text_analytics.R

Fiona

Fri Oct 13 18:45:40 2017

```
#
skills <- read.csv("C:/Users/Fiona/Desktop/Business Analytics and Decision
Sciences/Forecasting and Business Analytics/FABA-L9-Notes/skills.csv",
header= TRUE, stringsAsFactors = FALSE)
# The argument stringsAsFactors=FALSE keeps the text as a character type

# Next let's explore the data:
head(skills)

##
skillsneeded
## 1
How to apply tests of significance to my findings\nGeneral coding best
practices for legibility and reproducibility
## 2                                                    Just need to increase
comfort with core concepts (why researchers would need to do some
programming) in order to assist researchers with data curation. Not an active
researcher myself.
## 3
?
## 4
Programming in R and Python (little to no experience at this time)\nAdvanced
commands in Stata, particularly with regards to visualizing data\nGIS
## 5 I'd like to learn more programming languages (currently I know Stata and
a teeny bit of R and SQL). Potentially also survey data collection and
analysis, and it's always good to go over more statistics, conceptually and
pragmatically.
## 6
Learn the basics of browsing and manipulating data using Stata.

summary(skills)

##  skillsneeded
##  Length:72
##  Class :character
##  Mode  :character

# Let's install and add the tm package to your R library:
#install.packages("tm")
library(tm)

## Warning: package 'tm' was built under R version 3.3.3
```

```
## Loading required package: NLP

# Now we can convert our data to a document term matrix which reflects the
number of times each token (word) is used.
# First we convert the data to a corpus (a collection of documents containing
text in the same format) using:
corpus <- Corpus(DataframeSource(skills))
# Then to a document-term matrix with:
dtm <- DocumentTermMatrix(corpus)

# You can see the number of tokens in the data, referred to as terms and the
number of characters in the longest token under maximal term length:
dtm

## <<DocumentTermMatrix (documents: 72, terms: 511)>>
## Non-/sparse entries: 919/35873
## Sparsity           : 98%
## Maximal term length: 20
## Weighting          : term frequency (tf)

# You can also see a count of how many times each of the tokens are used in
the 72 responses:
inspect(dtm)

## <<DocumentTermMatrix (documents: 72, terms: 511)>>
## Non-/sparse entries: 919/35873
## Sparsity           : 98%
## Maximal term length: 20
## Weighting          : term frequency (tf)
## Sample             :
##     Terms
## Docs analysis and data for like more software the with would
##   19        0   4    0   0    1    0        0   2    3     0
##   35        0   1    0   1    0    1        0   0    0     1
##   36        2   2    0   0    1    0        2   0    1     2
##   38        2   3    1   0    1    0        0   0    1     2
##   39        0   1    0   1    0    2        0   0    0     1
##   40        0   2    2   0    0    0        0   0    0     0
##   44        0   1    2   1    1    1        1   0    0     0
##   5         0   5    1   0    1    2        0   0    0     0
##   54        0   1    0   3    0    0        2   0    0     0
##   9         0   3    1   0    0    1        0   0    0     1

# You will see that many of the tokens are uninteresting and uncommon and
therefore can be removed from the analysis.
# This is easily achieved with the tm package using the following code:
dtm <- DocumentTermMatrix(corpus, control = list(removePunctuation = TRUE,
stripWhitespace = TRUE, removeNumbers = TRUE, stopwords =  TRUE, tolower =
TRUE, wordLengths=c(1,Inf)))

# You can see punctuation, unnecessary spaces, numbers and useless words such
```

```
as "a" and "the" (called stopwords) are removed:
dtm

## <<DocumentTermMatrix (documents: 72, terms: 407)>>
## Non-/sparse entries: 719/28585
## Sparsity           : 98%
## Maximal term length: 19
## Weighting          : term frequency (tf)

# The code also converts all tokens to lower case and makes sure tokens have
at least 1 character (in many applications you may want to remove tokens with
less than 3 characters but in this case tokens/words such as R with one
character will be of interest)
Terms(dtm)

##   [1] "able"               "accessing"          "across"
##   [4] "active"             "advanced"           "advantage"
##   [7] "aligns"             "allowing"           "allows"
##  [10] "almost"             "also"               "always"
##  [13] "amounts"            "analyses"           "analysis"
##  [16] "analysisspatial"    "analyze"            "analyzing"
##  [19] "andor"              "answer"             "anthropac"
##  [22] "apply"              "appropriate"        "arcgis"
##  [25] "around"             "assist"             "atlasti"
##  [28] "aws"                "background"         "basic"
##  [31] "basics"             "basicsphilosophy"   "bayesian"
##  [34] "beatarea"           "become"             "beneficial"
##  [37] "best"               "better"             "beyond"
##  [40] "big"                "bit"                "books"
##  [43] "brain"              "broader"            "browsing"
##  [46] "build"              "building"           "call"
##  [49] "campus"             "can"                "capabilities"
##  [52] "capacity"           "career"             "catalyst"
##  [55] "certainly"          "class"              "classes"
##  [58] "clinical"           "closely"            "cluster"
##  [61] "coding"             "collection"         "comfort"
##  [64] "commands"           "competent"          "computing"
##  [67] "concepts"           "conceptually"       "conditions"
##  [70] "conducting"         "confusing"          "contained"
##  [73] "content"            "core"               "created"
##  [76] "css"                "curation"           "currently"
##  [79] "customize"          "data"               "database"
##  [82] "databases"          "department"         "design"
##  [85] "develop"            "diary"              "different"
##  [88] "digital"            "display"            "displays"
##  [91] "done"               "dont"               "draw"
##  [94] "drupal"             "early"              "easier"
##  [97] "ecological"         "effectively"        "efficient"
## [100] "eg"                 "enough"             "environments"
## [103] "etc"                "ethnographic"       "even"
```

```
## [106] "excel"               "except"              "exclusively"
## [109] "existing"            "expand"              "experience"
## [112] "explore"             "facility"            "familiar"
## [115] "familiarity"         "fields"              "findings"
## [118] "flash"               "flexibility"         "forms"
## [121] "fortran"             "forward"             "frequencyoccurances"
## [124] "frequently"          "friendly"            "fulltext"
## [127] "gain"                "general"             "generally"
## [130] "geospatial"          "getting"             "gis"
## [133] "go"                  "going"               "good"
## [136] "graduate"            "graph"               "great"
## [139] "group"               "growing"             "guess"
## [142] "handle"              "happy"               "hard"
## [145] "havent"              "help"                "helpful"
## [148] "hierarchical"        "humanities"          "id"
## [151] "identifying"         "im"                  "improve"
## [154] "including"           "incorporating"       "increase"
## [157] "info"                "information"         "instruction"
## [160] "interactionsaxure"   "interactionsphp"     "interested"
## [163] "interesting"         "interests"           "intersection"
## [166] "introduction"        "java"                "javascript"
## [169] "jobs"                "just"                "knitr"
## [172] "know"                "knowing"             "knowledge"
## [175] "lang"                "language"            "languages"
## [178] "large"               "larger"              "learn"
## [181] "learned"             "learning"            "legibility"
## [184] "legible"             "lets"                "level"
## [187] "libraries"           "license"             "like"
## [190] "linear"              "linguistic"          "link"
## [193] "listed"              "literary"            "little"
## [196] "longterm"            "lot"                 "luckily"
## [199] "mac"                 "machine"             "machines"
## [202] "make"                "making"              "manage"
## [205] "management"          "managing"            "manipulatable"
## [208] "manipulate"          "manipulating"        "many"
## [211] "mapping"             "maps"                "media"
## [214] "mediators"           "mediawiki"           "methods"
## [217] "mgmt"                "microsoft"           "ml"
## [220] "modeling"            "modelling"           "models"
## [223] "moderators"          "mostly"              "move"
## [226] "mplus"               "much"                "multilevel"
## [229] "multiple"            "multithreading"      "multivariate"
## [232] "mysql"               "names"               "natural"
## [235] "need"                "needed"              "needs"
## [238] "network"             "normalization"       "number"
## [241] "numpyscikitlearn"    "object"              "omeka"
## [244] "one"                 "ones"                "online"
## [247] "order"               "organization"        "orientated"
## [250] "outcomes"            "output"              "package"
## [253] "page"                "papers"              "particularly"
```

```
## [256] "patterns"           "people"             "php"
## [259] "places"             "plate"              "please"
## [262] "plone"              "plot"               "plots"
## [265] "point"              "pointed"            "police"
## [268] "possible"           "potentially"        "practices"
## [271] "pragmatically"      "presentation"       "primary"
## [274] "processing"         "program"            "programing"
## [277] "programmed"         "programming"        "programs"
## [280] "prose"              "provided"           "purpose"
## [283] "purposes"           "python"             "qual"
## [286] "qualitative"        "quantitative"       "queries"
## [289] "query"              "question"           "questions"
## [292] "quite"              "r"                  "rails"
## [295] "rcharts"            "rct"                "re"
## [298] "reader"             "refresher"          "regards"
## [301] "regression"         "relational"         "representations"
## [304] "reproducibility"    "required"           "research"
## [307] "researcher"         "researchers"        "results"
## [310] "revisit"            "ruby"               "running"
## [313] "said"               "sas"                "say"
## [316] "saying"             "scheme"             "scrape"
## [319] "scraping"           "scripting"          "search"
## [322] "see"                "series"             "shiny"
## [325] "shorterterm"        "significance"       "sites"
## [328] "skill"              "skills"             "social"
## [331] "software"           "something"          "spacetimerhyme"
## [334] "spatial"            "specialized"        "specific"
## [337] "spreading"          "spss"               "spssx"
## [340] "sql"                "start"              "stata"
## [343] "statistical"        "statistics"         "structure"
## [346] "structured"         "structures"         "student"
## [349] "sure"               "survey"             "surveys"
## [352] "syntax"             "sys"                "systems"
## [355] "table"              "tableau"            "tables"
## [358] "tabulating"         "take"               "techniques"
## [361] "teeny"              "tei"                "tests"
## [364] "text"               "textual"            "theme"
## [367] "theory"             "think"              "time"
## [370] "tool"               "tools"              "topic"
## [373] "transcripts"        "understand"         "understanding"
## [376] "update"             "use"                "used"
## [379] "useful"             "usemanipulation"    "user"
## [382] "using"              "variables"          "varying"
## [385] "video"              "visual"             "visualization"
## [388] "visualizations"     "visualizing"        "viz"
## [391] "want"               "web"                "webbased"
## [394] "websites"           "well"               "whatever"
## [397] "wikipedia"          "will"               "within"
## [400] "wordpress"          "work"               "workflow"
```

```
## [403] "working"             "works"               "xmltei"
## [406] "xslt"                "year"

# You will see that many of the tokens are quite obscure words and are
probably sparsely used in the data. The sparse terms can be removed using the
following code:
dtms <- removeSparseTerms(dtm, 0.98)
# The value of 0.98 was used here but you can experiment with different
values to see how many words remain (between 0 and 1; closer to 0 = less
words)

# You can see the remaining tokens with:
Terms(dtms)

##   [1] "able"          "advanced"      "also"          "analysis"
##   [5] "analyze"       "answer"        "basic"         "basics"
##   [9] "become"        "best"          "better"        "building"
##  [13] "coding"        "competent"     "concepts"      "currently"
##  [17] "data"          "database"      "databases"     "design"
##  [21] "different"     "digital"       "dont"          "eg"
##  [25] "enough"        "etc"           "excel"         "experience"
##  [29] "facility"      "familiar"      "gis"           "great"
##  [33] "id"            "interesting"   "know"          "knowing"
##  [37] "knowledge"     "language"      "learn"         "learning"
##  [41] "level"         "libraries"     "like"          "linear"
##  [45] "little"        "making"        "manage"        "management"
##  [49] "modeling"      "models"        "natural"       "need"
##  [53] "needs"         "object"        "order"         "patterns"
##  [57] "plot"          "practices"     "processing"    "program"
##  [61] "programming"   "programs"      "python"        "qualitative"
##  [65] "quantitative"  "question"      "r"             "regression"
##  [69] "research"      "researchers"   "sas"           "scripting"
##  [73] "search"        "sites"         "social"        "software"
##  [77] "something"     "spatial"       "specific"      "spss"
##  [81] "sql"           "stata"         "statistical"   "statistics"
##  [85] "survey"        "systems"       "time"          "tool"
##  [89] "tools"         "topic"         "understand"    "understanding"
##  [93] "use"           "used"          "useful"        "using"
##  [97] "visualization" "web"           "well"          "will"
## [101] "work"

# Now we can find the tokens that have a frequency of 5 or more:
findFreqTerms(dtms, 5)

##  [1] "able"          "also"          "analysis"      "analyze"
##  [5] "basic"         "better"        "data"          "databases"
##  [9] "design"        "gis"           "id"            "know"
## [13] "language"      "learn"         "like"          "management"
## [17] "programming"   "python"        "r"             "software"
## [21] "spss"          "stata"         "tools"         "use"
## [25] "using"         "visualization"
```

```r
# find tokens/words associated with "data":
findAssocs(dtms, 'data', corlimit = 0.2)
```

```
## $data
##         excel       digital      modeling visualization         learn
##          0.47          0.40          0.40          0.39          0.31
##       analyze      concepts            eg        manage      specific
##          0.30          0.27          0.27          0.27          0.27
##        survey         topic        useful          need        social
##          0.27          0.27          0.27          0.21          0.21
##       spatial         using
##          0.21          0.21
```

```r
# Try different tokens/words and changing the corlimit (the lower limit of
the correlation value between our word of interest and the rest of the words
in our data).

# create a vector of associated tokens/words for all frequency used
tokens/words:
findAssocs(dtms, findFreqTerms(dtms, 5), 0.25)
```

```
## $able
## advanced    level   little  spatial
##     0.52     0.52     0.52     0.31
##
## $also
##      models   competent interesting        will      become    building
##        0.58        0.45        0.45        0.45        0.35        0.35
##   currently   knowledge statistical
##        0.35        0.35        0.29
##
## $analysis
##       level statistical
##        0.34        0.28
##
## $analyze
##          eg     spatial interesting      linear      manage    modeling
##        0.49        0.38        0.29        0.29        0.29        0.29
##        plot       sites        well
##        0.29        0.29        0.29
##
## $basic
##     level  advanced   digital    little  modeling  programs something
##      0.62      0.29      0.29      0.29      0.29      0.29      0.29
##     topic
##      0.29
##
## $better
## understanding      facility     practices       program         topic
##          0.62          0.29          0.29          0.29          0.29
```

```
##          used
##          0.29
##
## $data
##    excel  digital modeling concepts       eg   manage specific   survey
##     0.47     0.40     0.40     0.27     0.27     0.27     0.27     0.27
##    topic   useful
##     0.27     0.27
##
## $databases
##    making  building    models   spatial competent      plot    search
##      0.62      0.49      0.38      0.38      0.29      0.29      0.29
##      well      will
##      0.29      0.29
##
## $design
##          web          etc      digital     modeling       object
##         0.49         0.41         0.29         0.29         0.29
##      program quantitative    something        topic
##         0.29         0.29         0.29         0.29
##
## $gis
##  advanced    little   digital     level  modeling something      time
##      0.62      0.62      0.29      0.29      0.29      0.29      0.29
##     topic
##      0.29
##
## $id
##       work understand     basics   familiar
##       0.45       0.34       0.28       0.26
##
## $know
##     answer       dont     enough   specific statistics        web
##       0.68       0.41       0.40       0.30       0.30       0.30
##
## $language
##    natural   building processing   database    digital   modeling
##       0.62       0.49       0.49       0.41       0.29       0.29
##     search  something      topic
##       0.29       0.29       0.29
##
## $learn
##  modeling    basics    linear currently
##      0.57      0.52      0.36      0.27
##
## $like
##      level  currently   research understand
##       0.36       0.25       0.25       0.25
##
## $management
```

```
##  database competent   digital  modeling  programs    search something
##      0.41      0.29      0.29      0.29      0.29      0.29      0.29
##     topic
##      0.29
##
## $programming
##  experience       order        need researchers
##        0.40        0.40        0.31        0.27
##
## $python
## experience   advanced  libraries     little    natural        sql
##       0.49       0.29       0.29       0.29       0.29       0.29
##       time
##       0.29
##
## $r
##   competent statistical    familiar      models      become
##        0.41        0.35        0.33        0.33        0.30
##
## $software
##   knowledge       great     knowing    patterns        plot    question
##        0.36        0.30        0.30        0.30        0.30        0.30
##        will    database statistical
##        0.30        0.29        0.29
##
## $spss
##  used excel
##  0.52  0.33
##
## $stata
##  currently experience
##        0.4        0.4
##
## $tools
##      plot scripting       web
##      0.38      0.29      0.29
##
## $use
##  spatial     plot  program       eg research
##     0.61     0.36     0.36     0.27     0.27
##
## $using
## currently        eg    basics    models
##      0.37      0.37      0.30      0.28
##
## $visualization
##   spatial   digital  modeling   natural      plot       sas something
##      0.38      0.29      0.29      0.29      0.29      0.29      0.29
##     topic      used
##      0.29      0.29
```