Task1 : Getting and Cleaning the Data

Jean-Luc BELLIER

23 avril 2017

General purpose

The goal of this task is to load a set of files and getting famililar with some cleaning functionalities.

Preliminaries : load libraries

```
library("clue")
```

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

```
library("tm")
```

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

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

```
## Warning: package 'NLP' was built under R version 3.3.2
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
script.dir <- "C:\\perso\\Culture generale\\Coursera\\Data_Science\\Cours_10_Data_Science_Capstone"
```

File load

The files come from a dataset of files provided by the Capstone Dataset. This dataset gathers 3 types of data : blogs, tweets and news, for 4 different languages : German, Finnish, English and Russian.

For this file, we only focus on the english dataset.

```
# First we get the folder of the current script
#script.dir <- dirname(sys.frame(1)$ofile)
#print(script.dir)
inputfiles.dir <- file.path(script.dir,"final","en_US")
print(inputfiles.dir)
```

```
## [1] "C:\\perso\\Culture generale\\Coursera\\Data_Science\\Cours_10_Data_Science_Capstone/final/en_US
```

```
displaySize <- 20

# Here we define a function to get the general properties of the file. The full path to the file is an

CreateCorpus <- function(filepath)
{
    # First we build a connection to th input file
    print(filepath)
    con = file(filepath,open="rb")

    fileRead <- readLines(filepath,skipNul = TRUE,encoding="UTF-16LE")
    close(con)

    # define a sample for the current dataset (1000 elements)
    #set.seed(1234)
    #rand_idx <- sample(length(fileRead),size=1000,replace=FALSE)
    #mysample <- fileRead[rand_idx]

    mysample <- fileRead

    vs <- VectorSource(mysample)
    Corpus( vs,readerControl = list(language="lat"))

}
```

We can now use this function to determine the parameters of the set of files.

```
#twitter_corpus = CreateCorpus(paste(inputfiles.dir,"en_US.twitter.txt",sep="/"))
```

Create functions to tokenize words, punctuation signs and digits

Here we will focus on the tweets, to get the tweets corresponding to input keywords. These functions will be used in the data cleaning of the different corpora.

```
getDigits <- function(string)
{
    trimws(gsub("\\D+"," ",string))
}

replacePunctuation <- function(string)
{
    trimws(gsub("[[:punct:]]"," ",string))
}

removeDigits <- function(string)
{
    gsub("\\s+"," ",gsub("\\d+"," ",string))
}

getWords <- function(string)
{
    strsplit(string," ")
}
```

```
removeNonASCII <- function(string)
{

    Encoding(string)="latin1"
    res = iconv(string,"latin1","ASCII",sub=" ")
    res
}

cleanStopWords <- function(string)
{
    # The standard reading gave some unexpected results (characters that looked like single quotes, but
# This function proposes an alternative way to remove the stopwords. It also removes the"'s" of possess

    WordsToRemove <- c(stopwords('english'),"'s")
    # Here we reorder the stopwords in order to get the words having quotes at the beginning. This will
    reorder.stoplist <- c(grep("[']", WordsToRemove, value = TRUE), WordsToRemove[!(1:length(WordsToRem

    # Creation of a vector containing all the words to remove after removal of single quotes
    stoplist2 <- unlist(lapply(gsub("'"," ",reorder.stoplist), function(x) strsplit(x," ")))
    # Note that the iput string has been already cleaned from all punctuation signs (quotes included)
    split1 <- strsplit(string," ")
    sapply(lapply(split1, function(x) x[!(x %in% stoplist2)]),paste0, collapse=" ")

}
cleanCorpus <- function(inputCorpus)
{
    # This function will clean the text defined in the input corpus
    # We apply different transformations
    # NB : the order of the transformations is important, because they may affect the data for the next

    WordstoRemove <- c(stopwords("en"),"'s")
    #WordstoRemove <- stopwords("en")
    reorder.stoplist <- c(grep("[']", WordstoRemove, value = TRUE), WordstoRemove[!(1:length(WordstoRem
    stoplist_quote <- grep("[']", WordstoRemove, value = TRUE)
    stoplist_noquote <- WordstoRemove[!(WordstoRemove %in% stoplist_quote)]

    myCorpus <- inputCorpus

    myCorpus <- tm_map(myCorpus,content_transformer(tolower))
    myCorpus <- tm_map(myCorpus, function(x) iconv(x,from='UTF-8', to="ASCII", sub=' '))
    myCorpus <- tm_map(myCorpus,stripWhitespace)
    myCorpus <- tm_map(myCorpus,replacePunctuation)
    myCorpus <- tm_map(myCorpus,cleanStopWords)
    myCorpus <- tm_map(myCorpus,stripWhitespace)
    myCorpus <- tm_map(myCorpus,removeNumbers)
    myCorpus <- tm_map(myCorpus,removeNonASCII)

    myCorpus
}

DisplayCorpusText <- function(inputCorpus, n)
{
```

```
    for (i in 1:n)
    {
       print(inputCorpus[[i]])

    }

}
```

Creation of the corpora

In this section, we will create corpora based on the input files (blogs, tweets and news) for english language.They will then be cleaned using the above functions. At last, the N first texts of the cleaned corpora will be displayed (N is an input parameter).

```
blogsCorpus <- CreateCorpus(paste(inputfiles.dir,"en_US.blogs.txt",sep="/"))
```

```
## [1] "C:\\perso\\Culture generale\\Coursera\\Data_Science\\Cours_10_Data_Science_Capstone/final/en_US/
```

```
show (blogsCorpus)
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 899288
```

```
newsCorpus <- CreateCorpus(paste(inputfiles.dir,"en_US.news.txt",sep="/"))
```

```
## [1] "C:\\perso\\Culture generale\\Coursera\\Data_Science\\Cours_10_Data_Science_Capstone/final/en_US/
```

```
## Warning in readLines(filepath, skipNul = TRUE, encoding = "UTF-16LE"):
## ligne finale incomplète trouvée dans 'C:\perso\Culture generale\Coursera
## \Data_Science\Cours_10_Data_Science_Capstone/final/en_US/en_US.news.txt'
```

```
show (newsCorpus)
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 77259
```

```
twitterCorpus <- CreateCorpus(paste(inputfiles.dir,"en_US.twitter.txt",sep="/"))
```

```
## [1] "C:\\perso\\Culture generale\\Coursera\\Data_Science\\Cours_10_Data_Science_Capstone/final/en_US/
```

```
show (twitterCorpus)
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 2360148
```

```
# Transform corpus elements

blogsCorpusClean <- cleanCorpus(blogsCorpus)
newsCorpusClean <- cleanCorpus(newsCorpus)
twitterCorpusClean <- cleanCorpus(twitterCorpus)

blogsList <- lapply(blogsCorpusClean,as.character)
twitterList <- lapply(twitterCorpusClean,as.character)
newsList <- lapply(newsCorpusClean,as.character)
```

```
DisplayCorpusText(blogsList,displaySize)
```

```
## [1] "years thereafter oil fields platforms named pagan gods"
## [1] "love mr brown"
## [1] "chad awesome kids holding fort work later usual kids busy together playing skylander xbox togetl
## [1] "anyways going share home decor inspiration storing folder puter amazing images stored away ready
## [1] "graduation season right around corner nancy whipped fun set help graduation cards gifts occasion
## [1] "alternative argument hear"
## [1] "bear"
## [1] "friends similar stories treated brusquely laurelwood staff often names keep coming half dozen fi
## [1] "although beloved cantab claim international recognition afforded station inn otherwise two joint
## [1] "peter schiff hard tell will look pretty bad americans prices will go way afford buy stuff also g
## [1] " winter time sleeps level time turning within make contact innermost winter best time result get
## [1] "en route cornwall   months slog sun sea always job needs tourist town bringing stuff back  frida
## [1] "pure large leaf assam waffling leaf thank want strong dark herby frills goodness sake fruit mixe
## [1] "also mention effect night sky issue raised report ignore lancing parish council use lighting inc
## [1] "state contracts worth  million ringgit"
## [1] "people like unknowing transformers things protected fairy tale love"
## [1] "one thing astounding though support marshals phenomenal rain long remaining cheery supportive ar
## [1] "attend friend kendra wedding joy watch people see end year waiting"
## [1] "neighbor recommended chasing fireflies charles martin never read anything martin although neighl
## [1] "fallout ellen hopkins p  "
```

```
DisplayCorpusText(twitterList,displaySize)
```

```
## [1] "btw thanks rt gonna dc anytime soon love see way way long"
## [1] "meet someone special know heart will beat rapidly smile reason"
## [1] "decided fun"
## [1] "tired played lazer tag ran lot ughh going sleep like  minutes"
## [1] "words complete stranger made birthday even better"
## [1] "first cubs game ever wrigley field gorgeous perfect go cubs go"
## [1] "get another day skool due wonderful snow wakes damn thing"
## [1] "coo jus work hella tired r u ever cali"
## [1] "new sundrop commercial hehe love first sight"
## [1] "need reconnect week"
## [1] "always wonder guys auctions shows learned talk fast hear djsosnekspqnslanskam"
## [1] "dammnnnnn catch"
## [1] "great picture green shirt totally brings eyes"
## [1] "desk put together room set oh boy oh boy"
## [1] ""
## [1] "beauty brainstorming alchemy office sally walker"
## [1] "looking new band blog month anyone interested"
## [1] "packing quick move street movers"
## [1] "ford focus hatchback"
## [1] "rt according national retail federation   billion spent mothersday last year"
```

```
DisplayCorpusText(newsList,displaySize)
```

```
## [1] "home alone apparently"
## [1] "st louis plant close die old age workers making cars since onset mass automotive production s"
## [1] "wsu plans quickly became hot topic local online sites though people applauded plans new biomedic
## [1] "alaimo group mount holly contract last fall evaluate suggest improvements trenton water works ca
## [1] "often difficult predict law impact legislators think twice carrying bill absolutely necessary is
## [1] "certain amount scoffing going around years ago nfl decided move draft weekend prime time eventua
## [1] " charlevoix detroit"
```

```
## [1] "just another long line failed attempts subsidize atlantic city said americans prosperity new je
## [1] "time report sullivan called cps correct problems improve employee accountability saying example
## [1] "just trying hit hard someplace said rizzo hit pitch opposite field left center just trying make
## [1] "mhta president ceo margaret anderson kelliher said construction likely begin soon suite offices
## [1] "absurdity attempting bottle news magnitude apparent later write"
## [1] "gm labor relations vice president diana tremblay said deal will enable gm fully competitive eli
## [1] "wandry matters current system imposes gift tax  taxpayers give assets away exceptions individua
## [1] "cheap said hit hard enough really feel batum rose feet squared elbows performed best elbow throu
## [1] "andrade children erin son patrick adults said intrigued opportunity play mentoring role life hi
## [1] "hair looks better"
## [1] "born april   pozzuoli naples italy resided north plainfield since marriage "
## [1] "house minority leader nancy pelosi top democrats called weiner resignation"
## [1] "first love self second love money two greek words philautos philargyros succinctly express conc
```

Interpretation of the results

We can notice, from the displayed data, that the cleaning functions are efficient : all the stopwords, digits and numbers have been removed. We can notice some noise, e.g. abbreviated words (like 'td' for 'towards'). This may affect the efficiency of the text analysis.