say', 19201),
('well', 18618), ('see', 18581),
('upon', 18458), ('think', 18419),
('never', 17690), ('good', 17660),
('old', 17067), ('hand', 16370),
('day', 16361), ('must', 16010),
('much', 15980), ('ladi', 15807),
('thought', 15509), ('even',
14693), ('make', 14407), ('love',
14159)]
```

It's difficult to determine the topic or descriptor of the corpus, but it could potentially be a collection of literary works or fictional stories given the presence of words like 'said', 'would', and 'could'. There are also words that suggest a possible historical or old-fashioned context such as 'old' and 'lady'.

For bigrams in Corpus1, we have

```
[(('project', 'gutenberg'), 2959),
(('dont', 'know'), 2188),
(('gutenberg', 'tm'), 1938),
(('said', 'mr'), 1682), (('brer',
'rabbit'), 1612), (('could',
'see'), 1541), (('old', 'man'),
1503), (('dont', 'think'), 1309),
(('come', 'back'), 1213), (('two',
'three'), 1188), (('sir', 'said'),
1091), (('let', 'us'), 1084),
(('good', 'deal'), 1042), (('ye',
'said'), 1000), (('brer', 'fox'),
997), (('half', 'hour'), 980),
(('dont', 'want'), 978), (('next',
'day'), 977), (('said', 'captain'),
971), (('one', 'day'), 960),
(('next', 'morn'), 959), (('let',
'go'), 951), (('everi', 'one'),
933), (('electron', 'work'), 918),
(('young', 'man'), 907), (('said',
'doctor'), 884), (('littl',
'girl'), 867), (('go', 'back'),
855), (('said', 'uncl'), 827),
(('great', 'deal'), 806)]
```

The presence of "project gutenberg" and "gutenberg tm" suggests that the corpus may be a collection of texts from the Project Gutenberg website. Several word pairs involve characters from children's stories, such as "brer rabbit" and "brer fox". Other common word pairs include phrases like "dont know", "dont think", "come back", "let us", "next day", and "go back". These phrases are relatively common in everyday

language and do not necessarily provide specific information about the topic or descriptor of the corpus.

Similarly in Corpus2:

```
[(('said', 'mr'), 2431), (('young',
'man'), 2385), (('dont', 'know'),
2263), (('cloth', 'extra'), 2073),
(('old', 'man'), 1847), (('crown',
'8vo'), 1787), (('young', 'ladi'),
1516), (('8vo', 'cloth'), 1475),
(('dont', 'think'), 1364),
(('crown', 'svo'), 1364), (('vol',
'ii'), 1357), (('let', 'us'),
1258), (('draw', 'room'), 1188),
(('great', 'deal'), 1107), (('svo',
'cloth'), 1036), (('look', 'upon'),
1000), (('come', 'back'), 960),
(('good', 'bye'), 950), (('would',
'like'), 943), (('first', 'time'),
925), (('good', 'deal'), 916),
(('vol', 'iii'), 902), (('would',
'never'), 869), (('one', 'day'),
816), (('could', 'see'), 811),
(('mr', 'gayr'), 786), (('go',
'away'), 784), (('must', 'go'),
779), (('7s', '6d'), 777),
(('could', 'help'), 769)]
```

The presence of phrases like "cloth extra", "crown 8vo", "8vo cloth", and "crown svo" suggests that the corpus may contain books or written materials. Phrases like "young man", "young lady", "old man", and "first time" suggest that the corpus may contain fictional or narrative works. The phrases "said mr" and "dont know" suggest that dialogue or conversations may be present in the corpus.

When we experimented with different values of k, we found that the most common words change, but the overall topic or descriptor of each corpus remained the same. For example, when we set k=50, the most common words in Corpus 1 include:

```
[('said', 60602), ('one', 36983),
('would', 33896), ('go', 26525),
('could', 23917), ('time', 23335),
('look', 23039), ('come', 22228),
('littl', 21884), ('like', 20824),
('see', 20258), ('well', 19761),
('say', 18226), ('mr', 17882),
('know', 17659), ('man', 17427),
('two', 17114), ('day', 16481),
('good', 16148), ('boy', 15763),
('upon', 15418), ('get', 15227),
('hand', 15221), ('way', 15208),
('men', 14695), ('dont', 14685),
('back', 14420), ('think', 14352),
('make', 14217), ('take', 14054),
('us', 14009), ('made', 13443),
('came', 13042), ('place', 12816),
('long', 12795), ('much', 12710),
('great', 12558), ('old', 12196),
('must', 12161), ('seem', 11972),
('work', 11771), ('away', 11764),
('went', 11625), ('never', 10744),
('ask', 10683), ('even', 10565),
('ye', 10485), ('first', 10430),
('cri', 10272), ('may', 10128)]
```

Some words in the list such as "mr," "boy," and "hand" might suggest a narrative or dialogue involving characters. Additionally, the presence of words like "place," "way," and "came" may suggest a setting or location.

When we use stopwords list In Corpus2

```
[('said', 48312), ('mr', 40730),
('would', 40337), ('one', 37019),
('could', 27344), ('look', 24409),
('like', 24099), ('know', 23026),
('littl', 21960), ('man', 21223),
('time', 21104), ('come', 19969),
('go', 19538), ('say', 19210),
('well', 18634), ('see', 18582),
('upon', 18458), ('think', 18421),
('never', 17692), ('good', 17661),
('old', 17069), ('hand', 16370),
('day', 16364), ('must', 16011),
('much', 15981), ('ladi', 15809),
('thought', 15513), ('even',
14693), ('make', 14407), ('love',
14161), ('made', 14013), ('eye',
13872), ('way', 13397), ('seem',
12991), ('dont', 12913), ('take',
12643), ('might', 12563), ('young',
12510), ('life', 12387), ('miss',
12154), ('thing', 12007), ('face',
11713), ('ask', 11695), ('long',
11432), ('great', 11285), ('tell',
11142), ('noth', 11022), ('two',
10995), ('sir', 10935), ('first',
10853)]
```

These top 50 words also seem to be from a corpus of fiction or literature, as we

can see the words "mr", "lady", "love", "life", "miss", "sir", and "fictional" words appearing frequently. It is difficult to guess the exact topic of the corpus, but we can assume that it is a work of fiction with a focus on relationships, as words like "love", "miss", and "young" appear frequently. There also seem to be words related to emotion and perception, such as "thought", "even", "seem", and "eye".

4. Conclusion

This assignment provided us an opportunity to analyze word distributions in two different corpora using Python and NLTK. The results of the analysis showed that the top 30 words in each corpus were different, indicating that the corpora had different characteristics. This project allowed us to see how text mining is applied in real-world scenarios and how it can be beneficial in analyzing large amounts of data with little human interaction. Additionally, we learned that there are multiple ways that this task could have been done just by seeing all the different ideas each group member came up with. This assignment was able to help us better understand the topics covered in class and allowed us to learn more about the Natural Language Toolkit and the many useful methods it has, especially for text-mining solutions.

Appendix A. Stop Words List

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