Homework 2C - DATA-312

Jeffrey Williams

09, April 2022

Abstract

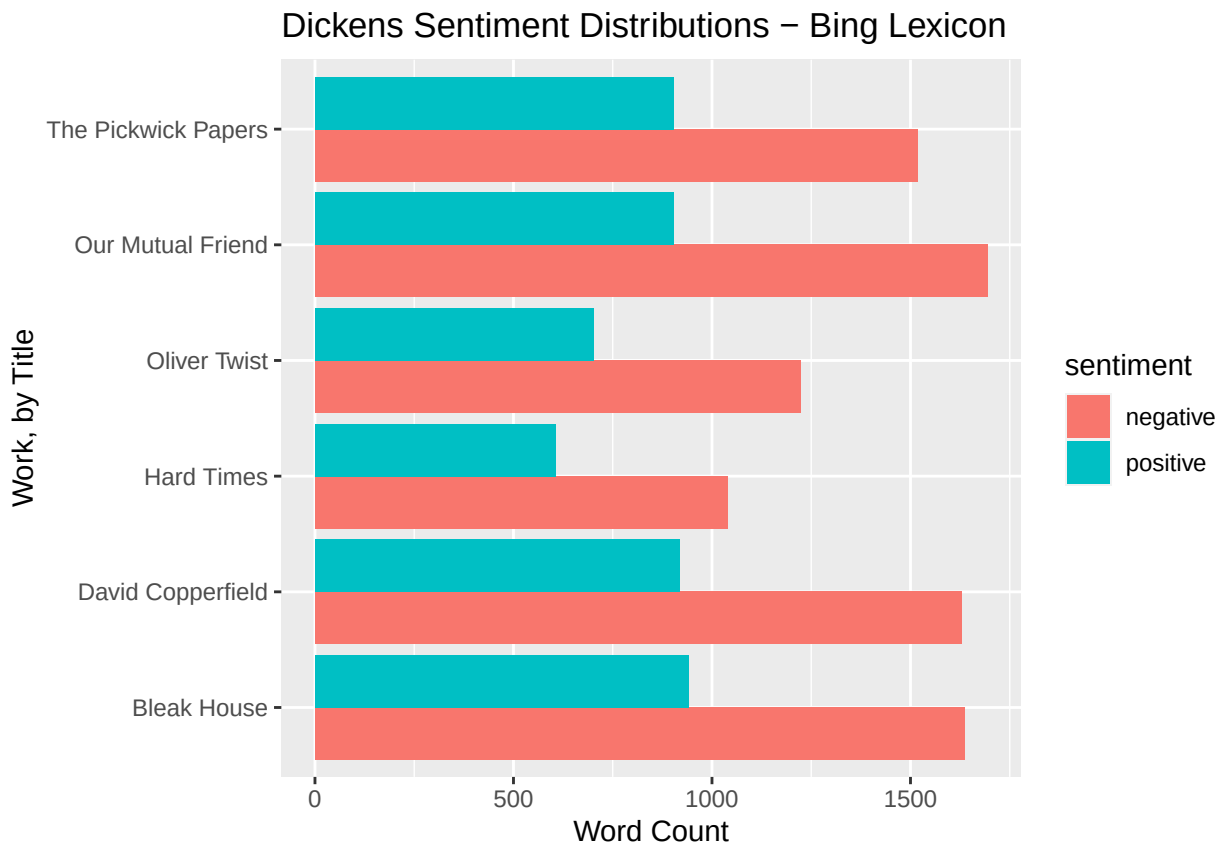
This writeup explores the

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 proportion 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.

```
## Joining, by = "word"
```



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-square test.

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

##                     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

##                     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
```

```
##                        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 numerial insight, we once again conduct a Chi-Squared test to allow for better comprehension of the above graph.

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

```
##                     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
```

```
##                      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
```

```
##                       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 0.541393, 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.

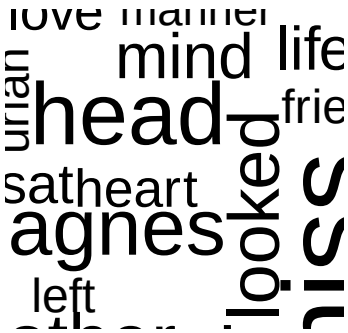
Though no clear classification could be found for the genre of these works with respect to sentiment (ex. Tragedy vs. Comedy) in the midst of searching various platforms, it is evident that Dickens' work, particularly the work included in the sample, are likely to be considered tragedies. Moreover, it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

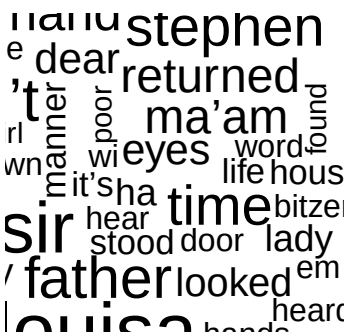
ANALYSIS OF WORD COMMONALITIES IN DICKENS - WORDCLOUDS

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.

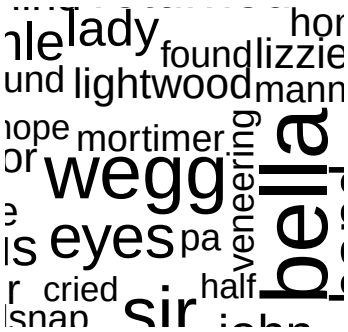
Below are a series of wordclouds for each work of Dickens.


lady
round
oliver
boy
words
noah
don’t
day
doctor
inquired
master
looked
hands
fagin
it’s
girl
that’s
house
left
bumble
table
bill
brownlow
oliver’s
monks
hand
cried
woman
rejoined
voice
time
hear
life
half
night
giles
sir
mind
child
jew
head
dear
replied
door


mind
life
head
looked
satheart
agnes
left


stephen
dear
returned
ma’am
manner
poor
found
eyes
word
it’s
life
ha
time
sir
hear
stood
door
lady
father
looked

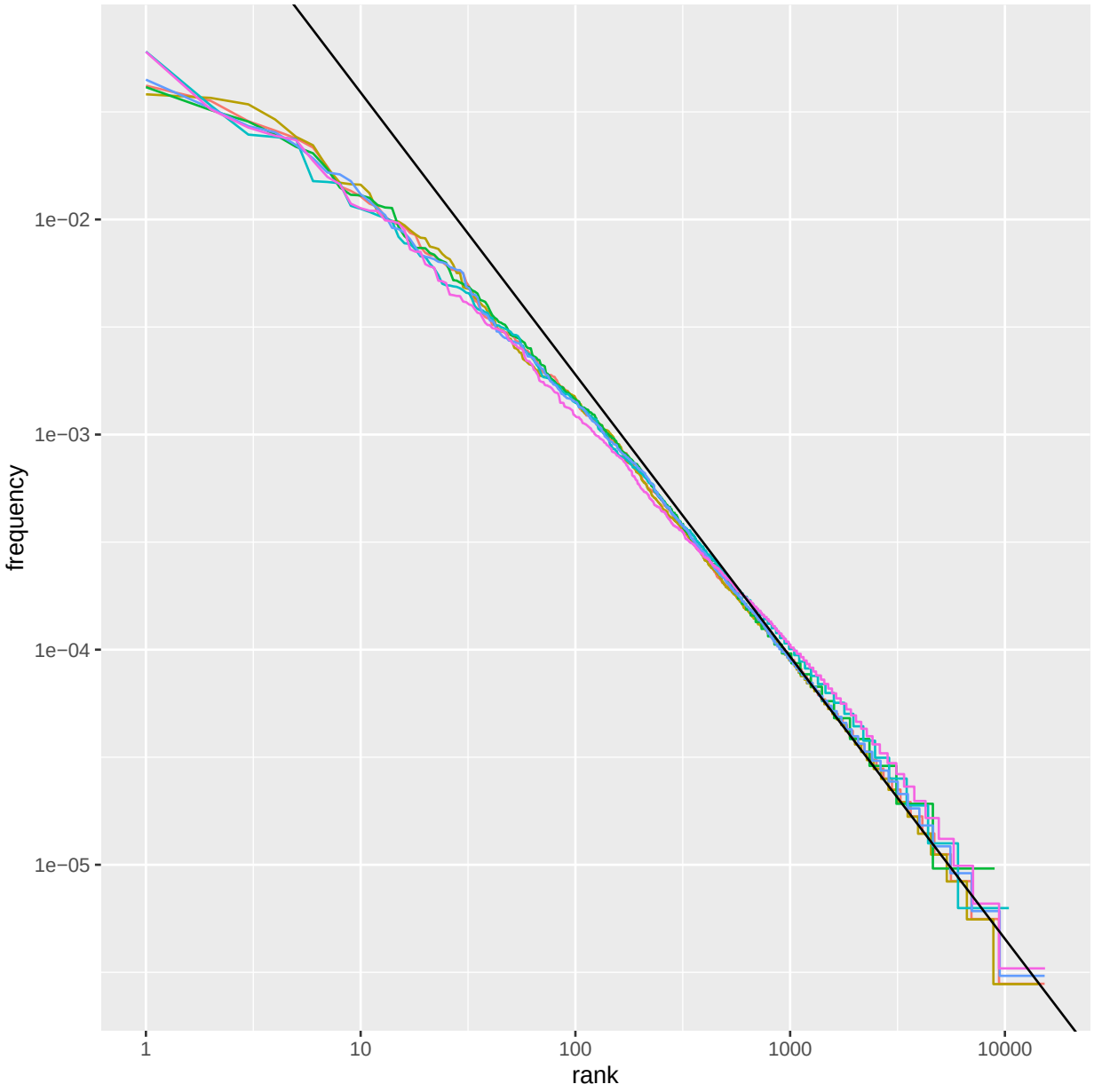

day
hand
charley
motherd
agsby
smallweed
heard
jarndyce
life
round
home
kinghorn

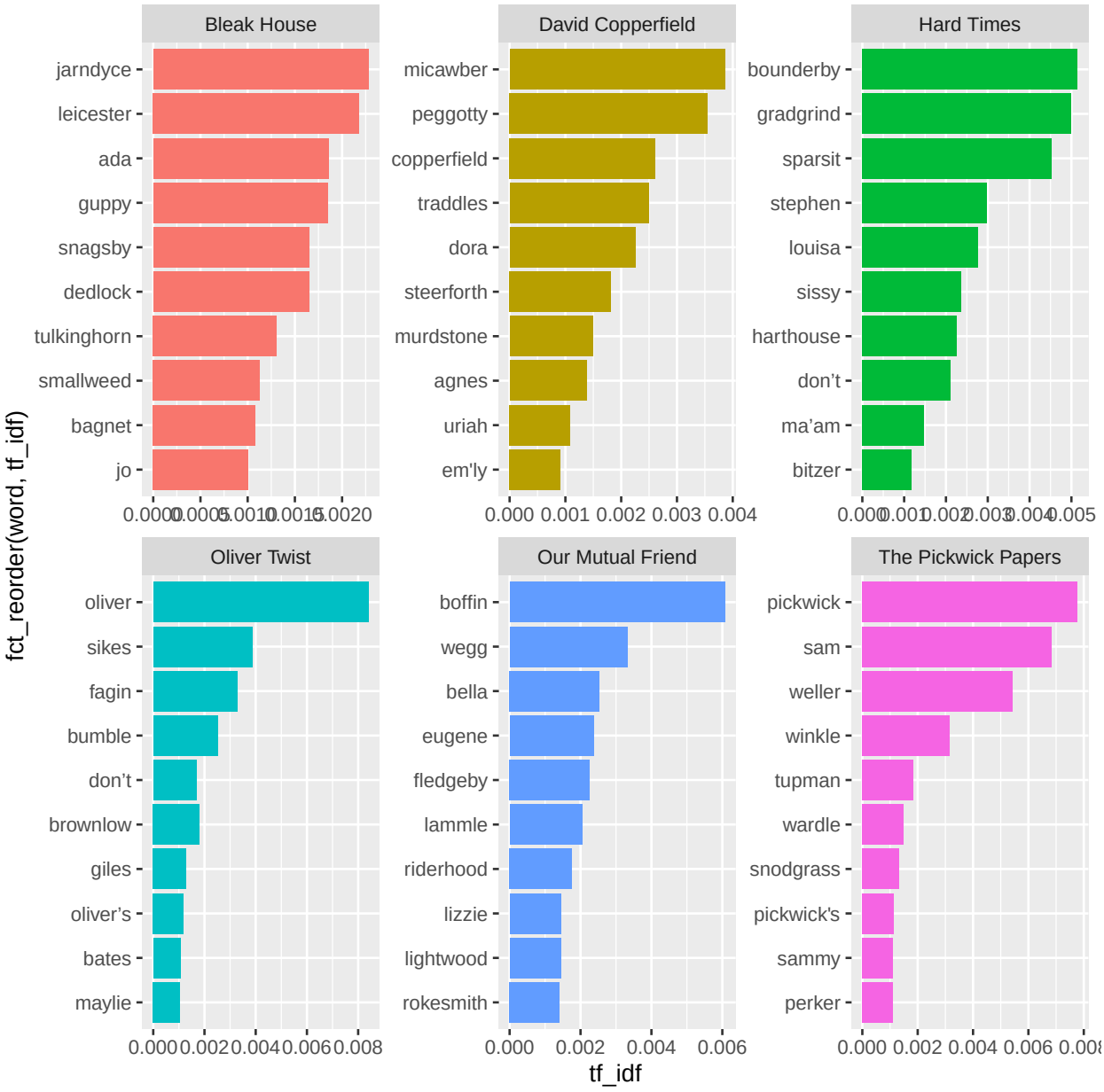


Some observations worth noting are the immediately evident dominance of the names of main characters and elements from the titles of the story. For example, Oliver Twist is well-implied to place emphasis on the main character, Oliver, based on the largest size

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

ANALYSIS OF WORD COMMONALITIES IN DICKENS - ZIPFS LAW


1e−02
1e−03
1e−04
1e−05
frequency
1
10
100
1000
10000
rank


Bleak House
jarndyce
leicester
ada
guppy
snagsby
dedlock
tulkinghorn
smallweed
bagnet
jo
0.0000 0.0005 0.0010 0.0015 0.0020
David Copperfield
micawber
peggotty
copperfield
traddles
dora
steerforth
murdstone
agnes
uriah
em'ly
0.000 0.001 0.002 0.003 0.004
Hard Times
bounderby
gradgrind
sparsit
stephen
louisa
sissy
harthouse
don't
ma'am
bitzer
0.000 0.001 0.002 0.003 0.004 0.005
Oliver Twist
oliver
sikes
fagin
bumble
don't
brownlow
giles
oliver's
bates
maylie
0.000 0.002 0.004 0.006 0.008
Our Mutual Friend
boffin
wegg
bella
eugene
fledgeby
lammle
riderhood
lizzie
lightwood
rokesmith
0.000 0.002 0.004 0.006
The Pickwick Papers
pickwick
sam
weller
winkle
tupman
wardle
snodgrass
pickwick's
sammy
perker
0.000 0.002 0.004 0.006 0.008
fct_reorder(word, tf_idf)
tf_idf