STAR TREK

THE NEXT GENERATION

PREDICTING IMDB RATINGS USING NLP & MACHINE LEARNING

CAPSTONE PROJECT BY KATYA KOGAN

WHAT IS STAR TREK: THE NEXT GENERATION?

AIRED FROM SEPTEMBER 28, 1987 TO MAY 23, 1994

SPANS 174 EPISODES OVER SEVEN SEASONS, PLUS FOUR FILMS

AVERAGE OF 20 MILLION VIEWERS

RECEIVED MANY ACCOLADES, INCLUDING 19 EMMY AWARDS, TWO HUGO AWARDS, FIVE SATURN AWARDS, AND A PEABODY AWARD

THE MISSION

PREDICT THE IMDB RATINGS FROM THE SCRIPT BY EACH CHARACTER'S LINES

REAL WORLD APPLICATIONS IN THE FILM INDUSTRY

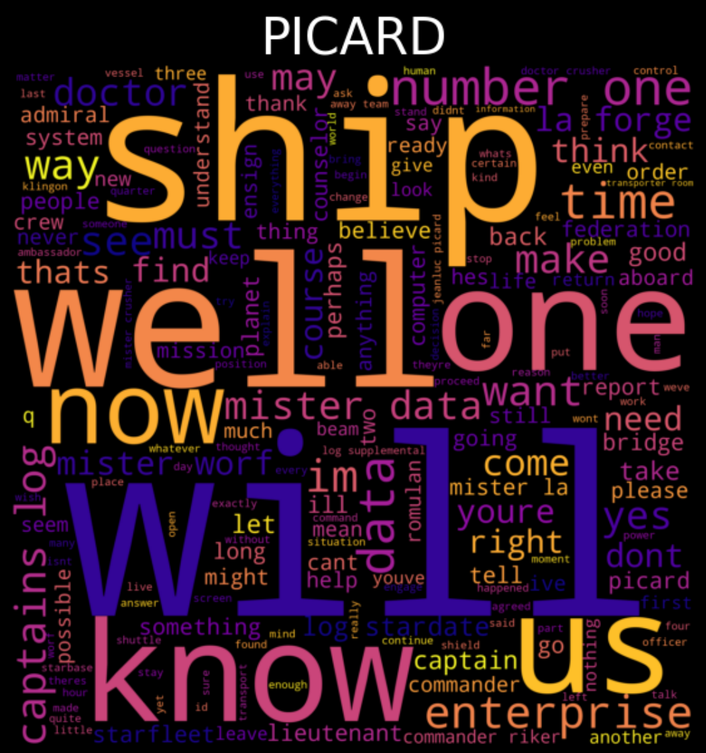
DATA ACQUISITION & PREPARATION

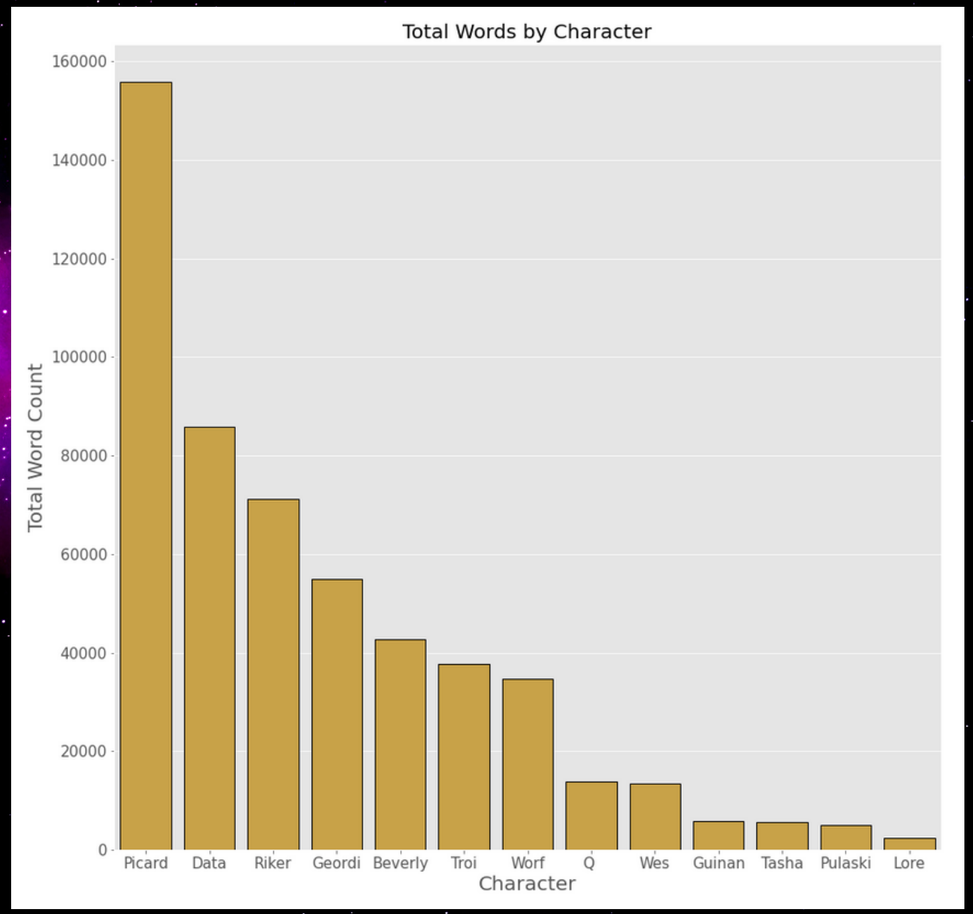
CLEANED THE DATASET

- REMOVE DUPLICATE ROWS
- CLEANED TEXT DATA
- IMPUTE/REMOVE MISSING VALUES

EXPLORED THE DATA BY THROUGH VISUALIZATIONS

VISUALIZATIONS





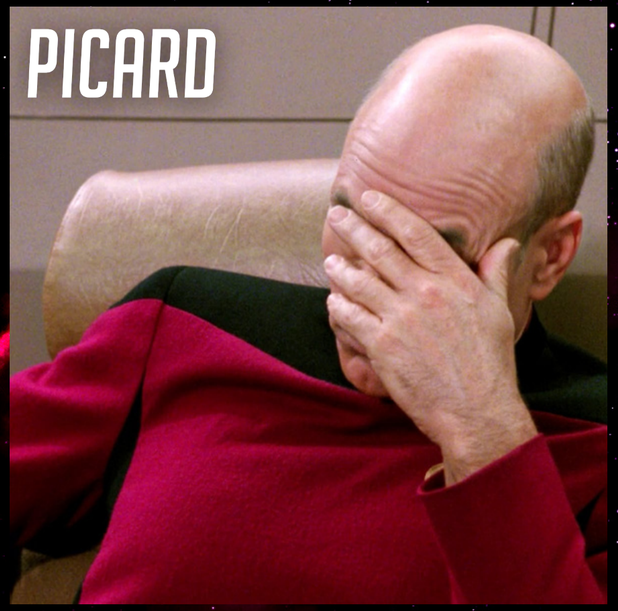
MODELLING

TESTED SEVERAL MODELS:

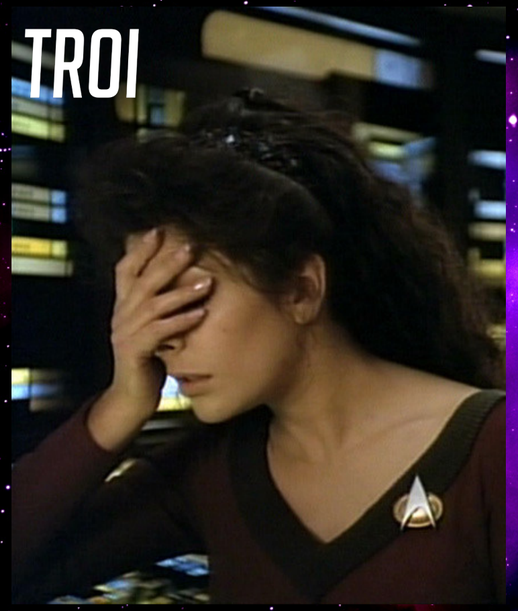
- LINEAR REGRESSION
- DECISION TREE REGRESSOR
- RANDOM FOREST REGRESSOR
- SVM REGRESSOR

PICARD_PCT	RIKER_PCT	DATA_PCT	WORF_PCT	TROI_PCT	BEVERLY_PCT	GEORDI_PCT	Q_PCT	LORE_PCT	WESLEY_PCT	GUINAN_PCT	TASHA_PCT	PULASKI_PCT	RATING
33.46	31.58	10.44	1.69	0.0	4.26	5.21	6.05	0.0	4.52	0.0	2.79	0.0	8.0
11.47	3.01	26.08	25.44	14.94	4.23	14.82	0.0	0.0	0.0	0.0	0.0	0.0	8.0
27.32	42.75	7.88	8.6	0.0	0.0	0.0	0.0	0.0	10.2	0.0	0.0	3.25	8.0
39.13	24.34	11.6	1.94	3.09	0.92	16.36	0.0	0.0	2.63	0.0	0.0	0.0	7.0
41.64	9.67	18.77	5.32	2.31	5.44	16.85	0.0	0.0	0.0	0.0	0.0	0.0	7.0
39.27	6.1	14.12	4.62	4.9	11.55	7.46	9.44	0.0	0.0	0.0	2.54	0.0	8.0
42.47	18.29	8.61	2.45	3.18	9.29	2.8	10.24	0.0	2.67	0.0	0.0	0.0	7.0
12.8	26.11	18.57	2.77	10.22	13.69	6.21	0.0	0.0	2.23	0.0	7.39	0.0	6.0
8.72	15.11	0.94	4.65	0.03	12.74	25.84	31.97	0.0	0.0	0.0	0.0	0.0	6.0
30.28	26.48	3.75	3.8	1.84	33.85	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.0
8.81	1.26	41.32	29.01	3.88	2.45	13.28	0.0	0.0	0.0	0.0	0.0	0.0	8.0
9.55	1.78	1.53	67.03	1.04	0.59	1.53	16.93	0.0	0.0	0.0	0.0	0.0	7.0
53.02	4.1	15.48	3.45	3.0	10.6	10.35	0.0	0.0	0.0	0.0	0.0	0.0	6.0
24.63	13.12	12.72	3.24	0.4	1.74	39.52	0.0	0.0	2.28	2.36	0.0	0.0	7.0
15.48	18.46	22.96	4.92	0.9	8.86	10.66	0.0	15.06	2.7	0.0	0.0	0.0	8.0

MODEL RESULTS



WORD COUNT
RATING



WORD COUNT
RATING

WORD COUNT
RATING

NEXT STEPS

USE NATURAL LANGUAGE PROCESSING TO FIND MORE CONTEXT OF THE SCRIPTS, AND HOW IT AFFECTS THE RATING

TEST WITH A NEURAL NETWORK


python
matplotlib
SHAP
seaborn
NumPy
□□TRACK MY PROGRESS □□
GITHUB.COM/KATYAZEROSS
GitHub
jupyter
scikit
learn
□□ADD ME ON LINKEDIN □□
LINKEDIN.COM/IN/KATYAKOGAN
pandas