Data-Science-Capstone

Carlos Araya

14 de mayo de 2017

Data Science Capstone

This report is about the capstone on data science specialization. And it corresponds to the first view.

The Data

The data is in https://d396qusza40orc.cloudfront.net/dsscapstone/dataset/Coursera-SwiftKey.zip

Its content are three files:

- en_US.blogs.txt
- en_US.news.txt
- en_US.twitter.txt

Load the Data:

We used the command.

```
blogs <- data_frame(text = read_lines("E:/CARAYA/II - GitRepos/en_US/en_US.blogs.txt"))
news <- data_frame(text = read_lines("E:/CARAYA/II - GitRepos/en_US/en_US.news.txt"))
twitter <- data_frame(text = read_lines("E:/CARAYA/II - GitRepos/en_US/en_US.twitter.txt"))

blogs.len <- nchar(blogs$text)
news.len <- nchar(news$text)
twitter.len <- nchar(twitter$text)
```

To load the data and get file's length.

File	N lines	Min	Median	Mean	Max
en_US.blogs.txt	899,288	1	156	229.98	40,833
en_US.news.txt	1,010,242	1	185	201.16	11,384
en_US.twitter.txt	2,360,148	2	64	68.68	140

Or plot the length of size of the lines.