Wikpedia Clusters

Using the k-means algorithm test separating the dataset into multiple clusters, inspect your results. Are the resulting clusters useful, do their members share characteristics? Use hierarchical clustering on the same data. Discuss your programs execution time, compare the run time of hierarchical clustering to k-means.

The code below is to create a sparse matrix of the

```
# install.packages("tm")
library(tm)
```

```
## Loading required package: NLP
```

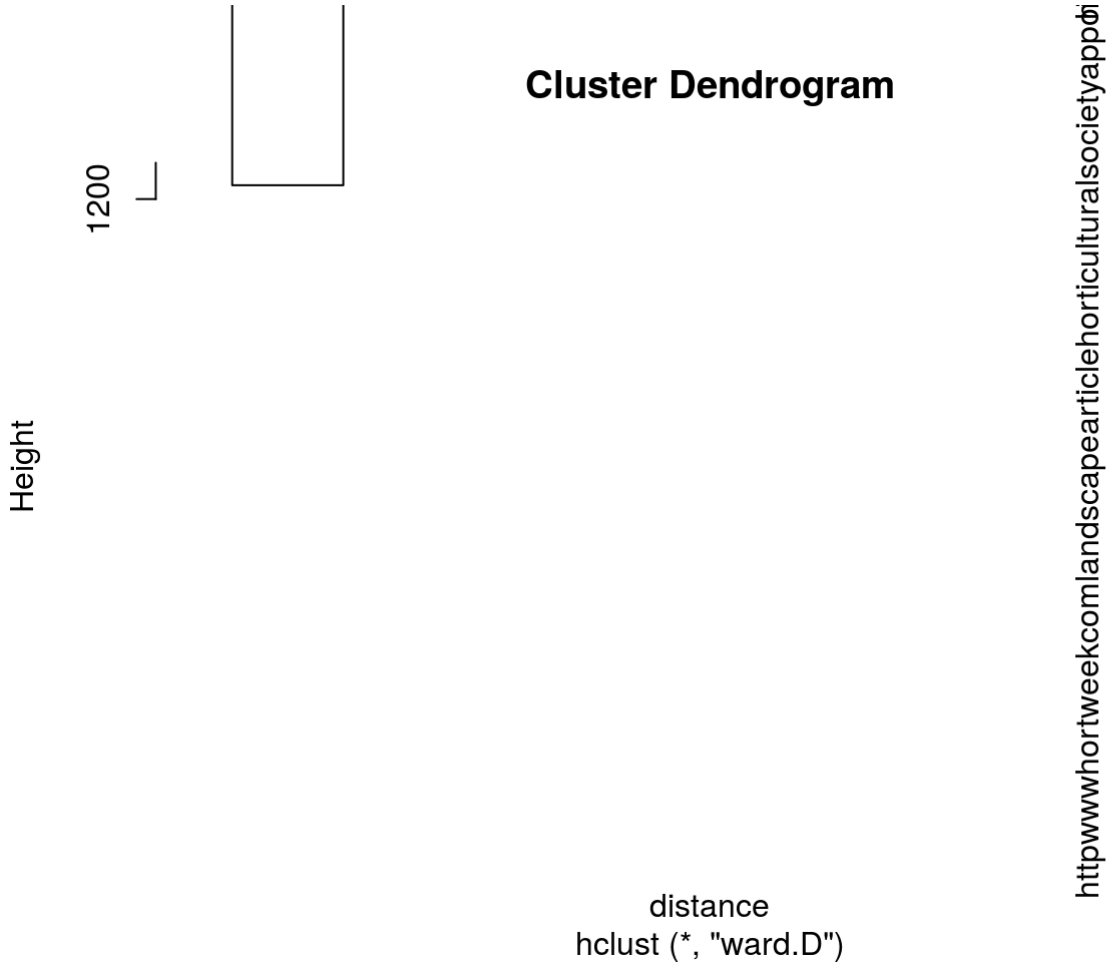
```
library(NLP)
d = read.csv("enwiki-sample.csv", sep=";", quote="", header=F,
stringsAsFactors=FALSE)
colnames(d) = c("title","content")
corpus = Corpus(VectorSource(d$content))
#
# # We can set metadata, e.g. title:
i=0
corpus = tm_map(corpus, function(x) {
i = i + 1
  meta(x, "title") = d$title[i]
x
})
#
# # you can do:
# inspect(corpus)
#
# # To access an element in the corpus use:
# # corpus[[X]]su
#
# # we apply functions to clean the data
corpus = tm_map(corpus, removePunctuation)
corpus = tm_map(corpus, stripWhitespace)
corpus = tm_map(corpus, removeNumbers)
corpus = tm_map(corpus, content_transformer(tolower))
corpus = tm_map(corpus, removeWords, stopwords("english"))
# Create a sparse matrix for each document the word frequency
dtm = DocumentTermMatrix(corpus)
# To see data: e.g. call inspect(dtm[1:5,1:50] )

distance <- dist(t(dtm), method="euclidian")
# ### cluster into 10 clusters
# kmeans10 <- kmeans(dtm, 5)
#
# kw_with_cluster <- as.data.frame(cbind(corpus, kmeans10$cluster))
# names(kw_with_cluster) <- c("keyword", "kmeans10")
#
# #Make df for each cluster result, quickly "eyeball" results
# cluster1 <- subset(kw_with_cluster, subset=kmeans10 == 1)
# cluster2 <- subset(kw_with_cluster, subset=kmeans10 == 2)
# cluster3 <- subset(kw_with_cluster, subset=kmeans10 == 3)
# cluster4 <- subset(kw_with_cluster, subset=kmeans10 == 4)
# cluster5 <- subset(kw_with_cluster, subset=kmeans10 == 5)
#
#  table(cl$cluster)
# Inspect results....
```

Due to the big amount of time consumed running the distance count for th 4248 line of articles on the original file. A new file was created using the linux command head -500 enwiki-clean-10mb > enwiki-sample.csv.

The processing of the smaller file made it hard to find any relevant data for the clustering

```
# install.packages('cluster')
library(fpc)
library(cluster)
fit <- hclust(d=distance, method="ward.D")
plot(fit, hang=-1)
```



Kmeans gave results that are better to be described. Below is the centers of the clusters. The clustering of each entry can be showen but it was commented out for the ease of reading.

```
# install.packages('fpc')
library(fpc)
library(cluster)
kfit <- kmeans(distance, 2)
kfit$withinss
```

```
## [1] 113399192  70890278
```

```
kfit$totss
```

```
## [1] 362177866
```

```
# kfit$centers
# kfit$cluster
# clusplot(as.matrix(distance), kfit$cluster, color=T, shade=T, labels=2, lines=0)
```