Capstone - Sampling the data

Kristian Gårdhus Wichmann

10 feb 2017

Handling large datasets - sampling

The three data files are huge. Holding all of them in active memory at the same time will be different for most computers. To handle this, a random sample of 5% is chosen from each.

Blogs

Load the blogs data:

```
blogs <- readLines("en_US/en_US.blogs.txt")
```

Take sample:

```
samplesize <- floor(length(blogs) * 0.05)
blogs_sample <- sample(blogs, samplesize)
```

To free memory, delete the blogs:

```
rm(blogs)
```

News

Load the blogs data:

```
news <- readLines("en_US/en_US.news.txt")
```

```
## Warning in readLines("en_US/en_US.news.txt"): ufuldstændig endelig linje
## fundet på 'en_US/en_US.news.txt'
```

The warning simply means that the file has no line break at the end. Take sample:

```
samplesize <- floor(length(news) * 0.05)
news_sample <- sample(news, samplesize)
```

Delete news:

```
rm(news)
```

Tweets

Load the tweets data:

```
tweets <- readLines("en_US/en_US.twitter.txt")
```

```
## Warning in readLines("en_US/en_US.twitter.txt"): linje 167155 ser ud til at
## indeholde en indlejret nul
```

```
## Warning in readLines("en_US/en_US.twitter.txt"): linje 268547 ser ud til at
## indeholde en indlejret nul
```

```
## Warning in readLines("en_US/en_US.twitter.txt"): linje 1274086 ser ud til
## at indeholde en indlejret nul
```

```
## Warning in readLines("en_US/en_US.twitter.txt"): linje 1759032 ser ud til
## at indeholde en indlejret nul
```

Inspect the lines mentioned in warnings:

```
tweets[c(167155, 268547, 1274086, 1759032)]
```

```
## [1] "Outlaw Deluxe Demo rel. with TED RUSSELL KAMP, after \"Honky Tonk Happy Hour\" with Matt Monta
## [2] "You're welcome, Jason. TY for the retweets. "
## [3] "We're building our o"
## [4] "There are 2 authorities when it comes to high security locks: "
```

There's nothing appearantly suspicious about these. Now a sample is chosen:

```
set.seed(42)
samplesize <- floor(length(tweets) * 0.05)
tweets_sample <- sample(tweets, samplesize)
```

Delete tweets:

```
rm(tweets)
```

Save the data

Finally, the three data sets are saved:

```
write(blogs_sample, "blogs_sample")
write(news_sample, "news_sample")
write(tweets_sample, "tweets_sample")
```