Wk2_SLP1_Nikki Fitzherbert

Nikki Fitzherbert 13848336

05 March 2021

Note: The code was copied from https://www.analytics-tuts.com/zipfs-law-introduction-text-analytics/

1. Install and Load Packages

```
library(tm)
library(SnowballC)
library(wordcloud)
library(RColorBrewer)
```

2. Load, Transform and Clean Data

```
filePath <- "https://archive.org/stream/AnneFrankTheDiaryOfAYoungGirl_201606/
Anne-Frank-The-Diary-Of-A-Young-Girl_djvu.txt"

text <- readLines(filePath)
docs <- Corpus(VectorSource(text))
```

```
toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
docs <- tm_map(docs, toSpace, "/")
docs <- tm_map(docs, toSpace, "@")
docs <- tm_map(docs, toSpace, "\\|")

docs <- tm_map(docs, content_transformer(tolower))
docs <- tm_map(docs, removeNumbers)
docs <- tm_map(docs, removePunctuation)
docs <- tm_map(docs, stripWhitespace)
```

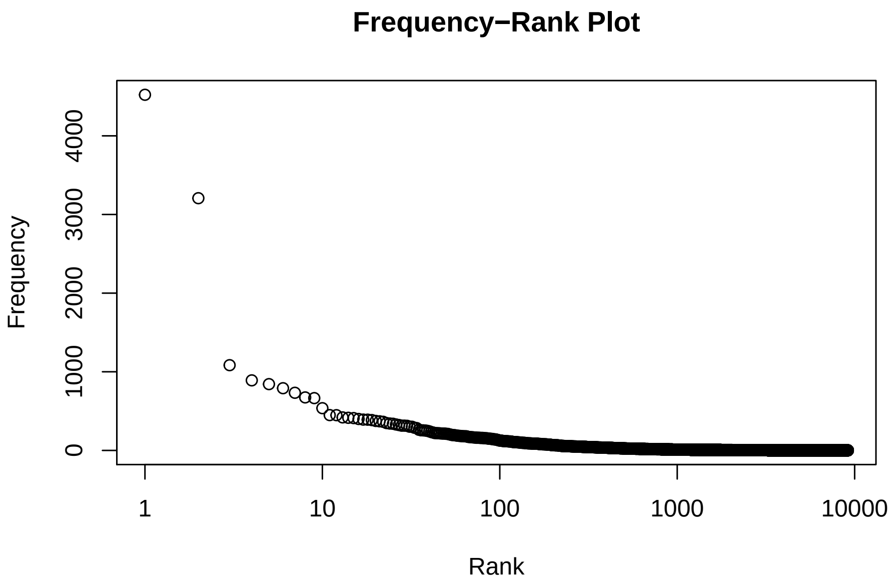
```
dtm <- TermDocumentMatrix(docs)

m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
```

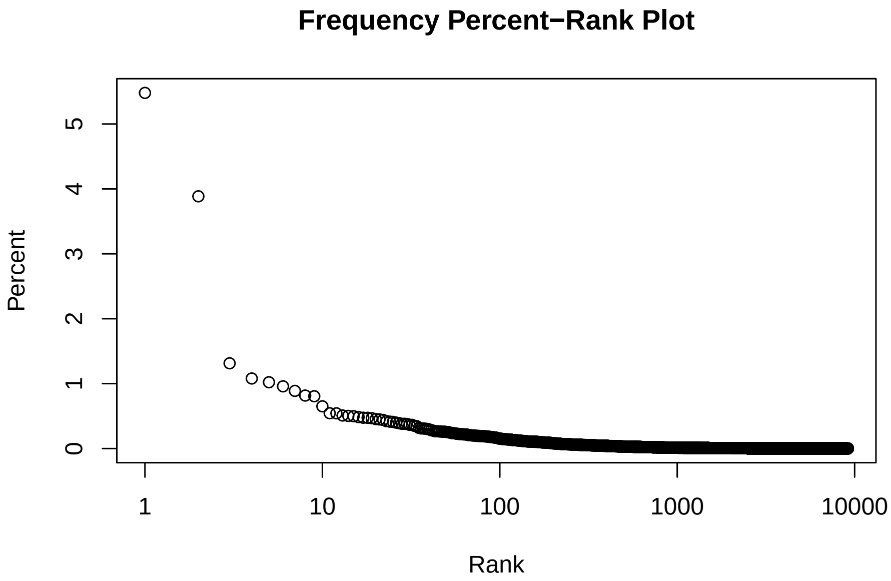
```
zipf <- cbind(d, Rank = 1:nrow(d), per=100 * d$freq / sum(d$freq))
```

3. Visualising Zipf's Law

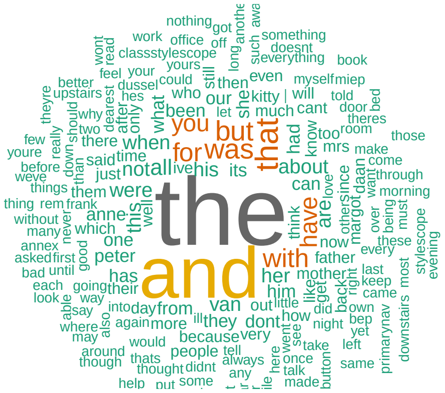
```
plot(zipf$Rank, zipf$freq,
     xlab = "Rank", ylab = "Frequency", main = "Frequency-Rank Plot",
     log = "x")
```



```r
plot(zipf$Rank, zipf$per,
     xlab = "Rank", ylab = "Percent", main = "Frequency Percent-Rank Plot",
     log = "x")
```



```r
wordcloud(words = zipf$word, freq = d$freq, min.freq = 1, max.words=200,
random.order=FALSE, rot.per=0.35, colors=brewer.pal(8, "Dark2"))
```


nothing
got
anothe
awa
work
office
off
something
wont
read
classstylescope
long
such
doesnt
everything
book
yours
feel
your
still
then
even
myself
miep
better
dussel
could
theyre
upstairs
hes
who
our
she
kitty
will
told
bed
dearest
after
only
what
been
let
much
cant
door
why
should
two
you
but
that
theres
room
those
really
few
there
when
had
know
too
youre
down
said
time
for
was
mrs
make
before
weve
than
just
notall
ive
his
its
about
daan
come
things
them
were
can
love
since
want
through
thing
rem
frank
the
morning
without
anne
this
well
think
have
are
others
margot
over
being
must
many
never
which
one
now
these
annex
good
peter
every
stylescope
asked
first
with
father
evening
bad
until
has
her
mother
right
last
most
each
going
their
and
him
like
get
back
keep
look
able
way
came
say
into
day
from
van
out
little
did
own
where
also
again
more
ill
they
dont
how
night
bep
primarynav
downstairs
may
because
very
went
see
yet
would
take
left
around
people
tell
here
once
though
thats
always
talk
button
same
thought
didnt
any
made
help
put
some