lab6_class.R

sathvik

2021-02-27

```
library(wordcloud)
```

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

```
library(tm)
```

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

```
sms_raw <-read.csv("/home/sathvik/EC8/ML/Lab/Lab6/sms_spam.csv")
str(sms_raw)
```

```
## 'data.frame':    5559 obs. of  2 variables:
##  $ type: Factor w/ 2 levels "ham","spam": 1 1 1 2 2 1 1 1 2 1 ...
##  $ text: Factor w/ 5156 levels " #  in mca. But not conform.",..: 1651 2566 257 626 3308 190 357 339
```

```
sms_raw[1,1]
```

```
## [1] ham
## Levels: ham spam
```

```
sms_raw[1,2]
```

```
## [1] Hope you are having a good week. Just checking in
## 5156 Levels:  #  in mca. But not conform. ...
```

```
sms_raw[1:3,]
```

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

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

```
## 'data.frame':    5559 obs. of  2 variables:
##  $ type: Factor w/ 2 levels "ham","spam": 1 1 1 2 2 1 1 1 2 1 ...
##  $ text: Factor w/ 5156 levels " #  in mca. But not conform.",..: 1651 2566 257 626 3308 190 357 339
```

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

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

```
sms_corpus <- Corpus(VectorSource(sms_raw$text))
print(sms_corpus)
```

```
## <<SimpleCorpus>>
```

```
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 5559
```

```
inspect(sms_corpus[1:3])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 3
##
## [1] Hope you are having a good week. Just checking in
## [2] K..give back my thanks.
## [3] Am also doing in cbe only. But have to pay.
```

```
sms_raw[1:3,2]
```

```
## [1] Hope you are having a good week. Just checking in
## [2] K..give back my thanks.
## [3] Am also doing in cbe only. But have to pay.
## 5156 Levels:  #  in mca. But not conform. ...
```

```
corpus_clean <- tm_map(sms_corpus, tolower)
```

```
## Warning in tm_map.SimpleCorpus(sms_corpus, tolower): transformation drops
## documents
```

```
corpus_clean <- tm_map(corpus_clean, removeNumbers)
```

```
## Warning in tm_map.SimpleCorpus(corpus_clean, removeNumbers): transformation
## drops documents
```

```
inspect(corpus_clean[1:3])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 3
##
## [1] hope you are having a good week. just checking in
## [2] k..give back my thanks.
## [3] am also doing in cbe only. but have to pay.
```

```
corpus_clean <- tm_map(corpus_clean, removeWords, stopwords())
```

```
## Warning in tm_map.SimpleCorpus(corpus_clean, removeWords, stopwords()):
## transformation drops documents
```

```
corpus_clean <- tm_map(corpus_clean, removePunctuation)
```

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

```
corpus_clean <- tm_map(corpus_clean, stripWhitespace)
```

```
## Warning in tm_map.SimpleCorpus(corpus_clean, stripWhitespace): transformation
## drops documents
```

```
inspect(corpus_clean[1:3])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 3
##
```

```
## [1] hope good week just checking  kgive back thanks
## [3]  also cbe pay
```

```
sms_dtm <- DocumentTermMatrix(corpus_clean)
sms_dtm
```

```
## <<DocumentTermMatrix (documents: 5559, terms: 7921)>>
## Non-/sparse entries: 42657/43990182
## Sparsity           : 100%
## Maximal term length: 40
## Weighting          : term frequency (tf)
```

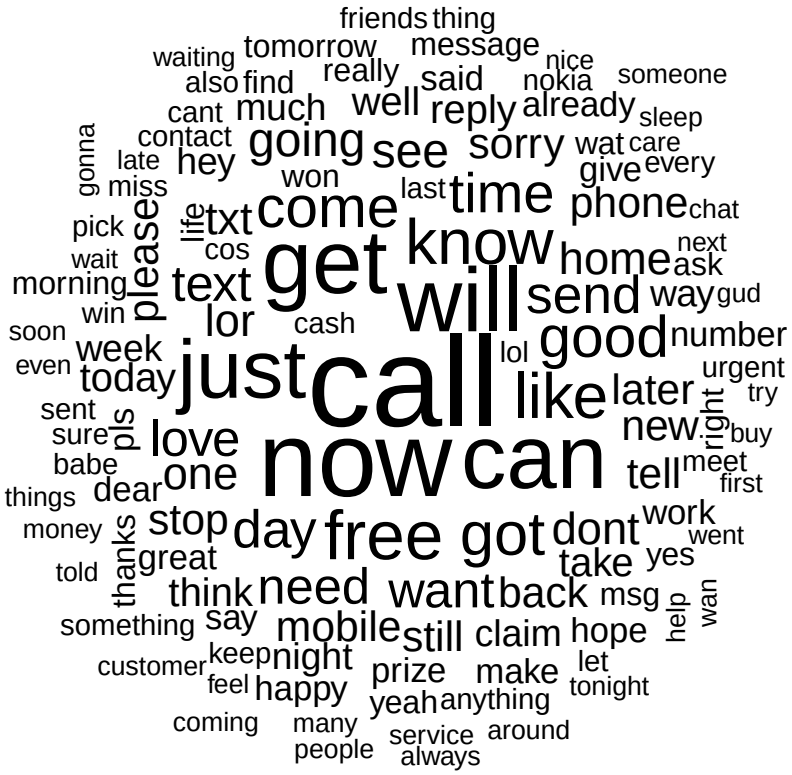
```
# Data preparation: Training and testing sets
sms_raw_train <- sms_raw[1:4169, ]
sms_raw_test<- sms_raw[4170:5559, ]
sms_dtm_train <- sms_dtm[1:4169, ]
sms_dtm_test<- sms_dtm[4170:5559, ]
sms_corpus_train <- corpus_clean[1:4169]
sms_corpus_test<- corpus_clean[4170:5559]

prop.table(table(sms_raw_train$type))
```

```
##
##       ham      spam
## 0.8647158 0.1352842
```

```
prop.table(table(sms_raw_test$type))
```

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

```
# Visualizing text data - word clouds
wordcloud(sms_corpus_train, min.freq = 40, random.order = FALSE)
```




```
spam <- subset(sms_raw_train, type == "spam")
ham <- subset(sms_raw_train, 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
```


```
#Data preparation - creating indicator features for frequent words
findFreqTerms(sms_dtm_train, 5)
```

```
##    [1] "checking"           "good"               "hope"
##    [4] "just"               "week"               "back"
##    [7] "thanks"             "also"               "pay"
##   [10] "cash"               "collection"         "complimentary"
##   [13] "holiday"            "landline"           "lose"
##   [16] "needs"              "now"                "urgent"
##   [19] "award"              "box"                "call"
##   [22] "collect"            "dear"               "final"
##   [25] "notice"             "ppm"                "sae"
##   [28] "tcs"                "tenerife"           "lar"
##   [31] "later"              "pick"               "much"
##   [34] "ask"                "father"             "please"
##   [37] "free"               "game"               "mobile"
##   [40] "official"           "play"               "right"
##   [43] "send"               "text"               "room"
##   [46] "swing"              "usf"                "whenever"
##   [49] "big"                "hour"               "longer"
##   [52] "man"                "sure"               "thing"
##   [55] "though"             "anything"           "lor"
##   [58] "ending"             "far"                "march"
##   [61] "never"              "problem"            "ready"
##   [64] "will"               "work"               "hmm"
##   [67] "night"              "well"               "get"
##   [70] "noon"               "see"                "chikku"
##   [73] "cool"               "cos"                "darren"
##   [76] "dat"                "den"                "dinner"
##   [79] "dun"                "feel"               "leave"
##   [82] "lunch"              "meet"               "meeting"
##   [85] "saying"             "angry"              "tell"
##   [88] "told"               "come"               "din"
##   [91] "wan"                "can"                "draw"
##   [94] "every"              "gift"               "music"
##   [97] "starting"           "tscs"               "txt"
```

```
##  [100] "vouchers"          "win"                "word"
##  [103] "coming"            "goodnight"          "gym"
##  [106] "birthday"          "today"              "wish"
##  [109] "gud"               "reading"            "contact"
##  [112] "cost"              "joke"               "less"
##  [115] "one"               "ones"               "school"
##  [118] "sent"              "think"              "thinking"
##  [121] "love"              "read"               "bank"
##  [124] "money"             "say"                "need"
##  [127] "stop"              "sup"                "weather"
##  [130] "thanx"             "plan"               "todays"
##  [133] "cancer"            "care"               "cup"
##  [136] "doctor"            "fat"                "sorry"
##  [139] "morning"           "sleep"              "train"
##  [142] "wine"              "auction"            "camera"
##  [145] "digital"           "nokia"              "part"
##  [148] "plus"              "take"               "won"
##  [151] "cant"              "fri"                "leh"
##  [154] "make"              "said"               "wait"
##  [157] "carlos"            "hear"               "might"
##  [160] "phones"            "texts"              "got"
##  [163] "job"               "whats"              "chat"
##  [166] "cheap"             "cheaper"            "gay"
##  [169] "hot"               "national"           "pmin"
##  [172] "rate"              "sale"               "anyway"
##  [175] "help"              "know"               "let"
##  [178] "use"               "yeah"               "centre"
##  [181] "like"              "liked"              "road"
##  [184] "something"         "min"                "awarded"
##  [187] "claim"             "guaranteed"         "hrs"
##  [190] "land"              "line"               "number"
##  [193] "prize"             "valid"              "mah"
##  [196] "tomorrow"          "believe"            "wat"
##  [199] "hey"               "sat"                "sec"
##  [202] "somebody"          "want"               "izzit"
##  [205] "mrt"               "outside"            "raining"
##  [208] "still"             "yup"                "going"
##  [211] "away"              "family"             "last"
##  [214] "mother"            "pray"               "bcoz"
##  [217] "enjoy"             "loved"              "world"
##  [220] "fact"              "gets"               "minutes"
##  [223] "really"            "bill"               "decimal"
##  [226] "give"              "congratulations"    "entry"
##  [229] "wkly"              "already"            "mum"
##  [232] "meh"               "fine"               "awesome"
##  [235] "deal"              "kind"               "loves"
##  [238] "thats"             "thought"            "watch"
##  [241] "company"           "happiness"          "nigeria"
##  [244] "safe"              "share"              "soon"
##  [247] "trip"              "fancy"              "hav"
##  [250] "pete"              "quite"              "ring"
##  [253] "tired"             "bed"                "bored"
##  [256] "car"               "checked"            "clean"
##  [259] "etc"               "feeling"            "hit"
```

```
##  [262] "least"              "left"             "many"
##  [265] "mean"               "phone"            "post"
##  [268] "sit"                "stuff"            "time"
##  [271] "times"              "drive"            "look"
##  [274] "lot"                "street"           "abt"
##  [277] "dont"               "even"             "others"
##  [280] "plz"                "things"           "wit"
##  [283] "words"              "pix"              "top"
##  [286] "beautiful"          "feels"            "heart"
##  [289] "heavy"              "leaves"           "light"
##  [292] "someone"            "truth"            "babe"
##  [295] "day"                "goes"             "long"
##  [298] "loving"             "miss"             "moment"
##  [301] "smile"              "together"         "loan"
##  [304] "dunno"              "mayb"             "remember"
##  [307] "tmr"                "opinion"          "wats"
##  [310] "liao"               "muz"              "sun"
##  [313] "bring"              "home"             "rest"
##  [316] "hmmm"               "players"          "selected"
##  [319] "yet"                "around"           "working"
##  [322] "cute"               "girls"            "local"
##  [325] "guy"                "nope"             "since"
##  [328] "maybe"              "called"           "dad"
##  [331] "oredi"              "alright"          "head"
##  [334] "deep"               "gone"             "net"
##  [337] "hospital"           "office"           "asked"
##  [340] "days"               "due"              "india"
##  [343] "mins"               "yes"              "bathe"
##  [346] "getting"            "message"          "chance"
##  [349] "happen"             "reached"          "bit"
##  [352] "gonna"              "probably"         "aha"
##  [355] "yesterday"          "friend"           "custcare"
##  [358] "shop"               "shopping"         "spree"
##  [361] "hair"               "whole"            "way"
##  [364] "evening"            "ttyl"             "house"
##  [367] "watching"           "€""               "died"
##  [370] "makes"              "went"             "ends"
##  [373] "first"              "life"             "people"
##  [376] "seems"              "sight"            "special"
##  [379] "till"               "knw"              "forgot"
##  [382] "lol"                "dog"              "else"
##  [385] "old"                "thk"              "wow"
##  [388] "sms"                "tried"            "may"
##  [391] "pub"                "laptop"           "movies"
##  [394] "eat"                "happy"            "joined"
##  [397] "met"                "thank"            "able"
##  [400] "late"               "plans"            "weeks"
##  [403] "lots"               "luv"              "poor"
##  [406] "xxx"                "sofa"             "facebook"
##  [409] "new"                "pictures"         "put"
##  [412] "show"               "always"           "better"
##  [415] "mom"                "boo"              "making"
##  [418] "moms"               "actually"         "boss"
##  [421] "fixed"              "kept"             "place"
```

```
##  [424] "talk"          "telling"         "year"
##  [427] "aint"          "comin"           "expecting"
##  [430] "pls"           "didnt"           "snow"
##  [433] "buy"           "early"           "gotta"
##  [436] "park"          "girl"            "thinks"
##  [439] "trouble"       "'s"              "interested"
##  [442] "msg"           "per"             "txts"
##  [445] "eatin"         "baby"            "shower"
##  [448] "num"           "wif"             "experience"
##  [451] "reply"         "spoke"           "asap"
##  [454] "found"         "jus"             "mrng"
##  [457] "pass"          "rose"            "sea"
##  [460] "tht"           "wake"            "woke"
##  [463] "flat"          "hello"           "college"
##  [466] "darlin"        "finish"          "finished"
##  [469] "ive"           "kate"            "break"
##  [472] "noe"           "promise"         "wonderful"
##  [475] "wont"          "midnight"        "bold"
##  [478] "saw"           "princess"        "bad"
##  [481] "dream"         "guess"           "tho"
##  [484] "drugs"         "learn"           "real"
##  [487] "using"         "admirer"         "looking"
##  [490] "rreveal"       "secret"          "specialcall"
##  [493] "ufind"         "wot"             "another"
##  [496] "try"           "competition"     "hurt"
##  [499] "england"       "euro"            "flag"
##  [502] "following"     "info"            "offer"
##  [505] "service"       "tone"            "driving"
##  [508] "two"           "worried"         "years"
##  [511] "test"          "apply"           "close"
##  [514] "pound"         "receive"         "specially"
##  [517] "boytoy"        "kiss"            "lesson"
##  [520] "missed"        "slept"           "children"
##  [523] "mob"           "poly"            "polys"
##  [526] "song"          "bag"             "find"
##  [529] "nice"          "check"           "pic"
##  [532] "cover"         "shd"             "month"
##  [535] "must"          "name"            "tel"
##  [538] "gettin"        "ure"             "dvd"
##  [541] "player"        "quiz"            "sony"
##  [544] "sunshine"      "eve"             "lover"
##  [547] "yahoo"         "lady"            "rain"
##  [550] "run"           "tomo"            "await"
##  [553] "booked"        "weekends"        "card"
##  [556] "credit"        "half"            "costa"
##  [559] "cost£pm"       "del"             "done"
##  [562] "maxmins"       "pobox"           "skxh"
##  [565] "sol"           "stockport"       "toclaim"
##  [568] "charges"       "mistake"         "boy"
##  [571] "gal"           "hand"            "hold"
##  [574] "jst"           "walking"         "wana"
##  [577] "slave"         "great"           "started"
##  [580] "change"        "update"          "xmas"
##  [583] "little"        "oso"             "plenty"
```

```
##  [586] "guys"          "aight"          "book"
##  [589] "orchard"       "food"           "friends"
##  [592] "important"     "stay"           "support"
##  [595] "system"        "weekend"        "discount"
##  [598] "offers"        "savamob"        "sub"
##  [601] "worth"         "keep"           "took"
##  [604] "apartment"     "chennai"        "age"
##  [607] "supposed"      "course"         "finally"
##  [610] "couple"        "pics"           "posted"
##  [613] "waking"        "calls"          "opt"
##  [616] "touch"         "waiting"        "minute"
##  [619] "lessons"       "sleeping"       "price"
##  [622] "used"          "case"           "hai"
##  [625] "replying"      "tonite"         "brother"
##  [628] "computer"      "mobiles"        "voucher"
##  [631] "caller"        "hurry"          "points"
##  [634] "goin"          "across"         "nothing"
##  [637] "worry"         "frm"            "taking"
##  [640] "advance"       "merry"          "wishing"
##  [643] "obviously"     "speak"          "wen"
##  [646] "came"          "cold"           "staying"
##  [649] "town"          "library"        "wanted"
##  [652] "wiv"           "delivery"       "sipix"
##  [655] "within"        "forwarded"      "friendship"
##  [658] "seeing"        "smiling"        "next"
##  [661] "kinda"         "attempt"        "type"
##  [664] "unlimited"     "haha"           "don"
##  [667] "dude"          "party"          "sunday"
##  [670] "cheers"        "boys"           "cause"
##  [673] "face"          "kids"           "mark"
##  [676] "omg"           "says"           "boost"
##  [679] "decided"       "energy"         "fast"
##  [682] "frnds"         "hell"           "mind"
##  [685] "replied"       "story"          "towards"
##  [688] "awake"         "valentines"     "reaching"
##  [691] "ticket"        "wanna"          "fault"
##  [694] "christmas"     "available"      "password"
##  [697] "ringtone"      "valentine"      "wed"
##  [700] "anyone"        "comes"          "cry"
##  [703] "end"           "god"            "start"
##  [706] "alone"         "person"         "sexy"
##  [709] "voice"         "brings"         "trying"
##  [712] "enough"        "ago"            "sell"
##  [715] "wants"         "cancel"         "cum"
##  [718] "fun"           "inc"            "msgs"
##  [721] "warm"          "air"            "hoping"
##  [724] "question"      "confirm"        "date"
##  [727] "join"          "activate"       "freemsg"
##  [730] "paid"          "earlier"        "mine"
##  [733] "order"         "dead"           "sweet"
##  [736] "everything"    "okay"           "vomit"
##  [739] "campus"        "messages"       "rental"
##  [742] "video"         "class"          "unsub"
##  [745] "club"          "terms"          "tones"
```

```
##  [748] "awaiting"          "add"                "ard"
##  [751] "haf"               "dogging"            "nite"
##  [754] "sex"               "sign"               "uks"
##  [757] "treat"             "okie"               "exactly"
##  [760] "knew"              "wife"               "arrive"
##  [763] "charged"           "customer"           "reference"
##  [766] "services"          "flower"             "picked"
##  [769] "cuz"               "ill"                "looks"
##  [772] "prob"              "drink"              "happened"
##  [775] "meds"              "pizza"              "sitting"
##  [778] "appreciate"        "sister"             "dreams"
##  [781] "made"              "askd"               "ever"
##  [784] "seen"              "bonus"              "reach"
##  [787] "meant"             "dating"             "account"
##  [790] "details"           "remove"             "sending"
##  [793] "trust"             "representative"     "wonder"
##  [796] "loads"             "neva"               "mad"
##  [799] "bday"              "bedroom"            "cabin"
##  [802] "colleagues"        "entered"            "felt"
##  [805] "invited"           "naked"              "parents"
##  [808] "sad"               "screaming"          "surprise"
##  [811] "wid"               "dis"                "ipod"
##  [814] "near"              "remind"             "shall"
##  [817] "abiola"            "either"             "charge"
##  [820] "credits"           "extra"              "goto"
##  [823] "unsubscribe"       "means"              "friday"
##  [826] "wrong"             "lets"               "cut"
##  [829] "hours"             "somewhere"          "true"
##  [832] "sir"               "information"        "listen"
##  [835] "full"              "missing"            "hiya"
##  [838] "busy"              "calling"            "hows"
##  [841] "instead"           "difficult"          "area"
##  [844] "numbers"           "code"               "shows"
##  [847] "laugh"             "havent"             "lei"
##  [850] "mon"               "without"            "holder"
##  [853] "til"               "contract"           "mobileupd"
##  [856] "motorola"          "optout"             "orange"
##  [859] "correct"           "reason"             "colour"
##  [862] "matches"           "network"            "reward"
##  [865] "valued"            "winner"             "content"
##  [868] "quick"             "photos"             "address"
##  [871] "set"               "black"              "realy"
##  [874] "wnt"               "bucks"              "happening"
##  [877] "search"            "asking"             "small"
##  [880] "side"              "become"             "takes"
##  [883] "camcorder"         "smoke"              "games"
##  [886] "wap"               "forget"             "mths"
##  [889] "latest"            "movie"              "urself"
##  [892] "spent"             "completely"         "malaria"
##  [895] "relax"             "self"               "thinkin"
##  [898] "worse"             "sense"              "tough"
##  [901] "created"           "fingers"            "gap"
##  [904] "holding"           "rite"               "btnationalrate"
##  [907] "easy"              "store"              "summer"
```

```
##  [910] "tuesday"            "heard"              "daddy"
##  [913] "grins"              "coz"                "january"
##  [916] "immediately"        "surely"             "willing"
##  [919] "comp"               "std"                "weekly"
##  [922] "pain"               "huh"                "private"
##  [925] "lect"               "download"           "situation"
##  [928] "online"             "quality"            "review"
##  [931] "enter"              "alex"               "knows"
##  [934] "torch"              "bslvyl"             "request"
##  [937] "ldn"                "pmsg"               "rcvd"
##  [940] "ish"                "drop"               "congrats"
##  [943] "mates"              "slowly"             "transaction"
##  [946] "childish"           "answer"             "hmv"
##  [949] "pounds"             "questions"          "barely"
##  [952] "howz"               "wil"                "mail"
##  [955] "wednesday"          "afternoon"          "talking"
##  [958] "ans"                "ull"                "excuse"
##  [961] "point"              "tampa"              "whatever"
##  [964] "planning"           "weed"               "definitely"
##  [967] "short"              "lovely"             "nature"
##  [970] "blue"               "saturday"           "texting"
##  [973] "second"             "sort"               "forward"
##  [976] "expires"            "identifier"         "statement"
##  [979] "unredeemed"         "inside"             "email"
##  [982] "leaving"            "crazy"              "oops"
##  [985] "moan"               "water"              "tonight"
##  [988] "freephone"          "drug"               "regards"
##  [991] "recently"           "south"              "hard"
##  [994] "workin"             "vikky"              "lmao"
##  [997] "bother"             "medical"            "track"
## [1000] "glad"               "exam"               "gas"
## [1003] "station"            "currently"          "best"
## [1006] "hee"                "move"               "nobody"
## [1009] "usual"              "everyone"           "lucky"
## [1012] "doesnt"             "double"             "sonyericsson"
## [1015] "slow"               "pretty"             "envelope"
## [1018] "paper"              "rock"               "walk"
## [1021] "tells"              "london"             "txting"
## [1024] "horny"              "live"               "ltd"
## [1027] "normptone"          "wwq"                "wwwgetzedcouk"
## [1030] "frnd"               "tour"               "choose"
## [1033] "rather"             "timing"             "write"
## [1036] "news"               "however"            "luck"
## [1039] "feb"                "sounds"             "waste"
## [1042] "open"               "goodmorning"        "uncle"
## [1045] "disturb"            "pin"                "lookin"
## [1048] "married"            "babes"              "months"
## [1051] "press"              "john"               "added"
## [1054] "fullonsmscom"       "visit"              "welcome"
## [1057] "round"              "putting"            "announcement"
## [1060] "spend"              "twice"              "tariffs"
## [1063] "anymore"            "exciting"           "past"
## [1066] "keeping"            "starts"             "almost"
## [1069] "partner"            "single"             "cashbalance"
```

```
## [1072] "hgsuitelands"       "maximize"           "rowwjhl"
## [1075] "role"               "exams"              "jay"
## [1078] "ten"                "simple"             "operator"
## [1081] "mate"               "doin"               "thurs"
## [1084] "normal"             "works"              "sis"
## [1087] "bus"                "tickets"            "battery"
## [1090] "unless"             "depends"            "anytime"
## [1093] "studying"           "€~s"                "marry"
## [1096] "relation"           "ave"                "stupid"
## [1099] "mode"               "smth"               "ahead"
## [1102] "aft"                "joking"             "bout"
## [1105] "nah"                "tear"               "wks"
## [1108] "paying"             "possible"           "fone"
## [1111] "mrw"                "bid"                "loyalty"
## [1114] "hungry"             "training"           "funny"
## [1117] "earth"              "yar"                "door"
## [1120] "receipt"            "registered"         "silent"
## [1123] "understand"         "vodafone"           "sick"
## [1126] "sucks"              "planned"            "serious"
## [1129] "sound"              "decide"             "happens"
## [1132] "rates"              "entitled"           "save"
## [1135] "access"             "lost"               "buying"
## [1138] "teach"              "excellent"          "hate"
## [1141] "lazy"               "direct"             "via"
## [1144] "coffee"             "yep"                "sch"
## [1147] "€~m"                "semester"           "otherwise"
## [1150] "yoga"               "lovable"            "juz"
## [1153] "dnt"                "seriously"          "copy"
## [1156] "aiyo"               "film"               "hurts"
## [1159] "thru"               "user"               "arcade"
## [1162] "crave"              "eating"             "dropped"
## [1165] "log"                "argument"           "kick"
## [1168] "wins"               "hop"                "die"
## [1171] "monday"             "tot"                "nyt"
## [1174] "link"               "future"             "sometimes"
## [1177] "match"              "girlfrnd"           "lift"
## [1180] "meaning"            "police"             "project"
## [1183] "different"          "member"             "callertune"
## [1186] "eyes"               "picking"            "fantastic"
## [1189] "space"              "rent"               "httpwwwurawinnercom"
## [1192] "onto"               "ugh"                "forever"
## [1195] "£wk"                "miracle"            "gave"
## [1198] "rply"               "king"               "pleasure"
## [1201] "uve"                "records"            "study"
## [1204] "frens"              "photo"              "bluetooth"
## [1207] "stand"              "persons"            "handset"
## [1210] "cell"               "catch"              "figure"
## [1213] "ice"                "tea"                "tick"
## [1216] "accept"             "naughty"            "worries"
```

```
Dictionary <- function(x) {
  if( is.character(x) ) {
    return (x)
  }
```

```r
  stop('x is not a character vector')
}
sms_dict <- Dictionary(findFreqTerms(sms_dtm_train, 5))

sms_train <- DocumentTermMatrix(sms_corpus_train, list(dictionary = sms_dict))
sms_test<- DocumentTermMatrix(sms_corpus_test,list(dictionary = sms_dict))

convert_counts <- function(x) {
  x <- ifelse(x > 0, 1, 0)
  x <- factor(x, levels = c(0, 1), labels = c("No", "Yes"))
  return(x)
}

sms_train <- apply(sms_train, MARGIN = 2, convert_counts)
sms_test<- apply(sms_test, MARGIN = 2, convert_counts)

# Step 3 - training a model on the data
library(e1071)
sms_classifier <- naiveBayes(sms_train, sms_raw_train$type)

# Step 4 - evaluating model performance
sms_test_pred <- predict(sms_classifier, sms_test)
library(gmodels)
CrossTable(sms_test_pred, sms_raw_test$type,
             prop.chisq = FALSE, prop.t = FALSE,
             dnn = c('predicted', 'actual'))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  1390
##
##
##              | actual
##    predicted |       ham |      spam | Row Total |
## -------------|-----------|-----------|-----------|
##          ham |      1203 |        32 |      1235 |
##              |     0.974 |     0.026 |     0.888 |
##              |     0.997 |     0.175 |           |
## -------------|-----------|-----------|-----------|
##         spam |         4 |       151 |       155 |
##              |     0.026 |     0.974 |     0.112 |
##              |     0.003 |     0.825 |           |
## -------------|-----------|-----------|-----------|
## Column Total |      1207 |       183 |      1390 |
##              |     0.868 |     0.132 |           |
## -------------|-----------|-----------|-----------|
```

```
##
##
```

```
predict(sms_classifier, "Marvel Mobile Play the official Ultimate Spider-man")
```

```
## [1] ham
## Levels: ham spam
```

```
# Step 5 - improving model performance
sms_classifier2 <- naiveBayes(sms_train, sms_raw_train$type,
                              laplace = 1)
sms_test_pred2 <- predict(sms_classifier2, sms_test)
CrossTable(sms_test_pred2, sms_raw_test$type,
           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 |      1204 |        31 |      1235 |
##              |     0.998 |     0.169 |           |
## -------------|-----------|-----------|-----------|
##         spam |         3 |       152 |       155 |
##              |     0.002 |     0.831 |           |
## -------------|-----------|-----------|-----------|
## Column Total |      1207 |       183 |      1390 |
##              |     0.868 |     0.132 |           |
## -------------|-----------|-----------|-----------|
##
##
```