Homework 2C - DATA-312

Jeffrey Williams

09, April 2022

Abstract

This writeup explores a sample of 6 works from author Charles Dickens with the objective of better understanding sentiment patterns, word commonalities, and the potential implications of such. In each book, who are the main characters, and what are the main themes? Perhaps, what kind of a mood did Charles Dickens typically have, based on the sentiment trends in his writing?

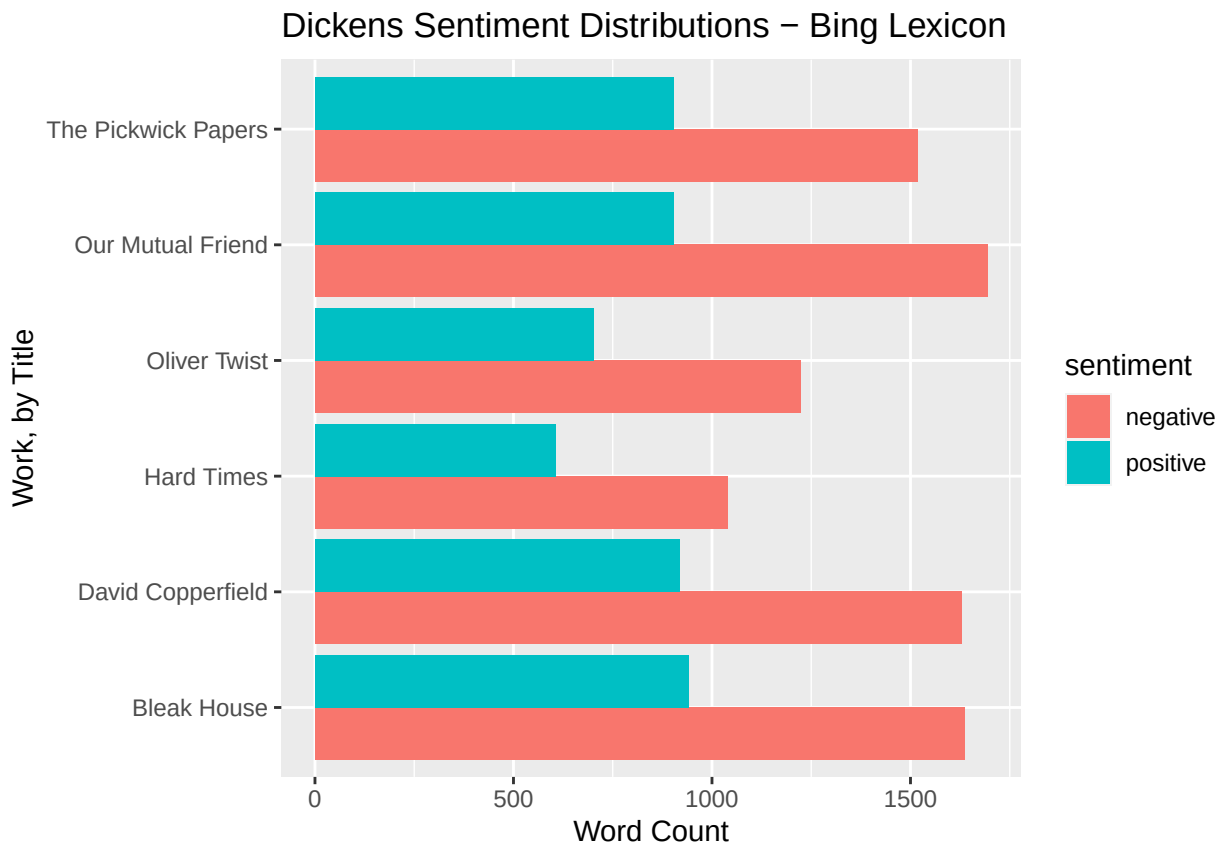
WORK OF CHARLES DICKENS INCLUDED IN SAMPLE

1. *Oliver Twist (1837)*
2. *David Copperfield (1849)*
3. *Hard Times (1854)*
4. *Bleak House (1852)*
5. *Our Mutual Friend (1865)*
6. *The Pickwick Papers (1836)*

ANALYSIS OF SENTIMENT IN DICKENS - USING THE BING LEXICON

The Bing lexicon consists of 6786 words, sorted into two categories, negative and positive, based on their perceived connotations. Such a lexicon was applied to the dataframe of the select works of Dickens in an effort to begin to understand the balance of sentiment (or lack thereof) thematic in his work, both in terms of individual works and, if possible, all of them. Is Dickens prone to writing generally negative works? A little more far-reaching, but can one imply the typicalities in his perspective and mood? What follows is an effort to arrive at more clear answers to such questions.

Bearing in mind the nature of Dickens as a serial novelist, which explains the similarities in word count for most of the works evaluated, it is obvious that there is a signifant overbearing of negative sentiment over positive. For each individual work, the count of words aligning with a negative sentiment per the Bing lexicon are significantly higher than words that are classified as negative.



It is strongly implied here the conclusion that sad themes are recurrent in the work of Dickens. However, this assertion could be even more strongly substantiated by a Chi-squared test.

```
chSq
```

```
##
##  Pearson's Chi-squared test
##
## data:  ct
## X-squared = 4.0559, df = 5, p-value = 0.5414
```

```
chSq$observed
```

```
##                    negative positive
## Bleak House            1637      941
## David Copperfield      1628      919
## Hard Times             1038      605
## Oliver Twist           1223      701
## Our Mutual Friend      1694      902
## The Pickwick Papers    1518      903
```

```
chSq$expected
```

```
##                       negative positive
## Bleak House          1643.195 934.8047
## David Copperfield    1623.436 923.5639
## Hard Times           1047.234 595.7658
## Oliver Twist         1226.341 697.6588
## Our Mutual Friend    1654.668 941.3317
## The Pickwick Papers  1543.125 877.8752
```

```
chSq$stdres
```

```
##                        negative   positive
## Bleak House          -0.2816649  0.2816649
## David Copperfield     0.2084625 -0.2084625
## Hard Times           -0.5051040  0.5051040
## Oliver Twist         -0.1708915  0.1708915
## Our Mutual Friend     1.7834267 -1.7834267
## The Pickwick Papers  -1.1705167  1.1705167
```

The Chi-squared test identifies a p-value of 0.541393, meaning that there is no level of significance here from a more numerical standpoint. With this lack of significance in mind, it can be concluded, therefore, that it is a typicality in Dickens to write generally negative pieces. What a sad individual he was!

ANALYSIS OF SENTIMENT IN DICKENS - USING THE AFINN LEXICON

Similarly to the Bing lexicon, the AFINN lexicon is used to evaluate the sentiment of a variety of words. In this case, the AFINN lexicon includes 2477 words from the English language. The key difference though is that rather than sorting individual words into different categories, AFINN instead assigns each included word an integer between -5 (most negative) and 5 (most positive). This is helpful in allowing us to understand the weight of a word's sentiment. In other words, in addition to showing that a word is negative or positive, it also helps us understand how negative or how positive a word is.

Here, we apply the AFINN lexicon to our dataframe consisting of all words from the selected work of Dickens, to better conceptualize the weight of the sentiment, in supplement to the overall sentiment implied in the previous analysis.

```
## # A tibble: 6 x 2
##   Name                 value
##   <chr>                <dbl>
## 1 Bleak House         -0.362
## 2 David Copperfield   -0.420
## 3 Hard Times          -0.355
## 4 Oliver Twist        -0.434
## 5 Our Mutual Friend   -0.451
## 6 The Pickwick Papers -0.395
```

It can be seen here that denser populations of words with negative connotations are present, particularly around -2. It can be seen that *Oliver Twist* and *Our Mutual Friend* both seem to contain exceptionally negative words with a sentiment of -5, the maximum negative value, whereas the latter works do not contain such words. *Hard Times* and *The Pickwick Papers* are the only works that evidently contain no words that equate to the maximum positive sentiment of 5, which is accomplished by the latter works. Generally, positive words have a sentiment around 2.

To give numerical insight, we once again conduct a Chi-squared test to allow for better comprehension of the above graph.

```
chSq2
```

```
##
##  Pearson's Chi-squared test
##
## data:  new_ct
## X-squared = 1.07, df = 5, p-value = 0.9567
```

```
chSq2$observed
```

```
##                     negative positive
## Bleak House              501      319
## David Copperfield        511      312
```

```
## Hard Times               359      233
## Oliver Twist             425      260
## Our Mutual Friend        519      305
## The Pickwick Papers      486      304
```

```
chSq2$expected
```

```
##                      negative positive
## Bleak House          506.5770 313.4230
## David Copperfield    508.4303 314.5697
## Hard Times           365.7239 226.2761
## Oliver Twist         423.1771 261.8229
## Our Mutual Friend    509.0481 314.9519
## The Pickwick Papers 488.0437 301.9563
```

```
chSq2$stdres
```

```
##                       negative   positive
## Bleak House         -0.4428301  0.4428301
## David Copperfield    0.2037524 -0.2037524
## Hard Times          -0.6099106  0.6099106
## Oliver Twist         0.1555638 -0.1555638
## Our Mutual Friend    0.7887197 -0.7887197
## The Pickwick Papers -0.1646629  0.1646629
```

Note the high p-value of 0.9567254, which is immediately indicative that, at the very least in the context of this sample size, there are no differences between these works of Dickens in terms of sentiment. This is rather consistent with the general uniformity in the "violin" graph above.

Though no clear classification could be found for the genre of these works with respect to sentiment (ex. Tragedy vs. Comedy) in the midst of searching various platforms, it is evident that Dickens' work, particularly the work included in the sample, are likely to be considered tragedies, or at least more broadly declared works of a sad nature. Moreover, it is inferrable that Dickens is prone to writing tragedies.

ANALYSIS OF WORD COMMONALITIES IN DICKENS - WORDCLOUDS

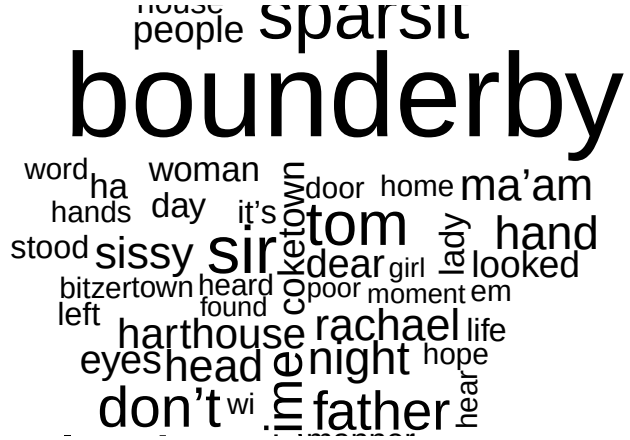
The frequencies of words can sometimes be indicative of major themes/characters/sentiment/etc. in a work. Here, we use wordclouds as a means of visualizing the most common words in Dickens' work.

Below are a series of wordclouds for each work of Dickens. The order of the wordclouds correspond to the order of the following list.

```
## [1] "Oliver Twist"
```


```
## [1] "David Copperfield"
```

```
## [1] "Hard Times"
```




```
## [1] "Bleak House"
```

```
## [1] "Our Mutual Friend"
```


```
## [1] "The Pickwick Papers"
```

Some observations worth noting are the immediately evident dominance of the names of main characters and elements from the titles of the story. For example, *Oliver Twist* is well-implied to place emphasis on the main character, Oliver, based on the large size of that particular word. A similar case is evident for the Boffins family in *Our Mutual Friend* and Samuel Pickwick in *The Pickwick Papers*. Upon further investigation in *Bleak House*, it is known beforehand that Lady Dedlock is a main character in the novel, explaining, at least in part, the dominance of the word "lady" in that particular cloud.

It is also suggestable that time bears significance as a theme in *Bleak House*, based on the frequency of such a word in that particular cloud.

It is known that there are several significant characters in *David Copperfield*, including Copperfield himself, his mother, his stepfather, Edward Murdstone, and Dora Spenlow, which may explain why words and names in this novel, albeit outstanding, are represented slightly less prominently in its respective wordcloud, opposed to other works.

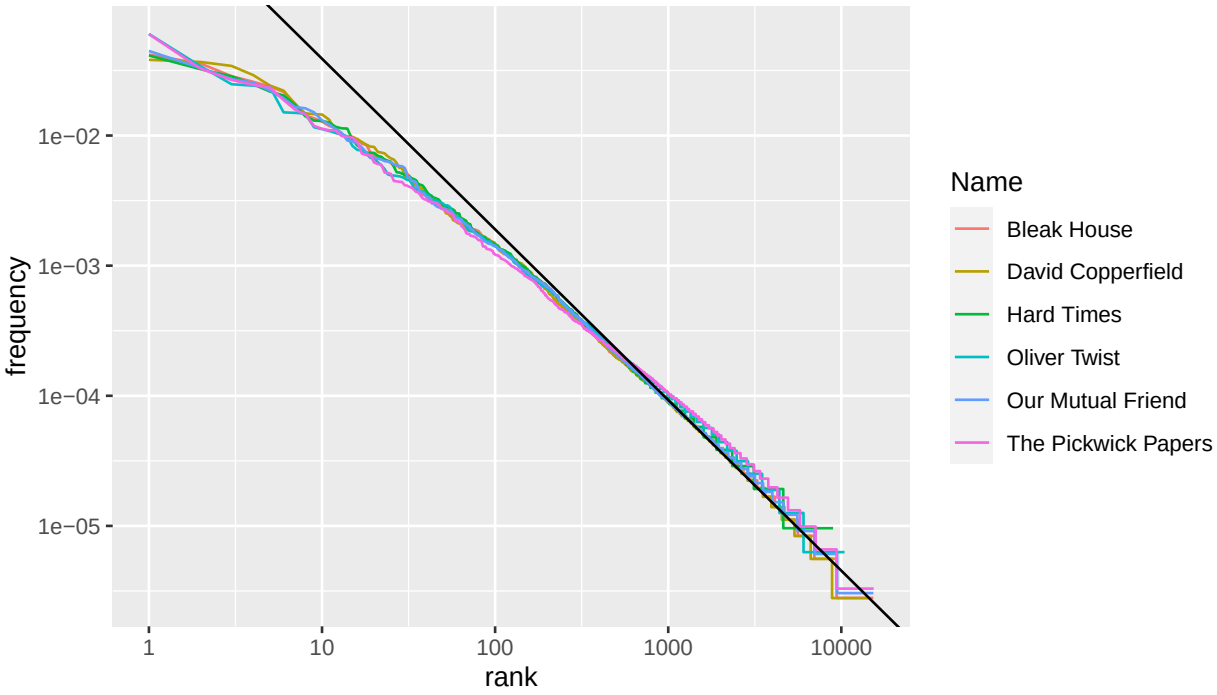
CUMULATIVE WORDCLOUD: Below is a wordcloud based on the master dataframe containing all words from all work from the sample.

The significance of looking at individual wordclouds comes to further fruition with this cumulative wordcloud. When creating a wordcloud based on a consolidated list of words across all works in the sample, there is a predominance in character names that significantly outweighs any other potential themes in the works, even though they are occasionally present. This wordcloud is rather effective in conveying to us important characters throughout Dickens' work. On a more secondary level, it is functional in implying recurring themes across all works that may be inferred through the presence of various words.

ANALYSIS OF WORD COMMONALITIES IN DICKENS - ZIPF'S LAW

Zipf's law is an empirical law that asserts that the frequency of an object, such as a common word like "the", is inversely related to that object's rank in a sorted list. To better understand word frequencies in Dickens' work, we apply Zipf's law to a custom dataframe containing every word across all of Dickens' work, including stop words, where a column is added which assigns each word their appropriate rank.

Below is a rank-frequency graph representative of all words in Dickens' work.

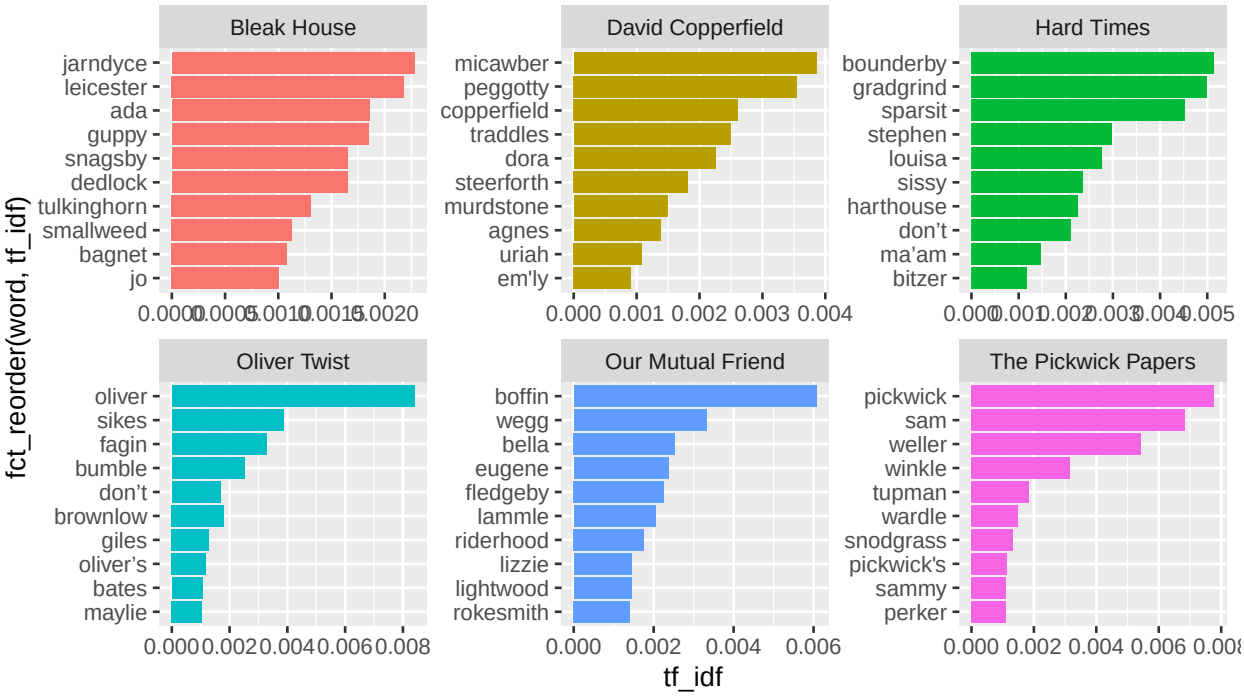


Evidently, word usage patterns across all works in the sample modestly comply with Zipf's Law, given the somewhat linear relationship with rank versus frequency. Upon investigation of the table, several outliers can be seen, particularly character names in their respective books.

ANALYSIS OF WORD COMMONALITIES IN DICKENS - tf-idf

tf-idf analysis entails "downweighting" common items that are recurring in more than just one dataframe. In this case, we downweight words that are common across all of Dickens' work from the sample, rather than ones that are common in an individual work. To do this, we multiply the frequency of the occurrence of a word in a document, "tf", with the inverse document frequency "idf".

The objective is to visualize commonalities in word-usage patterns in the context of individual works. The table below portrays the top ten words for each work in the sample on the basis of their frequency.



It can be seen across all six works that the most frequent words are the names of their respective main characters.

In the particular case of *Bleak House*, the highest word frequency implies the importance of the surrogate case of *Jarndyce v. Jarndyce*, with the word frequencies of other characters preceding each other down the line by modest margins (i.e. Leicester Dedlock's name is the second most frequent word by a small margin, followed by Ada Clare, and so on).

There are more significant margins between the top frequent word and the latter words in works such as *Our Mutual Friend* and *Oliver Twist*. Unusually, in *Hard Times* and *The Pickwick Papers*, the top frequent word is closely followed by two other words, with the latter words appearing much less frequently, implicative that there is a difference in importance or recurence in groups of characters in these works.

Another interesting observation is the presence of the words "don't" in *Oliver Twist* and *Hard Times*. This appears to be a stop word that was not enough present across all documents, thus was not included in the downweight procedure. Perhaps Dickens' writing patterns occasionally deviated.