Mapping All of South Park

by Fred

"All characters and events in this show –even those based on real people– are entirely fictional. All celebrity voices are impersonated ... poorly. The following program contains coarse language and due to its content it should not be viewed by anyone."

Step 1: Collect Underpants

```
##### Packages ####
# The usual suspects
library(magrittr)
library(tidyverse)

# Text analysis
library(text2vec)
library(stringr)
library(tidytext)
data("stop_words")
library(SnowballC)
library(wordcloud)

# Our very own
library(FredsVietorisRips)

# Graph adjacencies
library(igraph)
```

```
##### Import Data ####
df_original <- readr::read_csv("D:/southparklines/All-seasons.csv")
```

```
df <- df_original %>%
  unnest_tokens(word, Line) %>%
  anti_join(stop_words, by="word") %>%
  group_by(Season, Episode) %>%
  summarise(text = str_c(word, collapse = " ")) %>%
  ungroup()

df <- df[-258,]
```

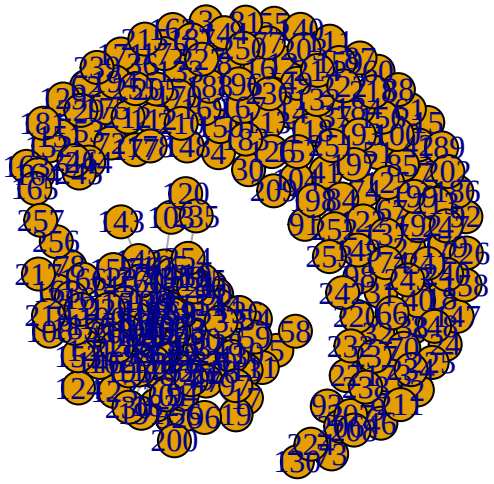
Step 2: ?

```r
##### Tokenize Sample ####
it_sample <- df$text %>%
  itoken(progressbar = FALSE)
```

```r
##### Vectorize (?) ####
vectorizer <- df$text %>%
  itoken(progressbar = FALSE) %>%
  create_vocabulary() %>%
  prune_vocabulary() %>%
  vocab_vectorizer()
```

```r
##### Document Term Matrix ####
dtm_sample <- create_dtm(it_sample, vectorizer)
# dim(dtm_sample)

##### Jaccard Similarity ####
jacsim <- sim2(dtm_sample, dtm_sample, method = "jaccard", norm = "none")
# dim(jacsim)

##### Convert fancy object to matrix ####
jsmat <- jacsim %>%
  as.matrix()

# Flip it around so 1 is max distance
# and 0 indicates two sets are the same
jsmat <- 1 - jsmat
```

```r
##### Create Adjacency Matrix ####
epsilon <- 0.85

adjmat <- jsmat %>%
  AdjacencyMatrix(epsilon)

##### Plot Adjacencies ####
graph_from_adjacency_matrix(adjmat, mode = "undirected") %>%
  plot(main=paste("Epsilon =", epsilon))
```



Step 3: Profit

Top Ten Episodes by Similarity

The episodes which appear most similar are all sequential, a quick googling after referencing them to our data set shows that they are all multi part episodes.

Table 1: Ten Closest Pairs of Episodes

Episode1	Episode2	Similarity
256	257	0.7955508
177	178	0.8018494
164	165	0.8034759
245	246	0.8050314
244	246	0.8067941
207	208	0.8087774
142	143	0.8089330
244	245	0.8093306
206	207	0.8114169
1	97	0.8221757

```
clist <- FredsDBSCAN(adjmat, 3, c(1:257))
cat("First Cluster: ", clist[[1]], "\n")
```

```
## First Cluster:  1 6 20 21 23 25 28 34 35 44 48 50 80 107 131 180
```

```
adjmat[clist[[1]], clist[[1]]] %>%
  graph_from_adjacency_matrix(mode = "undirected") %>%
  plot()
```