Data Science Project: Word Prediction

David Pham

07.10.2014

Introduction

This is the reproducible report of the data science cap stone project on 2014. The code is split into three main parts:

1. Data preprocessing, with installation of the required packages.
2. Modelling part, where all the transformation and adjustement are made in order to compute the probabilities.
3. Prediction part, where we try to predict the following word of a user input.

Data visualisation (with ggplot2) is left in annexe.

Data preprocessing

- Punctation;
- Number and all.

Modelling part

- Adjustement probability through bla bla smoothing;
- SessionInfo()
- Heavy rely on parallel computing. Kill cluster.
- Data.table are used extensively.

Prediction

Basically, the function cleans the string argument in order to have a common structure as the previous results. Then it splits the strings by spaces and find the last two words (or the last two word with a skip words in between). Then the function returns the 20 most probables following words as a table.

It works in 99% of the case as long as the user does not want to game the function.

Possible improvements

In the following, some possible ideas worth following are described.

- Levenstein distance could be used to predict unfinished words
- Faster implementation in c++ or java
- Use of maybe 4-gram and also skip 2 words.

- Better assesment on how precise the prediction is.

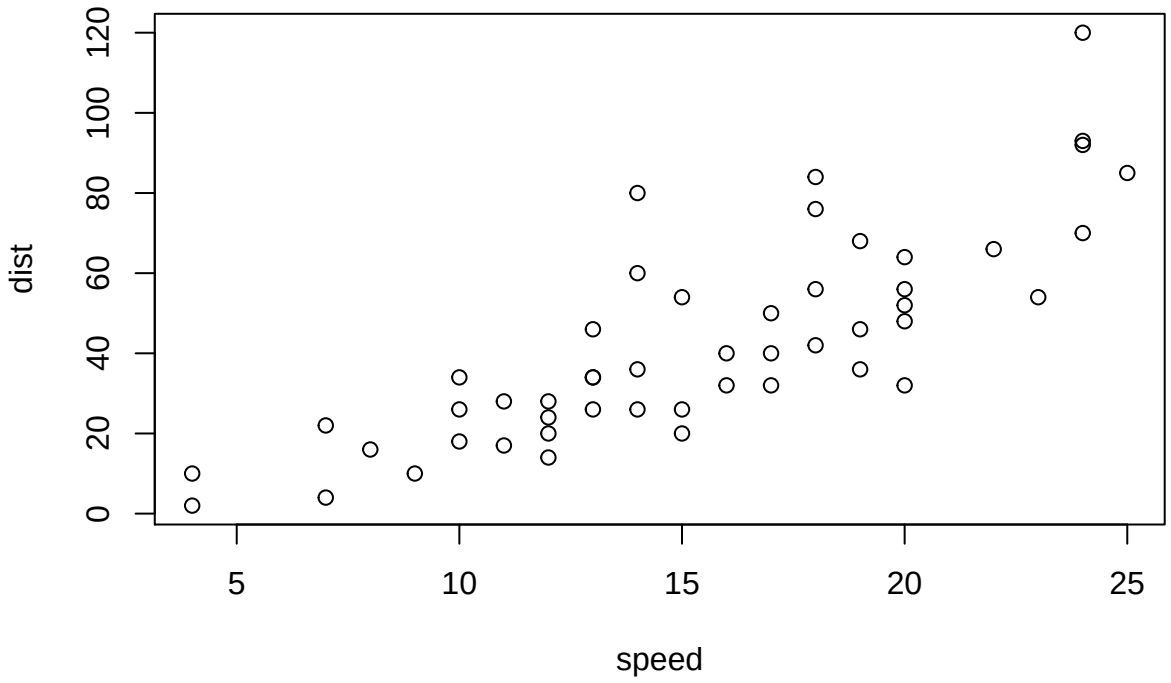
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2
##  1st Qu.:12.0   1st Qu.: 26
##  Median :15.0   Median : 36
##  Mean   :15.4   Mean   : 43
##  3rd Qu.:19.0   3rd Qu.: 56
##  Max.   :25.0   Max.   :120
```

You can also embed plots, for example:



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

Haha