PS6 Li

Donald Li

March 2018

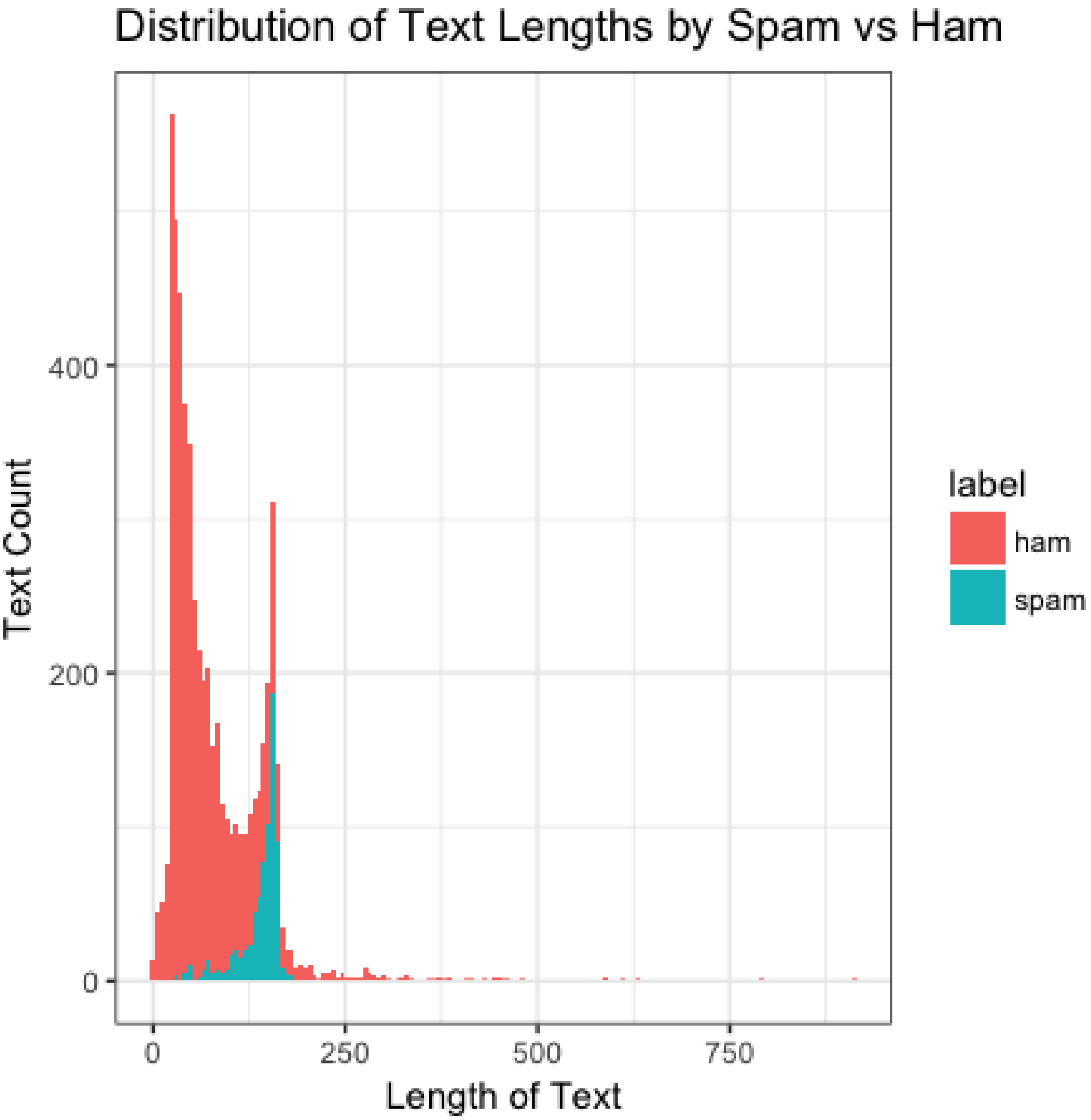
1 Introduction

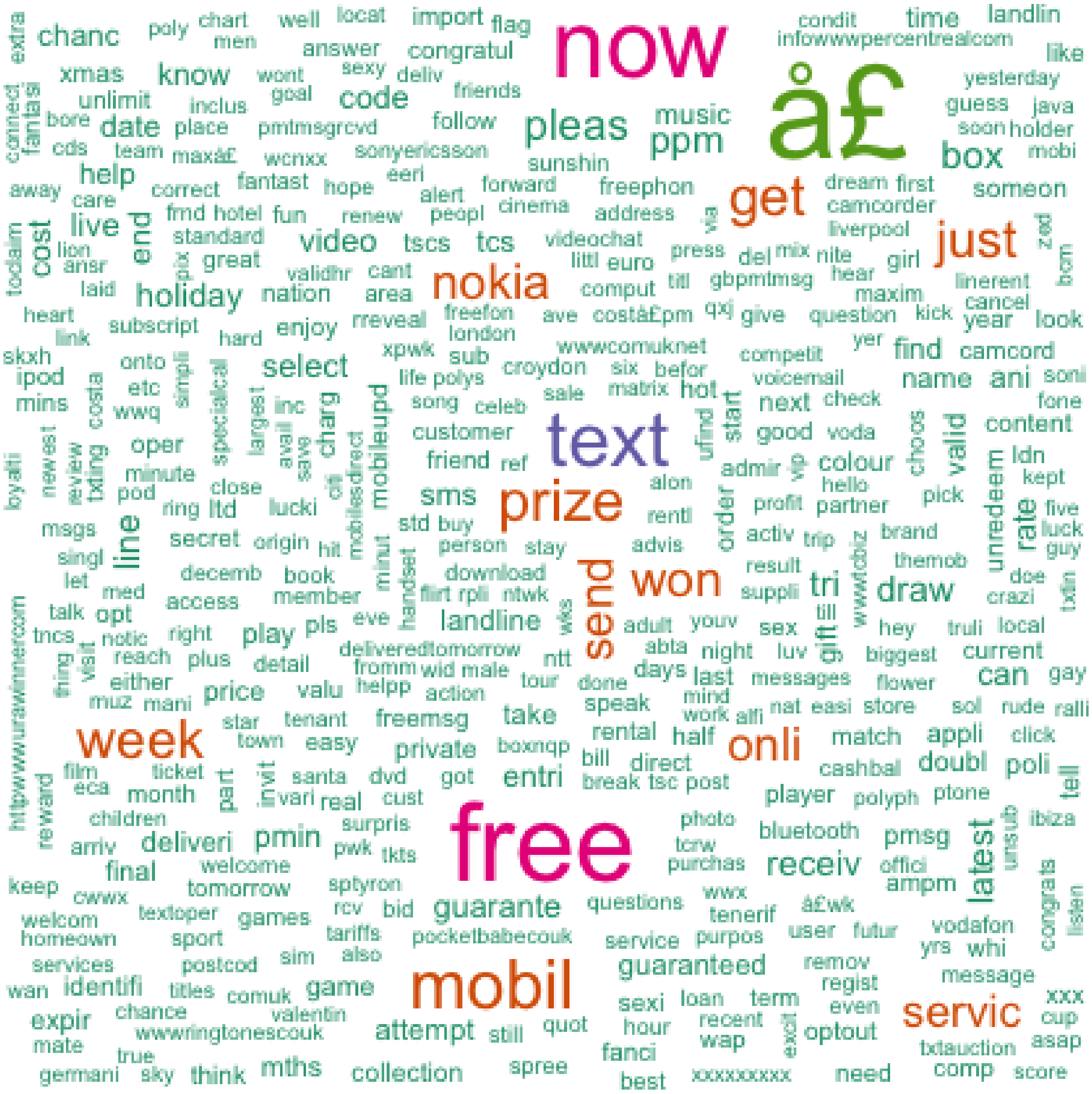
The data set I chose to visualize is from Kaggle. Natural Language processing is something that has always interested me. The data set from kaggle contains information over text messages. It then has a label stating if they are ”ham” real or spam ”fake”. My first step was to create column names for the labels. Then I set the label to factors and the texts to characters with the as.factor and as.character.

Going into it I knew I was going to create a bag of words model. However, I also knew the length of texts would be important to determine the classification of the texts. So as to not affect my future model I created a new data set called ”messages” to plot the spam and ham messages based off text length.

After that I used the ”TM” package to breakdown the texts. I changed all the texts to lower, stemming, removing white space, removing punctuation, removing stop words and remove numbers. At first I tried using the ”quanteda” package but I was having issues getting the word cloud visualizations to work.

From there on I created 3 word clouds. One was a word cloud of the entire corpus and the other two are the most common words in each class, ham/spam. The word clouds are important because it can give me clues as to which words are more predictive than others in the classification. Text analytics usually has high dimensionality so being able to reduce noise will help with my predicitions.


Distribution of Text Lengths by Spam vs Ham
Text Count
400
200
0
0
250
500
750
Length of Text
label
ham
spam



extra
chanc
poly
chart
men
well
locat
import
flag
now
condit
infowwwpercentrealcom
time
landlin
xmas
know
wont
answer
sexy
congratul
deliv
å£
like
yesterday
unlimit
inclus
goal
code
friends
guess
java
connect
fantasi
bore
date
place
pmtmsgrcvd
follow
pleas
music
soon
holder
cds
team
maxå£
wcnxx
sonyericsson
ppm
box
mobi
help
fantast
eeri
sunshin
away
care
correct
hope
alert
forward
freephon
dream
first
someon
live
fmd
hotel
fun
renew
peopl
cinema
address
get
camcorder
end
standard
video
tscs
tcs
videochat
via
liverpool
zed
just
todaim
cost
lion
ansr
great
validhr
cant
littl
euro
press
del
mix
nite
girl
bcm
pix
hear
inerent
laid
holiday
nation
area
nokia
comput
titl
gbpmtmsg
maxim
cancel
heart
subscript
mreveal
freefon
ave
costå£pm
qxj
give
question
kick
year
look
link
hard
enjoy
london
skxh
onto
xpwk
sub
wwwcomuknet
competit
yer
find
camcord
ipod
simpli
select
life
polys
croydon
six
befor
voicemail
name
ani
soni
costa
etc
largest
inc
song
celeb
sale
matrix
hot
check
fone
mins
wwq
specialcal
charg
mobileupd
customer
start
next
ufind
content
avail
text
good
voda
choos
valid
loyalti
newest
review
txting
oper
save
friend
ref
admir
vip
colour
ldn
minute
pod
close
cll
alon
hello
kept
ring
ltd
lucki
mobilesdirect
sms
prize
rentl
order
profit
partner
pick
unredeem
rate
five
msgs
line
secret
origin
hit
std
buy
person
stay
advis
activ
trip
wwwtcbiz
brand
luck
singl
minut
handset
result
themob
guy
let
decemb
book
download
send
won
suppli
tri
draw
doe
med
member
flirt
rpli
ntwk
crazi
txtin
talk
opt
access
landline
wks
gift
till
httpwwwurawinnercom
tncs
notic
right
play
pls
eve
adult
youv
sex
hey
truli
local
deliveredtomorrow
ntt
abta
night
luv
biggest
current
thing
visit
reach
plus
detail
fromm
wid
male
gb
days
last
messages
can
gay
either
tour
done
mind
flower
muz
mani
price
valu
helpp
action
speak
work
alfi
nat
easi
store
sol
rude
ralli
star
tenant
freemsg
take
rental
half
match
appli
click
week
town
easy
private
boxnqp
onli
bill
direct
cashbal
doubl
poli
film
ticket
part
invit
santa
dvd
got
entri
break
tsc
post
tell
eca
month
vari
real
cust
player
polyph
ptone
reward
children
surpris
photo
bluetooth
pmsg
ibiza
arriv
deliveri
pmin
pwk
tkts
free
tcrw
latest
unsub
final
welcome
purchas
receiv
offici
keep
cwwx
tomorrow
sptyron
wwx
ampm
congrats
listen
welcom
textoper
games
rcv
bid
guarante
questions
tenerif
å£wk
vodafon
homeown
sport
tariffs
pocketbabecouk
service
purpos
user
futur
yrs
whi
services
postcod
sim
also
guaranteed
remov
wan
identifi
titles
comuk
game
mobil
regist
message
chance
valentin
sexi
loan
term
even
xxx
expir
recent
cup
mate
wwwringtonescouk
attempt
still
quot
hour
wap
excit
optout
servic
asap
true
fanci
txtauction
germani
sky
think
mths
collection
spree
best
xxxxxxxxx
need
comp
score

Figure 2: Ham

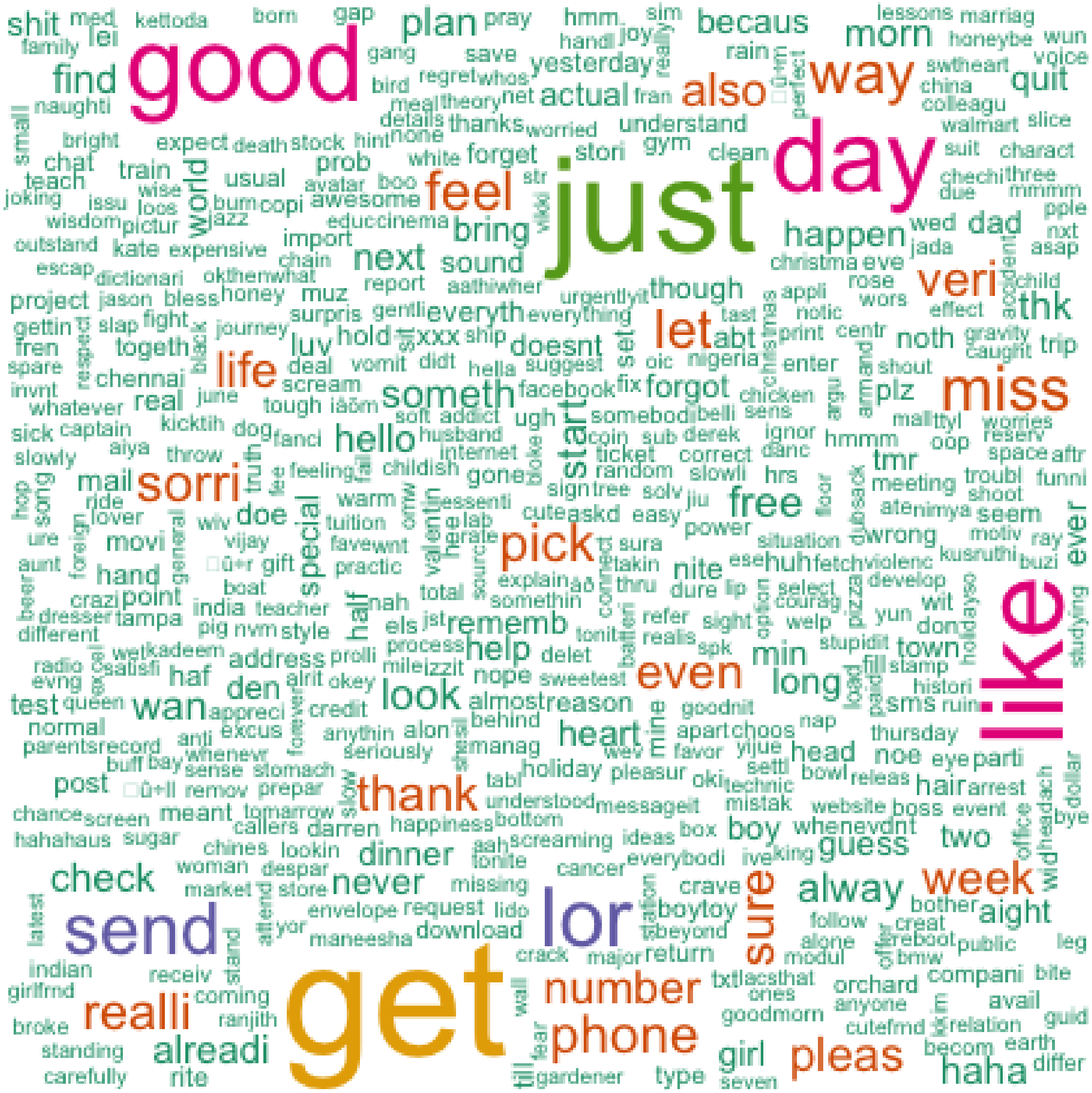


Figure 3: Spam