nlp.Rmd

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
- point
- serious
- seriously
- state

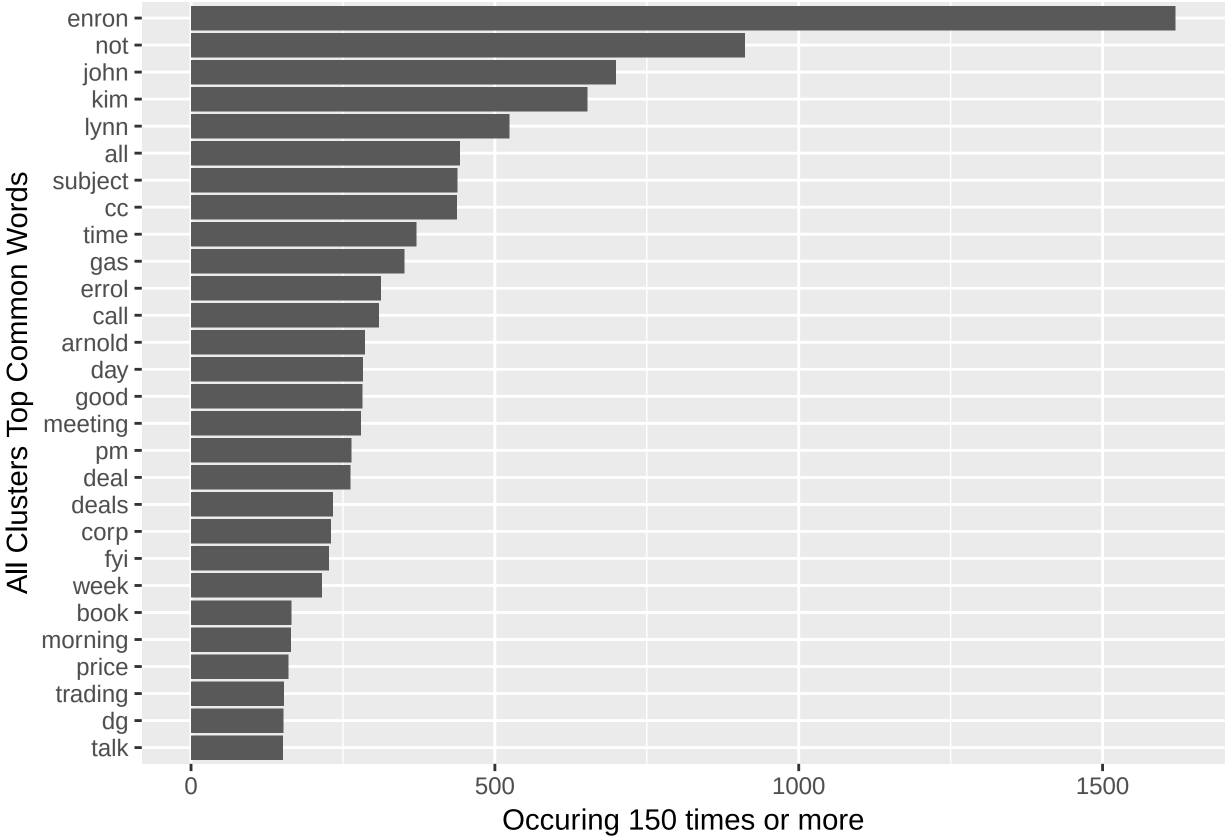
[1] The entire original contents are retained in the original `n_enron` object, however.

- states
- unfortunately
- value
- worked
- working
- zero

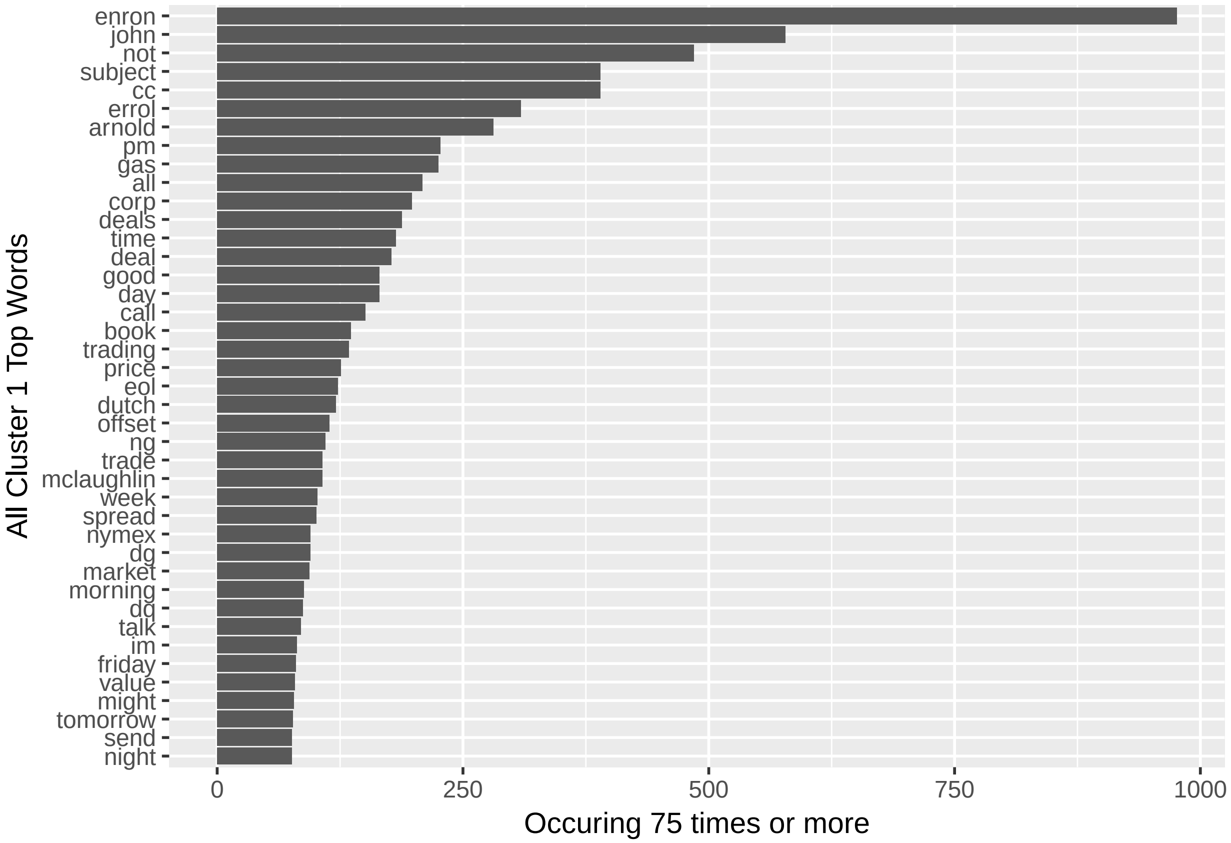
and to add ect, hou and enronxgate, which from data cleansing done originally were found to be email address components from an older email system.

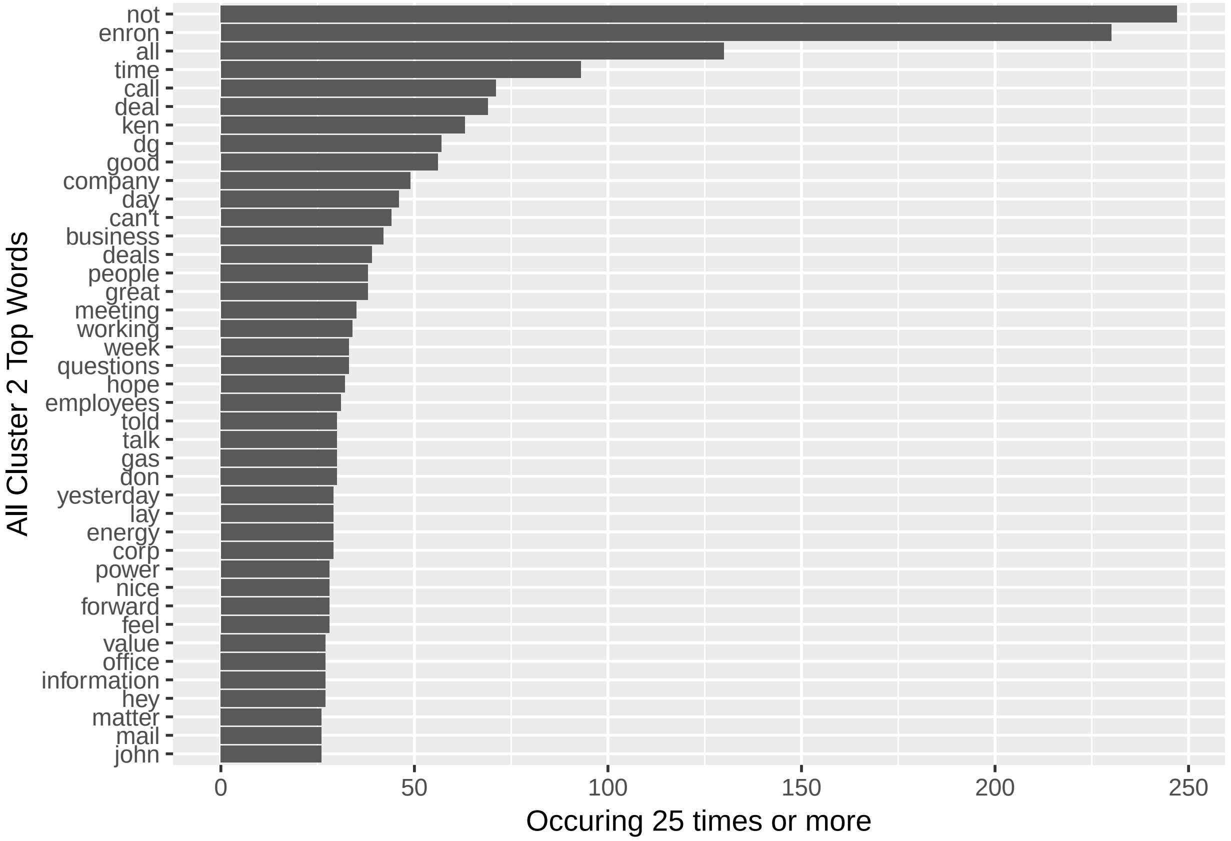
The 4765 mails contail a total of 68561 words, of which 10675 are distinct.

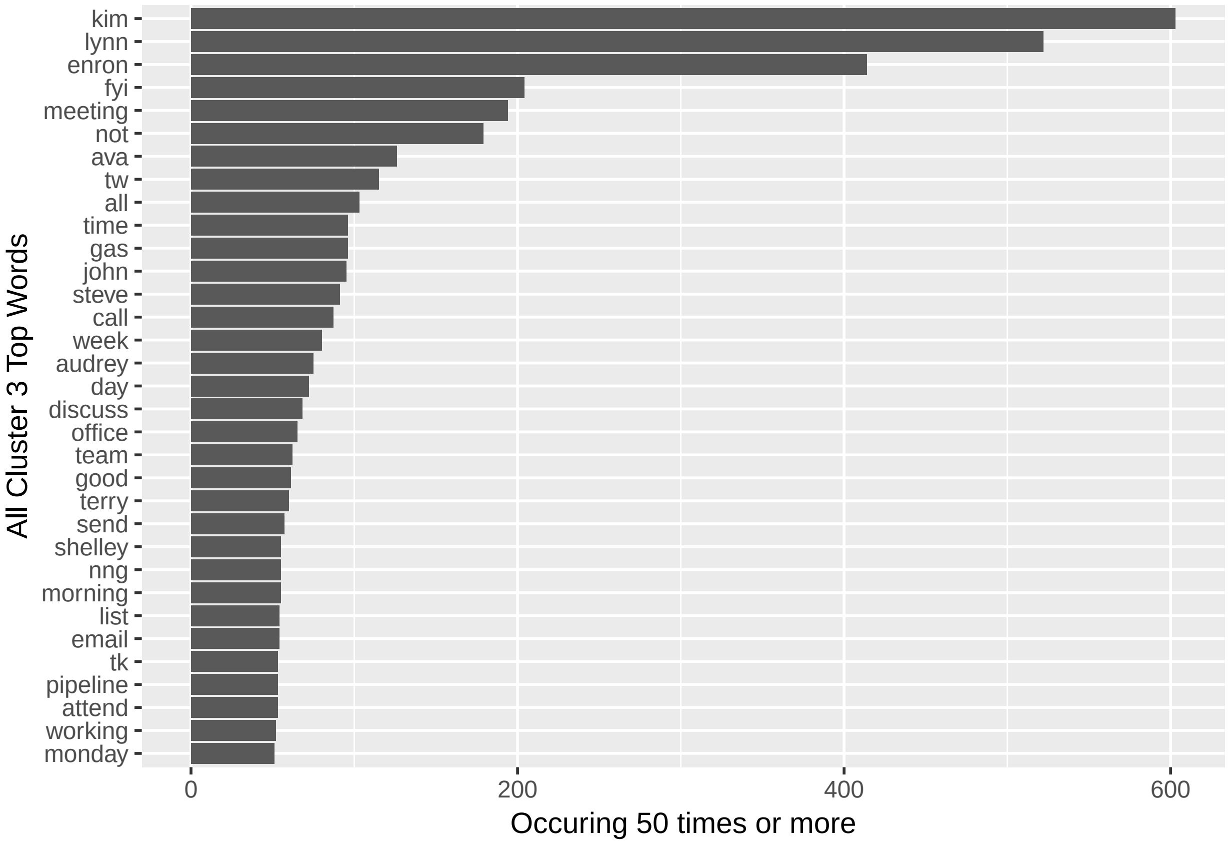
As a whole, the following chart shows the most frequent words in n_enron that occur 150 or more times.



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
dq
talk
im
friday
value
might
tomorrow
send
night
0
250
500
750
1000
Occuring 75 times or more

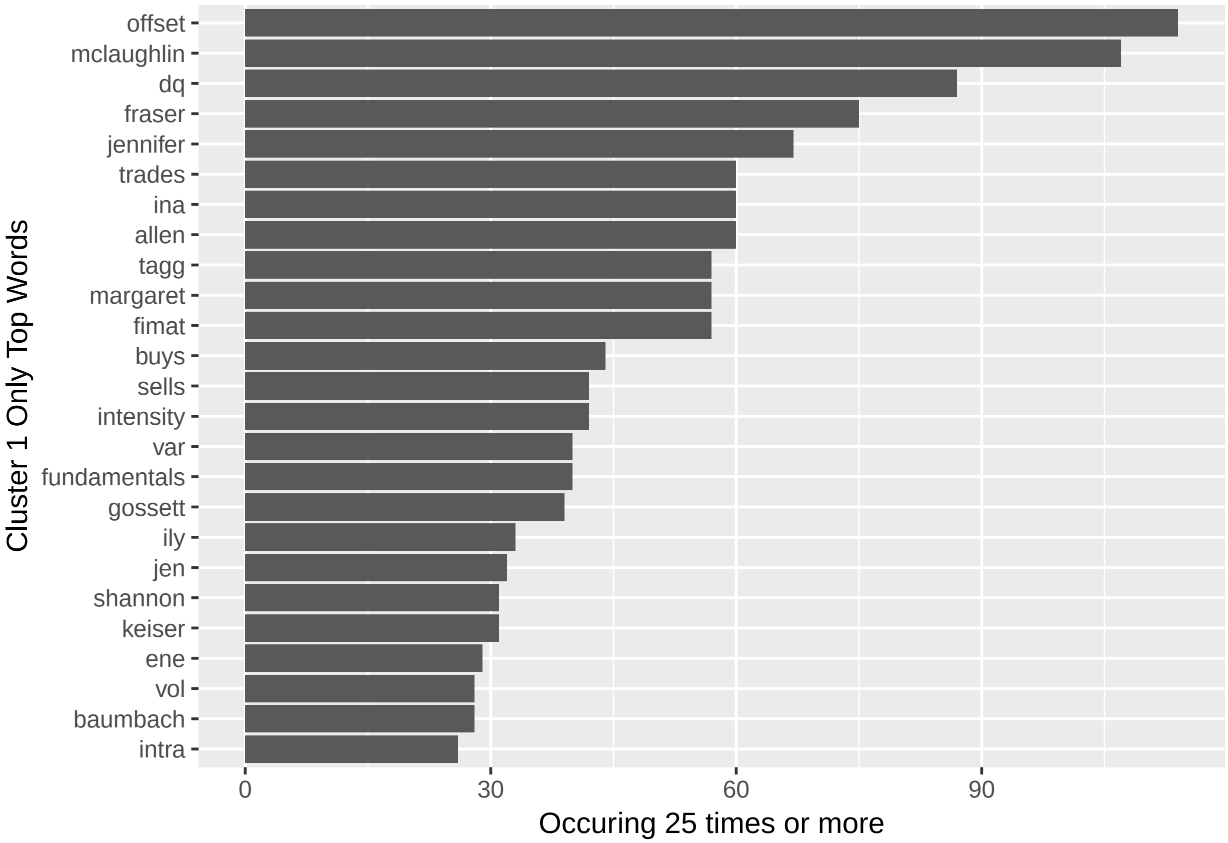

All Cluster 2 Top Words
not
enron
all
time
call
deal
ken
dg
good
company
day
can't
business
deals
people
great
meeting
working
week
questions
hope
employees
told
talk
gas
don
yesterday
lay
energy
corp
power
nice
forward
feel
value
office
information
hey
matter
mail
john
0
50
100
150
200
250
Occuring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

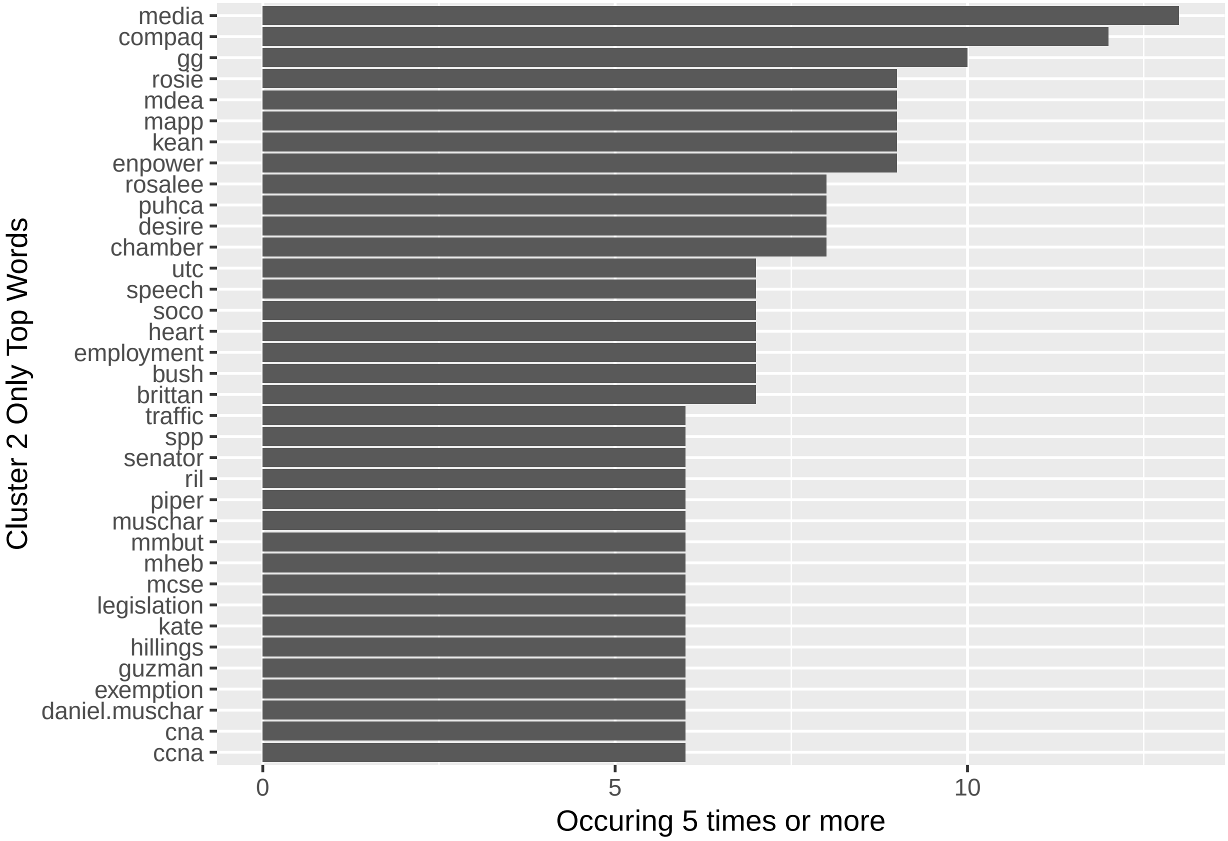


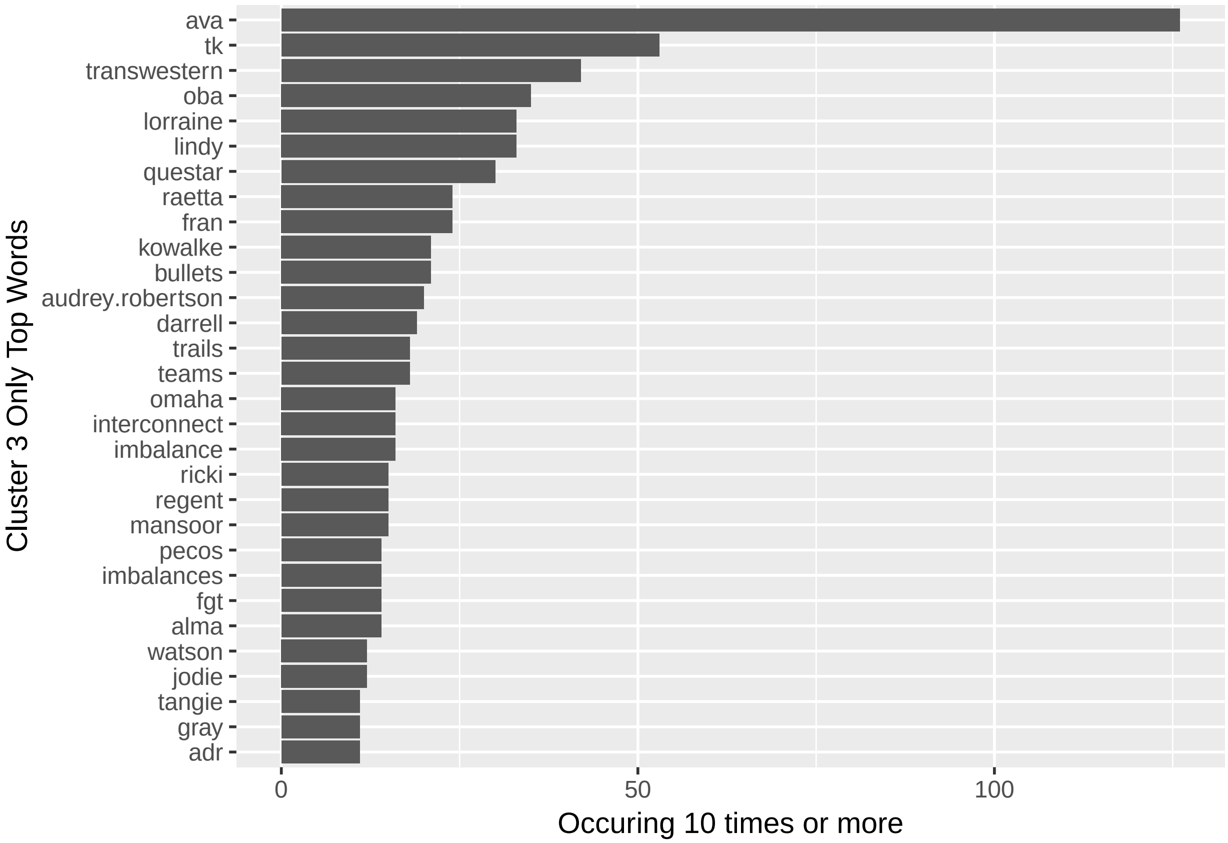
The clusters differ in the number of total words: Cluster 1 has 36653, Cluster 2, 14290 and Cluster 3, 17618.

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.

```
##
##  Pearson's product-moment correlation
##
## data:  n_freq$Cluster1 and n_freq$Cluster2
## t = -0.74671, df = 10673, p-value = 0.4553
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.02619435  0.01174417
## sample estimates:
##          cor
## -0.007227687
```

```
##
##  Pearson's product-moment correlation
##
## data:  n_freq$Cluster1 and n_freq$Cluster3
## t = -0.11397, df = 10673, p-value = 0.9093
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

[2] Pearson product-moment correlation

```
##  -0.02007298  0.01786748
## sample estimates:
##          cor
## -0.001103147
```

```
##
##  Pearson's product-moment correlation
##
## data:  n_freq$Cluster2 and n_freq$Cluster3
## t = -0.3436, df = 10673, p-value = 0.7312
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.02229475  0.01564533
## sample estimates:
##          cor
## -0.003325904
```

With respect to word frequencies, the three clusters are very weakly negatively correlated. The three clusters can be distinguished not only by volume, percentage of cluster-specif vocabulary, but also by the lack of intra-cluster correlations. Because the clusters derive from the machine learning latent network detection algorithm, there is now credible evidence that they can be further analyzed to identify collections of specific messages containing keywords of interest, to which we next turn.