Exploration of Wikipedia Data

Tobias Machnitzki and Finn Burgemeister

24 November 2017

```
library(knitr)
library(markdown)
library(scales)
library(ggplot2)
```

1. Read Dataset

```
fin = file("../../WORK/Blatt5/enwiki-clean-10MiB.csv", "r")

i=1

# Vectors for analysis
lengthArticle = NULL
meanLengthSentence = NULL
minLengthSentence = NULL
maxLengthSentence = NULL



while(TRUE){
  ##########################
  # Read file line per line
  line = readLines(fin, n = 1)

  if(length(line) == 0){
    break
  }


  ##########################
  # Process data
  data = read.csv(con <- textConnection(line), header=FALSE)

  oneID = data[[1]]
  oneAdress = data[[2]]
  oneTitle = data[[3]]
  oneArticle = data[[4]]
  oneCategories = data[[5]]


  ##########################
  # Break for testing
  #if(i == 6){break}


  ##########################
  # Exploration of one article
  lengthArticle[i] = sapply(gregexpr("\\w+", oneArticle), length)

  numberSentences = sapply(gregexpr('[[:alnum:] ][.!?]', oneArticle), length)

  Sentences = strsplit(toString(oneArticle), split="[\\.!?]+")
  lengthSentences = lapply(gregexpr("\\w+", Sentences[[1]]), length)
  meanLengthSentence[i] = mean(unlist(lengthSentences))
  minLengthSentence[i] = min(unlist(lengthSentences))
  maxLengthSentence[i] = max(unlist(lengthSentences))
```

```
  ##########################
  # End of while, do not modify
  i = i+1
}

close(fin)


dat = data.frame(lengthArticle, meanLengthSentence)
print("Correlation")
```

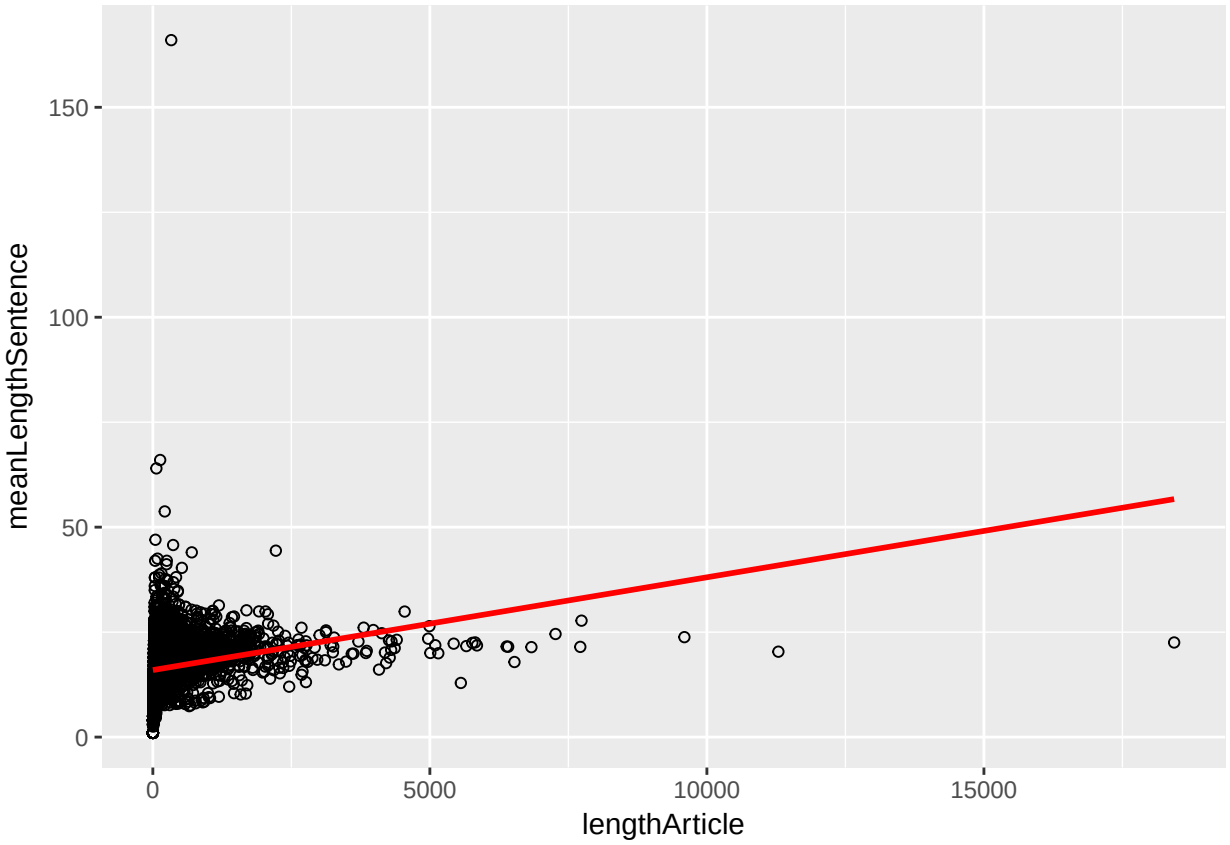
```
## [1] "Correlation"
```

```
print(cor(dat))
```

```
##                    lengthArticle meanLengthSentence
## lengthArticle            1.00000            0.24312
## meanLengthSentence       0.24312            1.00000
```

2 Plot Correlation of article length with the mean length of sentences within the article

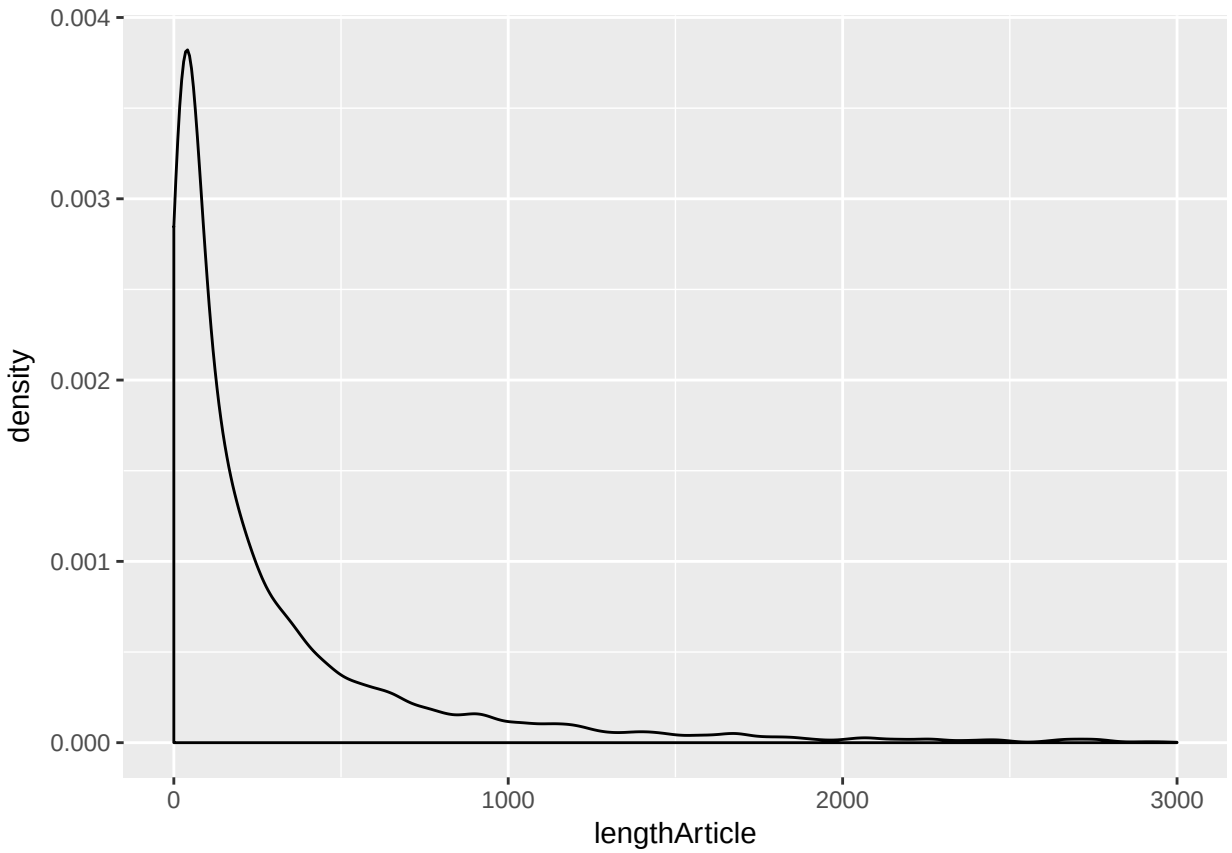
```
#########################
# Scatter Plot with regression
ggplot(dat, aes(lengthArticle, meanLengthSentence)) +
  geom_point(shape=1) +
  geom_smooth(method=lm , color="red", se=FALSE)  # Add linear regression line
```



3 Density Plot

```
#########################
# Density Plot
ggplot(dat, aes(lengthArticle)) +
  geom_density(alpha=0.55) +
  xlim(0, 3000)
```

```
## Warning: Removed 49 rows containing non-finite values (stat_density).
```



4 . . . time reason

“‘