Text Analysis: A Star Wars Story

Greg Eastman

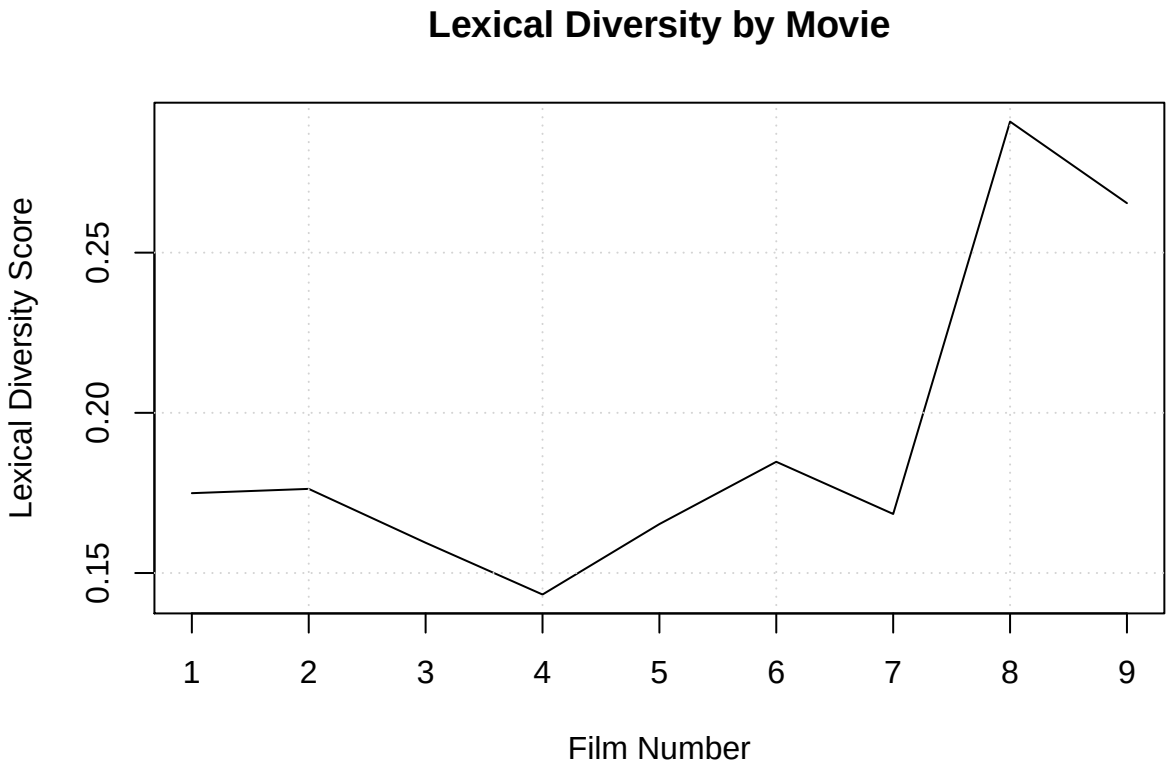
4/15/2021

WORD CLOUD:

Using a word cloud just to get a visually appealing idea of the most common themes in the films.

LEXICAL DIVERSITY:

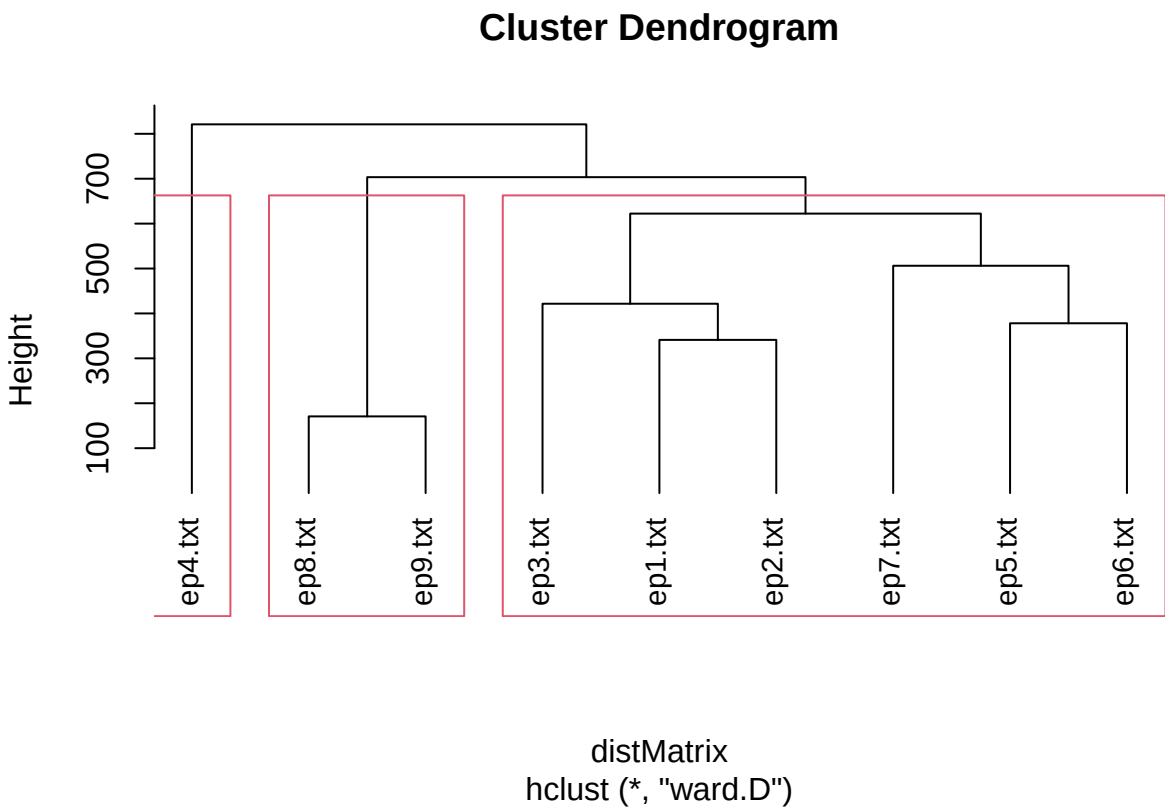
We are going to see how complex the language is in the films.

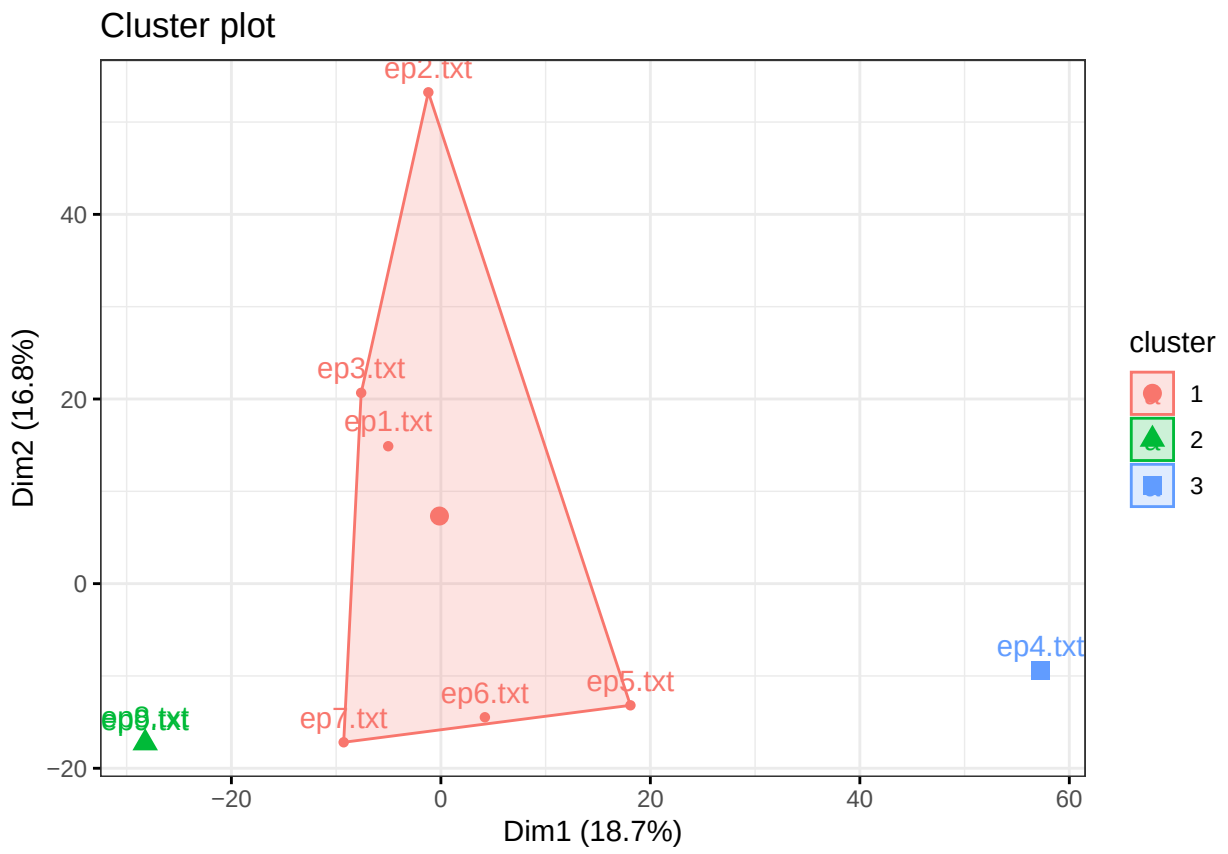



```
## numeric(0)
```

CLUSTERING:

We are going to see what films are considered “most similar” to the others.


Cluster Dendrogram
Height
100
300
500
700
ep4.txt
ep8.txt
ep9.txt
ep3.txt
ep1.txt
ep2.txt
ep7.txt
ep5.txt
ep6.txt
distMatrix
hclust (*, "ward.D")



Sentiment Analysis:

We are going to look at the overall sentiment in the films broken down by the clusters found in the previous part.

```
## Episodes 1, 2, 3, 5, 6, 7:
```

```
##   Category    x
## 1 negative 4951
## 2 positive 3044
```

```
## Episode 4:
```

```
##   Category    x
## 1 negative 1104
## 2 positive  574
```

```
## Episode 8, 9:
```

```
##   Category   x
## 1 negative 655
## 2 positive 407
```

Looking at the frequent word counts for episodes 1, 2, 3, 5, 6, 7.

```
## # A tibble: 10 x 3
##    document term    count
##    <chr>    <chr>   <dbl>
##  1 ep3.txt  jedi      233
##  2 ep7.txt  day       223
##  3 ep3.txt  droid     209
```

```
##  4 ep7.txt  continu   165
##  5 ep2.txt  jedi      162
##  6 ep1.txt  droid     149
##  7 ep3.txt  clone     130
##  8 ep2.txt  senat     128
##  9 ep2.txt  look      127
## 10 ep7.txt  back      124
```

Looking at the frequent word counts for episode 4.

```
## # A tibble: 10 x 3
##    document term    count
##    <chr>    <chr>   <dbl>
##  1 ep4.txt  star      236
##  2 ep4.txt  death     230
##  3 ep4.txt  cockpit   197
##  4 ep4.txt  fighter   180
##  5 ep4.txt  look      156
##  6 ep4.txt  red       152
##  7 ep4.txt  ship      122
##  8 ep4.txt  lukeâ€   110
##  9 ep4.txt  leader     92
## 10 ep4.txt  xwing      92
```

Looking at the frequent word counts for episodes 8, 9.

```
## # A tibble: 10 x 3
##    document term       count
##    <chr>    <chr>      <dbl>
##  1 ep9.txt  know          53
##  2 ep9.txt  ship          51
##  3 ep8.txt  ship          48
##  4 ep9.txt  come          48
##  5 ep8.txt  now           46
##  6 ep9.txt  stormtroop    44
##  7 ep9.txt  take          41
##  8 ep8.txt  resist        40
##  9 ep8.txt  just          39
## 10 ep9.txt  got           38
```

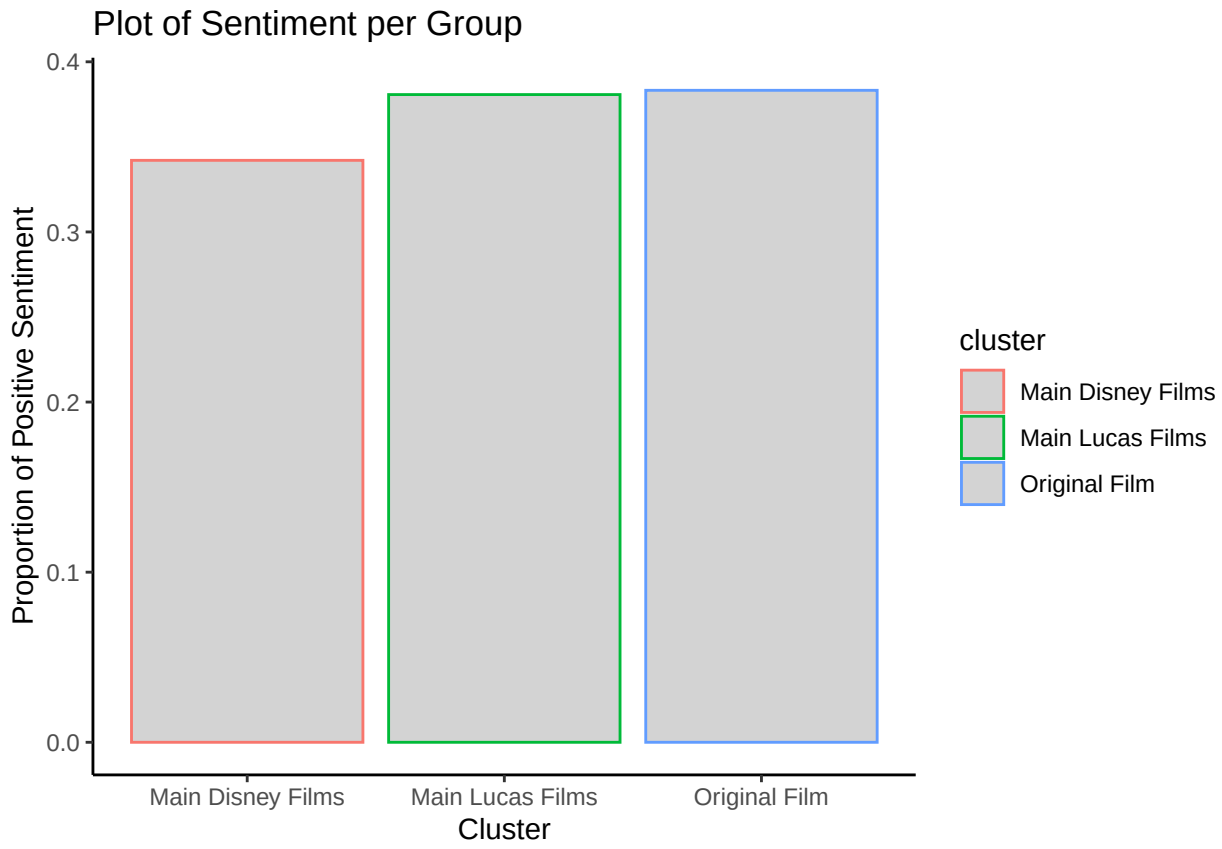
Proportion of "good" terms in each cluster.

```
## Proportion of positivity for episodes 1, 2, 3, 5, 6, 7:  0.380738
```

```
## Proportion of positivity for episodes 8,9:  0.3832392
```

```
## Proportion of positivity for episode 4:  0.3420739
```

```
##   sentiment           cluster
## 1 0.3807380  Main Lucas Films
## 2 0.3420739 Main Disney Films
## 3 0.3832392     Original Film
```



DIRICHELET ANALYSIS:

NOTE– I removed the code and analysis here because the work led nowhere. The large number and frequency of made up words made it impossible to get any meaningful knowledge from it and there is no reason to clutter my file with more useless code and diagrams.