Text Analytics - Investigation on emails

Data Analysis

Babak Barghi

December 28, 2020

Contents

First we load the libraries.

```
library(tidyverse)
library(tm)
library(SnowballC)
library(caTools)
library(rpart)
library(rpart.plot)
library(ROCR)
library(randomForest)
library(kableExtra)
```

1 Data description

We have in energy_bids.csv 855 mails and a binary data that relates (1) or not (0) with the investigation of the company in fraud in energy bids and scheduling.

2 Task 1: Load the data and have a look to one email related with the case and one that is not.

```
mydata <- read_csv("energy_bids.csv")
str(mydata)
```

```
## tibble [855 x 2] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ email     : chr [1:855] "North America's integrated electricity market requires cooperation on en
##  $ responsive: num [1:855] 0 1 0 1 0 0 1 0 0 0 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   email = col_character(),
##   ..   responsive = col_double()
##   .. )
```

```
#better look
strwrap(mydata$email[1],width = 0.9 * getOption("width"))
```

```
##  [1] "North America's integrated electricity market requires cooperation on"
##  [2] "environmental policies Commission for Environmental Cooperation"
##  [3] "releases working paper on North America's electricity market Montreal,"
##  [4] "27 November 2001 -- The North American Commission for Environmental"
##  [5] "Cooperation (CEC) is releasing a working paper highlighting the trend"
##  [6] "towards increasing trade, competition and cross-border investment in"
##  [7] "electricity between Canada, Mexico and the United States. It is hoped"
##  [8] "that the working paper, Environmental Challenges and Opportunities in"
##  [9] "the Evolving North American Electricity Market, will stimulate public"
## [10] "discussion around a CEC symposium of the same title about the need to"
```

```
## [11] "coordinate environmental policies trinationally as a North America-wide"
## [12] "electricity market develops. The CEC symposium will take place in San"
## [13] "Diego on 29-30 November, and will bring together leading experts from"
## [14] "industry, academia, NGOs and the governments of Canada, Mexico and the"
## [15] "United States to consider the impact of the evolving continental"
## [16] "electricity market on human health and the environment. \"Our goal [with"
## [17] "the working paper and the symposium] is to highlight key environmental"
## [18] "issues that must be addressed as the electricity markets in North"
## [19] "America become more and more integrated,\" said Janine Ferretti,"
## [20] "executive director of the CEC. \"We want to stimulate discussion around"
## [21] "the important policy questions being raised so that countries can"
## [22] "cooperate in their approach to energy and the environment.\" The CEC, an"
## [23] "international organization created under an environmental side"
## [24] "agreement to NAFTA known as the North American Agreement on"
## [25] "Environmental Cooperation, was established to address regional"
## [26] "environmental concerns, help prevent potential trade and environmental"
## [27] "conflicts, and promote the effective enforcement of environmental law."
## [28] "The CEC Secretariat believes that greater North American cooperation on"
## [29] "environmental policies regarding the continental electricity market is"
## [30] "necessary to: * protect air quality and mitigate climate change, *"
## [31] "minimize the possibility of environment-based trade disputes, * ensure"
## [32] "a dependable supply of reasonably priced electricity across North"
## [33] "America * avoid creation of pollution havens, and * ensure local and"
## [34] "national environmental measures remain effective. The Changing Market"
## [35] "The working paper profiles the rapid changing North American"
## [36] "electricity market. For example, in 2001, the US is projected to export"
## [37] "13.1 thousand gigawatt-hours (GWh) of electricity to Canada and Mexico."
## [38] "By 2007, this number is projected to grow to 16.9 thousand GWh of"
## [39] "electricity. \"Over the past few decades, the North American electricity"
## [40] "market has developed into a complex array of cross-border transactions"
## [41] "and relationships,\" said Phil Sharp, former US congressman and chairman"
## [42] "of the CEC's Electricity Advisory Board. \"We need to achieve this new"
## [43] "level of cooperation in our environmental approaches as well.\" The"
## [44] "Environmental Profile of the Electricity Sector The electricity sector"
## [45] "is the single largest source of nationally reported toxins in the"
## [46] "United States and Canada and a large source in Mexico. In the US, the"
## [47] "electricity sector emits approximately 25 percent of all NOx emissions,"
## [48] "roughly 35 percent of all CO2 emissions, 25 percent of all mercury"
## [49] "emissions and almost 70 percent of SO2 emissions. These emissions have"
## [50] "a large impact on airsheds, watersheds and migratory species corridors"
## [51] "that are often shared between the three North American countries. \"We"
## [52] "want to discuss the possible outcomes from greater efforts to"
## [53] "coordinate federal, state or provincial environmental laws and policies"
## [54] "that relate to the electricity sector,\" said Ferretti. \"How can we"
## [55] "develop more compatible environmental approaches to help make domestic"
## [56] "environmental policies more effective?\" The Effects of an Integrated"
## [57] "Electricity Market One key issue raised in the paper is the effect of"
## [58] "market integration on the competitiveness of particular fuels such as"
## [59] "coal, natural gas or renewables. Fuel choice largely determines"
## [60] "environmental impacts from a specific facility, along with pollution"
## [61] "control technologies, performance standards and regulations. The paper"
## [62] "highlights other impacts of a highly competitive market as well. For"
## [63] "example, concerns about so called \"pollution havens\" arise when"
## [64] "significant differences in environmental laws or enforcement practices"
```

```
## [65] "induce power companies to locate their operations in jurisdictions with"
## [66] "lower standards. \"The CEC Secretariat is exploring what additional"
## [67] "environmental policies will work in this restructured market and how"
## [68] "these policies can be adapted to ensure that they enhance"
## [69] "competitiveness and benefit the entire region,\" said Sharp. Because"
## [70] "trade rules and policy measures directly influence the variables that"
## [71] "drive a successfully integrated North American electricity market, the"
## [72] "working paper also addresses fuel choice, technology, pollution control"
## [73] "strategies and subsidies. The CEC will use the information gathered"
## [74] "during the discussion period to develop a final report that will be"
## [75] "submitted to the Council in early 2002. For more information or to view"
## [76] "the live video webcast of the symposium, please go to:"
## [77] "http://www.cec.org/electricity. You may download the working paper and"
## [78] "other supporting documents from:"
## [79] "http://www.cec.org/programs_projects/other_initiatives/electricity/docs.cfm?varlan=english."
## [80] "Commission for Environmental Cooperation 393, rue St-Jacques Ouest,"
## [81] "Bureau 200 Montréal (Québec) Canada H2Y 1N9 Tel: (514) 350-4300; Fax:"
## [82] "(514) 350-4314 E-mail: info@ccemtl.org **********"
```

According to result, The data set contains just two fields:

email: the text of the email in question, responsive: a binary (0/1) variable telling whether the email relates to energy schedules or bids.

3 Task 2: How many mails of each class you have?

In order to see count of emails *table* function is used.

```
table(mydata$responsive)
```

```
##
##   0   1
## 716 139
```

There are 716 emails that is not relevant, and there are 139 emails that are relevant.

4 Task 3: Create the corpus. How many characters does the first mail have?

```
corpus = VCorpus(VectorSource(mydata$email))
corpus
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 855
```

```
corpus[[1]]
```

```
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 5605
```

According to results, first mail have 5605 characters.

5 Task 4: Preprocess the corpus (lowercase, remove punctuation, remove stop words and stem) count the characters after each step for the first email.

Below function are used to get required information, and their answers also is shown step by step. The purpose of this process is making data more efficient to use. We see each step result in the email 11th

```
corpus <- tm_map(corpus, tolower)
corpus[[1]]
```

```
## [1] "north america's integrated electricity market requires cooperation on environmental policies co
```

```
corpus <- tm_map(corpus, PlainTextDocument)
corpus[[1]]
```

```
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 5605
```

```
corpus <- tm_map(corpus, removePunctuation)
corpus[[1]]
```

```
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 5448
```

```
corpus <- tm_map(corpus, removeWords, c(stopwords("english")))
corpus[[1]]
```

```
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 4711
```

```
corpus <- tm_map(corpus, stemDocument)
corpus[[1]]
```

```
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 3714
```

```
corpus[[1]]$content
```

```
## [1] "north america integr electr market requir cooper environment polici commiss environment cooper
```

It can be seen that, after the following corpus the number of characters in the first email has dropped from 5605 to 3714.

6 Task 5: Using a frequency matrix, how many words appear at least 20, 200 and 1000 times in any email?

In order to find frequencies following code is used. Data will be prepared for estimation by considering these probabilities.

```
frequencies = DocumentTermMatrix(corpus)
```

To find words appear at leas 20 times:

```
findFreqTerms(frequencies, lowfreq=20)
```

```
##    [1] "\230\230\230\230"                    "\230\230\230\230\230\230\230"
##    [3] "100"                      "10242000"
##    [5] "11012000"                 "12042000"
##    [7] "1400"                     "150"
##    [9] "1996"                     "1998"
##   [11] "1999"                     "200"
##   [13] "2000"                     "2001"
##   [15] "2002"                     "2003"
##   [17] "212"                      "300"
##   [19] "500"                      "713"
##   [21] "7136463490"               "77002"
##   [23] "916"                      "abil"
##   [25] "abl"                      "accept"
##   [27] "access"                   "accord"
##   [29] "account"                  "accrual"
##   [31] "achiev"                   "acquisit"
##   [33] "across"                   "act"
##   [35] "action"                   "activ"
##   [37] "actual"                   "add"
##   [39] "addit"                    "address"
##   [41] "adequ"                    "adjust"
##   [43] "administr"                "adopt"
##   [45] "advanc"                   "advantag"
##   [47] "advis"                    "affair"
##   [49] "affect"                   "affili"
##   [51] "afternoon"                "agenc"
##   [53] "agenda"                   "ago"
##   [55] "agre"                     "agreement"
##   [57] "ahead"                    "air"
##   [59] "alan"                     "alberta"
##   [61] "alex"                     "alleg"
```

```
##   [63] "allen"                "alloc"
##   [65] "allow"                "almost"
##   [67] "along"                "alreadi"
##   [69] "also"                 "altern"
##   [71] "although"             "alway"
##   [73] "amend"                "america"
##   [75] "american"             "among"
##   [77] "amount"               "analysi"
##   [79] "analyst"              "andi"
##   [81] "andor"                "andrew"
##   [83] "angel"                "ann"
##   [85] "announc"              "annual"
##   [87] "anoth"                "answer"
##   [89] "anthoni"              "anticip"
##   [91] "anyon"                "anyth"
##   [93] "apolog"               "appar"
##   [95] "appeal"               "appear"
##   [97] "appli"                "applic"
##   [99] "appoint"              "appreci"
##  [101] "approach"             "appropri"
##  [103] "approv"               "approxim"
##  [105] "april"                "arbitr"
##  [107] "area"                 "argu"
##  [109] "argument"             "arm"
##  [111] "around"               "arrang"
##  [113] "articl"               "asap"
##  [115] "ask"                  "assembl"
##  [117] "assess"               "asset"
##  [119] "assign"               "assist"
##  [121] "associ"               "assum"
##  [123] "assur"                "attach"
##  [125] "attempt"              "attend"
##  [127] "attent"               "attorney"
##  [129] "auction"              "audit"
##  [131] "august"               "author"
##  [133] "avail"                "averag"
##  [135] "avoid"                "awar"
##  [137] "award"                "away"
##  [139] "back"                 "bad"
##  [141] "bahama"               "balanc"
##  [143] "bank"                 "bankruptci"
##  [145] "barbara"              "base"
##  [147] "basi"                 "basic"
##  [149] "becam"                "becom"
##  [151] "began"                "begin"
##  [153] "behalf"               "believ"
##  [155] "ben"                  "benefit"
##  [157] "best"                 "better"
##  [159] "beyond"               "bid"
##  [161] "big"                  "biggest"
##  [163] "bill"                 "billion"
##  [165] "bit"                  "blackout"
##  [167] "blame"                "block"
##  [169] "blue"                 "board"
```

```
##  [171] "bob"                 "bodi"
##  [173] "bond"                "book"
##  [175] "border"              "bought"
##  [177] "box"                 "bradley"
##  [179] "brant"               "brattl"
##  [181] "brazil"              "break"
##  [183] "brent"               "brian"
##  [185] "brief"               "bring"
##  [187] "brown"               "brownout"
##  [189] "bruce"               "build"
##  [191] "built"               "bureau"
##  [193] "burton"              "bush"
##  [195] "busi"                "buy"
##  [197] "buyer"               "ca\230"
##  [199] "caiso"               "cal"
##  [201] "calcul"              "calendar"
##  [203] "calif"               "california"
##  [205] "californian"         "call"
##  [207] "calpin"              "calpx"
##  [209] "came"                "campaign"
##  [211] "can"                 "canada"
##  [213] "canadian"            "candid"
##  [215] "cant"                "cap"
##  [217] "capabl"              "capac"
##  [219] "capit"               "captur"
##  [221] "care"                "carlo"
##  [223] "carol"               "carri"
##  [225] "case"                "cash"
##  [227] "caus"                "cec"
##  [229] "cent"                "center"
##  [231] "central"             "ceo"
##  [233] "certain"             "certif"
##  [235] "cftc"                "chair"
##  [237] "chairman"            "chanc"
##  [239] "chang"               "charg"
##  [241] "charl"               "check"
##  [243] "cheryl"              "chief"
##  [245] "choic"               "chris"
##  [247] "christi"             "christoph"
##  [249] "circumst"            "cite"
##  [251] "citi"                "claim"
##  [253] "class"               "clay"
##  [255] "clean"               "clear"
##  [257] "click"               "client"
##  [259] "close"               "cng"
##  [261] "coal"                "code"
##  [263] "collater"            "collect"
##  [265] "combin"              "come"
##  [267] "comment"             "commerci"
##  [269] "commiss"             "commission"
##  [271] "commit"              "committe"
##  [273] "commod"              "common"
##  [275] "communic"            "communicationsenron"
##  [277] "compani"             "compar"
```

```
##  [279] "competit"               "complaint"
##  [281] "complet"                "complex"
##  [283] "complianc"              "comprehens"
##  [285] "comput"                 "concern"
##  [287] "conclud"                "condit"
##  [289] "conduct"                "confer"
##  [291] "confid"                 "confidenti"
##  [293] "confirm"                "congest"
##  [295] "congress"               "connect"
##  [297] "conserv"                "consid"
##  [299] "consider"               "consist"
##  [301] "constraint"             "construct"
##  [303] "consult"                "consum"
##  [305] "contact"                "contain"
##  [307] "content"                "continu"
##  [309] "contract"               "contribut"
##  [311] "control"                "convers"
##  [313] "cooper"                 "coordin"
##  [315] "copi"                   "copyright"
##  [317] "core"                   "corp"
##  [319] "corpor"                 "correct"
##  [321] "cost"                   "council"
##  [323] "counsel"                "count"
##  [325] "counterparti"           "counti"
##  [327] "countri"                "coupl"
##  [329] "cours"                  "court"
##  [331] "cover"                  "cpuc"
##  [333] "craig"                  "creat"
##  [335] "credit"                 "creditor"
##  [337] "crisi"                  "critic"
##  [339] "crude"                  "current"
##  [341] "curtail"                "curv"
##  [343] "custom"                 "cut"
##  [345] "cynthia"                "dabhol"
##  [347] "daili"                  "damag"
##  [349] "dan"                    "daniel"
##  [351] "dasovich"               "dasovichnaenron"
##  [353] "dasovichnaenronenron"   "data"
##  [355] "date"                   "dave"
##  [357] "davi"                   "david"
##  [359] "day"                    "deadlin"
##  [361] "deal"                   "dear"
##  [363] "debat"                  "debbi"
##  [365] "debra"                  "debt"
##  [367] "dec"                    "decemb"
##  [369] "decid"                  "decis"
##  [371] "declin"                 "default"
##  [373] "defend"                 "defin"
##  [375] "definit"                "degre"
##  [377] "delainey"               "delay"
##  [379] "delet"                  "deliv"
##  [381] "deliveri"               "demand"
##  [383] "democrat"               "deni"
##  [385] "depart"                 "depend"
```

```
##  [387] "deregul"                "deriv"
##  [389] "describ"                "design"
##  [391] "desk"                   "despit"
##  [393] "detail"                 "determin"
##  [395] "develop"                "didnt"
##  [397] "diego"                  "differ"
##  [399] "difficult"              "direct"
##  [401] "director"               "disclos"
##  [403] "disclosur"              "discount"
##  [405] "discuss"                "disput"
##  [407] "dissemin"               "distribut"
##  [409] "district"               "document"
##  [411] "doesnt"                 "dollar"
##  [413] "dolphin"                "domest"
##  [415] "don"                    "done"
##  [417] "donna"                  "dont"
##  [419] "doug"                   "dougla"
##  [421] "dow"                    "dpc"
##  [423] "draft"                  "drew"
##  [425] "drive"                  "drop"
##  [427] "due"                    "duke"
##  [429] "dwr"                    "dynegi"
##  [431] "earli"                  "earlier"
##  [433] "earn"                   "east"
##  [435] "eastern"                "econom"
##  [437] "economi"                "economist"
##  [439] "edison"                 "edit"
##  [441] "edward"                 "effect"
##  [443] "effici"                 "effort"
##  [445] "either"                 "eix"
##  [447] "ekrapelsesaiboscom"     "elect"
##  [449] "electr"                 "electron"
##  [451] "elimin"                 "elizabeth"
##  [453] "els"                    "email"
##  [455] "emerg"                  "emiss"
##  [457] "employe"                "ena"
##  [459] "enclos"                 "encourag"
##  [461] "end"                    "energi"
##  [463] "engin"                  "enhanc"
##  [465] "enough"                 "enron"
##  [467] "enrononlin"             "ensur"
##  [469] "enter"                  "entir"
##  [471] "entiti"                 "environ"
##  [473] "environment"            "eol"
##  [475] "epa"                    "epmi"
##  [477] "equiti"                 "ercot"
##  [479] "eric"                   "error"
##  [481] "especi"                 "essenti"
##  [483] "establish"              "estim"
##  [485] "etc"                    "europ"
##  [487] "european"               "even"
##  [489] "event"                  "ever"
##  [491] "everi"                  "everyon"
##  [493] "everyth"                "evid"
```

```
##  [495] "exampl"                "exceed"
##  [497] "excel"                 "except"
##  [499] "excess"                "exchang"
##  [501] "execut"                "exempt"
##  [503] "exercis"               "exist"
##  [505] "expand"                "expans"
##  [507] "expect"                "expens"
##  [509] "experi"                "explain"
##  [511] "explor"                "export"
##  [513] "exposur"               "express"
##  [515] "extend"                "extens"
##  [517] "extent"                "extrem"
##  [519] "face"                  "facil"
##  [521] "fact"                  "factor"
##  [523] "fail"                  "fair"
##  [525] "fall"                  "far"
##  [527] "farmout"               "favor"
##  [529] "fax"                   "feb"
##  [531] "februari"              "feder"
##  [533] "fee"                   "feel"
##  [535] "ferc"                  "field"
##  [537] "figur"                 "file"
##  [539] "final"                 "financ"
##  [541] "financi"               "find"
##  [543] "fine"                  "firm"
##  [545] "first"                 "fit"
##  [547] "five"                  "fix"
##  [549] "flexibl"               "floor"
##  [551] "flow"                  "focus"
##  [553] "folder"                "folk"
##  [555] "follow"                "forc"
##  [557] "forecast"              "foreign"
##  [559] "form"                  "formal"
##  [561] "format"                "former"
##  [563] "forward"               "found"
##  [565] "four"                  "francisco"
##  [567] "frank"                 "free"
##  [569] "fri"                   "friday"
##  [571] "friend"                "front"
##  [573] "fuel"                  "full"
##  [575] "fulli"                 "function"
##  [577] "fund"                  "fundament"
##  [579] "futur"                 "fyi"
##  [581] "gain"                  "game"
##  [583] "gari"                  "gas"
##  [585] "gather"                "general"
##  [587] "generat"               "georg"
##  [589] "gerald"                "germani"
##  [591] "get"                   "give"
##  [593] "given"                 "global"
##  [595] "goe"                   "good"
##  [597] "got"                   "gov"
##  [599] "govern"                "governor"
##  [601] "grand"                 "grant"
```

```
##  [603] "gray"                    "great"
##  [605] "greater"                 "green"
##  [607] "greg"                    "grid"
##  [609] "ground"                  "group"
##  [611] "grow"                    "growth"
##  [613] "guarante"                "guaranti"
##  [615] "guy"                     "half"
##  [617] "hall"                    "hand"
##  [619] "handl"                   "happen"
##  [621] "happi"                   "hard"
##  [623] "harri"                   "harrisetsenronenron"
##  [625] "hartsoecorpenronenron"   "havent"
##  [627] "head"                    "hear"
##  [629] "heard"                   "heat"
##  [631] "hedg"                    "held"
##  [633] "help"                    "here"
##  [635] "herein"                  "high"
##  [637] "higher"                  "highlight"
##  [639] "hilton"                  "histor"
##  [641] "hit"                     "hold"
##  [643] "home"                    "honor"
##  [645] "hope"                    "hot"
##  [647] "hour"                    "hous"
##  [649] "houston"                 "howev"
##  [651] "hpl"                     "hub"
##  [653] "huge"                    "hunter"
##  [655] "hydro"                   "idea"
##  [657] "identifi"                "iep"
##  [659] "ill"                     "imag"
##  [661] "imageimag"               "immedi"
##  [663] "impact"                  "implement"
##  [665] "import"                  "impos"
##  [667] "improv"                  "inc"
##  [669] "incent"                  "includ"
##  [671] "incorpor"                "increas"
##  [673] "independ"                "index"
##  [675] "india"                   "indian"
##  [677] "indic"                   "individu"
##  [679] "industri"                "info"
##  [681] "inform"                  "infrastructur"
##  [683] "initi"                   "inject"
##  [685] "innov"                   "input"
##  [687] "instal"                  "instead"
##  [689] "institut"                "instrument"
##  [691] "insur"                   "integr"
##  [693] "intend"                  "intent"
##  [695] "interconnect"            "interest"
##  [697] "intern"                  "internet"
##  [699] "interst"                 "interview"
##  [701] "introduc"                "invest"
##  [703] "investig"                "investor"
##  [705] "invit"                   "involv"
##  [707] "ipp"                     "isda"
##  [709] "island"                  "isnt"
```

```
##  [711] "iso"                   "issu"
##  [713] "item"                  "ive"
##  [715] "jack"                  "jame"
##  [717] "jan"                   "jane"
##  [719] "janet"                 "januari"
##  [721] "japan"                 "jason"
##  [723] "jay"                   "jeff"
##  [725] "jeffrey"               "jennif"
##  [727] "jim"                   "job"
##  [729] "joe"                   "john"
##  [731] "join"                  "joint"
##  [733] "jonathan"              "jone"
##  [735] "joneshouect"           "joneshouectect"
##  [737] "jorgensen"             "jose"
##  [739] "juli"                  "june"
##  [741] "jurisdict"             "just"
##  [743] "kaminski"              "kaminskihouect"
##  [745] "kaminskihouectect"     "kaplan"
##  [747] "karen"                 "kate"
##  [749] "kati"                  "kay"
##  [751] "kean"                  "keannaenron"
##  [753] "keannaenronenron"      "keep"
##  [755] "keith"                 "kelley"
##  [757] "kelli"                 "ken"
##  [759] "kent"                  "kevin"
##  [761] "key"                   "kim"
##  [763] "kimber"                "kind"
##  [765] "know"                  "known"
##  [767] "lack"                  "lamberthouectect"
##  [769] "languag"               "larg"
##  [771] "largest"               "larri"
##  [773] "last"                  "late"
##  [775] "later"                 "latest"
##  [777] "launch"                "law"
##  [779] "lawmak"                "lawsuit"
##  [781] "lawyer"                "lay"
##  [783] "lead"                  "leader"
##  [785] "learn"                 "least"
##  [787] "leav"                  "lee"
##  [789] "left"                  "legal"
##  [791] "legisl"                "legislatur"
##  [793] "lender"                "length"
##  [795] "lesli"                 "less"
##  [797] "let"                   "letter"
##  [799] "level"                 "liabil"
##  [801] "light"                 "like"
##  [803] "limit"                 "linda"
##  [805] "line"                  "link"
##  [807] "liquid"                "lisa"
##  [809] "list"                  "litig"
##  [811] "littl"                 "live"
##  [813] "llc"                   "lloyd"
##  [815] "llp"                   "lng"
##  [817] "load"                  "local"
```

```
##  [819] "locat"                "london"
##  [821] "long"                 "longer"
##  [823] "longterm"             "look"
##  [825] "lorrain"              "los"
##  [827] "lose"                 "loss"
##  [829] "lost"                 "lot"
##  [831] "louis"                "low"
##  [833] "lower"                "lynn"
##  [835] "made"                 "maharashtra"
##  [837] "mail"                 "main"
##  [839] "maintain"             "mainten"
##  [841] "major"                "make"
##  [843] "manag"                "mani"
##  [845] "manipul"              "mara"
##  [847] "march"                "margin"
##  [849] "mari"                 "maria"
##  [851] "mark"                 "market"
##  [853] "marti"                "master"
##  [855] "match"                "materi"
##  [857] "matt"                 "matter"
##  [859] "max"                  "maximum"
##  [861] "may"                  "mayb"
##  [863] "mean"                 "measur"
##  [865] "mechan"               "meet"
##  [867] "megawatt"             "melissa"
##  [869] "member"               "memo"
##  [871] "mention"              "merchant"
##  [873] "messag"               "met"
##  [875] "metal"                "meter"
##  [877] "method"               "mexico"
##  [879] "michael"              "michell"
##  [881] "middl"                "midwest"
##  [883] "might"                "mike"
##  [885] "mile"                 "mill"
##  [887] "miller"               "milleretsenronenron"
##  [889] "million"              "mind"
##  [891] "minimum"              "minut"
##  [893] "mirant"               "miss"
##  [895] "mitig"                "model"
##  [897] "modif"                "modifi"
##  [899] "molli"                "mona"
##  [901] "monday"               "money"
##  [903] "monitor"              "month"
##  [905] "morgan"               "morn"
##  [907] "motion"               "mou"
##  [909] "move"                 "mseb"
##  [911] "much"                 "must"
##  [913] "name"                 "nation"
##  [915] "natur"                "nda"
##  [917] "near"                 "necessari"
##  [919] "need"                 "needl"
##  [921] "negoti"               "nerc"
##  [923] "net"                  "network"
##  [925] "never"                "new"
```

```
##  [927] "news"                "newslett"
##  [929] "newswir"             "next"
##  [931] "ngx"                 "night"
##  [933] "nom"                 "nomin"
##  [935] "noon"                "normal"
##  [937] "north"               "northern"
##  [939] "northwest"           "note"
##  [941] "notic"               "notifi"
##  [943] "nov"                 "novemb"
##  [945] "now"                 "nrg"
##  [947] "number"              "numerix"
##  [949] "nymex"               "object"
##  [951] "oblig"               "obtain"
##  [953] "obvious"             "occur"
##  [955] "octob"               "offer"
##  [957] "offic"               "offici"
##  [959] "oil"                 "old"
##  [961] "one"                 "onlin"
##  [963] "open"                "oper"
##  [965] "opinion"             "opportun"
##  [967] "oppos"               "option"
##  [969] "order"               "oregon"
##  [971] "organ"               "origin"
##  [973] "other"               "otherwis"
##  [975] "outag"               "outlin"
##  [977] "output"              "outsid"
##  [979] "outsourc"            "outstand"
##  [981] "overal"              "owe"
##  [983] "own"                 "owner"
##  [985] "ownership"           "oxi"
##  [987] "pacif"               "packag"
##  [989] "page"                "paid"
##  [991] "panel"               "paper"
##  [993] "park"                "part"
##  [995] "parti"               "particip"
##  [997] "particular"          "partner"
##  [999] "paso"                "pass"
## [1001] "past"                "patricia"
## [1003] "paul"                "pay"
## [1005] "payment"             "peak"
## [1007] "penalti"             "pend"
## [1009] "peopl"               "per"
## [1011] "percent"             "perform"
## [1013] "period"              "permit"
## [1015] "person"              "peter"
## [1017] "petit"               "pge"
## [1019] "pges"                "phase"
## [1021] "phil"                "phillip"
## [1023] "phoenix"             "phone"
## [1025] "physic"              "pick"
## [1027] "pipelin"             "pjm"
## [1029] "place"               "plaintiff"
## [1031] "plan"                "plant"
## [1033] "platform"            "play"
```

```
## [1035] "player"                "pleas"
## [1037] "plus"                  "point"
## [1039] "polici"                "polit"
## [1041] "pollut"                "pool"
## [1043] "portfolio"             "portion"
## [1045] "portland"              "posit"
## [1047] "possibl"               "post"
## [1049] "potenti"               "power"
## [1051] "powerex"               "ppa"
## [1053] "practic"               "predict"
## [1055] "prefer"                "preliminari"
## [1057] "prepar"                "present"
## [1059] "presid"                "press"
## [1061] "pressur"               "pretti"
## [1063] "prevent"               "previous"
## [1065] "price"                 "primari"
## [1067] "princip"               "print"
## [1069] "prior"                 "privat"
## [1071] "privileg"              "probabl"
## [1073] "problem"               "procedur"
## [1075] "proceed"               "process"
## [1077] "procur"                "produc"
## [1079] "product"               "profil"
## [1081] "profit"                "program"
## [1083] "progress"              "prohibit"
## [1085] "project"               "promot"
## [1087] "prompt"                "properti"
## [1089] "propos"                "prospect"
## [1091] "protect"               "provid"
## [1093] "provis"                "public"
## [1095] "publish"               "puc"
## [1097] "pull"                  "purchas"
## [1099] "purpos"                "pursuant"
## [1101] "push"                  "put"
## [1103] "qualiti"               "quantiti"
## [1105] "quarter"               "question"
## [1107] "quick"                 "quit"
## [1109] "quot"                  "rais"
## [1111] "randi"                 "rang"
## [1113] "rate"                  "ratepay"
## [1115] "rather"                "ray"
## [1117] "reach"                 "read"
## [1119] "readi"                 "real"
## [1121] "realli"                "realtim"
## [1123] "reason"                "rebecca"
## [1125] "receipt"               "receiv"
## [1127] "recent"                "recipi"
## [1129] "recogn"                "recommend"
## [1131] "record"                "recov"
## [1133] "recoveri"              "redlin"
## [1135] "reduc"                 "reduct"
## [1137] "refer"                 "referenc"
## [1139] "reflect"               "reform"
## [1141] "refund"                "refus"
```

```
## [1143] "reg"                       "regard"
## [1145] "region"                    "regul"
## [1147] "regular"                   "regulatori"
## [1149] "relat"                     "releas"
## [1151] "relev"                     "reliabl"
## [1153] "reliant"                   "relief"
## [1155] "remain"                    "remedi"
## [1157] "remov"                     "replac"
## [1159] "repli"                     "report"
## [1161] "repres"                    "republican"
## [1163] "request"                   "requir"
## [1165] "research"                  "reserv"
## [1167] "resolut"                   "resolv"
## [1169] "resourc"                   "respect"
## [1171] "respond"                   "respons"
## [1173] "restructur"                "result"
## [1175] "retail"                    "retain"
## [1177] "return"                    "reuter"
## [1179] "revenu"                    "review"
## [1181] "revis"                     "richard"
## [1183] "rick"                      "right"
## [1185] "rise"                      "risk"
## [1187] "river"                     "road"
## [1189] "rob"                       "robert"
## [1191] "roger"                     "role"
## [1193] "roll"                      "ron"
## [1195] "room"                      "rto"
## [1197] "rule"                      "run"
## [1199] "russel"                    "sacramento"
## [1201] "said"                      "sale"
## [1203] "sampl"                     "samuel"
## [1205] "san"                       "sandra"
## [1207] "sara"                      "sarah"
## [1209] "sarashackletonenroncom" "saudi"
## [1211] "save"                      "say"
## [1213] "schedul"                   "schenk"
## [1215] "school"                    "scott"
## [1217] "screen"                    "sdge"
## [1219] "season"                    "second"
## [1221] "secretari"                 "section"
## [1223] "sector"                    "secur"
## [1225] "see"                       "seek"
## [1227] "seem"                      "seen"
## [1229] "select"                    "sell"
## [1231] "seller"                    "sempra"
## [1233] "sen"                       "senat"
## [1235] "send"                      "sender"
## [1237] "senior"                    "sens"
## [1239] "sent"                      "separ"
## [1241] "septemb"                   "serious"
## [1243] "serv"                      "servic"
## [1245] "session"                   "set"
## [1247] "settl"                     "settlement"
## [1249] "seven"                     "sever"
```

```
## [1251] "shackleton"             "shackletonhouect"
## [1253] "shackletonhouectect"    "shall"
## [1255] "shapironaenronenron"    "share"
## [1257] "sheet"                  "sheila"
## [1259] "shipper"                "shirley"
## [1261] "short"                  "shortag"
## [1263] "show"                   "shut"
## [1265] "side"                   "sign"
## [1267] "signific"               "similar"
## [1269] "simpli"                 "sinc"
## [1271] "singl"                  "site"
## [1273] "situat"                 "six"
## [1275] "size"                   "skeanenroncom"
## [1277] "skill"                  "slide"
## [1279] "small"                  "smaraenroncom"
## [1281] "smith"                  "socal"
## [1283] "societi"                "solar"
## [1285] "sold"                   "sole"
## [1287] "solut"                  "someon"
## [1289] "someth"                 "sometim"
## [1291] "soon"                   "sorri"
## [1293] "sound"                  "sourc"
## [1295] "south"                  "southern"
## [1297] "speak"                  "special"
## [1299] "specif"                 "specifi"
## [1301] "spike"                  "split"
## [1303] "spoke"                  "spokesman"
## [1305] "spot"                   "spread"
## [1307] "spreadsheet"            "spring"
## [1309] "stabil"                 "stacey"
## [1311] "staff"                  "stage"
## [1313] "stake"                  "stand"
## [1315] "standard"               "start"
## [1317] "state"                  "statement"
## [1319] "status"                 "stay"
## [1321] "steel"                  "steffesnaenronenron"
## [1323] "step"                   "stephani"
## [1325] "stephen"                "steve"
## [1327] "steven"                 "still"
## [1329] "stock"                  "stop"
## [1331] "storag"                 "stori"
## [1333] "strategi"               "street"
## [1335] "strict"                 "strong"
## [1337] "structur"               "studi"
## [1339] "subcommitte"            "subject"
## [1341] "submit"                 "subpoena"
## [1343] "subscrib"               "subsequ"
## [1345] "subsidiari"             "substanti"
## [1347] "success"                "sue"
## [1349] "suggest"                "suit"
## [1351] "summari"                "summer"
## [1353] "sunday"                 "suppli"
## [1355] "supplier"               "support"
## [1357] "surcharg"               "sure"
```

```
## [1359] "surpris"                "susan"
## [1361] "swap"                   "switch"
## [1363] "synchron"               "system"
## [1365] "tabl"                   "take"
## [1367] "taken"                  "talk"
## [1369] "tana"                   "tanya"
## [1371] "target"                 "tariff"
## [1373] "tax"                    "taylor"
## [1375] "taylorhouectect"        "team"
## [1377] "technic"                "technolog"
## [1379] "ted"                    "tel"
## [1381] "telephon"               "tell"
## [1383] "term"                   "termin"
## [1385] "terri"                  "test"
## [1387] "testimoni"              "texa"
## [1389] "thank"                  "that"
## [1391] "therefor"               "theyr"
## [1393] "thing"                  "think"
## [1395] "third"                  "thoma"
## [1397] "though"                 "thought"
## [1399] "three"                  "throughout"
## [1401] "thursday"               "ticket"
## [1403] "tim"                    "time"
## [1405] "today"                  "togeth"
## [1407] "told"                   "tom"
## [1409] "tomorrow"               "ton"
## [1411] "toni"                   "took"
## [1413] "top"                    "topic"
## [1415] "total"                  "toward"
## [1417] "track"                  "trade"
## [1419] "trader"                 "transact"
## [1421] "transfer"               "transmiss"
## [1423] "transport"              "travel"
## [1425] "tri"                    "true"
## [1427] "trust"                  "tuesday"
## [1429] "turn"                   "two"
## [1431] "type"                   "ultim"
## [1433] "unabl"                  "unbundl"
## [1435] "understand"             "unfair"
## [1437] "unit"                   "univers"
## [1439] "unlaw"                  "unless"
## [1441] "unlik"                  "unreason"
## [1443] "updat"                  "upon"
## [1445] "upstream"               "urg"
## [1447] "use"                    "user"
## [1449] "usual"                  "util"
## [1451] "valu"                   "valuat"
## [1453] "var"                    "various"
## [1455] "ventur"                 "version"
## [1457] "via"                    "vice"
## [1459] "view"                   "vinc"
## [1461] "vincejkaminskienroncom" "violat"
## [1463] "visit"                  "voic"
## [1465] "volatil"                "volum"
```

```
## [1467] "vote"                  "wait"
## [1469] "want"                  "washington"
## [1471] "watch"                 "water"
## [1473] "way"                   "weather"
## [1475] "web"                   "websit"
## [1477] "wed"                   "wednesday"
## [1479] "week"                  "weekend"
## [1481] "well"                  "went"
## [1483] "west"                  "western"
## [1485] "what"                  "whether"
## [1487] "white"                 "wholesal"
## [1489] "will"                  "william"
## [1491] "winter"                "wisconsin"
## [1493] "within"                "without"
## [1495] "wonder"                "wont"
## [1497] "wood"                  "word"
## [1499] "work"                  "world"
## [1501] "worth"                 "write"
## [1503] "writer"                "written"
## [1505] "wscc"                  "year"
## [1507] "yes"                   "yesterday"
## [1509] "yet"                   "york"
## [1511] "zone"
```

To find words appear at leas 200 times:

```
findFreqTerms(frequencies, lowfreq=200)
```

```
##   [1] "2000"       "2001"       "agreement"  "also"       "attach"
##   [6] "bill"       "busi"       "california" "call"       "can"
##  [11] "cap"        "chang"      "comment"    "commiss"    "compani"
##  [16] "contract"   "corp"       "cost"       "credit"     "current"
##  [21] "custom"     "david"      "day"        "deal"       "demand"
##  [26] "discuss"    "document"   "draft"      "electr"     "email"
##  [31] "energi"     "enron"      "fax"        "ferc"       "file"
##  [36] "first"      "follow"     "forward"    "gas"        "generat"
##  [41] "get"        "group"      "houston"    "includ"     "increas"
##  [46] "inform"     "iso"        "issu"       "jeff"       "john"
##  [51] "just"       "know"       "last"       "legal"      "let"
##  [56] "like"       "look"       "make"       "manag"      "mark"
##  [61] "market"     "may"        "meet"       "messag"     "month"
##  [66] "natur"      "need"       "new"        "now"        "one"
##  [71] "oper"       "order"      "origin"     "per"        "plan"
##  [76] "plant"      "pleas"      "point"      "power"      "price"
##  [81] "product"    "project"    "propos"     "provid"     "purchas"
##  [86] "question"   "rate"       "receiv"     "regard"     "report"
##  [91] "request"    "requir"     "respons"    "review"     "risk"
##  [96] "said"       "say"        "see"        "sent"       "servic"
## [101] "state"      "subject"    "suppli"     "system"     "take"
## [106] "term"       "thank"      "time"       "trade"      "transact"
## [111] "transmiss"  "two"        "use"        "util"       "want"
## [116] "week"       "will"       "work"       "year"
```

To find words appear at leas 1000 times:

```
findFreqTerms(frequencies, lowfreq=1000)
```

```
## [1] "email"  "enron"  "market" "power"  "will"
```

There are only five words with frequencies higher than 1000

7 Task 6: Remove those that appear less than 3%

For this part *removeSparseTerms* function is used. The purpose of this to eliminate unnecessary words to have a better model. Also, we can overcome sparsity with this method.

```
sparse = removeSparseTerms(frequencies, 0.97)
sparse
```

```
## <<DocumentTermMatrix (documents: 855, terms: 788)>>
## Non-/sparse entries: 51612/622128
## Sparsity           : 92%
## Maximal term length: 19
## Weighting          : term frequency (tf)
```

8 Task 7: Create the data frame with the outcome.

For the prediction analyses, data frame is needed to be created because we cannot know the effects of words if all of them are in one row. Thus, we need to separate all words in a correct order. Below codes are used to create the data frame.

```
emailSparse = as.data.frame(as.matrix(sparse))
colnames(emailSparse) = make.names(colnames(emailSparse))
emailSparse$responsive <- mydata$responsive
str(emailSparse, list.len = 20)
```

```
## 'data.frame':    855 obs. of  789 variables:
##  $ X100              : num  0 0 0 0 0 0 5 0 0 0 ...
##  $ X1400             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ X1999             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ X2000             : num  0 0 1 0 1 0 6 0 1 0 ...
##  $ X2001             : num  2 1 0 0 0 0 7 0 0 0 ...
##  $ X713              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ X77002            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ abl               : num  0 0 0 0 0 0 2 0 0 0 ...
##  $ accept            : num  0 0 0 0 0 0 1 0 0 0 ...
##  $ access            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ accord            : num  0 0 0 0 0 0 1 0 0 0 ...
##  $ account           : num  0 0 0 0 0 0 3 0 0 0 ...
##  $ act               : num  0 0 0 0 0 0 1 0 0 0 ...
##  $ action            : num  0 0 0 0 1 0 0 0 0 0 ...
##  $ activ             : num  0 0 1 0 1 0 1 0 0 0 ...
##  $ actual            : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ add                : num  0 0 0 0 0 0 1 0 0 0 ...
##  $ addit              : num  1 0 0 0 0 0 1 0 0 0 ...
##  $ address            : num  3 0 0 0 2 0 0 0 0 1 ...
##  $ administr          : num  0 0 0 0 0 0 1 0 0 0 ...
##   [list output truncated]
```

9 Task 8: Split in a training and a testing set

In order to create test and training data, below code is used, also split ratio is chosen as 0.75

```
set.seed(42)
split = sample.split(emailSparse$responsive, SplitRatio = 0.75)

trainSparse <- subset(emailSparse, split==TRUE)
testSparse <- subset(emailSparse, split==FALSE)
```

10 Task 9: Create a CART model

By using the train set the CART model will be built.

```
emailCART = rpart(responsive ~ ., data=trainSparse, method="class")
prp(emailCART)
```

11 Task 10: Predict the probability of each mail of being or not relevant in the investigation.

Now to predict probabilities, *predict* function is used without using type=”class” comment. These probabilities show the probability of being 0 or 1 for emails.

```
predictCART_prop = predict(emailCART, newdata=testSparse)
predictCART_prop[1:10,]
```

```
##                          0          1
## character.0.    0.9506903 0.04930966
## character.0..2  0.1250000 0.87500000
## character.0..5  0.9506903 0.04930966
## character.0..18 0.9506903 0.04930966
## character.0..23 0.9506903 0.04930966
## character.0..24 0.9506903 0.04930966
## character.0..29 0.9000000 0.10000000
## character.0..30 0.9506903 0.04930966
## character.0..32 0.9506903 0.04930966
## character.0..33 0.9506903 0.04930966
```

The first column is the predicted probability of the document being non-responsive. The second column is the predicted probability of the document being responsive. They sum to 1.

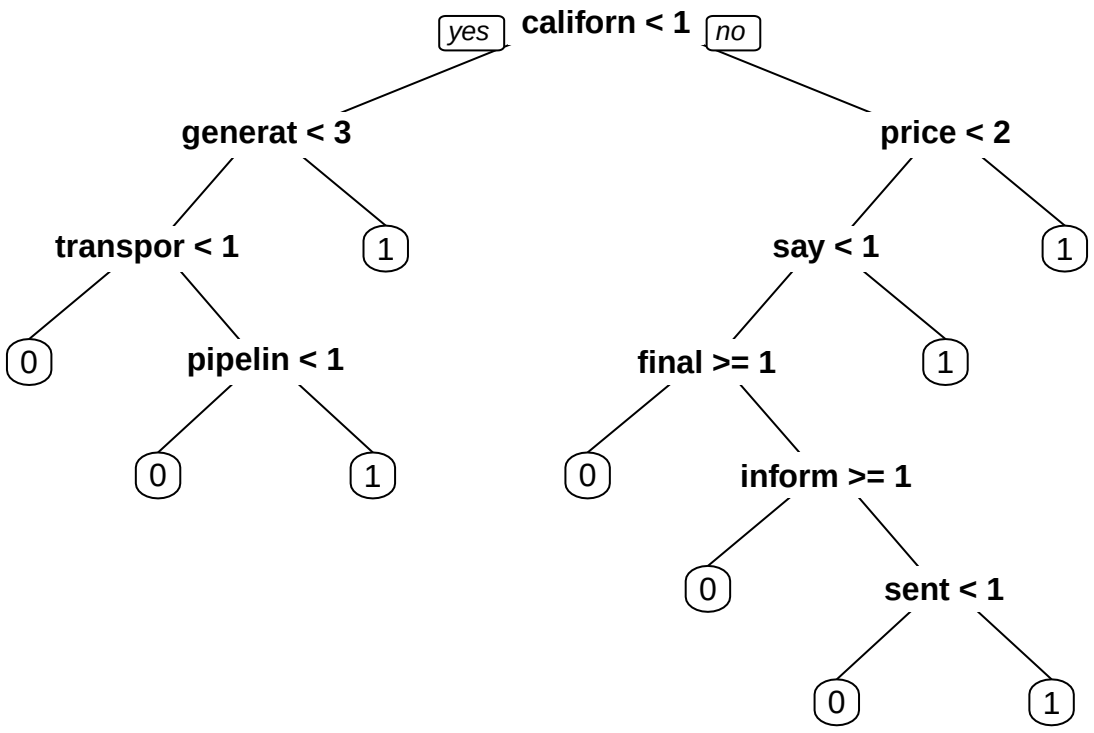

Figure 1: CART Model of trainSparse

12 Task 11: Use a threshold 0.5 to classify and compare its accuracy with the baseline.

For using a threshold 0.5 we need the type to be "class". Thus, the type will be added to the previous predict to convert the probilities obtaining results as 0 or 1.

```
predictCART = predict(emailCART, newdata=testSparse, type="class")
cmat <- table(testSparse$responsive, predictCART)
cmat
```

```
##    predictCART
##       0   1
##   0 172   7
##   1  16  19
```

```
accu_CART <- (cmat[1,1] + cmat[2,2])/sum(cmat)
accu_CART
```

```
## [1] 0.8925234
```

```
cmat_baseline <- table(testSparse$responsive)
cmat_baseline
```

```
##
##   0   1
## 179  35
```

```
accu_baseline <- max(cmat_baseline)/sum(cmat_baseline)
accu_baseline
```

```
## [1] 0.8364486
```

The found accuracy is acceptable as good due to being close to 1. Also, found accuracy is bigger than baseline accuracy as expected.

13 Add Task 12: Use ROC for selecting the threshold.

```
pred = prediction(predictCART_prop[,2],testSparse$responsive)
perf = performance(pred, "tpr", "fpr")
plot(perf, colorize=TRUE, print.cutoffs.at=seq(0,1,by=0.1), text.adj=c(-0.2,1.7))
```

According to ROC curve plot, threshold value can be chosen around 0.1.

14 Add Task 13: Which is the area under the curve.

In order to get area under curve below code is used.

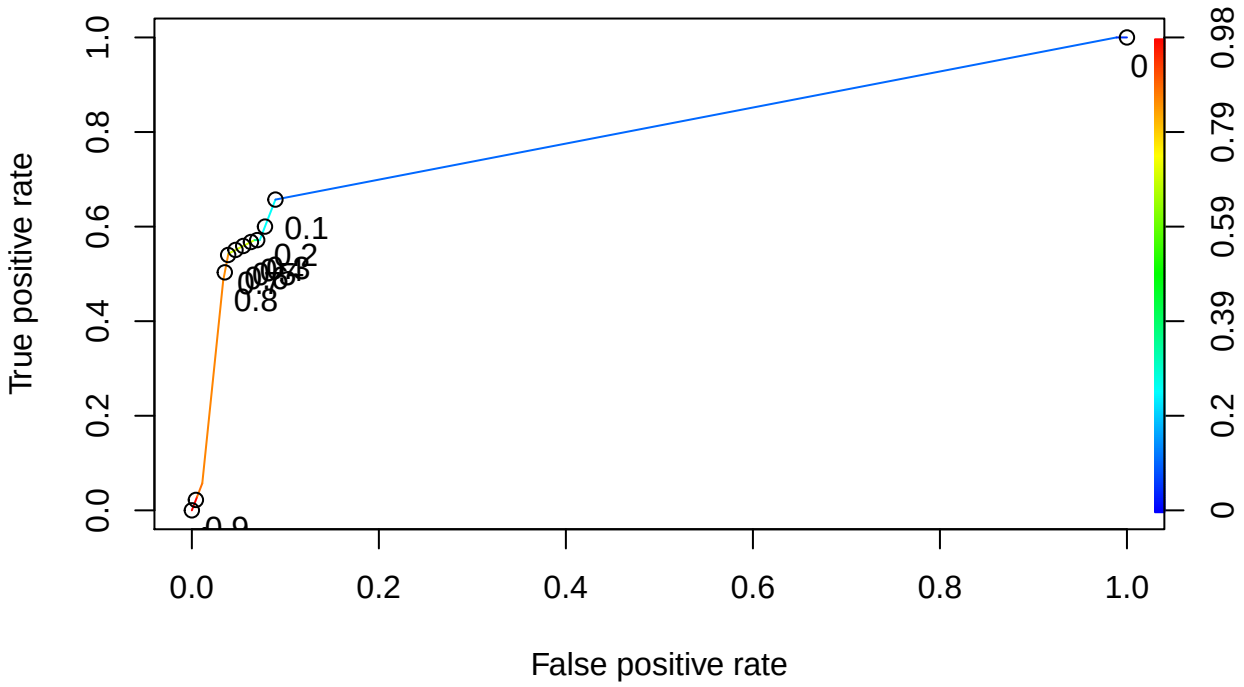

Figure 2: ROC Curve

```
auc_CART <- as.numeric(performance(pred, "auc")@y.values)
auc_CART
```

```
## [1] 0.7947326
```

The AUC of the CART models is 0.7947, which means that our model can differentiate between a randomly selected responsive and non-responsive document about 79.4% of the times.

15 Add Task 14: Use random forest for improving the accuracy.

In order to improve the accuracy of analysis, *random forest* can be used. This function improve accuracy by generating a large number of bootstrapped trees, classifying a case using each tree and deciding a final predicted outcome by combining all trees.

```
set.seed(422)
trainSparse_1 <- trainSparse
testSparse_1 <- testSparse
trainSparse_1$responsive = as.factor(trainSparse$responsive)
testSparse_1$responsive = as.factor(testSparse$responsive)


emailRF = randomForest(responsive ~ ., data=trainSparse_1)
predictRF = predict(emailRF, newdata=testSparse_1)
cmat1 <- table(testSparse_1$responsive, predictRF)

accu_CART1 <- (cmat1[1,1] + cmat1[2,2])/sum(cmat1)
accu_CART1
```

```
## [1] 0.8971963
```

The previous accuracy was 0.8925. Thus, we can observe a slightly improvement in the obtained accuracy.