Filtering mobile phone spam SMS with the Naive Bayes algorithm

Objective:

The objective is to build classification model using Naive Bayes algorithm to classify the SMS messages into Spam OR Not Spam.

Data Set Information:

This dataset includes the text of SMS messages along with a label indicating whether the message is unwanted. Junk messages are labeled spam, while legitimate messages are labeled ham

To develop the Naive Bayes classifier, the data adapted from the SMS Spam Collection at http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection

Data collection and Reading

```
# Download the data from above mentioned website and save in working director
y
# Read the data
sms_raw <- read.csv("sms_spam.csv", stringsAsFactors = FALSE)
str(sms_raw)
```

```
## 'data.frame':    5559 obs. of  2 variables:
##  $ type: chr  "ham" "ham" "ham" "spam" ...
##  $ text: chr  "Hope you are having a good week. Just checking in" "K..give
back my thanks." "Am also doing in cbe only. But have to pay." "complimentary
4 STAR Ibiza Holiday or Â£10,000 cash needs your URGENT collection. 090663643
49 NOW from Landline"| __truncated__ ...
```

Covert character vector (which is category) type into a factor

```
sms_raw$type <- factor(sms_raw$type)
str(sms_raw$type)
```

```
##  Factor w/ 2 levels "ham","spam": 1 1 1 2 2 1 1 1 2 1 ...
```

```
table(sms_raw$type)
```

```
##
##  ham spam
## 4812  747
```

Data preparation - cleaning and standardizing text data

```
# Build a corpus using the text mining (tm) package
library(tm)

## Warning: package 'tm' was built under R version 3.5.1

## Loading required package: NLP

## Warning: package 'NLP' was built under R version 3.5.1

sms_corpus <- VCorpus(VectorSource(sms_raw$text))
print(sms_corpus)

## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 5559

inspect(sms_corpus[1:2])

## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 2
##
## [[1]]
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 49
##
## [[2]]
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 23

as.character(sms_corpus[[1]])

## [1] "Hope you are having a good week. Just checking in"

lapply(sms_corpus[1:2], as.character)

## $`1`
## [1] "Hope you are having a good week. Just checking in"
##
## $`2`
## [1] "K..give back my thanks."

# Clean up the corpus-removing punctuation and convert characters to lowercas
e
sms_corpus_clean <- tm_map(sms_corpus, content_transformer(tolower))
# comparision original corpus and transformed corpus
as.character(sms_corpus[[1]])

## [1] "Hope you are having a good week. Just checking in"
```

```
as.character(sms_corpus_clean[[1]])

## [1] "hope you are having a good week. just checking in"

# Removing numbers,stop words and punctuation
sms_corpus_clean <- tm_map(sms_corpus_clean, removeNumbers)
sms_corpus_clean <- tm_map(sms_corpus_clean, removeWords, stopwords())
sms_corpus_clean <- tm_map(sms_corpus_clean, removePunctuation)

# Reducing words to their root form - Stemming
#Example- Stemming the variants of the word "learn""
library(SnowballC)
wordStem(c("learn", "learned", "learning", "learns"))

## [1] "learn" "learn" "learn" "learn"

#Apply stemming to the corpus
sms_corpus_clean <- tm_map(sms_corpus_clean, stemDocument)

# Eliminating unwanted whitespace
sms_corpus_clean <- tm_map(sms_corpus_clean, stripWhitespace)

# Examining the final clean corpus
lapply(sms_corpus[1:3], as.character)

## $`1`
## [1] "Hope you are having a good week. Just checking in"
##
## $`2`
## [1] "K..give back my thanks."
##
## $`3`
## [1] "Am also doing in cbe only. But have to pay."

lapply(sms_corpus_clean[1:3], as.character)

## $`1`
## [1] "hope good week just check"
##
## $`2`
## [1] "kgive back thank"
##
## $`3`
## [1] "also cbe pay"
```

Data preparation - splitting text documents into words

```
# creating a document-term sparse matrix
sms_dtm <- DocumentTermMatrix(sms_corpus_clean)

# Alternative solution: create a document-term sparse matrix directly from th
e SMS corpus
```

```
sms_dtm2 <- DocumentTermMatrix(sms_corpus, control = list(
  tolower = TRUE,
  removeNumbers = TRUE,
  stopwords = TRUE,
  removePunctuation = TRUE,
  stemming = TRUE
))
# Alternative solution: using custom stop words function ensures identical re
sult
sms_dtm3 <- DocumentTermMatrix(sms_corpus, control = list(
  tolower = TRUE,
  removeNumbers = TRUE,
  stopwords = function(x) { removeWords(x, stopwords()) },
  removePunctuation = TRUE,
  stemming = TRUE
))

# Comparing the result
sms_dtm

## <<DocumentTermMatrix (documents: 5559, terms: 6582)>>
## Non-/sparse entries: 42182/36547156
## Sparsity           : 100%
## Maximal term length: 40
## Weighting          : term frequency (tf)

sms_dtm2

## <<DocumentTermMatrix (documents: 5559, terms: 6971)>>
## Non-/sparse entries: 43240/38708549
## Sparsity           : 100%
## Maximal term length: 40
## Weighting          : term frequency (tf)

sms_dtm3

## <<DocumentTermMatrix (documents: 5559, terms: 6582)>>
## Non-/sparse entries: 42182/36547156
## Sparsity           : 100%
## Maximal term length: 40
## Weighting          : term frequency (tf)
```

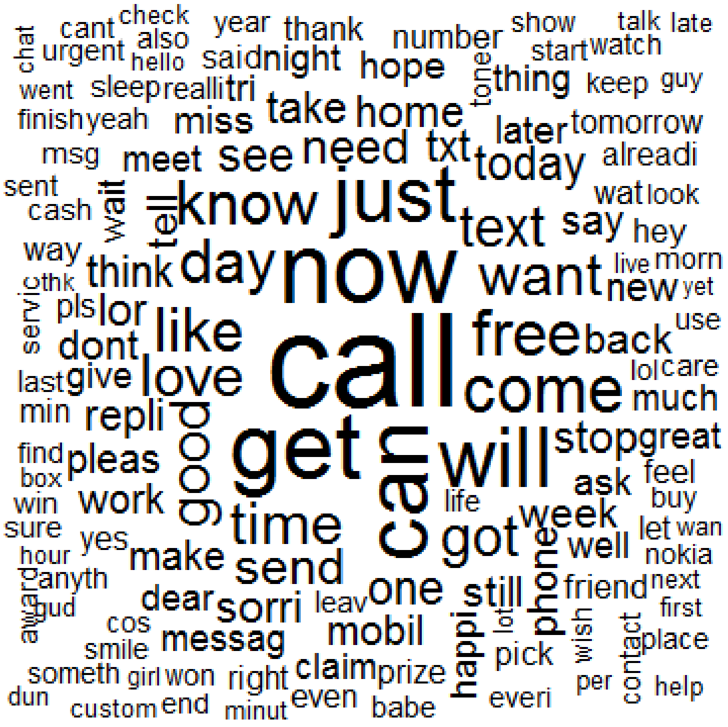
Data preparation - creating training and test datasets

```
sms_dtm_train <- sms_dtm[1:4169, ]
sms_dtm_test  <- sms_dtm[4170:5559, ]

# Saving the labels for training and testing matrices
sms_train_labels <- sms_raw[1:4169, ]$type
sms_test_labels  <- sms_raw[4170:5559, ]$type
```

```
# Checking the proportion of spam is similar
prop.table(table(sms_train_labels))

## sms_train_labels
##       ham      spam
## 0.8647158 0.1352842

prop.table(table(sms_test_labels))

## sms_test_labels
##       ham      spam
## 0.8683453 0.1316547
```

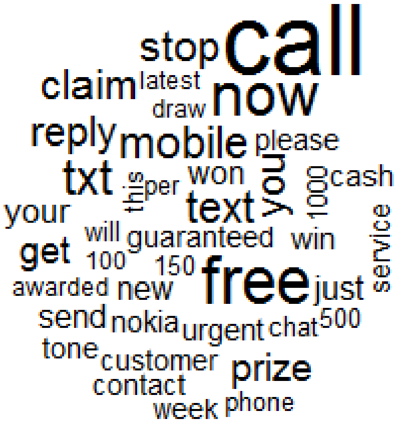
Visualizing text data - word clouds

```
# word cloud visualization
library(wordcloud)

## Warning: package 'wordcloud' was built under R version 3.5.1

## Loading required package: RColorBrewer

wordcloud(sms_corpus_clean, min.freq = 50, random.order = FALSE)
```

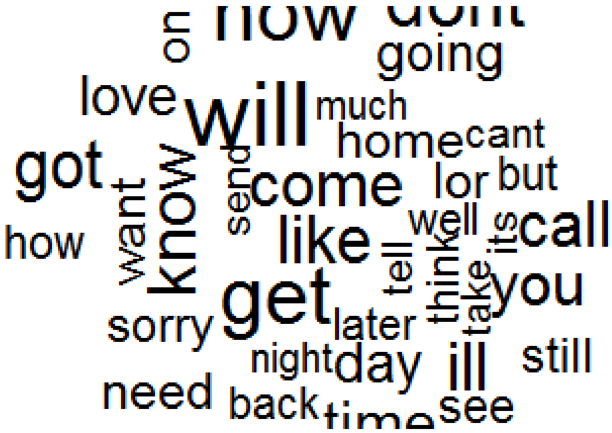


```
# World Cloud - Training data for spam and ham groups
spam <- subset(sms_raw, type == "spam")
ham  <- subset(sms_raw, type == "ham")
```

```
wordcloud(spam$text, max.words = 40, scale = c(3, 0.5))

## Warning in tm_map.SimpleCorpus(corpus, tm::removePunctuation):
## transformation drops documents

## Warning in tm_map.SimpleCorpus(corpus, function(x) tm::removeWords(x,
## tm::stopwords())): transformation drops documents
```



```
wordcloud(ham$text, max.words = 40, scale = c(3, 0.5))

## Warning in tm_map.SimpleCorpus(corpus, tm::removePunctuation):
## transformation drops documents

## Warning in tm_map.SimpleCorpus(corpus, tm::removePunctuation):
## transformation drops documents
```



```
sms_dtm_freq_train <- removeSparseTerms(sms_dtm_train, 0.999)
sms_dtm_freq_train

## <<DocumentTermMatrix (documents: 4169, terms: 1104)>>
## Non-/sparse entries: 24827/4577749
## Sparsity           : 99%
## Maximal term length: 19
## Weighting          : term frequency (tf)
```

Data preparation - creating indicator features for frequent words

```
findFreqTerms(sms_dtm_train, 5)

##    [1] "â?¬â\200?"              "â£wk"                    "abiola"
##    [4] "abl"                    "abt"                    "accept"
##    [7] "access"                 "account"                "across"
##   [10] "act"                    "activ"                  "actual"
##   [13] "add"                    "address"                "admir"
##   [16] "adult"                  "advanc"                 "aft"
##   [19] "afternoon"              "age"                    "ago"
##   [22] "aha"                    "ahead"                  "aight"
##   [25] "aint"                   "air"                    "aiyo"
##   [28] "alex"                   "almost"                 "alon"
##   [31] "alreadi"                "alright"                "also"
##   [34] "alway"                  "angri"                  "announc"
##   [37] "anoth"                  "answer"                 "anymor"
##   [40] "anyon"                  "anyth"                  "anytim"
##   [43] "anyway"               "apart"                "app"
##   [46] "appli"                "appreci"              "arcad"
##   [49] "ard"                  "area"                 "argu"
##   [52] "argument"             "armand"               "around"
##   [55] "arrang"               "arriv"                "asap"
##   [58] "ask"                  "askd"                 "attempt"
##   [61] "auction"              "avail"                "ave"
##   [64] "avoid"                "await"                "awak"
##   [67] "award"                "away"                 "awesom"
##   [70] "babe"                 "babi"                 "back"
##   [73] "bad"                  "bag"                  "bank"
##   [76] "bare"                 "basic"                "bath"
##   [79] "batteri"              "bcoz"                 "bday"
##   [82] "beauti"               "becom"                "bed"
##   [85] "bedroom"              "beer"                 "begin"
##   [88] "believ"               "best"                 "better"
##   [91] "bid"                  "big"                  "bill"
##   [94] "bird"                 "birthday"             "bit"
##   [97] "black"                "blank"                "bless"
##  [100] "blue"                 "bluetooth"            "bold"
##  [103] "bonus"                "boo"                  "book"
##  [106] "boost"                "bore"                 "boss"
##  [109] "bother"               "bout"                 "box"
##  [112] "boy"                  "boytoy"               "break"
##  [115] "breath"               "bring"                "brother"
##  [118] "bslvyl"               "btnationalr"          "buck"
##  [121] "bus"                  "busi"                 "buy"
##  [124] "cabin"                "call"                 "caller"
##  [127] "callertun"            "camcord"              "came"
##  [130] "camera"               "campus"               "can"
##  [133] "cancel"               "cancer"               "cant"
##  [136] "car"                  "card"                 "care"
##  [139] "carlo"                "case"                 "cash"
##  [142] "cashbal"              "catch"                "caus"
##  [145] "celebr"               "cell"                 "centr"
##  [148] "chanc"                "chang"                "charg"
##  [151] "chat"                 "cheap"                "cheaper"
##  [154] "check"                "cheer"                "chennai"
##  [157] "chikku"               "childish"             "children"
##  [160] "choic"                "choos"                "christma"
##  [163] "claim"                "class"                "clean"
##  [166] "clear"                "close"                "club"
##  [169] "code"                 "coffe"                "cold"
##  [172] "colleagu"             "collect"              "colleg"
##  [175] "colour"               "come"                 "comin"
##  [178] "comp"                 "compani"              "competit"
##  [181] "complet"              "complimentari"        "comput"
##  [184] "condit"               "confirm"              "congrat"
##  [187] "congratul"            "connect"              "contact"
##  [190] "content"              "contract"             "cook"
##  [193] "cool"                "copi"                "correct"
##  [196] "cos"                 "cost"                "costa"
##  [199] "costâ£pm"            "coupl"               "cours"
##  [202] "cover"               "coz"                 "crave"
##  [205] "crazi"               "creat"               "credit"
##  [208] "cri"                 "cross"               "cuddl"
##  [211] "cum"                 "cup"                 "current"
##  [214] "custcar"             "custom"              "cut"
##  [217] "cute"                "cuz"                 "dad"
##  [220] "daddi"               "darl"                "darlin"
##  [223] "darren"              "dat"                 "date"
##  [226] "day"                 "dead"                "deal"
##  [229] "dear"                "decid"               "decim"
##  [232] "decis"               "deep"                "definit"
##  [235] "del"                 "deliv"               "deliveri"
##  [238] "den"                 "depend"              "detail"
##  [241] "didnt"               "die"                 "diet"
##  [244] "differ"              "difficult"           "digit"
##  [247] "din"                 "dinner"              "direct"
##  [250] "dis"                 "discount"            "discuss"
##  [253] "disturb"             "dnt"                 "doc"
##  [256] "doctor"              "doesnt"              "dog"
##  [259] "doin"                "don"                 "done"
##  [262] "dont"                "door"                "doubl"
##  [265] "download"            "draw"                "dream"
##  [268] "drink"               "drive"               "drop"
##  [271] "drug"                "dude"                "due"
##  [274] "dun"                 "dunno"               "dvd"
##  [277] "earli"               "earlier"             "earth"
##  [280] "easi"                "eat"                 "eatin"
##  [283] "egg"                 "either"              "els"
##  [286] "email"               "embarass"            "end"
##  [289] "energi"              "england"             "enjoy"
##  [292] "enough"              "enter"               "entitl"
##  [295] "entri"               "envelop"             "etc"
##  [298] "euro"                "eve"                 "even"
##  [301] "ever"                "everi"               "everybodi"
##  [304] "everyon"             "everyth"             "exact"
##  [307] "exam"                "excel"               "excit"
##  [310] "excus"               "expect"              "experi"
##  [313] "expir"               "extra"               "eye"
##  [316] "face"                "facebook"            "fact"
##  [319] "fall"                "famili"              "fanci"
##  [322] "fantasi"             "fantast"             "far"
##  [325] "fast"                "fat"                 "father"
##  [328] "fault"               "feb"                 "feel"
##  [331] "felt"                "fetch"               "fight"
##  [334] "figur"               "file"                "fill"
##  [337] "film"                "final"               "find"
##  [340] "fine"                "finger"              "finish"
##  [343] "first"                "fix"                  "flag"
##  [346] "flat"                 "flight"               "flower"
##  [349] "follow"               "fone"                 "food"
##  [352] "forev"                "forget"               "forgot"
##  [355] "forward"              "found"                "freak"
##  [358] "free"                 "freemsg"              "freephon"
##  [361] "fren"                 "fri"                  "friday"
##  [364] "friend"               "friendship"           "frm"
##  [367] "frnd"                 "frnds"                "full"
##  [370] "fullonsmscom"         "fun"                  "funni"
##  [373] "futur"                "gal"                  "game"
##  [376] "gap"                  "gas"                  "gave"
##  [379] "gay"                  "gentl"                "get"
##  [382] "gettin"               "gift"                 "girl"
##  [385] "girlfrnd"             "give"                 "glad"
##  [388] "god"                  "goe"                  "goin"
##  [391] "gone"                 "gonna"                "good"
##  [394] "goodmorn"             "goodnight"            "got"
##  [397] "goto"                 "gotta"                "great"
##  [400] "grin"                 "guarante"             "gud"
##  [403] "guess"                "guy"                  "gym"
##  [406] "haf"                  "haha"                 "hai"
##  [409] "hair"                 "half"                 "hand"
##  [412] "handset"              "hang"                 "happen"
##  [415] "happi"                "hard"                 "hate"
##  [418] "hav"                  "havent"               "head"
##  [421] "hear"                 "heard"                "heart"
##  [424] "heavi"                "hee"                  "hell"
##  [427] "hello"                "help"                 "hey"
##  [430] "hgsuiteland"          "hit"                  "hiya"
##  [433] "hmm"                  "hmmm"                 "hmv"
##  [436] "hol"                  "hold"                 "holder"
##  [439] "holiday"              "home"                 "hook"
##  [442] "hop"                  "hope"                 "horni"
##  [445] "hospit"               "hot"                  "hotel"
##  [448] "hour"                 "hous"                 "how"
##  [451] "howev"                "howz"                 "hrs"
##  [454] "httpwwwurawinnercom"  "hug"                  "huh"
##  [457] "hungri"               "hurri"                "hurt"
##  [460] "iâ⍰¬ë⍰m"              "ice"                  "idea"
##  [463] "identifi"             "ignor"                "ill"
##  [466] "immedi"               "import"               "inc"
##  [469] "includ"               "india"                "info"
##  [472] "inform"               "insid"                "instead"
##  [475] "interest"             "invit"                "ipod"
##  [478] "irrit"                "ish"                  "island"
##  [481] "issu"                 "ive"                  "izzit"
##  [484] "januari"              "jay"                  "job"
##  [487] "john"                 "join"                 "joke"
##  [490] "joy"                  "jst"                  "jus"
##  [493] "just"                "juz"                 "kate"
##  [496] "keep"                "kept"                "kick"
##  [499] "kid"                 "kill"                "kind"
##  [502] "kinda"               "king"                "kiss"
##  [505] "knew"                "know"                "knw"
##  [508] "ladi"                "land"                "landlin"
##  [511] "laptop"              "lar"                 "last"
##  [514] "late"                "later"               "latest"
##  [517] "laugh"               "lazi"                "ldn"
##  [520] "lead"                "learn"               "least"
##  [523] "leav"                "lect"                "left"
##  [526] "leh"                 "lei"                 "less"
##  [529] "lesson"              "let"                 "letter"
##  [532] "liao"                "librari"             "lie"
##  [535] "life"                "lift"                "light"
##  [538] "like"                "line"                "link"
##  [541] "list"                "listen"              "littl"
##  [544] "live"                "lmao"                "load"
##  [547] "loan"                "local"               "locat"
##  [550] "log"                 "lol"                 "london"
##  [553] "long"                "longer"              "look"
##  [556] "lookin"              "lor"                 "lose"
##  [559] "lost"                "lot"                 "lovabl"
##  [562] "love"                "lover"               "loyalti"
##  [565] "ltd"                 "luck"                "lucki"
##  [568] "lunch"               "luv"                 "mad"
##  [571] "made"                "mah"                 "mail"
##  [574] "make"                "malaria"             "man"
##  [577] "mani"                "march"               "mark"
##  [580] "marri"               "match"               "mate"
##  [583] "matter"              "maxim"               "maxmin"
##  [586] "may"                 "mayb"                "meal"
##  [589] "mean"                "meant"               "med"
##  [592] "medic"               "meet"                "meetin"
##  [595] "meh"                 "member"              "men"
##  [598] "merri"               "messag"              "met"
##  [601] "mid"                 "midnight"            "might"
##  [604] "min"                 "mind"                "mine"
##  [607] "minut"               "miracl"              "miss"
##  [610] "mistak"              "moan"                "mob"
##  [613] "mobil"               "mobileupd"           "mode"
##  [616] "mom"                 "moment"              "mon"
##  [619] "monday"              "money"               "month"
##  [622] "morn"                "mother"              "motorola"
##  [625] "move"                "movi"                "mrng"
##  [628] "mrt"                 "mrw"                 "msg"
##  [631] "msgs"                "mths"                "much"
##  [634] "mum"                 "murder"              "music"
##  [637] "must"                "muz"                 "nah"
##  [640] "nake"                "name"                "nation"
##  [643] "natur"               "naughti"             "near"
##  [646] "need"                "net"                 "network"
##  [649] "neva"                "never"               "new"
##  [652] "news"                "next"                "nice"
##  [655] "nigeria"             "night"               "nite"
##  [658] "nobodi"              "noe"                 "nokia"
##  [661] "noon"                "nope"                "normal"
##  [664] "normpton"            "noth"                "notic"
##  [667] "now"                 "num"                 "number"
##  [670] "nyt"                 "obvious"             "offer"
##  [673] "offic"               "offici"              "okay"
##  [676] "oki"                 "old"                 "omg"
##  [679] "one"                 "onlin"               "onto"
##  [682] "oop"                 "open"                "oper"
##  [685] "opinion"             "opt"                 "optout"
##  [688] "orang"               "orchard"             "order"
##  [691] "oredi"               "oso"                 "other"
##  [694] "otherwis"            "outsid"              "pack"
##  [697] "page"                "paid"                "pain"
##  [700] "paper"               "parent"              "park"
##  [703] "part"                "parti"               "partner"
##  [706] "pass"                "passion"             "password"
##  [709] "past"                "pay"                 "peopl"
##  [712] "per"                 "person"              "pete"
##  [715] "phone"               "photo"               "pic"
##  [718] "pick"                "pictur"              "pin"
##  [721] "piss"                "pix"                 "pizza"
##  [724] "place"               "plan"                "play"
##  [727] "player"              "pleas"               "pleasur"
##  [730] "plenti"              "pls"                 "plus"
##  [733] "plz"                 "pmin"                "pmsg"
##  [736] "pobox"               "point"               "poli"
##  [739] "polic"               "poor"                "pop"
##  [742] "possess"             "possibl"             "post"
##  [745] "pound"               "power"               "ppm"
##  [748] "pray"                "present"             "press"
##  [751] "pretti"              "previous"            "price"
##  [754] "princess"            "privat"              "prize"
##  [757] "prob"                "probabl"             "problem"
##  [760] "project"             "promis"              "pub"
##  [763] "put"                 "qualiti"             "question"
##  [766] "quick"               "quit"                "quiz"
##  [769] "quot"                "rain"                "random"
##  [772] "rang"                "rate"                "rather"
##  [775] "rcvd"                "reach"               "read"
##  [778] "readi"               "real"                "reali"
##  [781] "realli"              "reason"              "receipt"
##  [784] "receiv"              "recent"              "record"
##  [787] "refer"               "regard"              "regist"
##  [790] "relat"               "relax"               "remain"
##  [793] "rememb"              "remind"              "remov"
##  [796] "rent"                "rental"              "repli"
##  [799] "repres"              "request"             "respond"
##  [802] "respons"             "rest"                "result"
##  [805] "return"              "reveal"              "review"
##  [808] "reward"              "right"               "ring"
##  [811] "rington"             "rite"                "road"
##  [814] "rock"                "role"                "room"
##  [817] "roommat"             "rose"                "round"
##  [820] "rowwjhl"             "rpli"                "rreveal"
##  [823] "run"                 "rush"                "sad"
##  [826] "sae"                 "safe"                "said"
##  [829] "sale"                "sat"                 "saturday"
##  [832] "savamob"             "save"                "saw"
##  [835] "say"                 "sch"                 "school"
##  [838] "scream"              "sea"                 "search"
##  [841] "sec"                 "second"              "secret"
##  [844] "see"                 "seem"                "seen"
##  [847] "select"              "self"                "sell"
##  [850] "semest"              "send"                "sens"
##  [853] "sent"                "serious"             "servic"
##  [856] "set"                 "settl"               "sex"
##  [859] "sexi"                "shall"               "share"
##  [862] "shd"                 "ship"                "shirt"
##  [865] "shop"                "short"               "show"
##  [868] "shower"              "sick"                "side"
##  [871] "sigh"                "sight"               "sign"
##  [874] "silent"              "simpl"               "sinc"
##  [877] "singl"               "sipix"               "sir"
##  [880] "sis"                 "sister"              "sit"
##  [883] "situat"              "skxh"                "skype"
##  [886] "slave"               "sleep"               "slept"
##  [889] "slow"                "slowli"              "small"
##  [892] "smile"               "smoke"               "sms"
##  [895] "smth"                "snow"                "sofa"
##  [898] "sol"                 "somebodi"            "someon"
##  [901] "someth"              "sometim"             "somewher"
##  [904] "song"                "soni"                "sonyericsson"
##  [907] "soon"                "sorri"               "sort"
##  [910] "sound"               "south"               "space"
##  [913] "speak"               "special"             "specialcal"
##  [916] "spend"               "spent"               "spoke"
##  [919] "spree"               "stand"               "start"
##  [922] "statement"           "station"             "stay"
##  [925] "std"                 "step"                "still"
##  [928] "stockport"           "stone"               "stop"
##  [931] "store"               "stori"               "street"
##  [934] "student"             "studi"               "stuff"
##  [937] "stupid"              "style"               "sub"
##  [940] "subscrib"            "success"             "suck"
##  [943] "suit"                "summer"              "sun"
##  [946] "sunday"              "sunshin"             "sup"
##  [949] "support"             "suppos"              "sure"
##  [952] "surf"                "surpris"             "sweet"
##  [955] "swing"               "system"              "take"
##  [958] "talk"                "tampa"               "tariff"
##  [961] "tcs"                 "tea"                 "teach"
##  [964] "tear"                "teas"                "tel"
##  [967] "tell"                "ten"                 "tenerif"
##  [970] "term"                "test"                "text"
##  [973] "thank"               "thanx"               "that"
##  [976] "thing"               "think"               "thinkin"
##  [979] "thk"                 "tho"                 "though"
##  [982] "thought"             "throw"               "thru"
##  [985] "tht"                 "thur"                "tick"
##  [988] "ticket"              "til"                 "till"
##  [991] "time"                "tire"                "titl"
##  [994] "tmr"                 "toclaim"             "today"
##  [997] "togeth"              "told"                "tomo"
## [1000] "tomorrow"            "tone"                "tonight"
## [1003] "tonit"               "took"                "top"
## [1006] "torch"               "tot"                 "total"
## [1009] "touch"               "tough"               "tour"
## [1012] "toward"              "town"                "track"
## [1015] "train"               "transact"            "travel"
## [1018] "treat"               "tri"                 "trip"
## [1021] "troubl"              "true"                "trust"
## [1024] "truth"               "tscs"                "ttyl"
## [1027] "tuesday"             "turn"                "twice"
## [1030] "two"                 "txt"                 "txting"
## [1033] "txts"                "type"                "ufind"
## [1036] "ugh"                 "ull"                 "uncl"
## [1039] "understand"          "unless"              "unlimit"
## [1042] "unredeem"            "unsub"               "unsubscrib"
## [1045] "updat"               "ure"                 "urgent"
## [1048] "urself"              "use"                 "user"
## [1051] "usf"                 "usual"               "uve"
## [1054] "valentin"            "valid"               "valu"
## [1057] "via"                 "video"               "vikki"
## [1060] "visit"               "vodafon"             "voic"
## [1063] "vomit"               "voucher"             "wait"
## [1066] "wake"                "walk"                "wan"
## [1069] "wana"                "wanna"               "want"
## [1072] "wap"                 "warm"                "wast"
## [1075] "wat"                 "watch"               "water"
## [1078] "way"                 "weak"                "wear"
## [1081] "weather"             "wed"                 "wednesday"
## [1084] "weed"                "week"                "weekend"
## [1087] "welcom"              "well"                "wen"
## [1090] "went"                "what"                "whatev"
## [1093] "whenev"              "whole"               "wid"
## [1096] "wif"                 "wife"                "wil"
## [1099] "will"                "win"                 "wine"
## [1102] "winner"              "wish"                "wit"
## [1105] "within"              "without"             "wiv"
## [1108] "wkli"                "wks"                 "wnt"
## [1111] "woke"                "won"                 "wonder"
## [1114] "wont"                "word"                "work"
## [1117] "workin"              "world"               "worri"
## [1120] "wors"                "worth"               "wot"
## [1123] "wow"                 "write"               "wrong"
## [1126] "wwq"                 "wwwgetzedcouk"       "xmas"
## [1129] "xxx"                 "yahoo"               "yar"
## [1132] "yeah"                "year"                "yep"
## [1135] "yes"                 "yesterday"           "yet"
## [1138] "yoga"                "yup"
```

```
sms_freq_words <- findFreqTerms(sms_dtm_train, 5)
str(sms_freq_words)
```

```
##  chr [1:1139] "â⍰¬â\200⍰" "â£wk" "abiola" "abl" "abt" "accept" ...
```

```
# creating DTMs with only the frequent terms
sms_dtm_freq_train <- sms_dtm_train[ , sms_freq_words]
sms_dtm_freq_test <- sms_dtm_test[ , sms_freq_words]

# converting counts to a factor
convert_counts <- function(x) {
  x <- ifelse(x > 0, "Yes", "No")
}

# apply() convert_counts() to columns of train/test data
sms_train <- apply(sms_dtm_freq_train, MARGIN = 2, convert_counts)
sms_test  <- apply(sms_dtm_freq_test, MARGIN = 2, convert_counts)
```

Training a model on the data

```
# Train the model
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 3.5.1
```

```
sms_classifier <- naiveBayes(sms_train, sms_train_labels)
```

Evaluating model performance

```
sms_test_pred <- predict(sms_classifier, sms_test)
library(gmodels)
```

```
## Warning: package 'gmodels' was built under R version 3.5.1
```

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.5.1
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.5.1
```

```
## 
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:NLP':
## 
##     annotate
```

```r
CrossTable(sms_test_pred, sms_test_labels,
           prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE,
           dnn = c('predicted', 'actual'))
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  1390 
## 
## 
##              | actual 
##    predicted |       ham |      spam | Row Total | 
## -------------|-----------|-----------|-----------|
##          ham |      1201 |        30 |      1231 | 
##              |     0.995 |     0.164 |           | 
## -------------|-----------|-----------|-----------|
##         spam |         6 |       153 |       159 | 
##              |     0.005 |     0.836 |           | 
## -------------|-----------|-----------|-----------|
## Column Total |      1207 |       183 |      1390 | 
##              |     0.868 |     0.132 |           | 
## -------------|-----------|-----------|-----------|
## 
## 
```

```r
confusionMatrix(sms_test_pred, sms_test_labels)
```

```
## Confusion Matrix and Statistics
## 
```

```
##           Reference
## Prediction  ham spam
##       ham  1201   30
##       spam    6  153
##
##                Accuracy : 0.9741
##                  95% CI : (0.9643, 0.9818)
##     No Information Rate : 0.8683
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8801
##  Mcnemar's Test P-Value : 0.0001264
##
##             Sensitivity : 0.9950
##             Specificity : 0.8361
##          Pos Pred Value : 0.9756
##          Neg Pred Value : 0.9623
##              Prevalence : 0.8683
##          Detection Rate : 0.8640
##    Detection Prevalence : 0.8856
##       Balanced Accuracy : 0.9155
##
##        'Positive' Class : ham
##
```

Improving model performance using Laplace estimator

```
sms_classifier2 <- naiveBayes(sms_train, sms_train_labels, laplace = 1)
sms_test_pred2 <- predict(sms_classifier2, sms_test)
CrossTable(sms_test_pred2, sms_test_labels,
           prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE,
           dnn = c('predicted', 'actual'))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  1390
##
##
##              | actual
##    predicted |       ham |      spam | Row Total |
## -------------|-----------|-----------|-----------|
##          ham |      1202 |        28 |      1230 |
##              |     0.996 |     0.153 |           |
## -------------|-----------|-----------|-----------|
##         spam |         5 |       155 |       160 |
##              |     0.004 |     0.847 |           |
## -------------|-----------|-----------|-----------|
## Column Total |      1207 |       183 |      1390 |
##              |     0.868 |     0.132 |           |
## -------------|-----------|-----------|-----------|
##
##
```

```
confusionMatrix(sms_test_pred, sms_test_labels)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  ham spam
##       ham  1201   30
##       spam    6  153
##
##                Accuracy : 0.9741
##                  95% CI : (0.9643, 0.9818)
##     No Information Rate : 0.8683
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8801
##  Mcnemar's Test P-Value : 0.0001264
##
##             Sensitivity : 0.9950
##             Specificity : 0.8361
##          Pos Pred Value : 0.9756
##          Neg Pred Value : 0.9623
##              Prevalence : 0.8683
##          Detection Rate : 0.8640
##    Detection Prevalence : 0.8856
##       Balanced Accuracy : 0.9155
##
##        'Positive' Class : ham
##
```