Analysis of “Star Trek: The Next Generation” Using NLP

John Michitsch

05-27-2021

Table Of Contents

01 Objectives

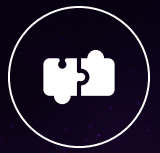
“Space... The final frontier. These are the voyages of the starship Enterprise. Its continuing mission, to explore strange new worlds. To seek out new life and new civilizations. To boldly go where no one has gone before.”

Intro narration

Goals

EDA

Understand the scripts / description data



NLP

Use NLP algorithms to draw insights from scripts

Clustering

Can we find groups in the data

Background

What?

- “Star Trek: The Next Generation” was an American science fiction TV show
- Aired from 1987-1994

Background

What?

- “Star Trek: The Next Generation” was an American science fiction TV show
- Aired from 1987-1994

When?

- Aired from 1987-1994
- 178 episodes

Background

What?

- “Star Trek: The Next Generation” was an American science fiction TV show
- Aired from 1987-1994

When?

- Aired from 1987-1994
- 178 episodes

Why?

- This could be interesting… maybe

Plot

“Set almost 100 years after Captain Kirk's five-year mission, a new **generation** of Starfleet officers set off in the U.S.S. ... Under the command of Captain Jean-Luc Picard, the all new Enterprise NCC 1701-D travels out to distant planets to seek out new life and to boldly go where no one has gone before”

Source : IMBD

02 Methodology

Data, Tools, etc.

Data

Scripts

“Star Trek Minutiae” (www.st-minutiae.com)

Descriptions

“Star Trek Guide” (startrekguide.com)

Python Libraries

SKLearn

NLP and modeling algorithms

TextBlob

Sentiment scoring

Textatistic

Reading Scores

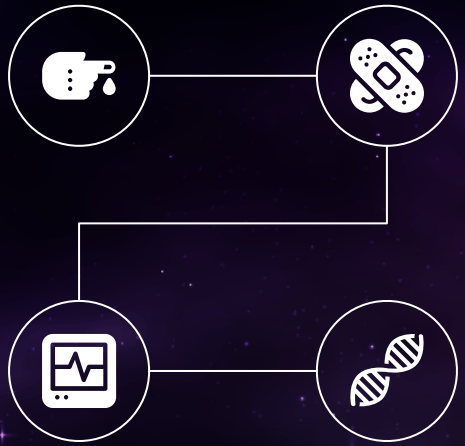
Data pipeline

Download scripts

Available as .zip file

Scrape Episode descriptions

Used BeautifulSoup to extract descriptions from HTML



Process scripts

Parse script into data frame capturing dialog by character

Data ready for NLP

03 Results Analysis

Episode Description - Topic Modeling

N-grams

1 - 3 grams

NMF

5 components model selected

Vectorizer

TFIDFVectorizer

Stopwords

English Stopwords + [‘Star’ ,’Trek’ , ‘Episode’ ,’Act’ ,..]

Topic Model - Episode Description

NMF Topic	NMF Topic	Most Common Terms	Example Episode (similar term)
1	5%	“weaponry” , “kidnaps”, “dying”	The High Ground (“kidnapping”)
2	6%	“fight”, “death”, “away mission”	Code of Honor (“fight to the death”)
3	42%	“holodeck”, “away team”, “mission”	Chain of Command (“mission”)
4	21%	“jean luc”, “picard”, “ensign ro”	Gambit, Part I (“picard”)
5	25%	“riker”, “time”, “head trip”	Hide and Q (“riker”)

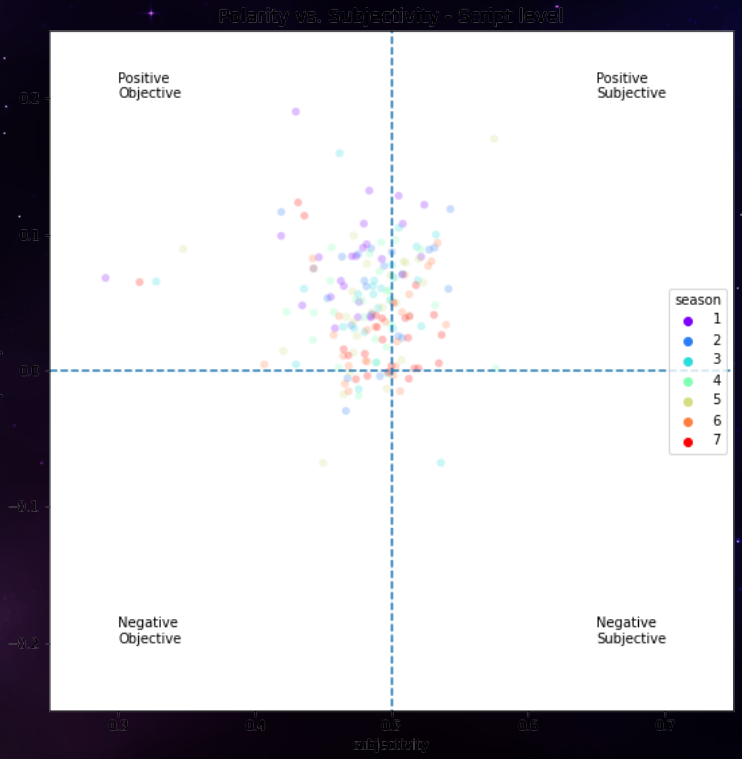
Topic Model - Episode Description

NMF Topic Label	NMF Topic	Most Common Terms	Example Episode (similar term)
Kidnap	5%	“weaponry” , “kidnaps”, “dying”	The High Ground (“kidnapping”)
Fight	6%	“fight”, “death”, “away mission”	Code of Honor (“fight to the death”)
MIssion	42%	“holodeck”, “away team”, “mission”	Chain of Command (“mission”)
Picard	21%	“jean luc”, “picard”, “ensign ro”	Gambit, Part I (“picard”)
Riker	25%	“riker”, “time”, “head trip”	Hide and Q (“riker”)

Script Level - Sentiment

Subjectivity

Objective to Subjective



Polarity

Positive or negative in tone

Script Level - Sentiment

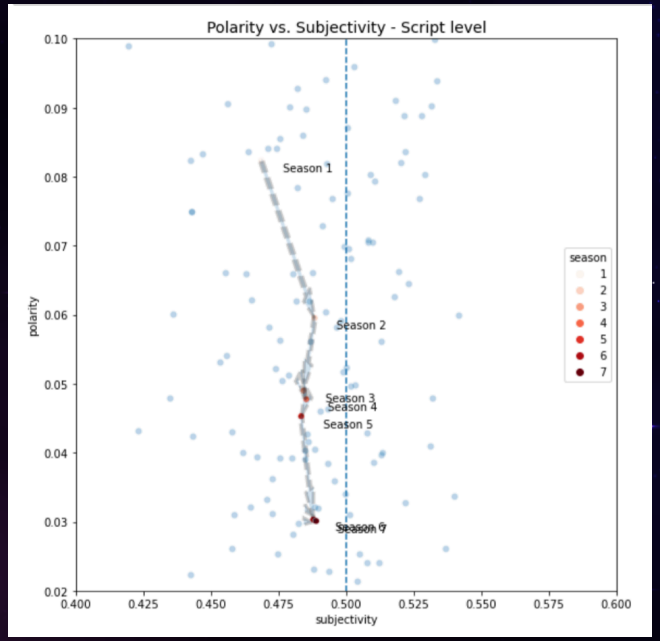
Subjectivity

Objective to Subjective

Polarity

Positive or negative in tone

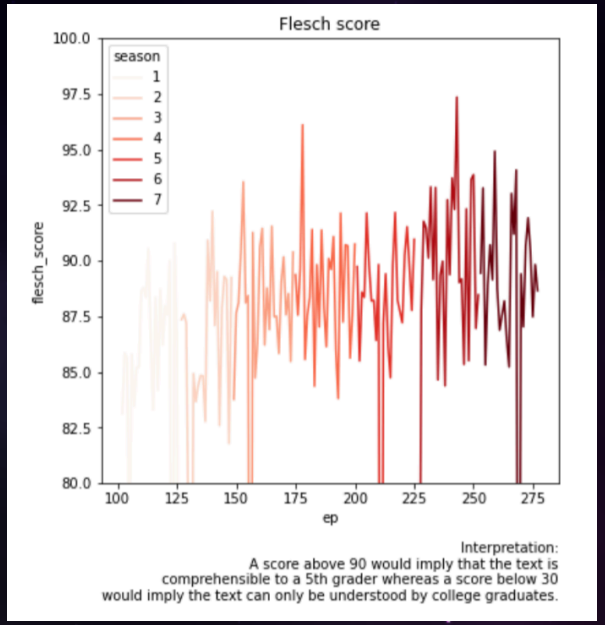
Script Level - Sentiment over time



Trend

Over the series the tone became marginally more negative

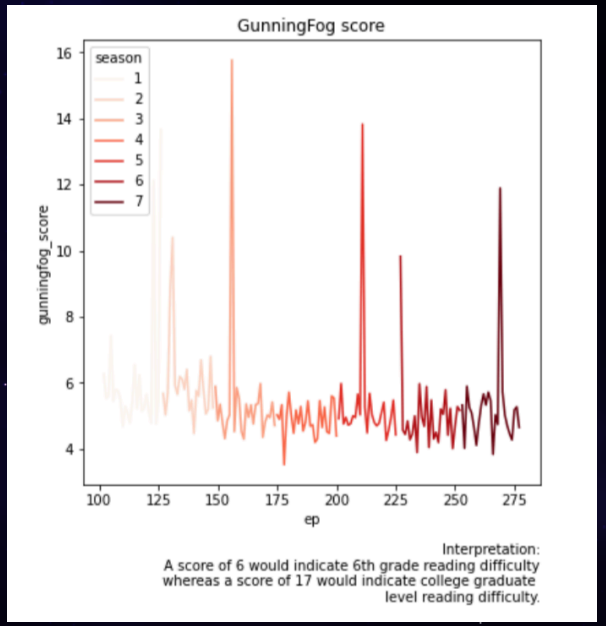
Script Level - Reading Level



Reading Level

Skill needed to comprehend the script

5th - 6th grade

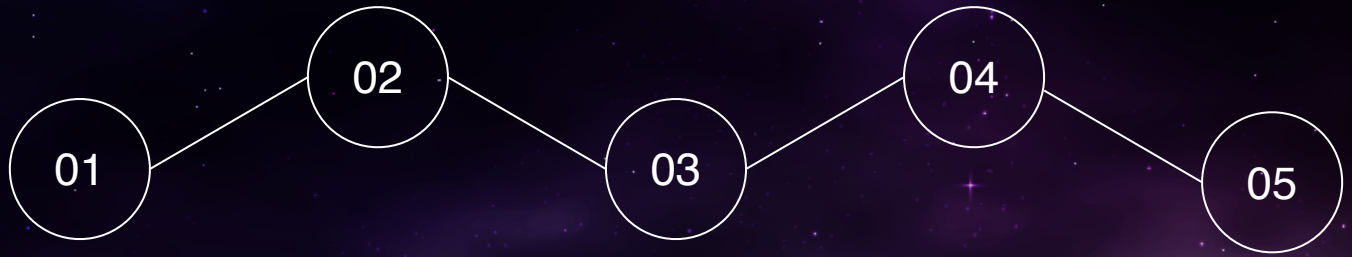


Who’s Talking?

7 Characters

Have the majority of the dialog

Episode Description - Topic Modeling



Vectorizer

TFIDFVectorizer

N-grams

1 - 3 grams

Stopwords

English Stopwords + [‘Star’ ,’Trek’ , ‘Episode’ ,’Act’ ,..]

TruncatedSVD

50 components model selected

K-means

6 cluster model selected

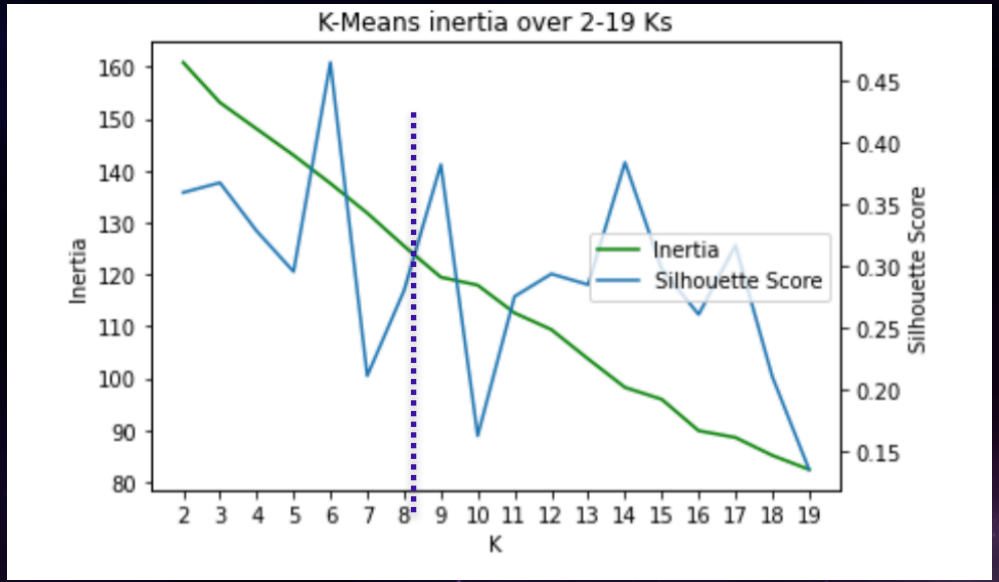
‘Picard’ vs. ‘Riker’ Top Topics

Cluster	%	Most Common Terms
1	90%	All else
3	3%	“sir”
4	3%	“Captain”
0	3%	“Yes”, “Yes, exactly”
2	1%	“Beverly…” or “Beverly?”

Cluster	%	Most Common Terms
0	88%	All else
4	3%	“Neutral Zone”
1	3%	“Captain”
2	3%	“Aye, Sir”, “No Sir”
3	3%	“Yes”, “Yes, sir”

Inertia & Silhouette Scores over 2-19 Ks

Choose K=6 For 'Picard' model



01 Conclusions

Here you can write a subtitle if you need it

Conclusions

- Finding the right chunks of text can be a challenge
  - Isolated lines of dialog maybe too short
  - Entire scripts with stage directions might be too messy

- Not surprisingly language used will differ by character

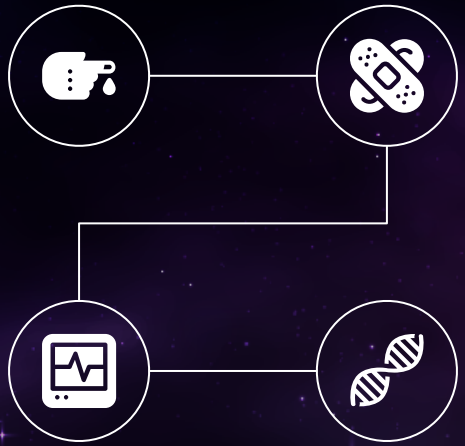
Future Work

Different Clustering

Mean shift, DBSCAN

Different Tokenization

Tweak Min_df and Max_df



Other NLP

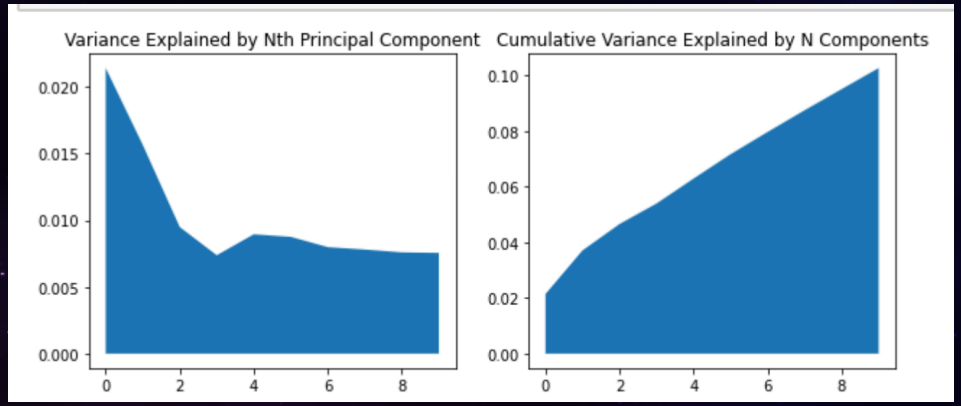
SpaCY, Gensim

Trends

Dialog shifts over time?

Appendix

TruncatedSVD Variance Explained -‘Picard’ model



Thanks