R Notebook

Code

Hide

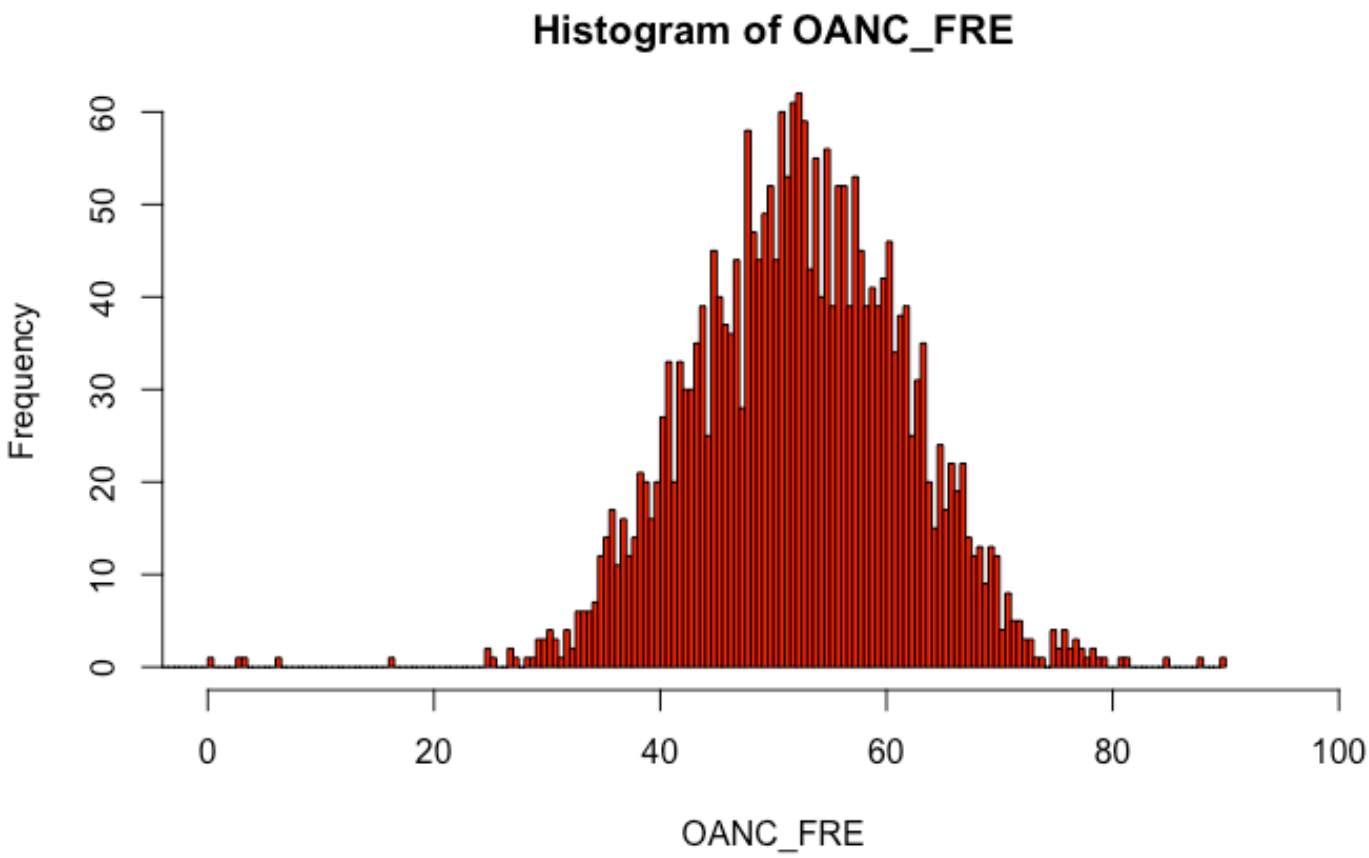
```
library(knitr)
library(readtext)
library(corpora)
library(quanteda)
library(dplyr)
```

Hide

```
#loading file
BNC <- readtext("BNC/*.txt", docvarsfrom=c("filenames"), docvarnames = "id")
BNC_corpus <- corpus(BNC)
bnc_docvars <- data.frame(id = docvars(BNC_corpus)$id,stringsAsFactors = F)
# left join to find meta info from BNCmeta
data(BNCmeta)
df_bnc_docvars <- data.frame(
  id = BNCmeta["id"],
  mode = BNCmeta["mode"],
  domain = BNCmeta["domain"],
  stringAsFactors = F)
docvar_info <- left_join(bnc_docvars,df_bnc_docvars,by = c("id" = "id"))
# add docvars
docvars(BNC_corpus)$mode <- df_bnc_docvars$mode
docvars(BNC_corpus)$domain <- df_bnc_docvars$domain
#unique(df_bnc_docvars$domain)
BNC_sub_corpus <- corpus_subset(BNC_corpus,mode == "written")
BNC_select_domain <- c("social_science","world_affairs","arts","leisure","commerce_fi
nance")
BNC_sub_corpus <- corpus_subset(BNC_sub_corpus,domain %in% BNC_select_domain)
BNC_dfm <- dfm(BNC_sub_corpus)
```

Hide

```
OANC <- readtext("OANC/*.txt", docvarsfrom=c("filenames"))
OANC_corpus <- corpus(OANC)
OANC_dfm <- dfm(OANC_corpus)
```



Hide

```
hist(OANC_FRE, breaks=1000, col="red",xlim = c(0,100))
print(length((OANC_FRE)))
```

```
[1] 2480
```

Hide

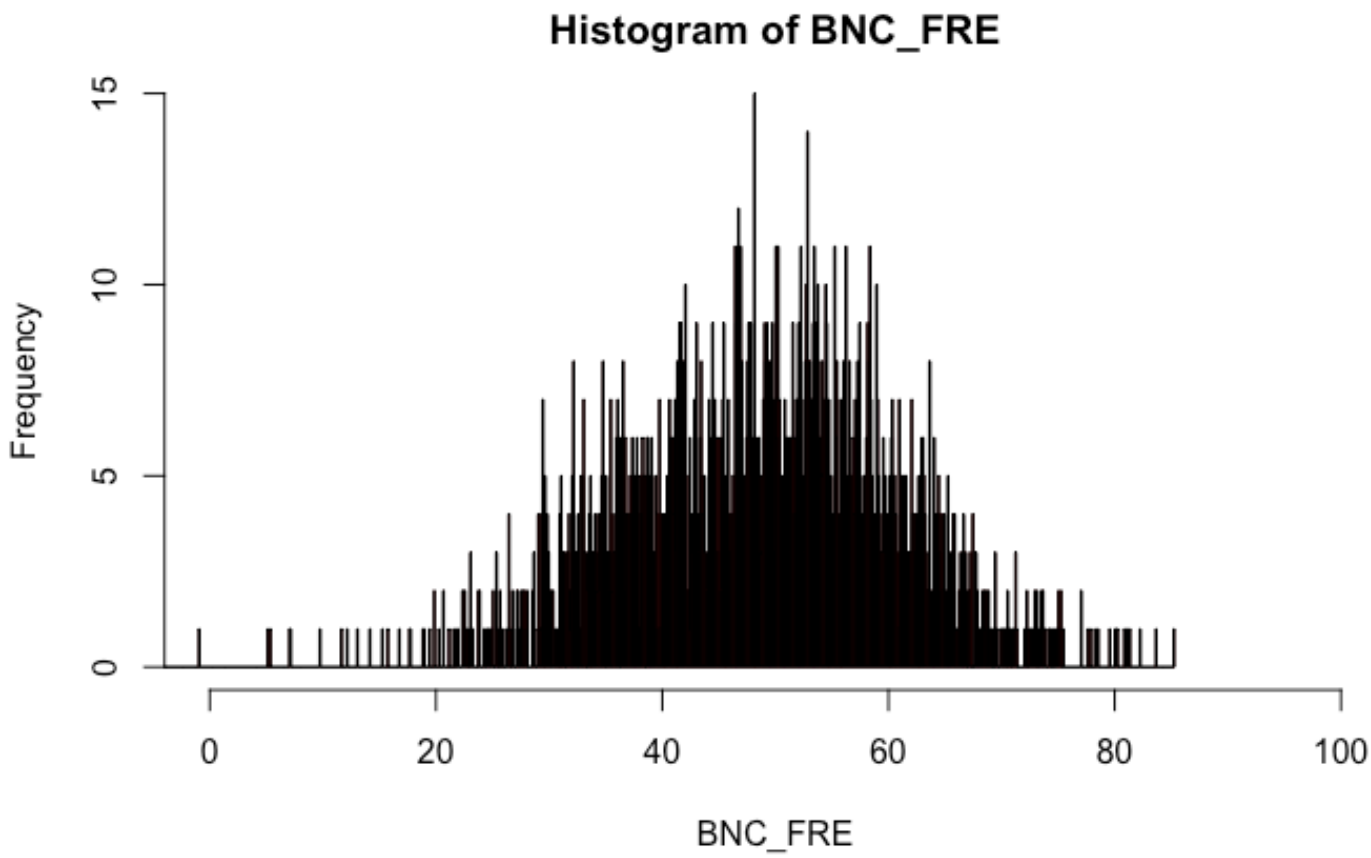
```
print(mean(OANC_FRE))
```

```
[1] 51.91524
```

Hide

```
print(var(OANC_FRE))
```

```
[1] 154.7285
```



Hide

```
hist(BNC_FRE, breaks=1000, col="red",xlim = c(0,100))
print(length((BNC_FRE)))
```

```
[1] 2002
```

Hide

```
print(mean(BNC_FRE))
```

```
[1] 48.8381
```

Hide

```
print(var(BNC_FRE))
```

```
[1] 141.0676
```

Hide

```
t.test(BNC_FRE,OANC_FRE)
```

```
    Welch Two Sample t-test

data:  BNC_FRE and OANC_FRE
t = -8.4423, df = 4356.4, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.791732 -2.362555
sample estimates:
mean of x mean of y
 48.83810  51.91524
```

Hide

```
BNC_dfm
```

```
Document-feature matrix of: 2,002 documents, 354,135 features (98.8% sparse).
```