Understand frequencies of words and word pairs

Alexander Alexandrov

Friday, May 06, 2016

The first step in building a predictive model for text is understanding the distribution and relationship between the words, tokens, and phrases in the text. The goal of this task is to understand the basic relationships you observe in the data and prepare to build your first linguistic models.

Questions to consider:

1. Some words are more frequent than others - what are the distributions of word frequencies?
2. What are the frequencies of 2-grams and 3-grams in the dataset?
3. How many unique words do you need in a frequency sorted dictionary to cover 50% of all word instances in the language? 90%?
4. How do you evaluate how many of the words come from foreign languages?
5. Can you think of a way to increase the coverage – identifying words that may not be in the corpora or using a smaller number of words in the dictionary to cover the same number of phrases?

```
## Loading required package: NLP
```

Load and clean english corpus:

```
set.seed(123)
corpus <- ReadAndCleanCorpus("./data/en_US/", prob=0.001)
```

Compute term's frequencies:

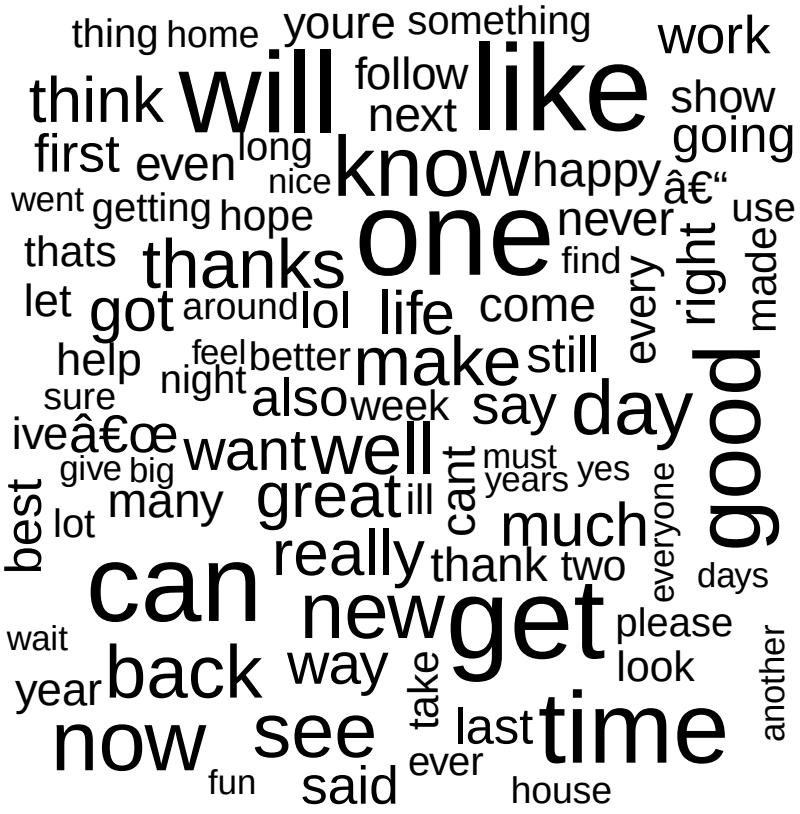
```
doc.term.matrix <- as.matrix(DocumentTermMatrix(corpus))
term.freq <- colSums(doc.term.matrix)
term.freq <- sort(term.freq, decreasing=TRUE)
head(term.freq)
```

```
## just like  one  get will  can
##  240  229  220  211  209  202
```

Show top100 terms:

```
## Loading required package: RColorBrewer
```

```
wordcloud(names(term.freq)[1:100], term.freq[1:100])
```


thing
home
youre
something
work
will
like
follow
think
next
show
going
first
even
long
know
happy
â€“
nice
went
getting
hope
one
never
use
thats
thanks
find
right
made
let
got
around
lol
life
come
every
good
help
feel
better
make
still
night
also
week
say
day
sure
iveâ€œ
want
well
give
big
must
years
yes
best
many
great
ill
cant
much
everyone
lot
really
thank
two
days
can
new
get
please
wait
back
way
take
look
year
now
see
last
time
another
ever
fun
said
house