Final Project

Christian Uriostegui

2022-12-04

Objective

The Sopranos is a critically acclaimed tv series that stars fictional mob boss Tony Soprano. The show takes viewers through a look at his internal struggles with balancing his family life and his illegal lifestyle.

As a fan of the show, I was particularly drawn to the topics that were covered in the series. There was sprinkles of political commentary about the growing xenophobia of that period during the war on terror. Viewers are also privy to Tony and his therapist's conversations about freewill and determinism. It also also questions about morale and whether violent individuals can be redeemed.

In my study, I will be performing sentiment/text analysis on the pilot episode of the Sopranos and see whether the text foreshadows any of the critical themes or events of season 1 and the rest of the show.

Given that Tony's therapy sessions with Dr. Melfi are a frequent occurrence in the beginning of the show and device that moves the plot, I expect to see the foreshadowing of events and conflicts in season 1.

Load Library

```
library(readr)
library(dplyr)
library(tidyverse)
library(stringr)
library(knitr)
library(tidytext)
library(tidyr)
library(magrittr)
library(quanteda)
library(tm)
library(e1071)
library(wordcloud)
library(quanteda.textplots)
library(quanteda.textstats)
library(SentimentAnalysis)
```

Load Data

```
# I will be downloading the pilot script directly from my github
data <- 'https://raw.githubusercontent.com/curiostegui/CUNY-SPS/main/Data%20607/Final%20Project/Season1-
sopranos_ep1 <- read.csv(file = data, header = TRUE, sep = ",")
glimpse(sopranos_ep1)
```

```
## Rows: 596
## Columns: 3
## $ Index     <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17~
## $ Character <chr> "MELFI", "MELFI", "MELFI", "TOMMY", "MELFI", "MELFI", "TOMMY~
## $ Script    <chr> "Mr. Soprano'?", "Have a seat.", "My understanding from your~
```

The dataset show above contains the following columns:

Index: Number of row

Character: The character the line of dialogue is for.

Script: Dialogue that is read by the character.

```
# Used the function to see the characters with dialogue in this episode
unique(sopranos_ep1$Character)
```

```
##  [1] "MELFI"            "TOMMY"            "HUNTER"           "CARMELA"
##  [5] "MEADOW"           "KIDS"             "CHRIS"            "MAHAFFEY"
##  [9] "BIG PUSSY"        "DICK BARONE"      "SILVIO"           "PAULIE WALNUTS"
## [13] "JUNIOR"           "ARTHUR"           "VOICE"            "LIVIA"
## [17] "FATHER PHIL"      "BOTH PARENTS"     "KOLAR"            "BEPPY"
## [21] "TOMMY JR"         "HERMAN"           "DIRECTOR"         "COMPUTER VOICE"
## [25] "CHARMAINE"        "NILS"             "HOSTESS"          "OWNER"
## [29] "IRINA"            "ANNOUNCER VOICE"  "YOUNG WOMAN"      "TOMMY AND SILVIO"
```

When checking the names of the characters in this script, I observed some typos. In this script, for some reason, Tony's name is spelled as "TOMMY". I also see a typo for lines involving Anthony Jr. His name is listed as "TOMMY JR".

```
sopranos_ep1$Character <- str_replace(sopranos_ep1$Character,"TOMMY JR","ANTHONY JR")
sopranos_ep1$Character <- str_replace(sopranos_ep1$Character,"TOMMY","TONY")
```

Overview of Character Lines

We can see in our table that Tony, Melfi and Carmela, land in the top 3 in characters with most lines this episode.

```
sort(desc(table(sopranos_ep1$Character)))
```

```
##
##           TONY          MELFI        CARMELA          CHRIS         MEADOW
##           -209            -68            -57            -42            -32
##          LIVIA       BIG PUSSY         HERMAN         ARTHUR       MAHAFFEY
##            -31            -19            -18            -15            -15
```

```
##          JUNIOR          SILVIO       CHARMAINE     FATHER PHIL          KOLAR
##             -14              -9              -8              -8              -7
##          HUNTER            NILS  PAULIE WALNUTS      ANTHONY JR          VOICE
##              -5              -5              -5              -4              -4
##     DICK BARONE        DIRECTOR           IRINA           BEPPY         HOSTESS
##              -3              -3              -3              -2              -2
##           OWNER ANNOUNCER VOICE    BOTH PARENTS  COMPUTER VOICE            KIDS
##              -2              -1              -1              -1              -1
## TONY AND SILVIO     YOUNG WOMAN
##              -1              -1
```

Data Cleaning

Before examining, we need to tokenize the script by breaking down the lines into a word per column as well as removing stop words.

```
clean_sop <- sopranos_ep1 %>%
  unnest_tokens(word, Script) %>% # lines of script are broken down to words
  mutate(linenumber = row_number()) %>% # added column to count each row
  select(4,2,3) # removed 'index' column it was no longer be accurate
```

```
# Used anti_join to remove stopwords
clean_sop <- clean_sop %>%
  anti_join(stop_words)
```

```
## Joining, by = "word"
```

Create Corpus

```
corp1 <- corpus(clean_sop, text_field = 'word')
corp1
```

```
## Corpus consisting of 2,094 documents and 2 docvars.
## text1 :
## "soprano"
##
## text2 :
## "seat"
##
## text3 :
## "understanding"
##
## text4 :
## "family"
##
## text5 :
## "physician"
##
```

```
## text6 :
## "dr"
##
## [ reached max_ndoc ... 2,088 more documents ]
```

```
corp2 <- corpus(sopranos_ep1, text_field = 'Script')
corp2
```

```
## Corpus consisting of 596 documents and 2 docvars.
## text1 :
## "Mr. Soprano'?"
##
## text2 :
## "Have a seat."
##
## text3 :
## "My understanding from your family physician, Dr. Cusamano, i..."
##
## text4 :
## "They said it was a panic attack -- because all the neurologi..."
##
## text5 :
## "You don't agree you had a panic attack?"
##
## text6 :
## "How are you feeling now?"
##
## [ reached max_ndoc ... 590 more documents ]
```

Thoughts Before Analysis

I expect that when performing my analysis, I will see words that tie to important thematic elements of the show. In season 1 the audience become aware of Tony's panic attacks. It explore his dissatisfaction with today's world, his father's mob lifestyle, and troubled relationships in his family, in particular with his mother and uncle - another mob boss with whom he has a power struggle with.

Tony's inability to cope with his depression and pressure as boss of his crew, also affects his children and wife, as he is often absent or too irritable to help with domestic conflicts. I also anticipate seeing lots of associations with family or conflict with family members.

Also, because this episode explores Tony's depression, I expect there to be more negative and positive associations in my analysis.

Document Term Matrix

In order to generate our visual insights , we now need to transform `corp1` into a document term matrix. While doing so, I continue to clean the data by making changes such as removing punctuation marks, symbols, and numbers.

```
dtm1 <- corp1 %>%
  tokens(remove_punct=T, remove_numbers=T, remove_separators=T,remove_symbols=T) %>% # further cleaning
  tokens_tolower() %>% # turned all words to lowercase
  tokens_remove(stopwords('en')) %>% # removed stopwords again just in case
  tokens_wordstem() %>% #reduced words to their stem root
  dfm() # document term matrix creation
dtm1
```

```
## Document-feature matrix of: 2,094 documents, 1,053 features (99.91% sparse) and 2 docvars.
##        features
## docs    soprano seat understand famili physician dr cusamano collaps unabl
##   text1       1    0          0      0         0  0        0       0     0
##   text2       0    1          0      0         0  0        0       0     0
##   text3       0    0          1      0         0  0        0       0     0
##   text4       0    0          0      1         0  0        0       0     0
##   text5       0    0          0      0         1  0        0       0     0
##   text6       0    0          0      0         0  1        0       0     0
##        features
## docs    breath
##   text1      0
##   text2      0
##   text3      0
##   text4      0
##   text5      0
##   text6      0
## [ reached max_ndoc ... 2,088 more documents, reached max_nfeat ... 1,043 more features ]
```

Most Frequent Words

```
textplot_wordcloud(dtm1, max_words = 50, colors = brewer.pal(8, "Dark2"))
```

write observations

```
topfeatures(dtm1, 20)
```

```
##   uncl fuckin   feel   talk    guy   fuck mother   duck   time   life   busi
##     22     21     20     20     18     18     16     15     15     14     14
##   home junior  pussi  peopl  gonna father  money    run friend
##     13     13     13     12     12     11     11     11     10
```

Family: The most frequent word is "uncle", which is not surprising, given their power struggle. Tony is conflicted in the episode because he views Uncle Junior as a father figure who respects, but doesn't feel he gets the same respect in return. Uncle Junior is planning to attack one of Tony's associates to the dismay of Tony.

We also see the word "junior" which is his uncle's name. Other mentions of his family includes "mother", and "father"

Ducks: Another interesting appearance is the word "duck". Tony in the show is attached to animals, particularly a flock of docks that frequently visits his pool but have now dissappeared. Dr. Melfi, suggests the ducks flying away symbolizes the turmoil in his house and the fear of his family leaving him too.

Feelings There are associations with feelings explored in this episode such as "feel", "life", and "home". Since Tony's therapy session is center in the episode, it is not surprising to see these words come up.

Words in Context

To further understand why these words appear frequently, I decide to explore their appearance through the function `kwic` or key words in context.

```
k = kwic(corp2, 'uncle', window = 5) # I utilize corp2 because it contains the script's full lines
as_tibble(head(k, 8))
```

```
## # A tibble: 8 x 7
##   docname  from    to pre                          keyword post           pattern
##   <chr>   <int> <int> <chr>                        <chr>   <chr>          <fct>
## 1 text104    10    10 "up . It involves my"        uncle   . I can't go ~ uncle
## 2 text106     7     7 "I'll say this - my"         uncle   adds to my ge~ uncle
## 3 text107    20    20 ", the word is your"         Uncle   Junior is goi~ uncle
## 4 text109     1     1 ""                           Uncle   Jun ' , how .  uncle
## 5 text116     7     7 "better sit down with your" uncle   .              uncle
## 6 text117     1     1 ""                           Uncle   Junior is my ~ uncle
## 7 text117    35    35 ". , -I love my"             uncle   .              uncle
## 8 text118     6     6 "At the same time ,"         Uncle   Junior also t~ uncle
```

The words "uncle" appear in Tony's conversations to Dr. Melfi and his crew regarding his Uncle Junior. It shows the contrast of his feeling towards his uncle. On one side he love him but also mentions that he is a source of his stress on a personal and professional level.

```
k1 = kwic(corp2, 'feel*', window = 5) # added * to account multiple variation of feel (for ex: feels an
as_tibble(head(k1, 8))
```

```
## # A tibble: 8 x 7
##   docname  from    to pre                         keyword post           pattern
##   <chr>   <int> <int> <chr>                       <chr>   <chr>          <fct>
## 1 text6       4     4 How are you                 feeling now ?          feel*
## 2 text15     42    42 But lately I'm getting the  feeling Imight be in a~ feel*
## 3 text16      7     7 Americans , I think ,       feel    this . *       feel*
## 4 text18      5     5 Did you have this           feeling of loss more a~ feel*
## 5 text39      3     3 My wife                     feels   this friend of~ feel*
## 6 text215     5     5 Vesuvio is where Pussy      feels   safe ! He's be~ feel*
## 7 text225     6     6 Vesuvio is where . Pussy    feels   safe ! He's be~ feel*
## 8 text252    10    10 Dr . Cusamano you were      feeling depressed ?    feel*
```

The word feel appears frequently in the context of Tony's therapy sessions with Dr. Melfi where is expressing his feelings about personal and political topics.

```
k2 = kwic(corp2, 'duck*', window = 5)
as_tibble(head(k2, 8))
```

```
## # A tibble: 8 x 7
##   docname  from    to pre                        keyword post             pattern
##   <chr>   <int> <int> <chr>                      <chr>   <chr>            <fct>
## 1 text19     12    12 all this these two wild    ducks   had landed in my~ duck*
## 2 text22      7     7 , your father with those   ducks   .                duck*
## 3 text24      5     5 The male and female        duck    just made a home~ duck*
## 4 text31      5     5 Him . With those           ducks   .                duck*
## 5 text262     3     3 Since the                  ducks   left , I guess .  duck*
## 6 text263     2     2 The                        ducks   that preceded yo~ duck*
## 7 text500     3     3 What about                 ducks   ?                duck*
## 8 text501     2     2 The                        ducks   . Those damn duc~ duck*
```

Dr. Melfi and Tony Sopranos explores the significance of the ducks that appear in his pool and how they are possibly relation his panic attack.

Tony Vs. The Rest

Tony has the most lines in this episode compared to the characters. In this section I will look into the words that Tony specifically uses as well as explore the difference in words used between Tony and the rest of the characters,

```
tony_words = docvars(dtm1)$Character == "TONY" # create dtm that only features tony's lines
tony_dtm = dtm1[tony_words,] # here I merge tony's lines alone and the rest of the script into one dtm
```

```
textplot_wordcloud(tony_dtm, max_words = 50, colors = brewer.pal(8, "Dark2"))
```


```
topfeatures(tony_dtm, 20)
```

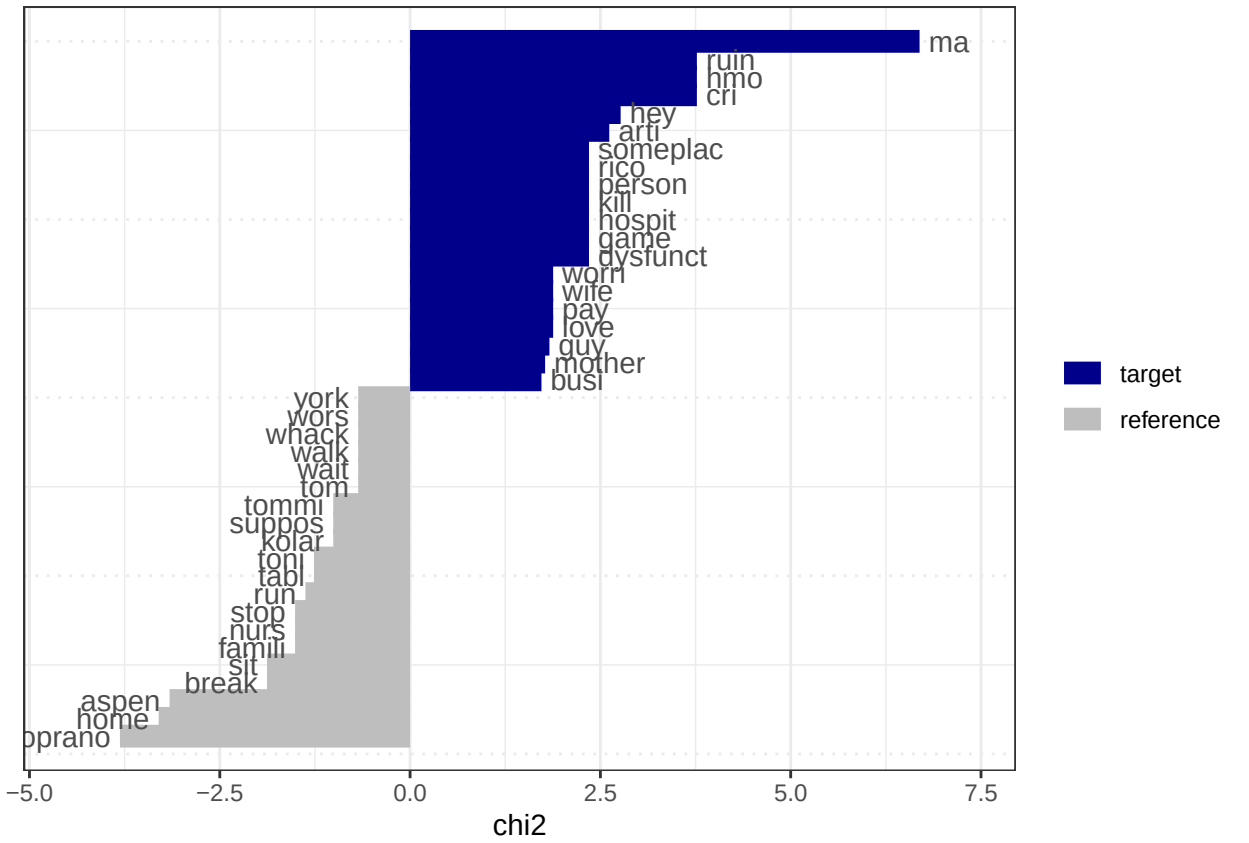
```
##     talk      guy   fuckin     uncl   mother     feel     life     busi
##       10       10       10       10        9        8        8        8
##     duck     fuck     time     arti    peopl      hey       ma      kid
##        7        7        7        7        6        6        6        5
##     told   junior daughter   friend
##        5        5        4        4
```

We can see a lot of the words that appeared most in the script overall, also make appearances in `tony_words`. We see “uncl”, “mother”, “feel”, “life”, “duck”, among other words.

```
tony_compare <- textstat_keyness(dtm1, tony_words)
head(tony_compare,20)
```

```
##     feature     chi2           p n_target n_reference
## 1        ma 6.692825 0.009680169        6           0
## 2       cri 3.767627 0.052253702        4           0
## 3       hmo 3.767627 0.052253702        4           0
## 4      ruin 3.767627 0.052253702        4           0
## 5       hey 2.765795 0.096298828        6           2
## 6      arti 2.616687 0.105745034        7           3
## 7  dysfunct 2.350596 0.125235096        3           0
## 8      game 2.350596 0.125235096        3           0
## 9    hospit 2.350596 0.125235096        3           0
## 10     kill 2.350596 0.125235096        3           0
## 11   person 2.350596 0.125235096        3           0
## 12     rico 2.350596 0.125235096        3           0
## 13 someplac 2.350596 0.125235096        3           0
## 14     love 1.879496 0.170391396        4           1
## 15      pay 1.879496 0.170391396        4           1
## 16     wife 1.879496 0.170391396        4           1
## 17    worri 1.879496 0.170391396        4           1
## 18      guy 1.830581 0.176058986       10           8
## 19   mother 1.773998 0.182888129        9           7
## 20     busi 1.725846 0.188942220        8           6
```

```
textplot_keyness(tony_compare)
```



The textplot_keyness function looks at both what words was most used by Tony compared to those in the show and the opposite end of that as well. We can see that this episode was centered around Tony.

In the target end we see themes visited in the show such as Tony's trouble relationships: "ma" (his mom), "arti" (artie, a friend), "mother", "wife", etc.

On the reference side we see "toni", "soprano" and "tommi" (typo, it's supposed to say Tony), which of course refer to Tony Soprano himself. Since he is at the center of this episode, his name is brought up frquently in dialogue.

Sentiment Analysis

For the sentiment analysis, I will be examining how negative words were used compared to positive words in the script. I will be utilizing the DictionaryGI lexicon from the `SentimentAnalysis` package. To look at the totals, I will first clean the dictionary, turn into a dataframe and then merge it with the `clean_sop` dataset where the script is.

Load Lexicon

```
data(DictionaryGI)
str(DictionaryGI)
```

```
## List of 2
##  $ negative: chr [1:2005] "abandon" "abandonment" "abate" "abdicate" ...
##  $ positive: chr [1:1637] "abide" "ability" "able" "abound" ...
```

Cleaning Lexicon Dataframe

To avoid errors when turning the list into a dataframe, I made the lengths of both rows the same.

```
length(DictionaryGI$positive) <- length(DictionaryGI$negative)
```

I then turned the list object into a dataframe.

```
DictionaryGI_df <- as.data.frame(DictionaryGI)
```

To create our tidy format, I need to join the columns vertically. Before doing so, I seperated the negative and positive columns into their own objects, and added a label that would identify it's sentiment. Lastly, when joining I also made sure to omit Null values that were added when I made the `positive` and `negative` column of equal length.

```
negative <- DictionaryGI_df$negative
negative <- as.data.frame(negative)
negative <- negative %>%
  mutate(sentiment = "negative") %>%
  rename("word"="negative")
```

```
positive <- DictionaryGI_df$positive
positive <- as.data.frame(positive)
positive <- positive %>%
  mutate(sentiment="positive") %>%
  rename("word"="positive")
```

```
Lex_DictionaryGI <- bind_rows(positive, negative)
Lex_DictionaryGI <- Lex_DictionaryGI %>%
  na.omit()
```

Merge to Soprano Dataset

```
sop_sentiment <- clean_sop %>%
  inner_join(Lex_DictionaryGI)
```

Total Sentiment Words

```
table(sop_sentiment$sentiment)
```

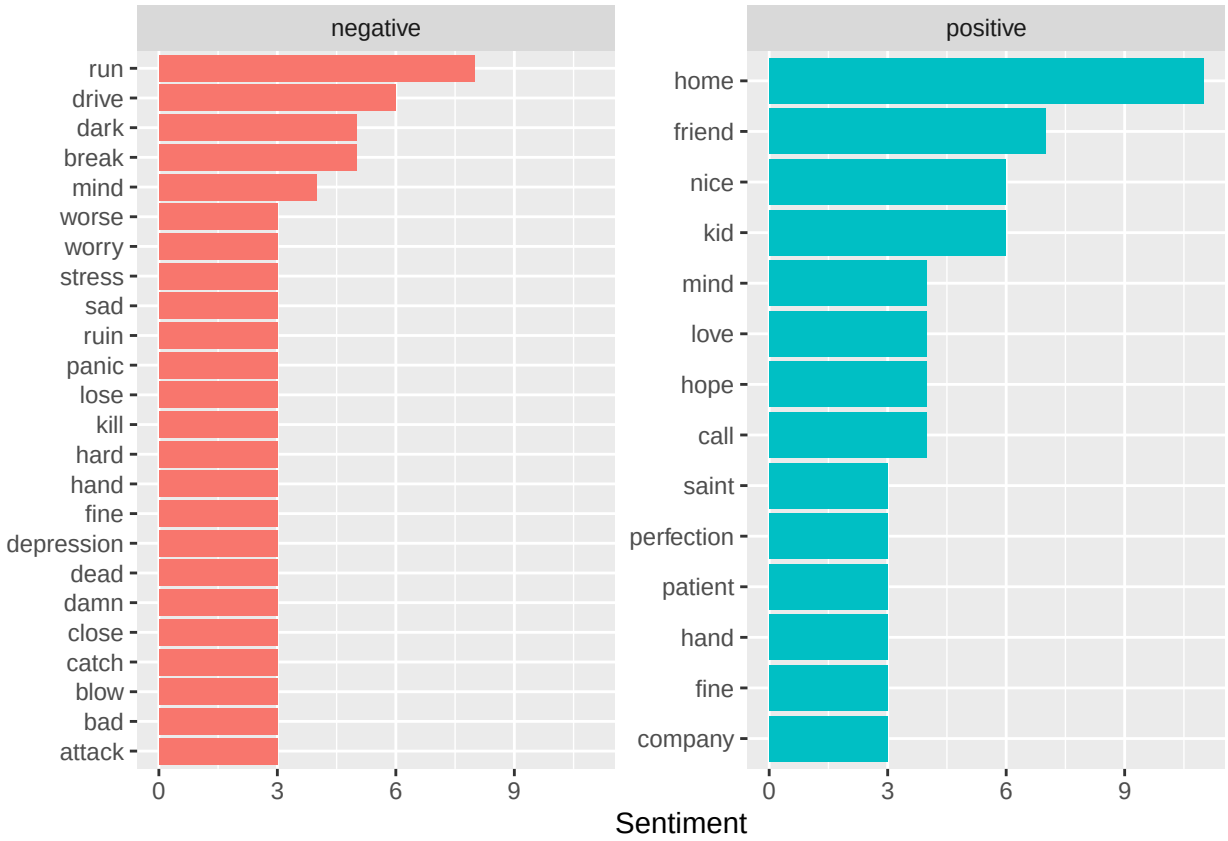
```
##
## negative positive
##      185      169
```

As expected we can see the the total number of negative numbers (52%) exceeds those that are positive. However, it is worth noting that in the join between the dictionary and sopranos dataset, that our the number of observations from our dataset shrunk from 2,094 to 354.

Top Negative & Positive Words

```
sop_sentiment_count <- sop_sentiment %>%
  inner_join(Lex_DictionaryGI) %>% # join with dictionary
  count(word, sentiment, sort = TRUE) %>% # count words
  ungroup()
```

```
sop_sentiment_count %>%
  group_by(sentiment) %>%
  slice_max(n, n = 10) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(n, word, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(x = "Sentiment",
       y = NULL)
```



Some of the classification is questionable such as "home" being labeled as positive. Home based on how Tony describes it, can be a very stressful place.

What stands out to me is the words under negative such as "mind", "worry", "stress", "sad", "ruin", "panic","depression", etc. This feels very accurate given that we the viewer are exploring the reason for his panic attack - which caused him to pass out.

Conclusion

Ultimately, in the pilot episode script we do end up seeing many recurring themes that are explored in this season. We see mentions of conflicts at home, with his mother, uncle and wife; internal disturbance, and feeling this family is falling apart.

Not only does this set up many obstacles for Tony this season, but for the remainder of the show. We will witness an attempt on Tony's life by someone in his family later this season. The discussion about his fear of losing his family foreshadows a separation between Tony and his wife Carmela that happens in the show.

If I had to revisit this study I would choosing a different dictionary. Though the sentiment analysis found that there were more negative words, a lot of the classification wasn't fully accurate. There were a lot of words removed as well when joining the dictionary and the script, which makes the results less reliable.