TextMining\_Report

Sarthak

October 2, 2016

# Introduction

In this report I analyzed the text of two text files "Iliyad" and "Odyssey". After data preprocessing, I plotted bar graphs for both the documents for word with frequency greater than 450. Then I plotted the word cloud for both the documents which gives us the idea about the words in terms of their frequency. The higher the word frequency, greater the size of the word in the plot. To analyze about how the words are associated with each other in both Iliyad and Odyssey, I performed hierarachy clustering on both of them seperately and then combined. To further analyze the similarities between the words in Iliyad and Odyssed, I plotted the word frequency graph which displays the common words between the two documents and their degree of commonness

A summary of the steps performed are:

1. Load the required library.
2. Data Preprocessing

* Remove Punctuation
* Remove numbers
* convert text to lowercase
* Remove stopwords
* Stemming
* Remove whitespace

1. Data Staging

* Convert the doc into DTM (Document Term matrix)

1. Data Exploration

* Plot the bar graph for both the documents for the words who have the frequency more than 450.

1. Hierarachy Clustering
2. Relationship between Iliyad and Odyssey

# Load the required libraries

#Needed <- c("tm", "SnowballCC", "RColorBrewer", "ggplot2", "wordcloud", "biclust", "cluster", "igraph", "fpc")  
  
#install.packages(Needed, dependencies=TRUE)  
#install.packages("Rcampdf", repos = "http://datacube.wu.ac.at/", type = "source")

# Load documents

cname <- file.path("D:/GRAD\_SCHOOL/Fall2016/Lecture3", "texts")  
dir(cname)

## [1] "iliad.txt" "odyssey.txt"

# Load text in R using TM package

library(tm)

## Loading required package: NLP

docs <- Corpus(DirSource(cname))  
summary(docs)

## Length Class Mode  
## iliad.txt 2 PlainTextDocument list  
## odyssey.txt 2 PlainTextDocument list

# Preprocessing data

## Removing punctuation

docs <- tm\_map(docs, removePunctuation)  
  
for(j in seq(docs))  
{  
 docs[[j]] <- gsub("/", " ", docs[[j]])  
 docs[[j]] <- gsub("@", " ", docs[[j]])  
 docs[[j]] <- gsub("\\|", " ", docs[[j]])  
}

## Remove numbers

docs <- tm\_map(docs, removeNumbers)

## Convert to lowercase

docs <- tm\_map(docs, tolower)

## Remove stopwords

docs <- tm\_map(docs, removeWords, stopwords("english"))

## Remove word endings - stemming

library(SnowballC)  
docs <- tm\_map(docs, stemDocument)

## Remove whitespaces

docs <- tm\_map(docs, stripWhitespace)

## Tell R to treat your pre-processed doc as text doc

docs <- tm\_map(docs, PlainTextDocument)

# Stage the data

## Create a document term matrix (DTM)

dtm <- DocumentTermMatrix(docs)  
dtms <- removeSparseTerms(dtm, 0.1)

## Data Exporation

In this section, we analyze the frequency of words in both the documents seperately and combinied.

freq.doc1 <- sort(colSums(as.matrix(dtm[1,])), decreasing=TRUE)  
freq.doc2 <- sort(colSums(as.matrix(dtm[2,])), decreasing=TRUE)

Below are the words with highest frequency for Iliyad.

wf.doc1 <- data.frame(word=names(freq.doc1), freq=freq.doc1)  
head(wf.doc1)

## word freq  
## thy thy 943  
## thus thus 621  
## now now 573  
## shall shall 510  
## great great 478  
## achilles achilles 385

Below are the words with highest frequency for Odyssey.

wf.doc2 <- data.frame(word=names(freq.doc2), freq=freq.doc2)  
head(wf.doc2)

## word freq  
## will will 686  
## ulysses ulysses 591  
## one one 516  
## said said 483  
## house house 375  
## men men 325

## Plot the word frequency as a bar graph

In this section I plot the word frequency bar graph for both the documents. I only plot for the words with frequency more than 450.

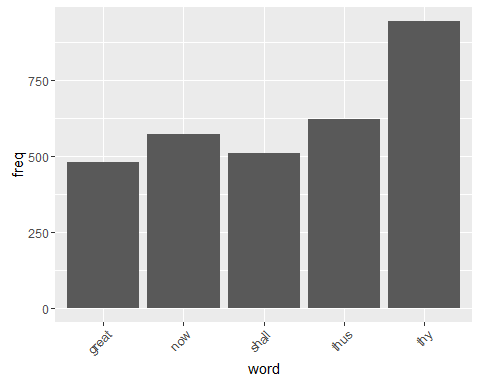
Bar graph for Iliyad:

library(ggplot2)

##   
## Attaching package: 'ggplot2'

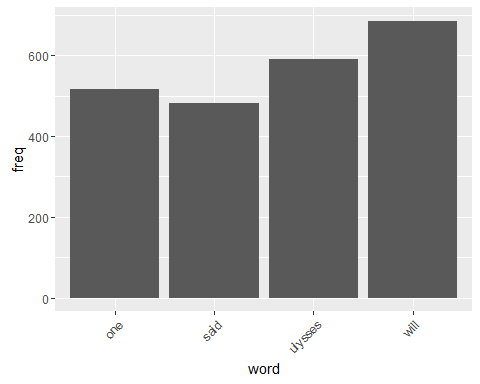
## The following object is masked from 'package:NLP':  
##   
## annotate

p.doc1 <- ggplot(subset(wf.doc1, freq>450), aes(word, freq)) +  
 geom\_bar(stat="identity") +  
 theme(axis.text.x=element\_text(angle=45, hjust=1))  
p.doc1



Bar graph for Odyssey:

p.doc2 <- ggplot(subset(wf.doc2, freq>450), aes(word, freq)) +  
 geom\_bar(stat="identity") +  
 theme(axis.text.x=element\_text(angle=45, hjust=1))  
p.doc2



# Plotting Word Cloud

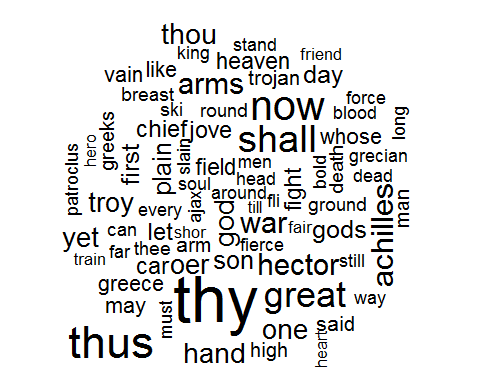
In this section, I further analyzed the word frequency by plotting Word cloud.

Word Cloud for Iliyad: In black and white:

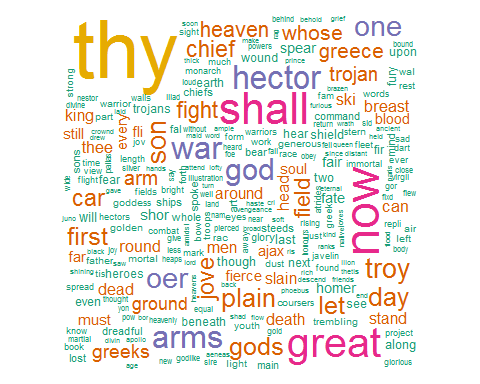
library(wordcloud)

## Loading required package: RColorBrewer

set.seed(142)  
wordcloud(names(freq.doc1), freq.doc1, min.freq=150)

 In color:

set.seed(142)  
wordcloud(names(freq.doc1), freq.doc1, min.freq=60, scale=c(5, .1), colors=brewer.pal(6, "Dark2"))

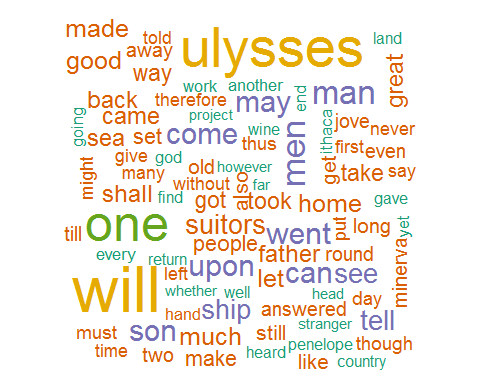


We can see that "thy" is the biggest word, hence it has the highest frequency.

Word Cloud for Odyssey:

In Color:

set.seed(142)  
dark2 <- brewer.pal(6, "Dark2")  
wordcloud(names(freq.doc2), freq.doc2, max.words=100, rot.per=0.2, colors=dark2)



We can see that "ulysses" and "will" are the biggest words and of same color, hence they have the highest frequency.

# Hierarachy Clustering

To further analyze the two documents, I performed hierarachy clustering to analyze the distance between the words. I wanted to analyze which words are more likely to appear together and which words are more likely to appear. First, I performed the clustering for Iliyad.

dtms <- removeSparseTerms(dtm, 0.15)  
library(cluster)  
# First calculate distance between words  
d.doc1 <- dist(t(dtms[1,as.vector(dtms[1,] > 200)]), method="euclidian")   
d.doc2 <- dist(t(dtms[2,as.vector(dtms[2,] > 200)]), method="euclidian")  
fit.doc1 <- hclust(d=d.doc1, method="ward")

## The "ward" method has been renamed to "ward.D"; note new "ward.D2"

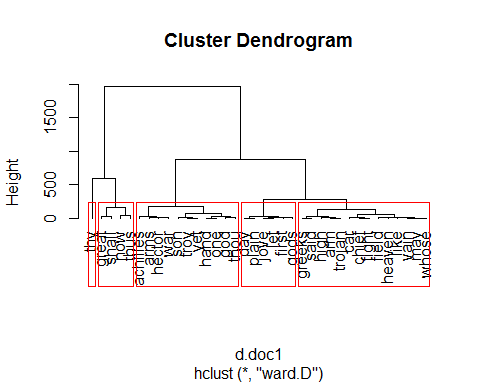
fit.doc2 <- hclust(d=d.doc2, method="ward")

## The "ward" method has been renamed to "ward.D"; note new "ward.D2"

### Plot the Dendrogram for Iliyad

k defines the number of clusters you are using. I choose 5

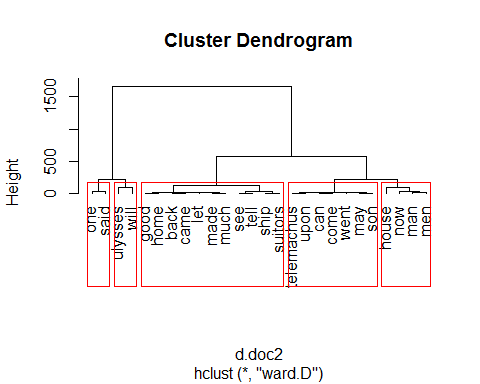
plot.new()  
plot(fit.doc1, hang=-1)  
groups <- cutree(fit.doc1, k=5)   
rect.hclust(fit.doc1, k=5, border="red")



We can see that, "thy" has its own group and so on.

### Plot the Dendrogram for Odyssey

plot.new()  
plot(fit.doc2, hang=-1)  
groups <- cutree(fit.doc2, k=5)   
rect.hclust(fit.doc2, k=5, border="red")



# Relationship between Illiad and Odyssey

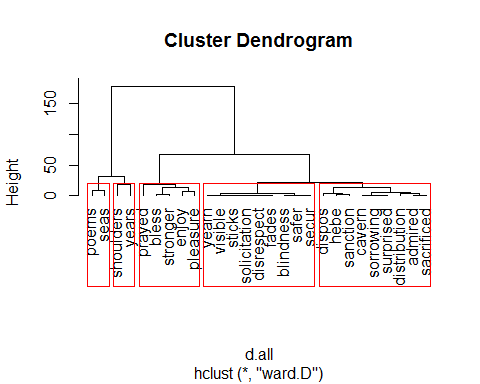
I studied the said relationship by two methods:

## 1. I performed the clustering on both the documents together.

d.all <- dist(t(dtms[,as.vector(dtms[,] > 200)]), method="euclidian")  
fit.all <- hclust(d=d.all, method="ward")

## The "ward" method has been renamed to "ward.D"; note new "ward.D2"

plot.new()  
plot(fit.all, hang= -1)  
groups <- cutree(fit.all, k=5)   
rect.hclust(fit.all, k=5, border="red")



Terms higher in the plot appear more frequently within the corpus. Terms grouped near to each other are more frequently found together.

## 2. Comparing the word frequency between Illiad and Odyssey.

COnvert the DTm to a matrix.

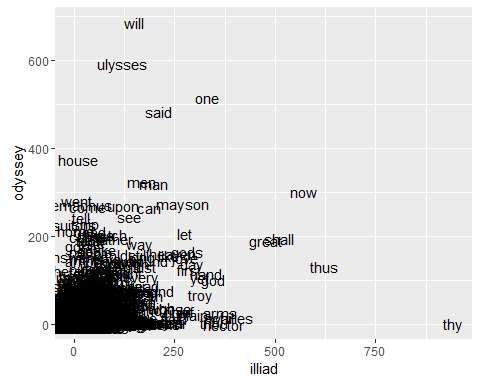
mat <- as.matrix(dtm)  
cb <- data.frame(illiad = mat[1,], odyssey = mat[2,])

Select the words which are common in both the documents.

cb1 <- subset(cb, cb$illiad != '0')  
cb2 <- subset(cb1, cb1$odyssey != '0')

Plot the graph between the two

ggplot(cb2, aes(illiad, odyssey)) +  
 geom\_text(label = rownames(cb2),  
 position=position\_jitter())



We can observe that the word "thy" in the extreme right is occurring more in Illiyad and significantly less in Odyssey. the words "house", "ulysses" are occuring more in odyssey and less in iliyad. The word jungle in the bottom left represents the words that are uslually common in both the documents.