Stock Prediction using Headline

Techtive

5 March 2018

Headlines data

Headlines and categories for 400k news items scraped from the web in 2014. Columns are:

- ID : the numeric ID of the article
- TITLE : the headline of the article
- URL : the URL of the article
- PUBLISHER : the publisher of the article
- CATEGORY : the category of the news item; one of: – b : business – t : science and technology – e : entertainment – m : health
- STORY : alphanumeric ID of the news story that the article discusses
- HOSTNAME : hostname where the article was posted
- TIMESTAMP : approximate timestamp of the article's publication, given in Unix time (seconds since midnight on Jan 1, 1970)

```
# remove all the variables in the environment
rm(list=ls())

data<-read.csv("uci-news-aggregator.csv",fill=T, sep=",", stringsAsFactors = FALSE)
dim(data)
```

```
## [1] 422419      8
```

```
names(data)
```

```
## [1] "ID"        "TITLE"     "URL"       "PUBLISHER" "CATEGORY"  "STORY"
## [7] "HOSTNAME"  "TIMESTAMP"
```

```
attach(data)
#
head(ID,5)
```

```
## [1] 1 2 3 4 5
```

```
head(TITLE,5)
```

```
## [1] "Fed official says weak data caused by weather, should not slow taper"
## [2] "Fed's Charles Plosser sees high bar for change in pace of tapering"
## [3] "US open: Stocks fall after Fed official hints at accelerated tapering"
## [4] "Fed risks falling 'behind the curve', Charles Plosser says"
## [5] "Fed's Plosser: Nasty Weather Has Curbed Job Growth"
```

```
head(URL,5)
```

```
## [1] "http://www.latimes.com/business/money/la-fi-mo-federal-reserve-plosser-stimulus-economy-2014031
## [2] "http://www.livemint.com/Politics/H2EvwJSK2VE6OF7iK1g3PP/Feds-Charles-Plosser-sees-high-bar-for-
## [3] "http://www.ifamagazine.com/news/us-open-stocks-fall-after-fed-official-hints-at-accelerated-tap
## [4] "http://www.ifamagazine.com/news/fed-risks-falling-behind-the-curve-charles-plosser-says-294430"
## [5] "http://www.moneynews.com/Economy/federal-reserve-charles-plosser-weather-job-growth/2014/03/10/
```

```r
head(PUBLISHER,5)
```

```
## [1] "Los Angeles Times" "Livemint"          "IFA Magazine"
## [4] "IFA Magazine"      "Moneynews"
```

```r
head(CATEGORY,5)
```

```
## [1] "b" "b" "b" "b" "b"
```

```r
head(STORY,5)
```

```
## [1] "ddUyU0VZz0BRneMioxUPQVP6sIxvM" "ddUyU0VZz0BRneMioxUPQVP6sIxvM"
## [3] "ddUyU0VZz0BRneMioxUPQVP6sIxvM" "ddUyU0VZz0BRneMioxUPQVP6sIxvM"
## [5] "ddUyU0VZz0BRneMioxUPQVP6sIxvM"
```

```r
head(HOSTNAME,5)
```

```
## [1] "www.latimes.com"     "www.livemint.com"    "www.ifamagazine.com"
## [4] "www.ifamagazine.com" "www.moneynews.com"
```

```r
head(TIMESTAMP,5)
```

```
## [1] 1.39447e+12 1.39447e+12 1.39447e+12 1.39447e+12 1.39447e+12
```

There are **422419** obervations in this dataset.

Time

The time of news range from **2014-03-10 16:52:50 GMT** to **2014-08-28 12:33:11 GMT**.

```r
mytime <- as.POSIXct(TIMESTAMP/1000, origin="1970-01-01", tz = "GMT")
range(mytime)
```

```
## [1] "2014-03-10 16:52:50 GMT" "2014-08-28 12:33:11 GMT"
```

Category

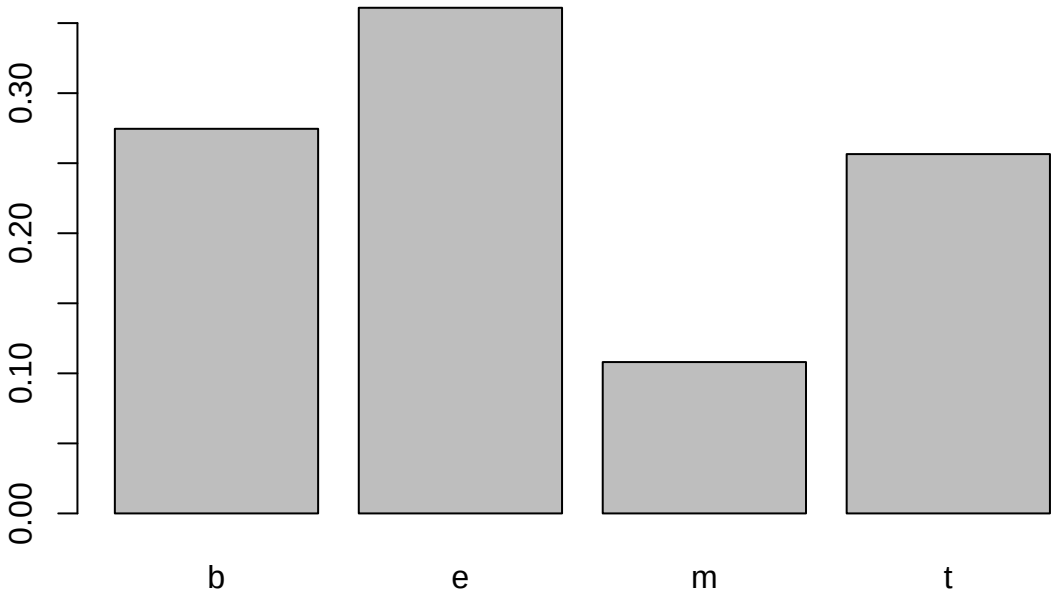
There are:

- 152746 news of business category
- 108465 news of science and technology category
- 115920 news of business category
- 45615 news of health category

```r
table(CATEGORY)
```

```
## CATEGORY
##      b      e      m      t
## 115967 152469  45639 108344
```

```r
# Freq plot
barplot(prop.table(table(CATEGORY)))
```



Story

There are:

- 2076 clusters of similar news for entertainment category
- 1789 clusters of similar news for science and technology category
- 2019 clusters of similar news for business category
- 1347 clusters of similar news for health category

```
# Business
story_b = STORY[CATEGORY == "b"]
tsb <- as.data.frame(table(story_b))
dim(tsb)
```

```
## [1] 2019    2
```

```
# entertainment
story_e = STORY[CATEGORY == "e"]
tse <- as.data.frame(table(story_e))
dim(tse)
```

```
## [1] 2075    2
```

```
# health
story_m = STORY[CATEGORY == "m"]
tsm <- as.data.frame(table(story_m))
dim(tsm)
```

```
## [1] 1347    2
```

```
# science and technology
story_t = STORY[CATEGORY == "t"]
tst <- as.data.frame(table(story_t))
dim(tst)
```

```
## [1] 1789    2
```

Dow Jones Industrial Average (DJIA)

We collect the DJIA data from **2008-08-08** to **2016-07-01**.

```
DJIA<-read.csv("DJIA_table.csv",fill=T, sep=",", stringsAsFactors = FALSE)
dim(DJIA)
```

```
## [1] 1989    7
```

```
names(DJIA)
```

```
## [1] "Date"      "Open"      "High"      "Low"       "Close"     "Volume"
## [7] "Adj.Close"
```

```
attach(DJIA)
```

Date

```
class(Date)
```

```
## [1] "character"
```

```
range(Date)
```

```
## [1] "2008-08-08" "2016-07-01"
```

Close Price

```
# Close Price
summary(Close)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6547   10913   13026   13463   16478   18312
```

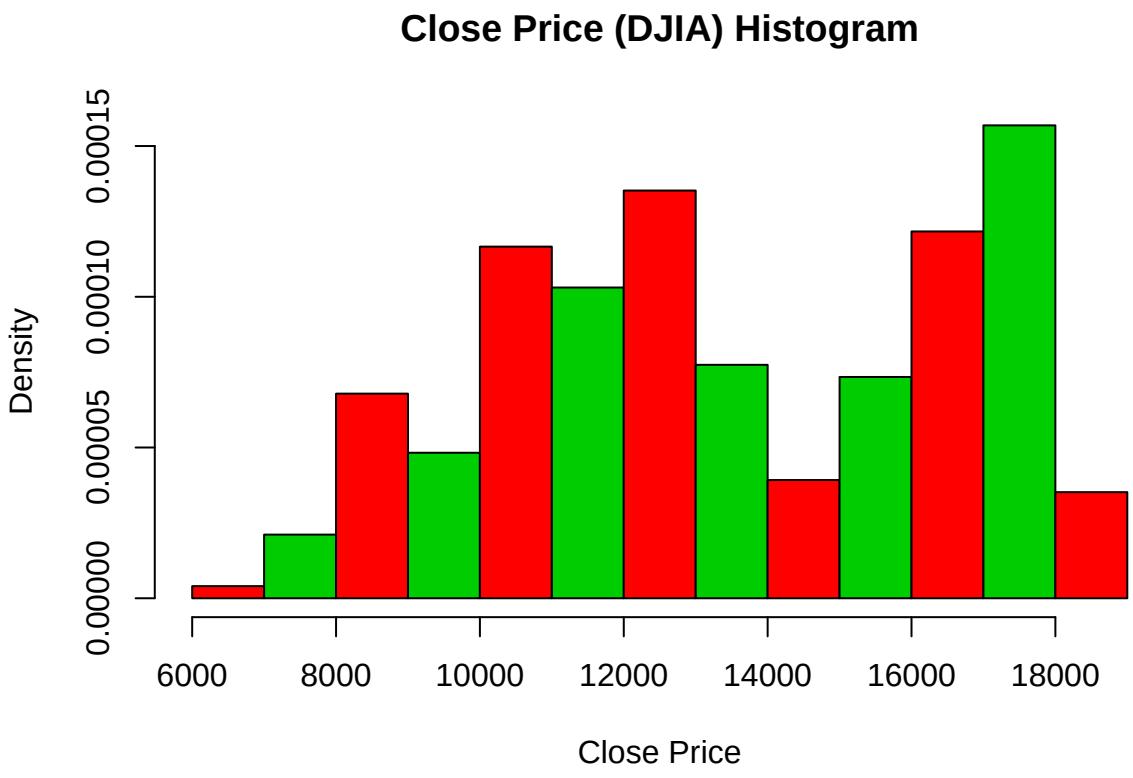
```
# Log Close Price
log_Close <- log(Close)
summary(log_Close)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.787   9.298   9.475   9.479   9.710   9.815
```

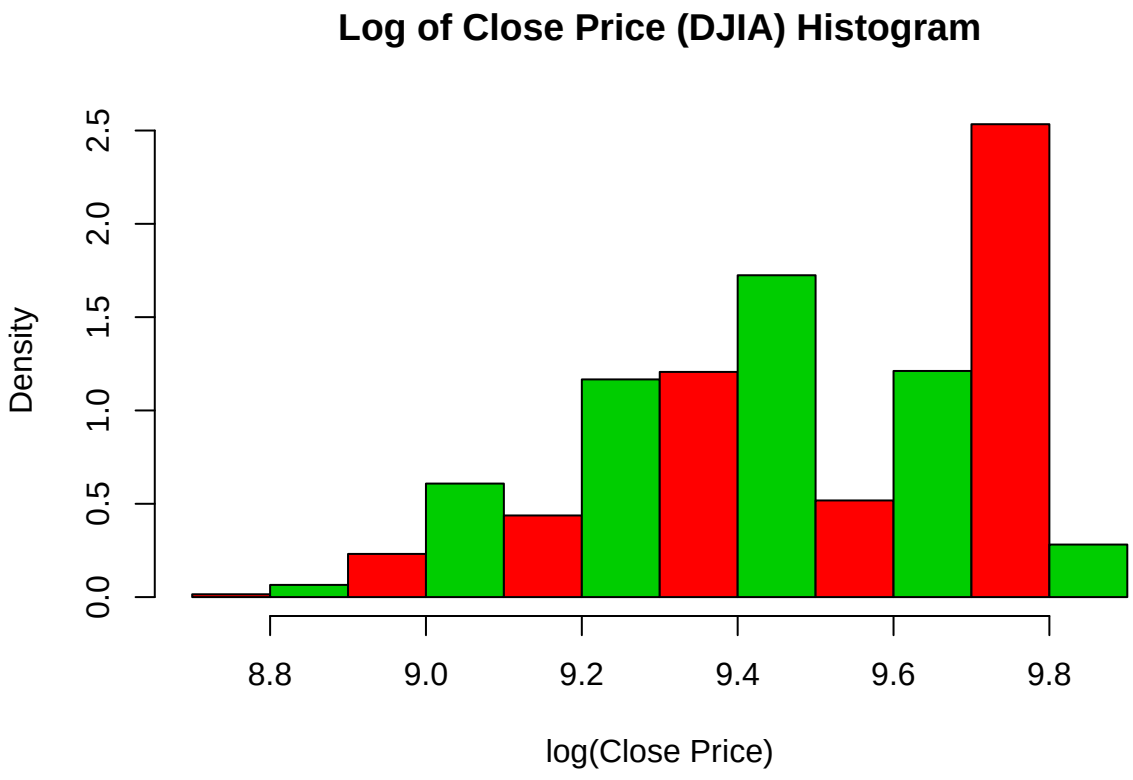
```
# Log Return
log_Return<-diff(log_Close, differences = 1)
summary(log_Return)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -0.1050835 -0.0057316 -0.0005430 -0.0002138  0.0045634  0.0820051
```

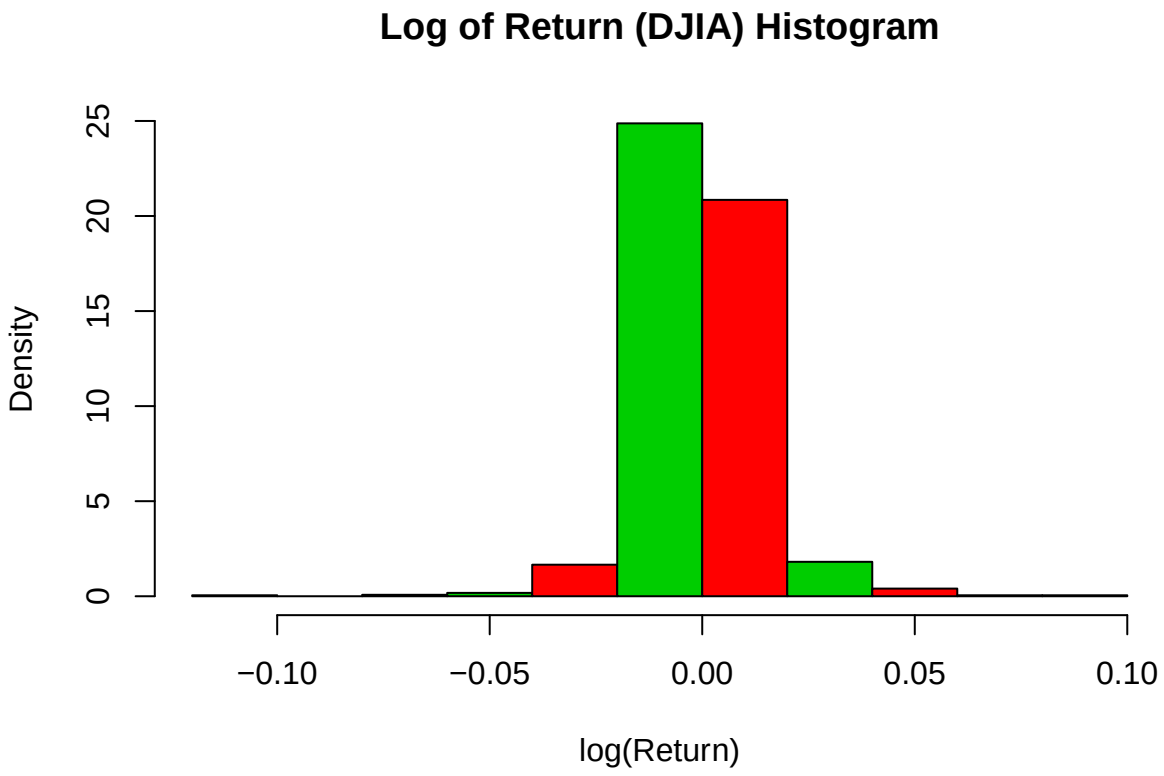
```
 # Histogram
hist(Close, freq=F, main="Close Price (DJIA) Histogram", col=c(2,3), xlab="Close Price")
```



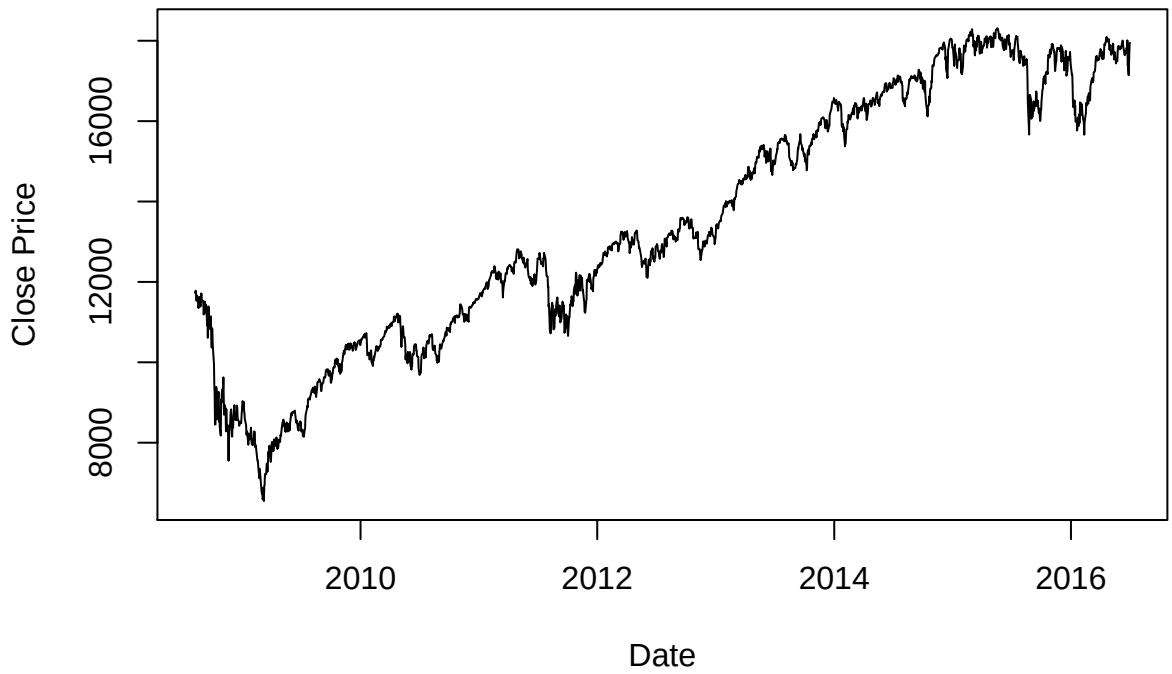

```
hist(log_Close, freq=F, main="Log of Close Price (DJIA) Histogram", col=c(2,3), xlab="log(Close Price)"
```



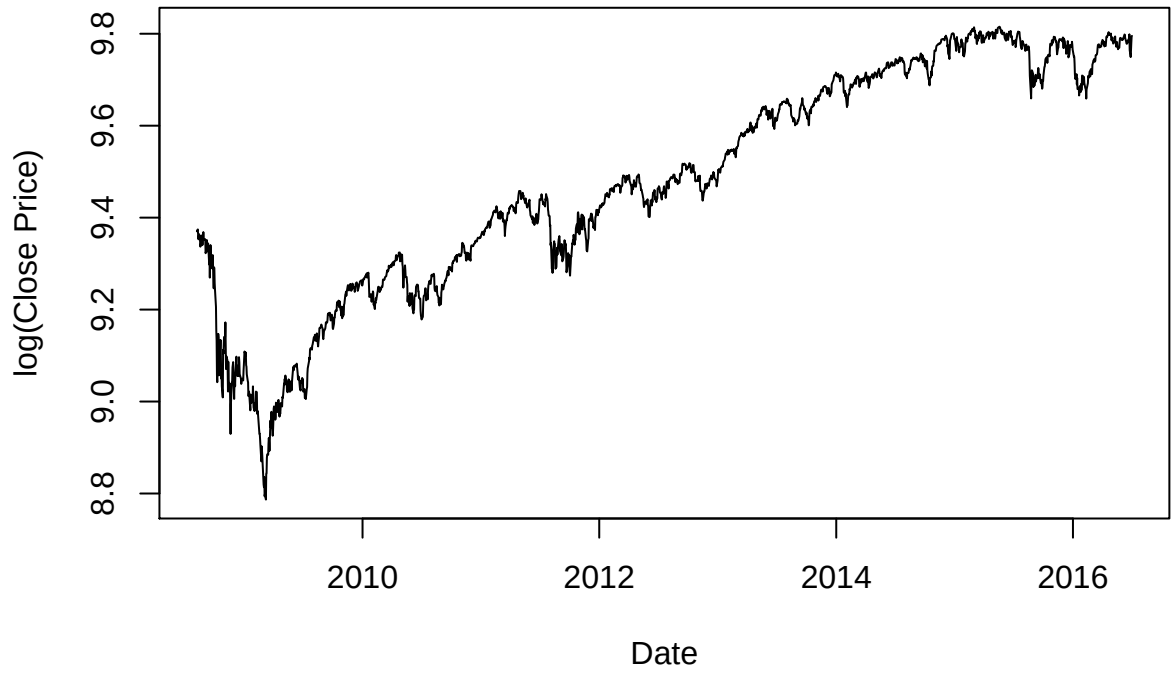

```
hist(log_Return, freq=F, main="Log of Return (DJIA) Histogram", col=c(2,3), xlab="log(Return)")
```




```
#
plot(as.Date(Date), Close, type = "l", xlab = "Date", ylab = "Close Price")
```




```
plot(as.Date(Date), log_Close, type = "l", xlab = "Date", ylab = "log(Close Price)")
```




```
plot(as.Date(Date[-1]), log_Return, type = "l", xlab = "Date", ylab = "log(Return)")
```