Clustering

```
library(quanteda)
library(tidytext)
library(topicmodels)
library(ldatuning)
library(stringi)
library(stm)
library(tm)
```

```
BI1_lyrics = read.csv('~/Documents/NYU/Text-as-Data/British_Invasion/Preprocessing/BI_Final.csv')
BI1_lyrics = BI1_lyrics[,-1][BI1_lyrics$Label == 1,]
BI1_lyrics$Lyrics = as.character(BI1_lyrics$Lyrics)
lyrics_corpus = corpus(BI1_lyrics, docnames = "Song", text_field = "Lyrics")
lyrics_dfm = dfm(lyrics_corpus,tolower = T, groups = "Artist")
lyrics_trim = dfm_trim(lyrics_dfm, min_count = 20, verbose = T)
```

```
## Removing features occurring:
##   - fewer than 20 times: 12,634
##   Total features removed: 12,634 (87.3%).
```

```
hc = hclust(dist(lyrics_trim, method = "euclidean"))
```

```
plot(hc, hang = -1)
```