nlp.Rmd

Natural language processing

Exploratory data analysis

The tibble extracted from the graph analysis contains two fields that will be used to determine word usage differences, if any, among the three graph clusters, payload (the original content of the email) and f_cluster (the cluster to which the email sender was assigned).

The dataset consists of 4765 records.

Any collection of natural language contains many high-frequency words that obscure differences. These words, such as *a, as, an, and, of, this, that, which, what* and the like are censored from analysis along with numerals.[1]

There are a variety of stopword lists. Some may be under-inclusive, removing only the most common parts of speech, while others may be over-inclusive, removing words that should be kept. For example, the word `not` often appears on lists of stopwords but is an important term in doing sentiment analysis using phrases – *not good* has a difference valence than *[blank] good*.

The `stop_word` data from the `tidytext` package was hand edited to exempt the following list of words:

- against
- all
- allow
- allows
- always
- awfully
- beforehand
- behind
- below
- better
- big
- can't
- cannot
- cant
- case
- cases
- downwards
- except
- gives
- good
- great
- highest
- hopefully
- immediate
- might
- necessary
- not
- nowhere
- numbers
- otherwise

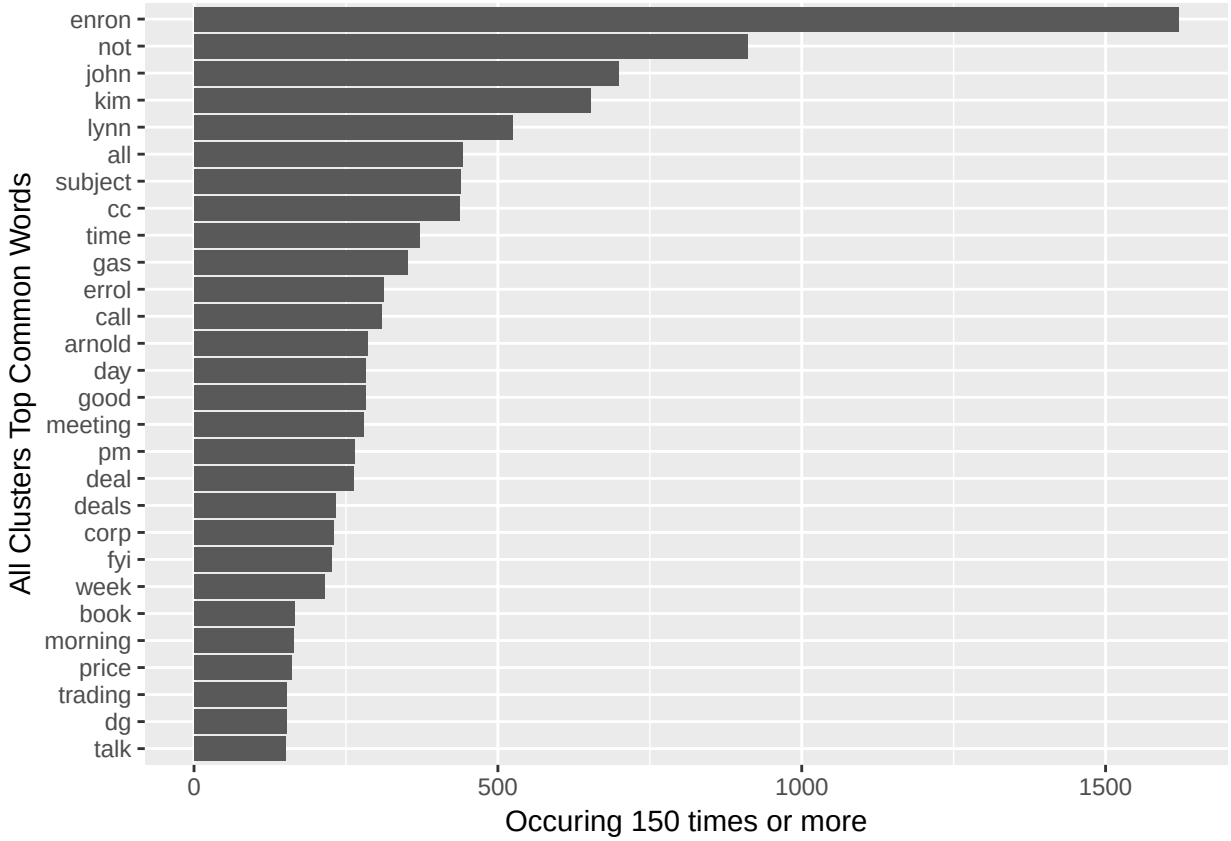
[1] The entire original contents are retained in the original `n_enron` object, however.

- point
- serious
- seriously
- state
- states
- unfortunately
- value
- worked
- working
- zero

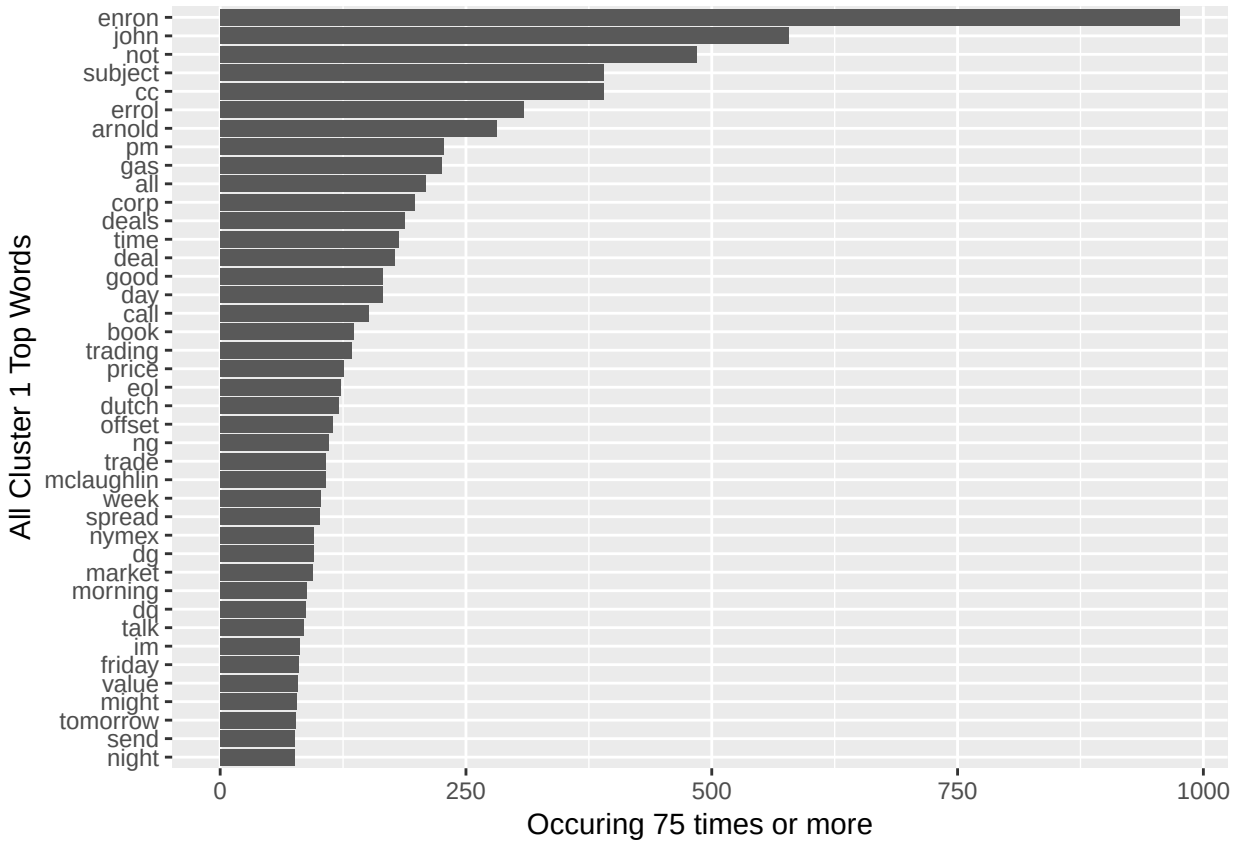
and to add `ect`, `hou` and `enronxgate`, which from data cleansing done originally were found to be email address components from an older email system.

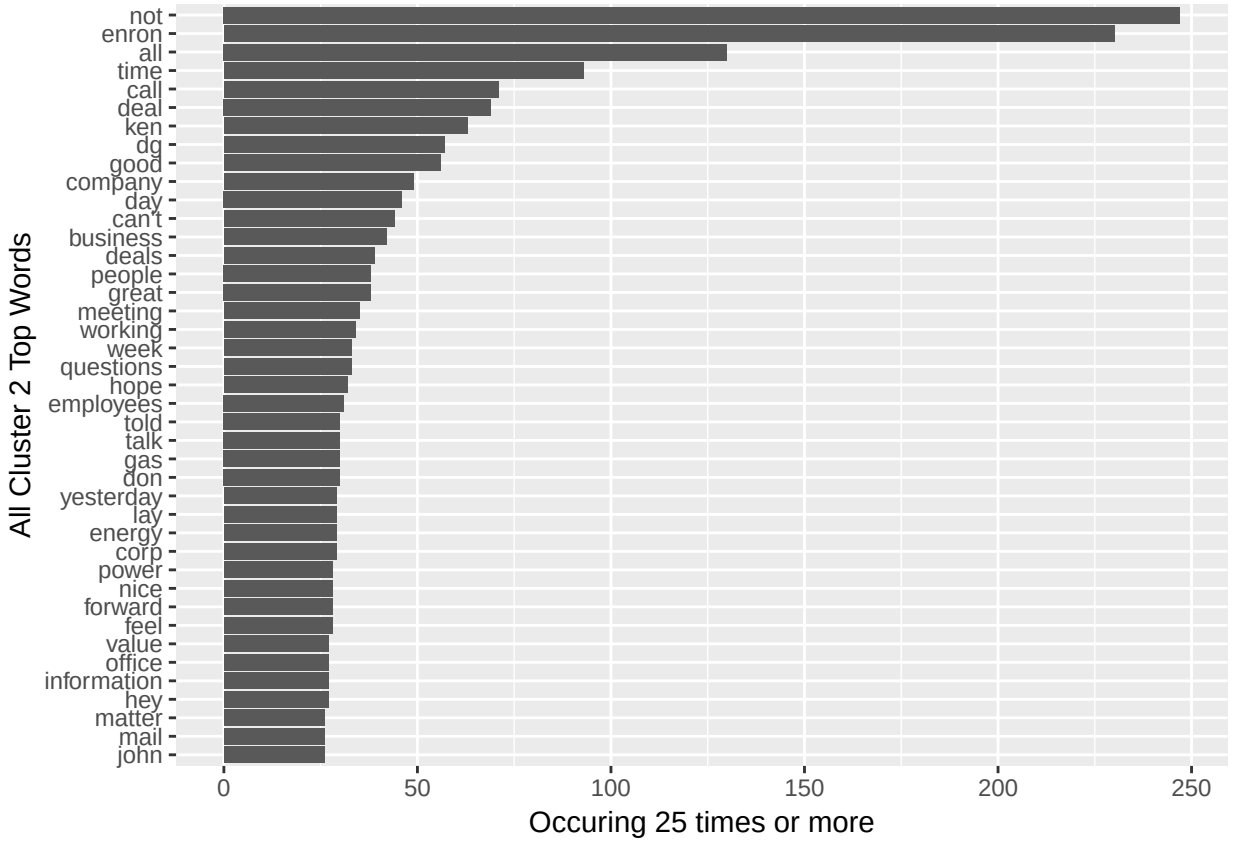
The 4765 mails contail a total of 68561 words, of which 10675 are distinct.

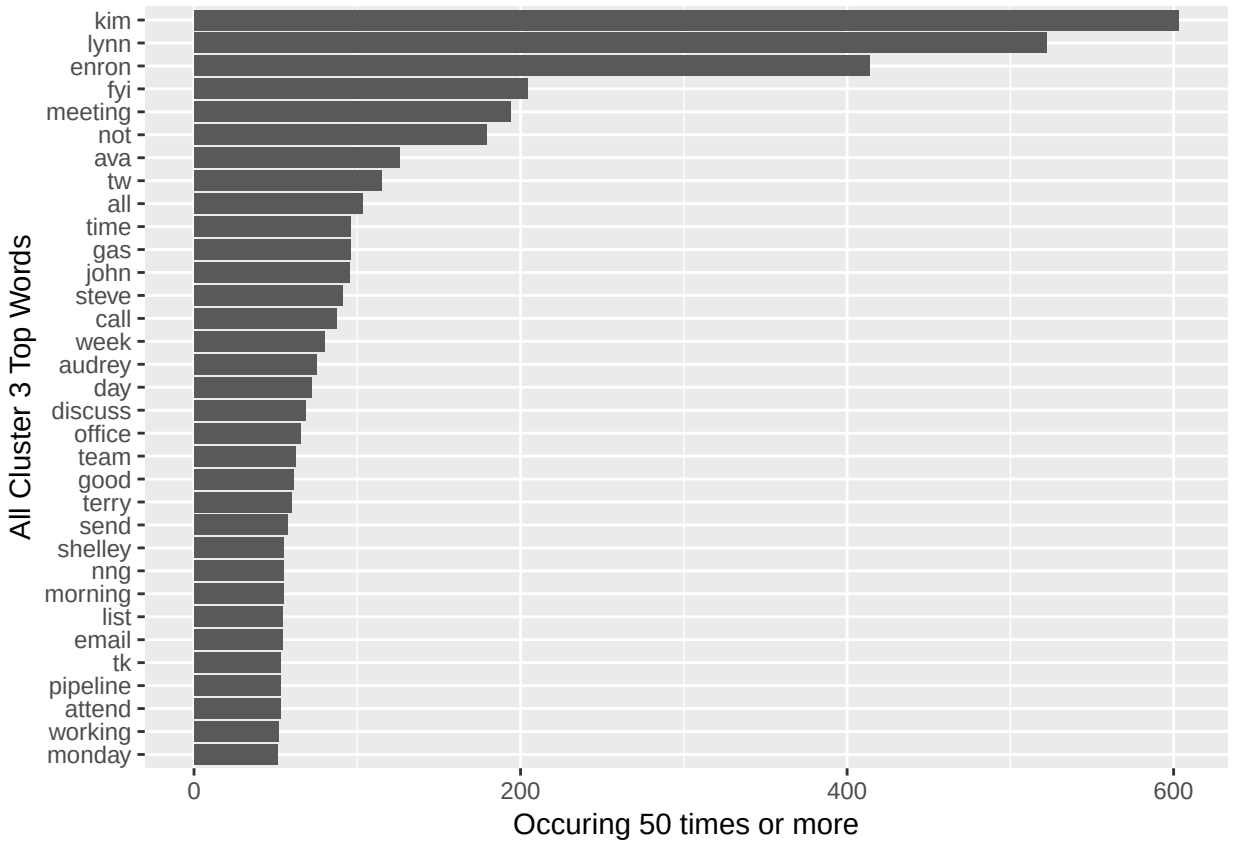
As a whole, the following chart shows the most frequent words in `n_enron` that occur 150 or more times.



Here are the top words in each cluster, after removing the adjusted stopwords and numbers.


All Cluster 1 Top Words
enron
john
not
subject
cc
errol
arnold
pm
gas
all
corp
deals
time
deal
good
day
call
book
trading
price
eol
dutch
offset
ng
trade
mclaughlin
week
spread
nymex
dg
market
morning
dd
talk
im
friday
value
might
tomorrow
send
night
0
250
500
750
1000
Occuring 75 times or more

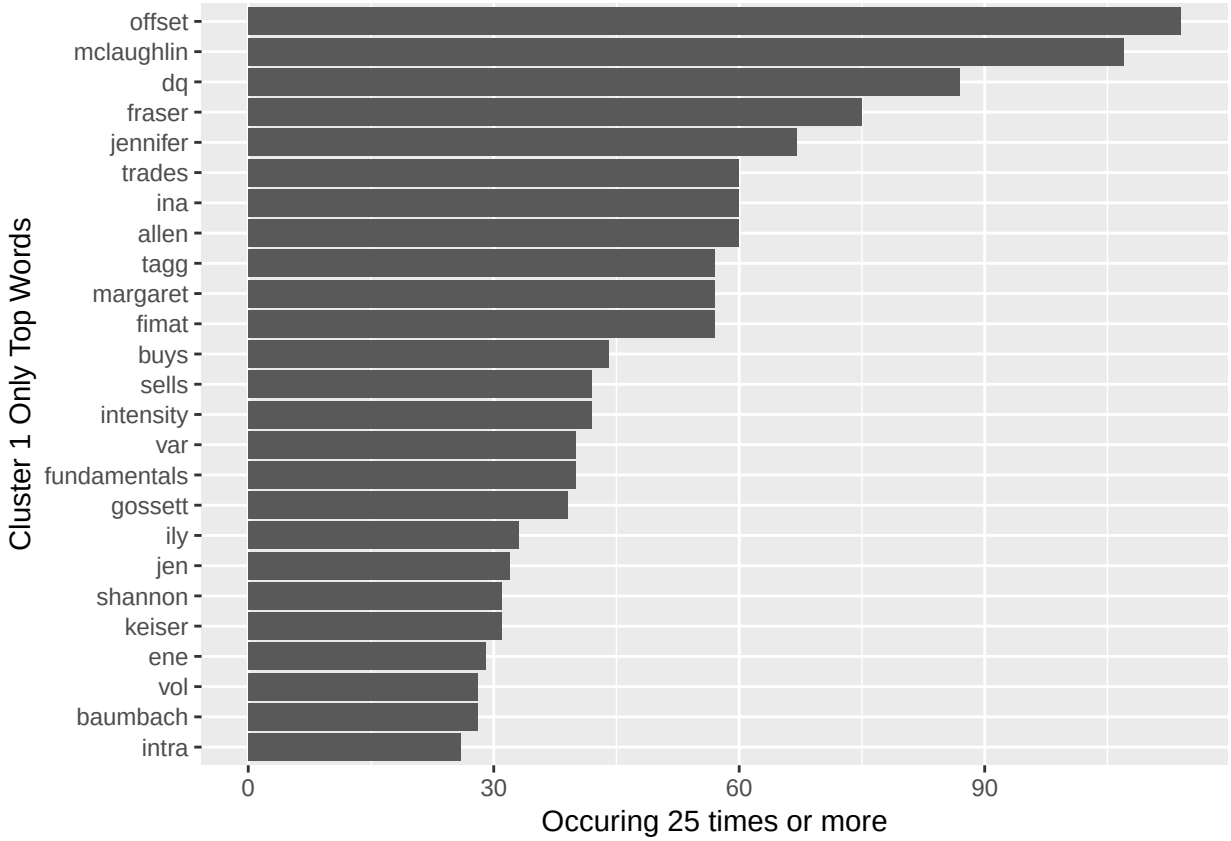

All Cluster 2 Top Words
not
enron
all
time
call
deal
ken
dg
good
company
day
can't
business
deals
people
great
meeting
working
week
questions
hope
employees
told
talk
gas
don
yesterday
lay
energy
corp
power
nice
forward
feel
value
office
information
hey
matter
mail
john
0
50
100
150
200
250
Occuring 25 times or more

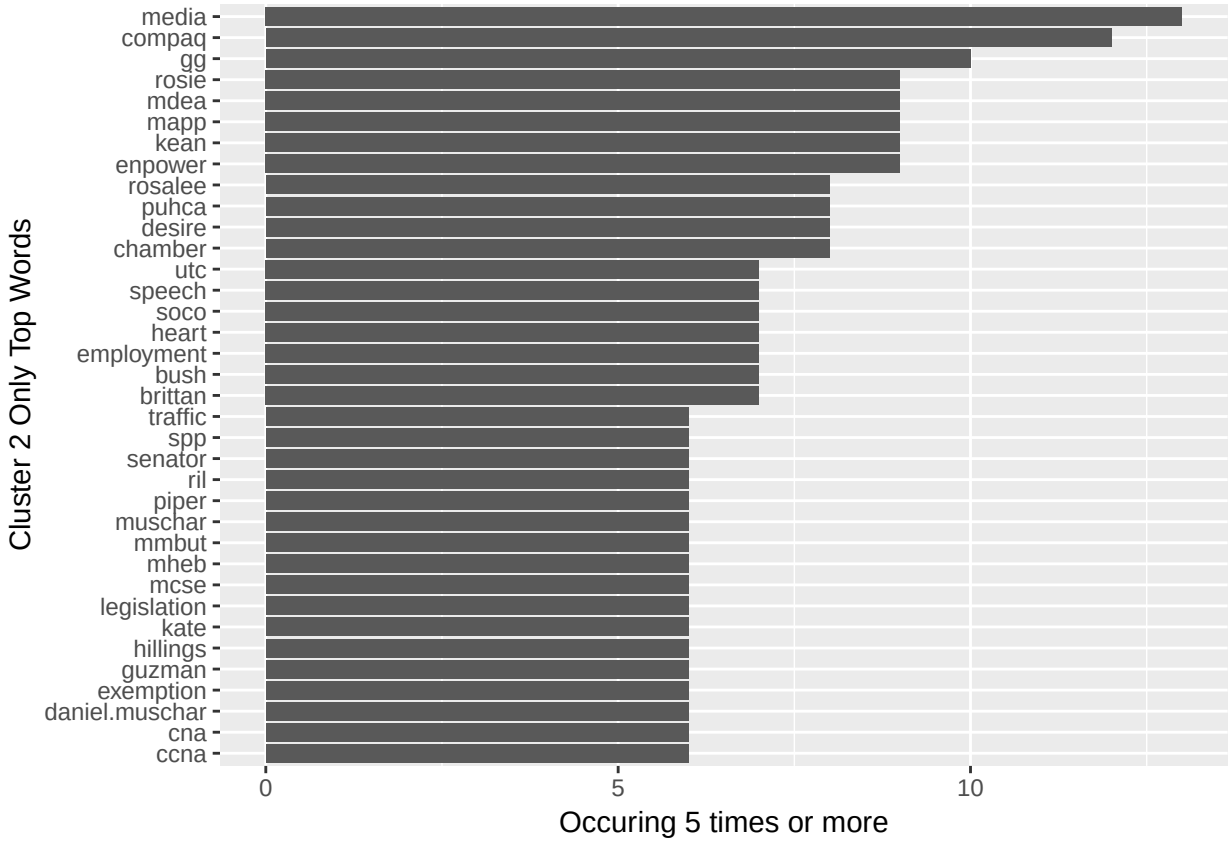


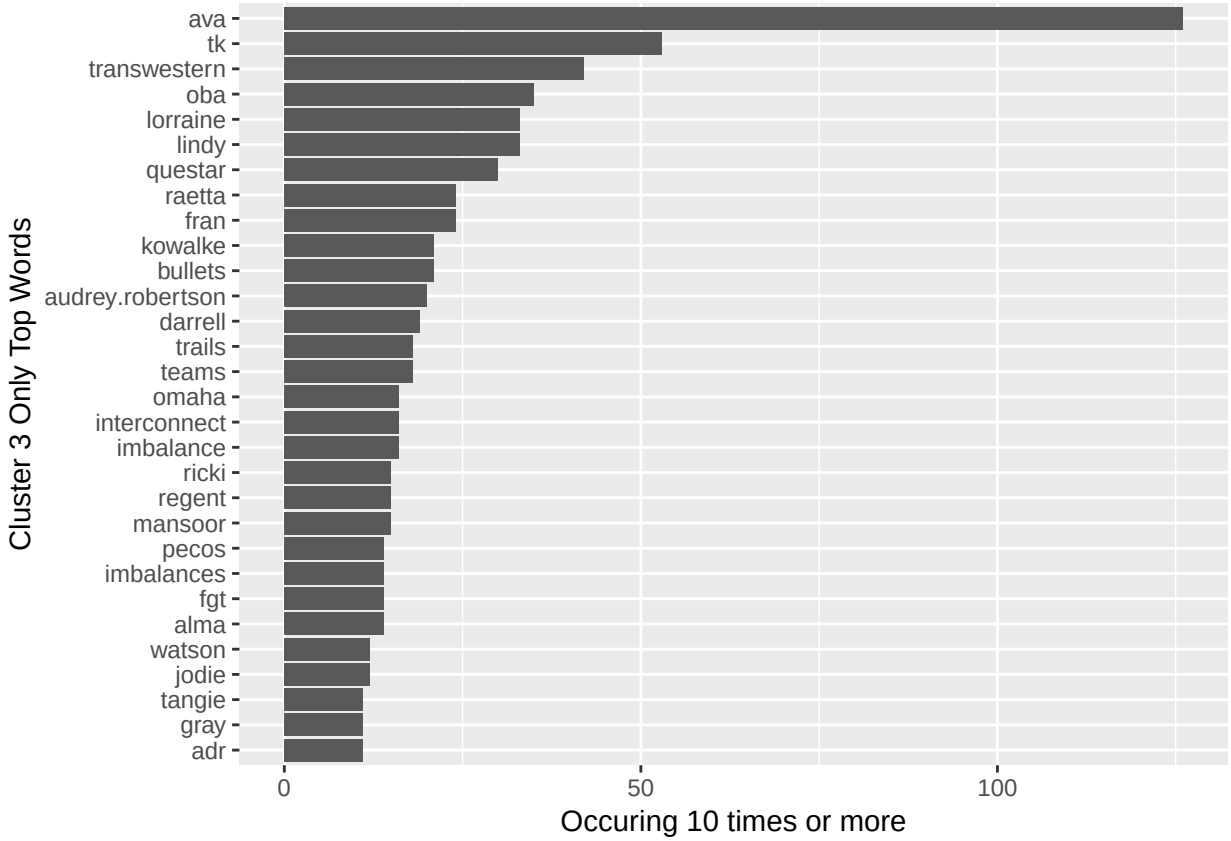
The clusters differ in the number of total words: Cluster 1 has 36653, Cluster 2, 14290 and Cluster 3, 17618.

The clusters also have distinct vocabularies. The three clusters all share 4849 terms in common. Cluster 1 includes 3934 terms that do not appear in Clusters 2 or 3. Cluster 2 includes 1892 terms that do not appear in Clusters 1 or 3. Cluster 3 includes 1502 terms that do not appear in Clusters 1 or 2.

The top unique terms in each cluster are shown in the following three charts.


Cluster 1 Only Top Words
offset
mclaughlin
dq
fraser
jennifer
trades
ina
allen
tagg
margaret
fimat
buys
sells
intensity
var
fundamentals
gossett
ily
jen
shannon
keiser
ene
vol
baumbach
intra
0
30
60
90
Occuring 25 times or more


Cluster 2 Only Top Words
media
compaq
gg
rosie
mdea
mapp
kean
enpower
rosalee
puhca
desire
chamber
utc
speech
soco
heart
employment
bush
brittan
traffic
spp
senator
ril
piper
muschar
mmbut
mheb
mcse
legislation
kate
hillings
guzman
exemption
daniel.muschar
cna
ccna
0
5
10
Occuring 5 times or more



The proportion of cluster-specific words to total words for each cluster are:

- Cluster 1: 0.240335
- Cluster 2: 0.193352
- Cluster 3: 0.1804972

To test whether the differences in vocabulary and frequently used terms, a correlation[2] test may be applied.

Table 1: Pearson's product-moment correlation: `n_freq$Cluster1` and `n_freq$Cluster2`

Test statistic	df	P value	Alternative hypothesis	cor
-0.7467	10673	0.4553	two.sided	-0.007228

Table 2: Pearson's product-moment correlation: `n_freq$Cluster1` and `n_freq$Cluster3`

Test statistic	df	P value	Alternative hypothesis	cor
-0.114	10673	0.9093	two.sided	-0.001103

[2]Pearson product-moment correlation

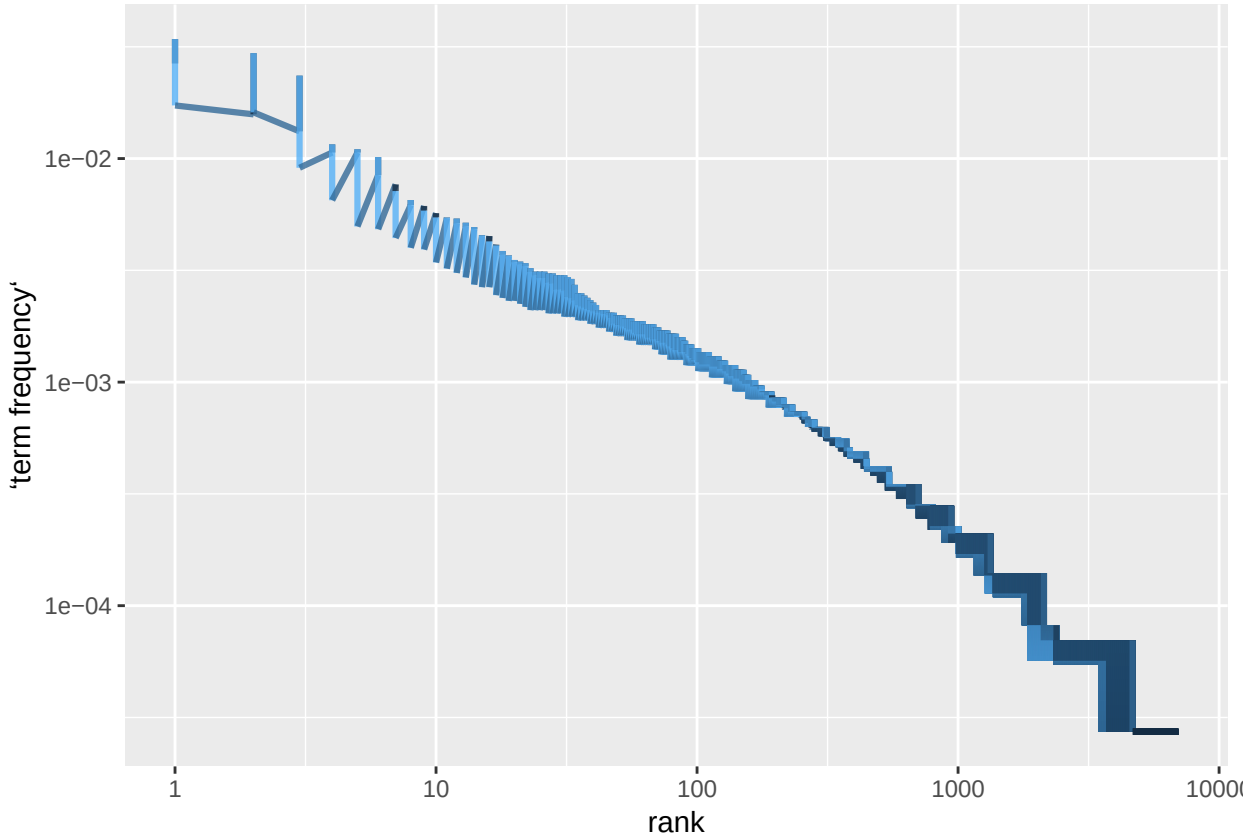
Table 3: Pearson's product-moment correlation: `n_freq$Cluster2` and `n_freq$Cluster3`

Test statistic	df	P value	Alternative hypothesis	cor
-0.3436	10673	0.7312	two.sided	-0.003326

With respect to word frequencies, the three clusters have very week negative coefficients, but the corresponding p-values require rejection of the null hypothosis that they are uncorrelated. The three clusters can be distinguished not only by volume, percentage of cluster-specific vocabulary, but also by the lack of intra-cluster correlations.

However, term frequency alone overweights frequently occurring words and should be supplemented by considering rarely occurring words in combination.[3] A plot of how often a term occurs within each cluster by its rank order shows the straight line log-log plot characteristic of a power-law distribution.

```
freq_by_rank %>%
 ggplot(aes(rank, `term frequency`, color = f_cluster))   +
 geom_line(size = 1.1, alpha = 0.8, show.legend = FALSE)  +
 scale_x_log10()                                          +
 scale_y_log10()
```



Regressing the log of term frequency on the log of term rank illustrates how closely the distribution follows that model.

```
pander(summary(fit))
```

[3]The analysis of `td-idf` here follows the examples given in chapter 3 of [Tidytext].

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.604	0.006926	-231.7	0
log10(rank)	-0.6547	0.002975	-220	0

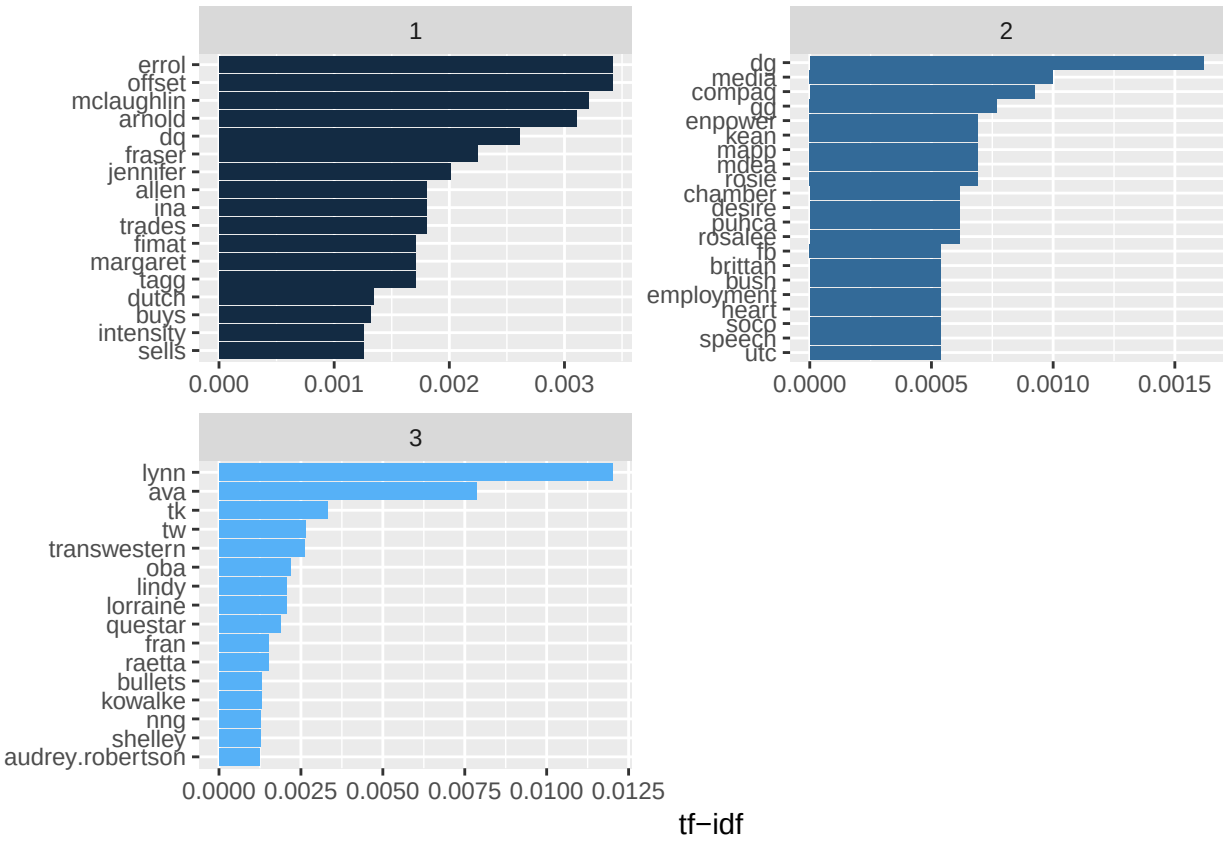
Table 5: Fitting linear model: log10(`term frequency`) ~ log10(rank)

Observations	Residual Std. Error	R^2	Adjusted R^2
1467	0.04059	0.9706	0.9706

The `td-idf` (term frequency/inverse term frequency) word table provides a more nuanced and useful depiction of differences in word usage among the graph clusters.

```
c_words                              %>%
  select(-total)                     %>%
  arrange(desc(tf_idf))              %>%
  mutate(word =
    factor(word, levels =
             rev(unique(word))))     %>%
  group_by(f_cluster)                %>%
  top_n(16)                          %>%
  ungroup                            %>%
  ggplot(aes(word, tf_idf, fill = f_cluster))       +
  geom_col(show.legend = FALSE)                     +
  labs(x = NULL, y = "tf-idf")                      +
  facet_wrap(~f_cluster, ncol = 2, scales = "free") +
  coord_flip()
```

```
## Selecting by tf_idf
```



In these charts, there is more specificity. Names are more prominent, both natural and business, in clusters 1 and 3; while cluster 2 also contains names, it has more abstract terms than the other clusters. "Media, puhca[4]. employment, and speech" are examples. The abstract terms "buys, sells, trades" in cluster 1 are suggestive of an operation that engages in frequent transactions. Cluster 3 has "transwestern, questar and nng", which were all natural gas pipeline distributors.

Although the word associations differences among are suggestive of distinct aspects of Enron's business, machine learning has the potential to further organize the email texts to find other connections and distinctions through a process of latent Dirichet allocation, or `LDA`[5]

Machine learning

[4]Public Utilities Holding Company Act
[5]Not to be confused with linear discriminant analysis.