Text Exploration

Types of corpus

- **VCorpus**: Volatile Corpus, stored in memory.
- **PCorpus**: Persistent Corpus, stored on disk.

Predefined Sources of in package

- **VectorSource**: Vector of text.
- **DataframeSource**: Data frame of text.
- **DirSource**: Directory of text files

Part 1: Extract Text from Files

```
eg1 <- read.table("dataset/GC.txt", fill=T,header=F) #Data CG.txt
eg1[1,]
```

```
##         V1   V2  V3    V4   V5 V6         V7 V8 V9 V10 V11 V12 V13 V14 V15 V16
## 1 Gertrude Cox, The First Lady Of Statistics
##   V17 V18 V19 V20 V21 V22 V23 V24 V25
## 1
```

```
eg2 <- read.csv("dataset/GC.csv",header=F) #Data CG.csv
eg2[1,]
```

```
## [1] "Gertrude Cox, The First Lady Of Statistics"
```

Using tm Package

```
library(tm)
```

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

```
eg3 <- c("Hi!","Welcome to STQD6114","Tuesday, 11-1pm")
mytext <-VectorSource(eg3)
mycorpus <- VCorpus(mytext)
#inspect(mycorpus)
as.character(mycorpus[[1]])
```

```
## [1] "Hi!"
```

Example using VectorSource

```
eg4<-t(eg1) #From example 1
a<-sapply(1:7,function(x) trimws(paste(eg4[,x],collapse=" "),"right"))
mytext<-VectorSource(a)
mycorpus<-VCorpus(mytext)
#inspect(mycorpus)
as.character(mycorpus[[1]])
```

```
## [1] "Gertrude Cox, The First Lady Of Statistics"
```

Example using DataFrameSource

```
eg5<-read.csv("dataset/doc6.csv",header=F) #Using doc6.csv
docs<-data.frame(doc_id=c("doc_1","doc_2"),
                 text=c(as.character(eg5[1,]),as.character(eg5[2,])),
                 dmeta1=1:2,dmeta2=letters[1:2],stringsAsFactors=F)
mytext<-DataframeSource(docs)
mycorpus<-VCorpus(mytext)
#inspect(mycorpus)
```

Example using DirSource

```
mytext<-DirSource("dataset/movies")
mycorpus<-VCorpus(mytext)
#inspect(mycorpus)
as.character(mycorpus[[1]])
```

```
## [1] "Spider-Man: Homecoming"
## [2] ""
## [3] "Thrilled by his experience with the Avengers, young Peter Parker returns home to live with his
```

Part 2: Web scrapping

```
eg6<-readLines("https://en.wikipedia.org/wiki/Data_science")
```

```
## Warning in readLines("https://en.wikipedia.org/wiki/Data_science"): incomplete
## final line found on 'https://en.wikipedia.org/wiki/Data_science'
```

```
#eg6[grep("\\h2",eg6)]
#eg6[grep("\\p",eg6)] #paragraph
```

Using library XML

```
library(XML)
doc<-htmlParse(eg6)
doc.text<-unlist(xpathApply(doc,'//p',xmlValue))
unlist(xpathApply(doc,'//h2',xmlValue))
```

```
## [1] "Contents"
## [2] "Foundations"
## [3] "Etymology"
## [4] "Data science and data analysis"
## [5] "Cloud computing for data science"
## [6] "Ethical consideration in data science"
## [7] "See also"
## [8] "References"
```

Using library httr

```
library(httr)
```

```
##
## Attaching package: 'httr'

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

```
eg7<-GET("https://www.edureka.co/blog/what-is-data-science/")
doc<-htmlParse(eg7)
doc.text<-unlist(xpathApply(doc,'//p',xmlValue))
```

Using library rvest

```
library(rvest)
eg8<-read_html("https://www.edureka.co/blog/what-is-data-science/")
nodes<-html_nodes(eg8,'.col-lg-9 :nth-child(1)')
texts<-html_text(nodes)
```

Selecting multiple pages

```
pages <- paste0("--url--&page=",0:9)
```

```
pages <- paste0("https://www.amazon.com/s?k=skincare&ref=nav_bb_sb&page=",0:9)
#pages<-paste0('https://www.amazon.co.jp/s?k=skincare&crid=28HIW1TYLV9UM&sprefix#=skincare%2Caps%2C268&
eg10<-read_html(pages[1])
nodes<-html_nodes(eg10,'.a-price-whole')
texts<-html_text(nodes)
texts
```

```
##  [1] "17."  "84."  "195." "7."   "17."  "8."   "7."   "5."   "8."   "9."
## [11] "14."  "9."   "5."   "3."   "17."  "15."  "29."  "18."  "9."   "14."
## [21] "4."   "13."  "0."   "14."  "4."   "9."   "8."   "49."  "8."   "13."
## [31] "8."   "2."   "7."   "11."  "14."  "89."  "18."  "79."  "39."  "6."
## [41] "8."   "5."   "9."   "8."   "4."   "16."  "202." "13."  "16."  "6."
## [51] "24."  "8."   "42."  "19."  "17."  "195." "32."  "9."   "8."
```

```
Price <- function(page) {
  url <- read_html(page)
  nodes <- html_nodes(url, '.a-price-whole')
  html_text(nodes)
}

sapply(pages,Price)
```

```
## $`https://www.amazon.com/s?k=skincare&ref=nav_bb_sb&page=0`
##  [1] "17."  "84."  "195." "7."   "9."   "31."  "9."   "8."   "9."   "3."
## [11] "9."   "14."  "9."   "5."   "17."  "15."  "9."   "7."   "29."  "12."
## [21] "13."  "4."   "14."  "9."   "0."   "8."   "49."  "13."  "8."   "7."
## [31] "4."   "8."   "11."  "2."   "6."   "39."  "8."   "5."   "14."  "8."
## [41] "79."  "16."  "18."  "13."  "109." "9."   "8."   "202." "4."   "16."
## [51] "89."  "24."  "6."   "84."  "195." "17."  "9."   "32."  "9."
##
## $`https://www.amazon.com/s?k=skincare&ref=nav_bb_sb&page=1`
##  [1] "17."  "84."  "9."   "7."   "9."   "9."   "31."  "9."   "9."   "19."
## [11] "17."  "14."  "9."   "5."   "3."   "4."   "7."   "12."  "29."  "195."
## [21] "49."  "15."  "9."   "4."   "0."   "39."  "8."   "13."  "14."  "14."
## [31] "8."   "8."   "2."   "202." "13."  "11."  "8."   "79."  "89."  "18."
## [41] "8."   "84."  "7."   "5."   "6."   "9."   "13."  "69."  "16."  "42."
## [51] "24."  "23."  "8."   "35."  "149." "24."  "8."   "8."   "71."
##
## $`https://www.amazon.com/s?k=skincare&ref=nav_bb_sb&page=2`
##  [1] "14."  "17."  "9."   "39."  "0."   "35."  "42."  "29."  "17."  "109."
## [11] "89."  "149." "19."  "23."  "70."  "29."  "195." "495." "109." "69."
## [21] "32."  "43."  "4."   "7."   "89."  "299." "19."  "7."   "529." "25."
## [31] "29."  "29."  "129." "17."  "0."   "47."  "27."  "0."   "6."   "0."
## [41] "21."  "0."   "49."  "4."   "0."   "124." "9."   "16."  "0."   "24."
## [51] "0."   "39."  "36."  "124." "5."   "29."  "0."   "0."   "24."  "0."
## [61] "31."  "9."   "9."   "69."  "16."  "32."  "7."   "99."  "19."  "13."
##
## $`https://www.amazon.com/s?k=skincare&ref=nav_bb_sb&page=3`
##  [1] "14."  "9."   "17."  "19."  "0."   "5."   "24."  "17."  "9."   "32."
## [11] "19."  "29."  "18."  "20."  "0."   "32."  "29."  "29."  "7."   "35."
## [21] "0."   "24."  "195." "18."  "8."   "8."   "18."  "49."  "13."  "29."
## [31] "0."   "7."   "28."  "26."  "15."  "0."   "5."   "0."   "5."   "0."
## [41] "25."  "0."   "49."  "14."  "0."   "3."   "19."  "20."  "16."  "134."
## [51] "50."  "0."   "0."   "0."
##
## $`https://www.amazon.com/s?k=skincare&ref=nav_bb_sb&page=4`
##  [1] "14."  "29."  "19."  "9."   "53."  "6."   "52."  "32."  "17."  "19."
## [11] "299." "9."   "96."  "8."   "51."  "195." "18."  "9."   "31."  "5."
## [21] "14."  "32."  "15."  "21."  "13."  "14."  "12."  "74."  "18."  "19."
## [31] "6."   "32."  "9."   "5."   "19."  "5."   "13."  "35."  "1."   "11."
## [41] "38."  "24."  "45."  "89."  "55."  "2."   "7."   "32."  "19."  "25."
```

```
## [51] "7."   "25."
##
## $`https://www.amazon.com/s?k=skincare&ref=nav_bb_sb&page=5`
##  [1] "14."  "9."   "195." "9."   "7."   "7."   "15."  "96."  "17."  "32."
## [11] "9."   "69."  "12."  "7."   "9."   "17."  "29."  "29."  "32."  "21."
## [21] "13."  "12."  "20."  "18."  "19."  "5."   "15."  "20."  "20."  "30."
## [31] "7."   "42."  "12."  "14."  "29."  "7."   "9."   "3."   "12."  "9."
## [41] "9."   "12."  "21."  "19."  "23."  "49."  "49."  "12."  "12."  "9."
## [51] "12."  "24."  "59."  "23."  "23."  "32."  "9."   "24."  "19."  "7."
##
## $`https://www.amazon.com/s?k=skincare&ref=nav_bb_sb&page=6`
##  [1] "14."  "122." "9."   "75."  "23."  "15."  "12."  "12."  "195." "17."
## [11] "29."  "32."  "23."  "20."  "38."  "16."  "18."  "29."  "23."  "19."
## [21] "31."  "23."  "23."  "23."  "79."  "84."  "14."  "13."  "69."  "8."
## [31] "29."  "31."  "29."  "6."   "55."  "35."  "39."  "5."   "130." "50."
## [41] "6."   "134." "124." "7."   "49."  "42."  "176." "2."   "42."  "49."
## [51] "51."  "4."   "33."  "32."  "16."  "7."   "19."  "45."
##
## $`https://www.amazon.com/s?k=skincare&ref=nav_bb_sb&page=7`
##  [1] "99."  "74."  "14."  "109." "5."   "130." "99."  "74."  "250." "130."
## [11] "17."  "50."  "6."   "130." "9."   "195." "85."  "134." "124." "7."
## [21] "49."  "17."  "250." "89."  "29."  "42."  "176." "42."  "49."  "4."
## [31] "99."  "32."  "45."  "24."  "7."
##
## $`https://www.amazon.com/s?k=skincare&ref=nav_bb_sb&page=8`
##  [1] "14."  "9."   "195." "29."  "15."  "29."  "18."  "5."   "19."  "19."
## [11] "32."  "19."  "31."  "32."  "17."  "19."  "9."   "32."  "31."  "19."
## [21] "14."  "19."
##
## $`https://www.amazon.com/s?k=skincare&ref=nav_bb_sb&page=9`
##  [1] "31." "17." "9."  "32." "19." "32." "19." "31." "14." "19."
```

```
do.call("c", lapply(pages,Price))
```

```
##   [1] "17."  "79."  "84."  "9."   "31."  "9."   "9."   "9."   "19."  "9."
##  [11] "14."  "195." "5."   "17."  "3."   "4."   "9."   "5."   "7."   "7."
##  [21] "15."  "9."   "13."  "49."  "14."  "8."   "7."   "8."   "0."   "13."
##  [31] "39."  "11."  "14."  "202." "79."  "8."   "18."  "5."   "89."  "6."
##  [41] "8."   "2."   "4."   "8."   "84."  "6."   "42."  "16."  "8."   "19."
##  [51] "9."   "69."  "24."  "70."  "195." "17."  "9."   "19."  "32."  "84."
##  [61] "12."  "19."  "9."   "9."   "31."  "9."   "9."   "19."  "3."   "9."
##  [71] "14."  "17."  "65."  "28."  "17."  "15."  "25."  "9."   "12."  "9."
##  [81] "13."  "4."   "14."  "9."   "0."   "8."   "49."  "13."  "8."   "7."
##  [91] "4."   "8."   "11."  "2."   "6."   "39."  "8."   "5."   "14."  "8."
## [101] "79."  "16."  "18."  "13."  "109." "9."   "8."   "202." "4."   "16."
## [111] "89."  "24."  "6."   "84."  "24."  "8."   "8."   "71."  "6."   "14."
## [121] "17."  "9."   "195." "35."  "109." "6."   "17."  "29."  "9."   "35."
## [131] "32."  "0."   "7."   "4."   "29."  "39."  "495." "299." "70."  "25."
## [141] "16."  "43."  "8."   "109." "19."  "7."   "29."  "32."  "529." "0."
## [151] "129." "27."  "17."  "4."   "0."   "9."   "29."  "21."  "0."   "0."
## [161] "124." "0."   "16."  "47."  "0."   "0."   "0."   "124." "49."  "0."
## [171] "6."   "0."   "36."  "4."   "22."  "16."  "9."   "39."  "24."  "9."
## [181] "69."  "22."  "29."  "39."  "32."  "16."  "99."  "19."  "7."   "14."
## [191] "17."  "9."   "195." "69."  "27."  "18."  "5."   "0."   "19."  "18."
```

```
## [201] "32."  "29."  "20."  "0."   "35."  "24."  "7."   "29."  "31."  "19."
## [211] "0."   "18."  "13."  "49."  "29."  "0."   "0."   "7."   "15."  "28."
## [221] "0."   "26."  "5."   "5."   "0."   "25."  "0."   "0."   "49."  "0."
## [231] "14."  "15."  "3."   "19."  "15."  "134." "50."  "32."  "19."  "25."
## [241] "7."   "6."   "14."  "195." "9."   "249." "53."  "6."   "52."  "32."
## [251] "17."  "35."  "79."  "9."   "8."   "96."  "5."   "28."  "29."  "29."
## [261] "17."  "14."  "51."  "5."   "32."  "19."  "79."  "14."  "12."  "74."
## [271] "6."   "18."  "19."  "32."  "9."   "5."   "19."  "5."   "13."  "35."
## [281] "1."   "11."  "38."  "24."  "45."  "89."  "55."  "2."   "7."   "32."
## [291] "16."  "19."  "7."   "26."  "14."  "29."  "9."   "19."  "7."   "7."
## [301] "15."  "32."  "17."  "19."  "299." "9."   "12."  "7."   "9."   "18."
## [311] "32."  "195." "299." "21."  "20."  "20."  "20."  "13."  "3."   "7."
## [321] "8."   "13."  "12."  "30."  "42."  "7."   "14."  "12."  "9."   "3."
## [331] "7."   "29."  "12."  "9."   "9."   "12."  "21."  "49."  "19."  "49."
## [341] "23."  "12."  "12."  "9."   "12."  "24."  "59."  "23."  "23."  "14."
## [351] "29."  "9."   "19."  "23."  "15."  "32."  "19."  "9."   "17."  "299."
## [361] "12."  "12."  "18."  "32."  "195." "19."  "23."  "20."  "38."  "16."
## [371] "13."  "8."   "3."   "7."   "31."  "23."  "23."  "23."  "79."  "84."
## [381] "14."  "13."  "69."  "8."   "29."  "31."  "29."  "6."   "55."  "35."
## [391] "39."  "5."   "130." "50."  "6."   "134." "124." "7."   "49."  "42."
## [401] "176." "2."   "42."  "49."  "51."  "4."   "33."  "32."  "24."  "19."
## [411] "7."   "25."  "99."  "74."  "14."  "9."   "39."  "5."   "99."  "74."
## [421] "250." "130." "17."  "130." "29."  "50."  "130." "85."  "195." "89."
## [431] "6."   "134." "124." "7."   "250." "32."  "29."  "17."  "49."  "42."
## [441] "176." "42."  "49."  "4."   "99."  "32."  "45."  "7."   "24."  "14."
## [451] "9."   "195." "29."  "29."  "18."  "5."   "19."  "32."  "19."  "7."
## [461] "31."  "32."  "31."  "19."  "14."  "7."   "14."  "9."   "195." "29."
## [471] "29."  "5."   "7."   "19."  "19."  "18."  "7."   "32."  "31."  "32."
## [481] "9."   "17."  "19."
```

```
pricelist <- do.call("c", lapply(pages,Price))
pricelist <- as.numeric(pricelist)
mean(pricelist)
```

```
## [1] 34.26059
```