Capstone Project - Exploratory Data Analysis

Pedro A. Alonso Baigorri

4 de julio de 2018

Objective

The objective of this project is to apply data science in the area of natural language processing to create a prototype of a Text Predictor application similar to what the mobile text editors are using when they suggest some text to introduce based on previous words.

To create this application we are going to use a dataset ("Corpora") that includes texts collected from twitter, blogs and news data sources.

In this document I will report the result of the exploratory data analysis of the dataset and the planned strategies to develop the text predictor.

Opening and cleaning the dataset

As I mentioned before the dataset is composed by 3 different text files:

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

I have built some R code to read these files and load them into R objects in a efficent way, saving the objects locally so, the next time I need to load the objects I can avoid to download and read the text files again.

The number of lines of each file can be obtained from:

```r
library(ngram)
wordcount(dataset$twitter)
```

```
## [1] 30373543
```

```r
wordcount(dataset$blogs)
```

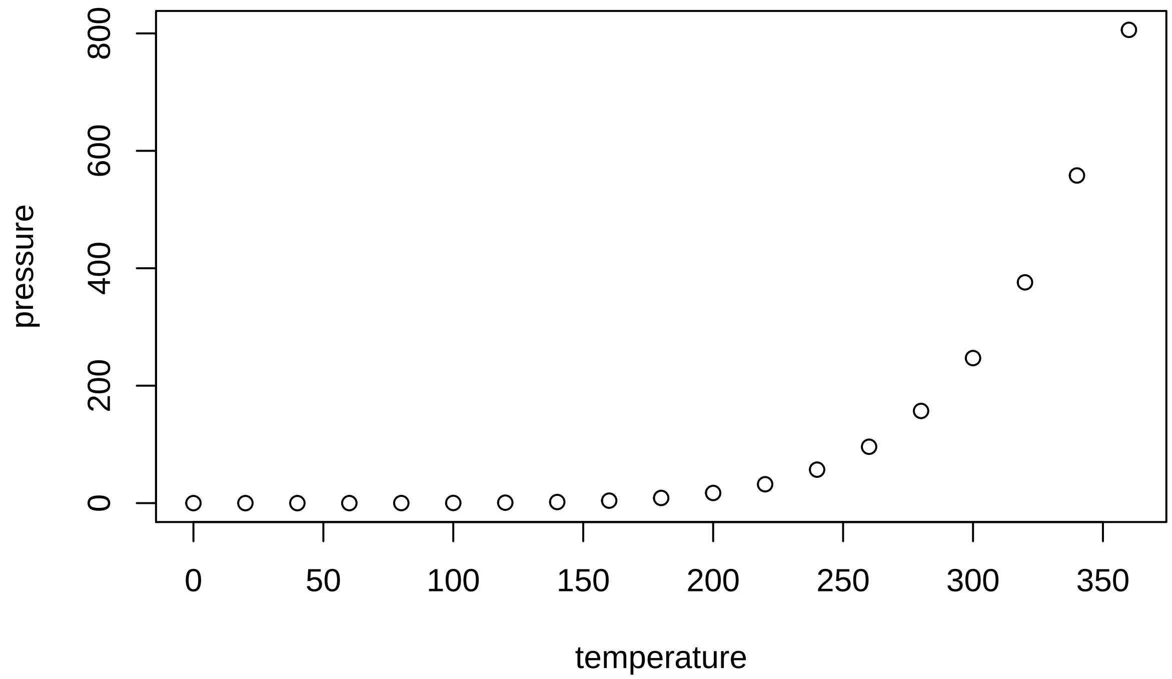
```
## [1] 37334131
```

```r
wordcount(dataset$news)
```

```
## [1] 2643969
```

Dataset Basic Summary

You can also embed plots, for example:



Plots and Features of the dataset

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Strategies to build the Text Predictor