Data Science Captsone Visualization

```
library(tm)
library(caret)
library(RWeka)
x <- file(file.choose(), "r")
  read <- readLines(x, 15000)
  close(x)
x <- file(file.choose(), "r")
  read2 <- readLines(x, 15000)
  close(x)
x <- file(file.choose(), "r")
  read3 <- readLines(x, 15000)
  close(x)
```

```
docs <- VCorpus(VectorSource(c(read, read2, read3)))

print("Read The Files")
```

```
## [1] "Read The Files"
```

```
toSpace <- content_transformer(function(x, pattern) { return (gsub(pattern, " ", x))})

docs <- tm_map(docs, toSpace, "-")
docs <- tm_map(docs, toSpace, ":")
docs <- tm_map(docs, toSpace, "'")
docs <- tm_map(docs, toSpace, " -")

print("Clearning a bit")
```

```
## [1] "Clearning a bit"
```

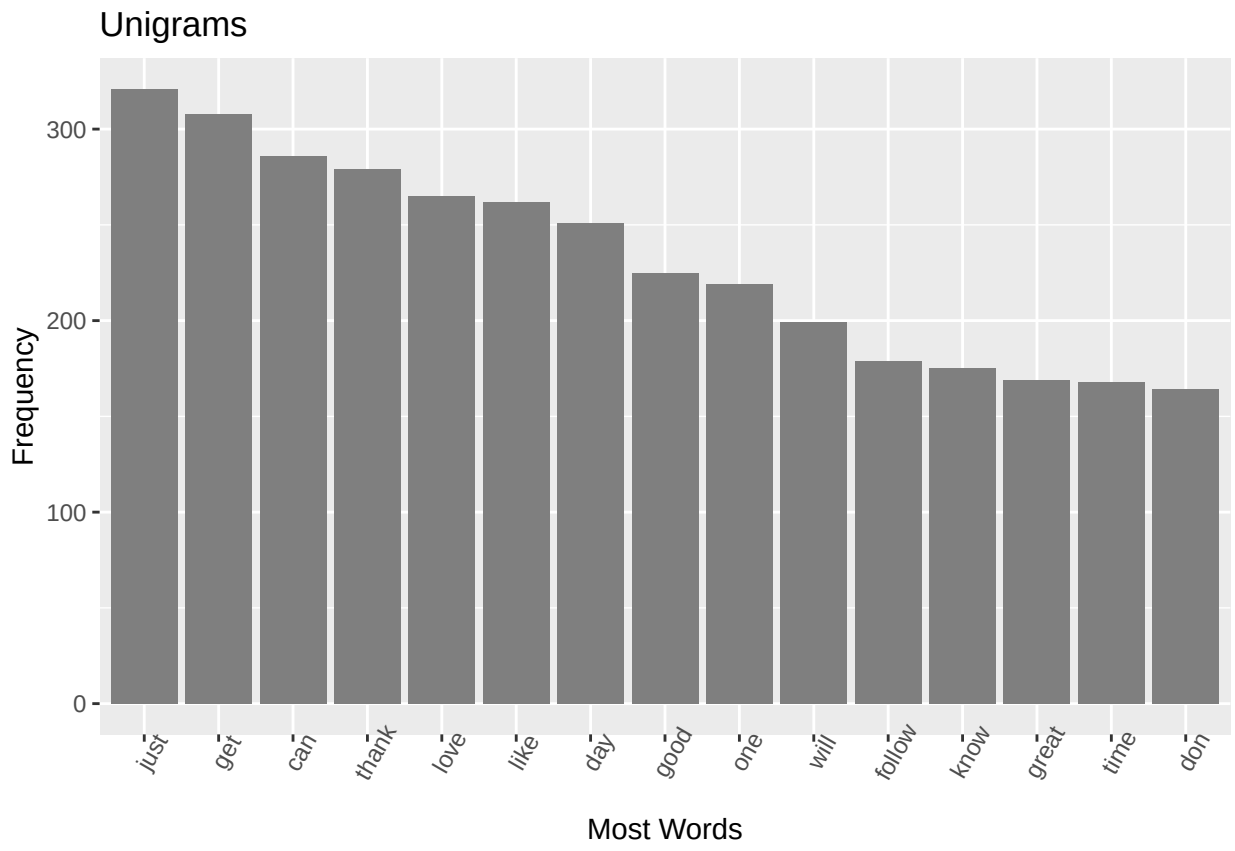
```
docs <- tm_map(docs, removePunctuation)
docs <- tm_map(docs, content_transformer(tolower))
docs <- tm_map(docs, removeNumbers)
docs <- tm_map(docs, stripWhitespace)
docs <- tm_map(docs, removeWords, stopwords("english"))
docs <- tm_map(docs, stemDocument)

print("Deep Clearning Time")
```

```
## [1] "Deep Clearning Time"
```

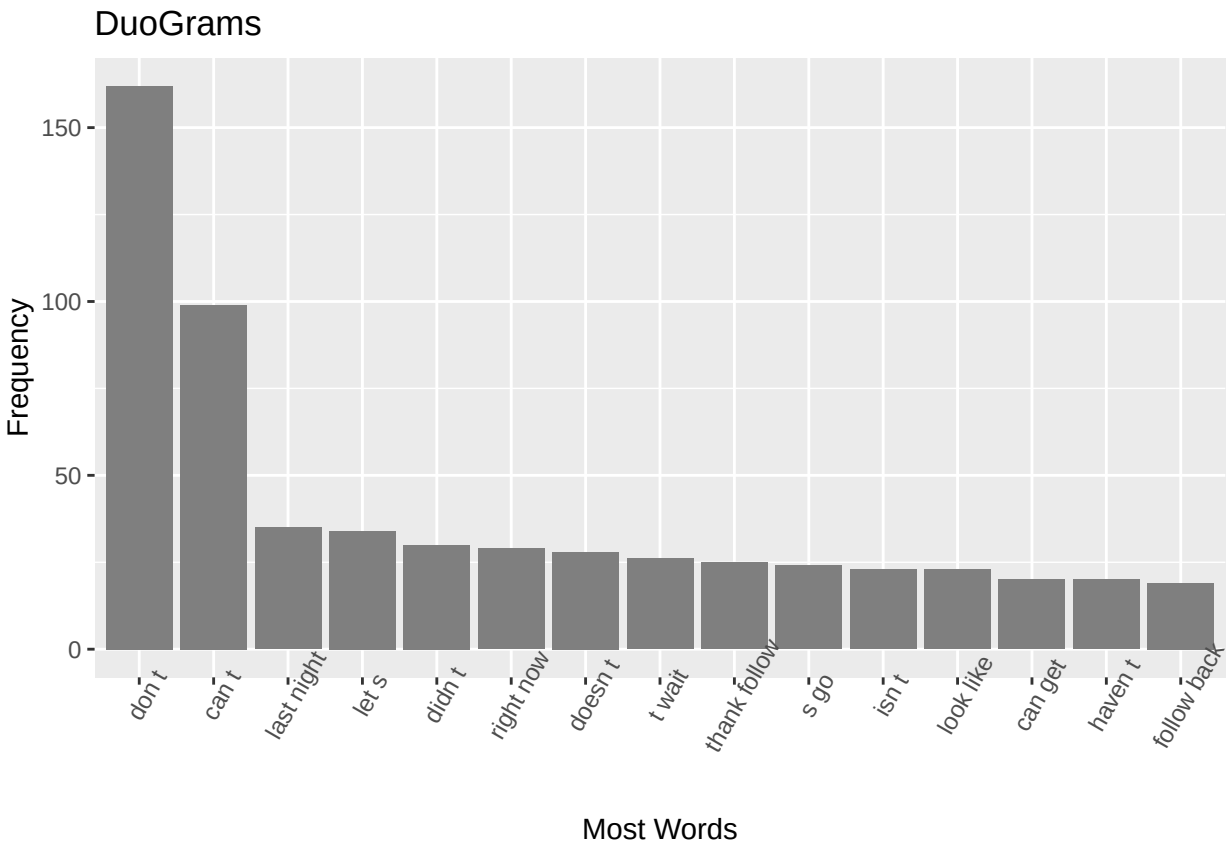
```
gram <- function(x) NGramTokenizer(x, Weka_control(min = 1, max = 1))
docsOne <- TermDocumentMatrix(docs, control = list(tokenize = gram))
freq <- findFreqTerms(docsOne)
freqsum <- rowSums(as.matrix(docsOne[freq,]))
dF <- data.frame(Word = names(freqsum), frequency = freqsum)
dF <- dF[order(dF$frequency, decreasing = T),]
```

```
ggplot(dF[1:15,],aes(x=reorder(Word,-frequency),y=frequency))+
geom_bar(stat="identity",fill = I("grey50"))+
labs(title="Unigrams",x="Most Words",y="Frequency")+
theme(axis.text.x=element_text(angle=60))
```



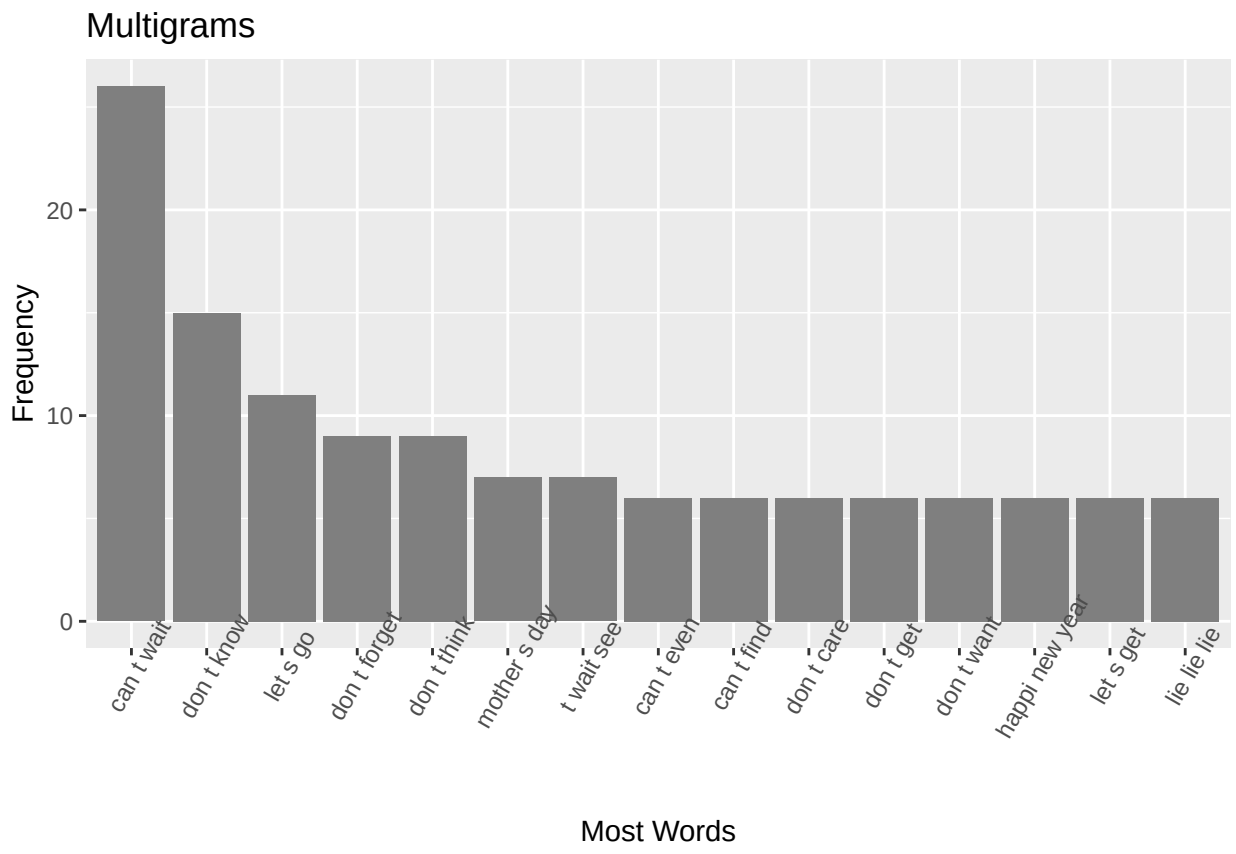

```
gram <- function(x) NGramTokenizer(x, Weka_control(min = 2, max = 2))
docsOne <- TermDocumentMatrix(docs, control = list(tokenize = gram))
freq <- findFreqTerms(docsOne)
freqsum <- rowSums(as.matrix(docsOne[freq,]))
dF <- data.frame(Word = names(freqsum), frequency = freqsum)
dF <- dF[order(dF$frequency, decreasing = T),]

ggplot(dF[1:15,],aes(x=reorder(Word,-frequency),y=frequency))+
geom_bar(stat="identity",fill = I("grey50"))+
labs(title="DuoGrams",x="Most Words",y="Frequency")+
theme(axis.text.x=element_text(angle=60))
```




```
gram <- function(x) NGramTokenizer(x, Weka_control(min = 3, max = 3))
docsOne <- TermDocumentMatrix(docs, control = list(tokenize = gram))
freq <- findFreqTerms(docsOne)
freqsum <- rowSums(as.matrix(docsOne[freq,]))
dF <- data.frame(Word = names(freqsum), frequency = freqsum)
dF <- dF[order(dF$frequency, decreasing = T),]

ggplot(dF[1:15,],aes(x=reorder(Word,-frequency),y=frequency))+
geom_bar(stat="identity",fill = I("grey50"))+
labs(title="Multigrams",x="Most Words",y="Frequency")+
theme(axis.text.x=element_text(angle=60))
```


Multigrams
Frequency
20
10
0
can t wait
don t know
let s go
don t forget
don t think
mother s day
t wait see
can t even
can t find
don t care
don t get
don t want
happi new year
let s get
lie lie lie
Most Words