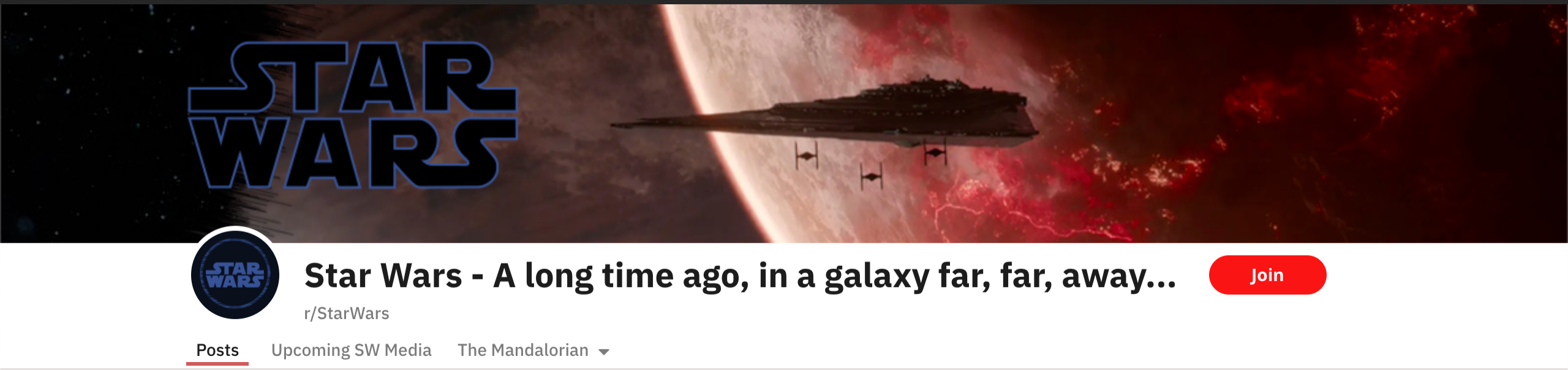
Project 3 : Web APIs & NLP.
Star Wars vs. Star Trek

Author: Artem Lukinov

Problem Statement

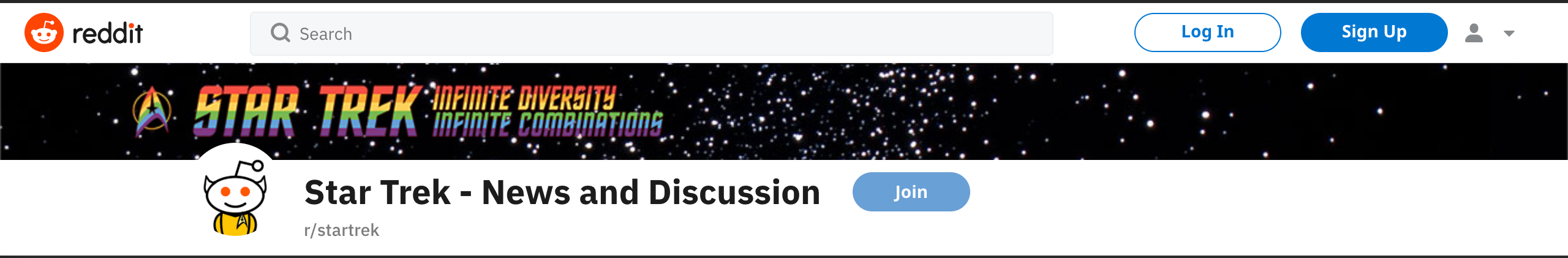
A glitch in Reddit's system caused the Star Wars and Star Trek subreddits to drop their names and scrambled all posts. The project is tasked with analyzing how different the subreddits are and developing a model to classify posts into correct subreddits.

What was looked at?



- 1.8 million users
- Created May 27, 2008

What was looked at?



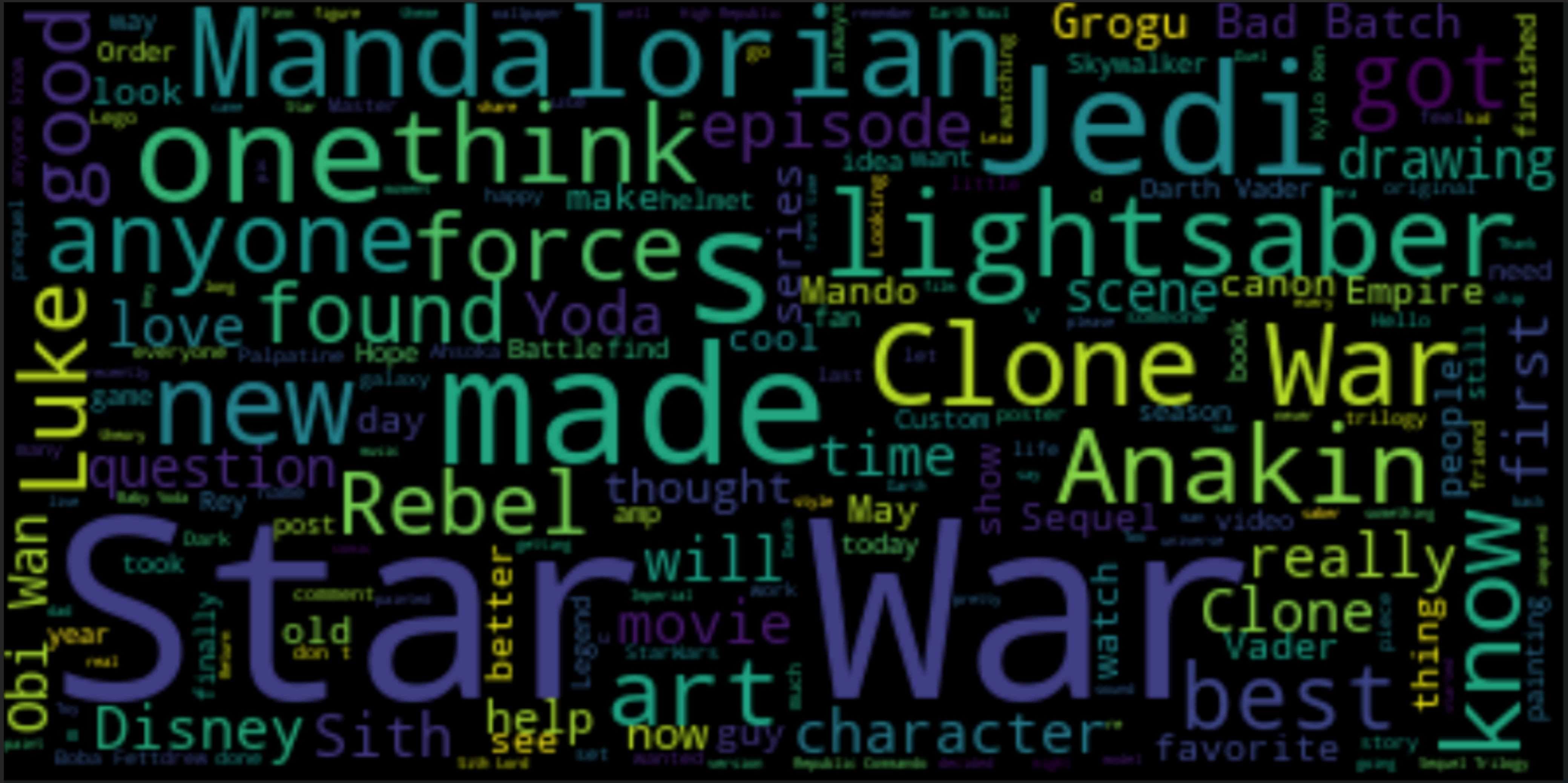
- 300,000+ users
- Created June 17, 2008

How was the data cleaned?

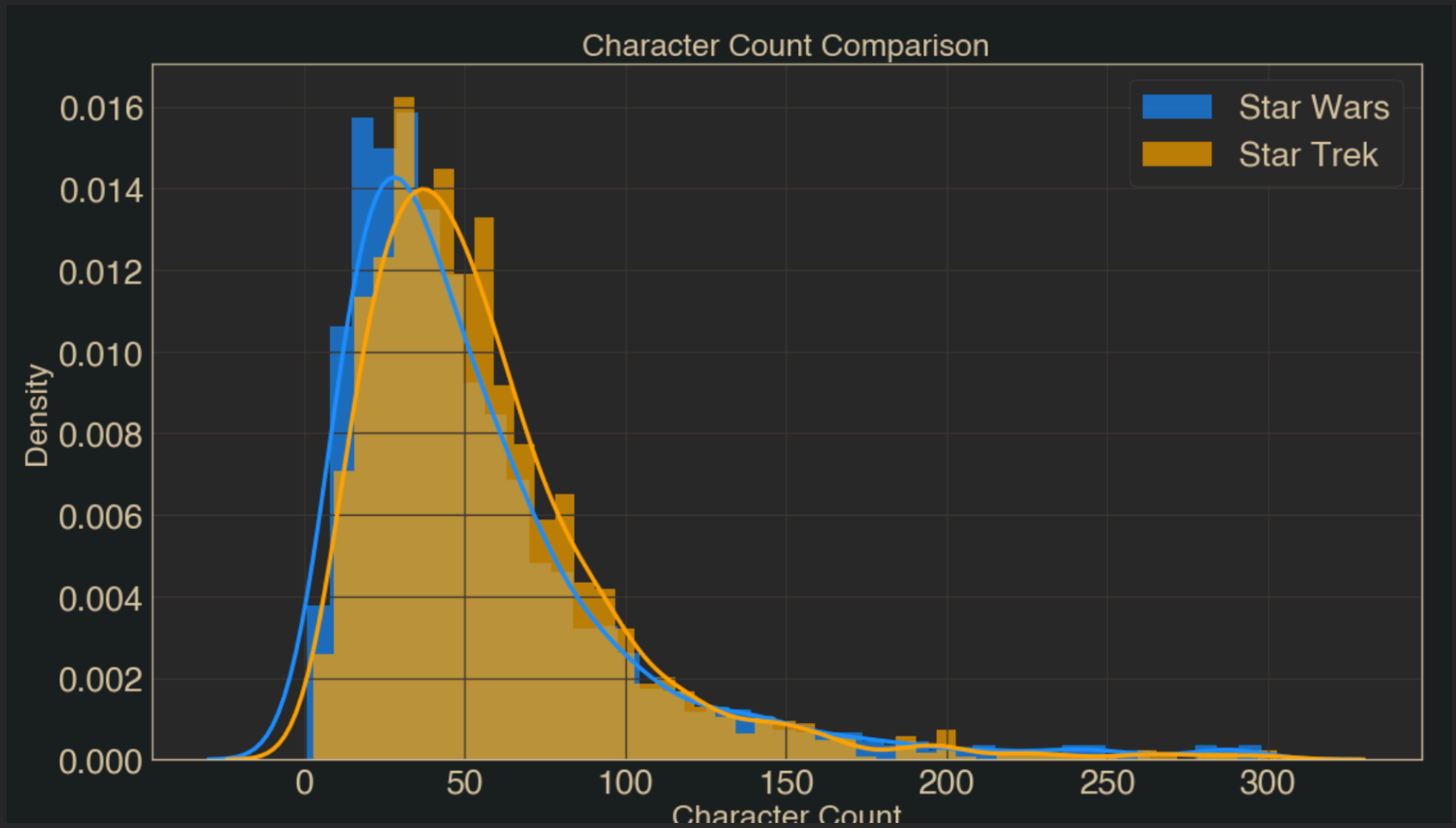
- two dataframes merged into one
- thorough analysis of 80+ columns
- three columns remained: title, content and subreddit's name
- rows with empty data were removed

Pre-processing

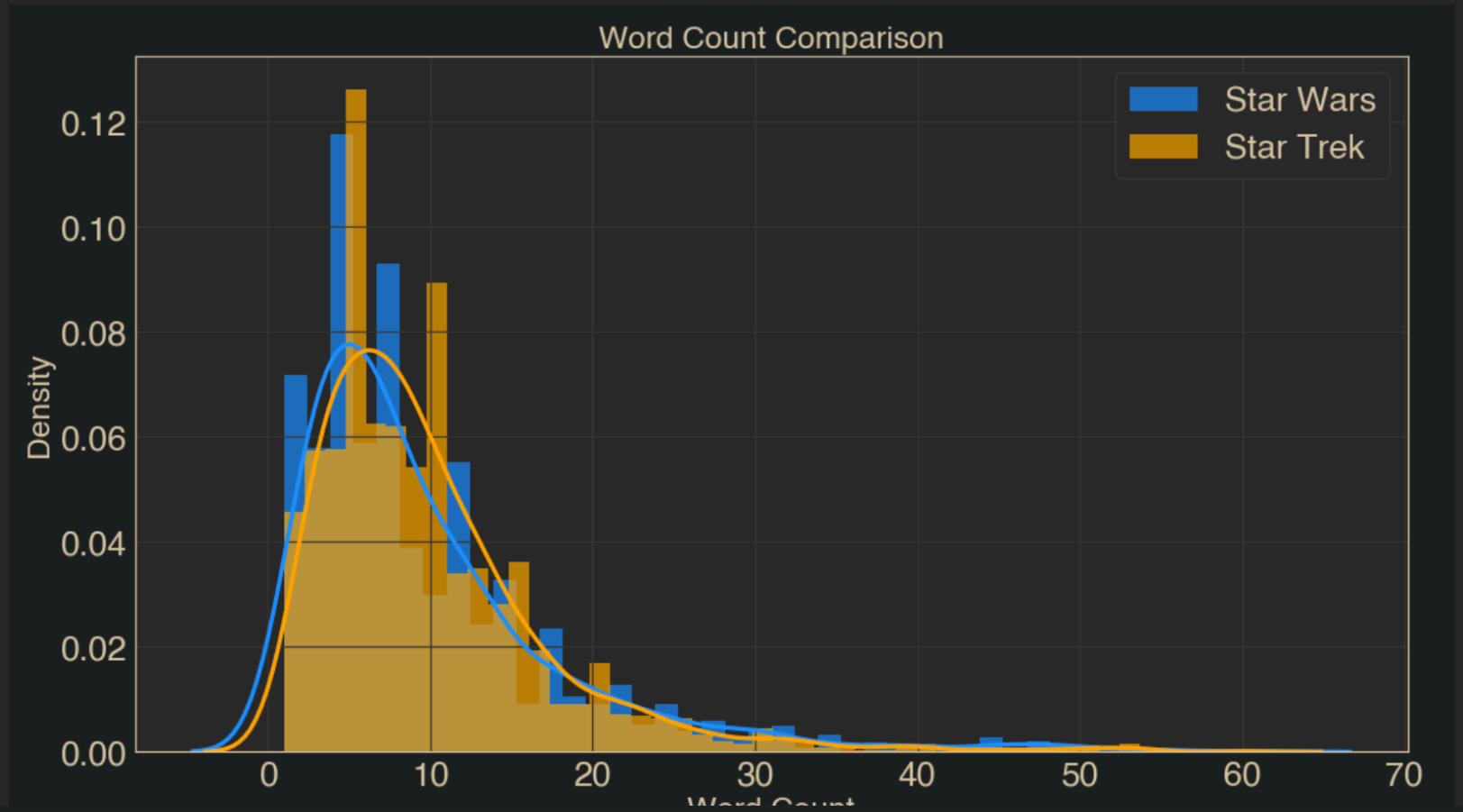
- emojis removed
- “Star Trek” and “Star Wars” as phrases removed
- URLs removed
- Lemmatization applied
- Stopwords list was augmented with additional elements


Star War
Mandalorian
Jedi
lightsaber
made
onethink
anyone
force
s
Clone War
Anakin
new
art
Luke
good
got
episode
drawing
found
Yoda
love
Rebel
question
Disney
Sith
character
best
know
first
really
Clone
Vader
time
thought
will
movie
help
see
now
Obi Wan
scene
canon
Empire
Mando
Grogu
Bad Batch
series
Sequel
favorite
people
watch
better
show
look
Order
make
helmet
cool
Darth Vader
Skywalker
day
May
today
old
year
post
game
took
idea
fan
season
trilogy
video
thing
book
story
guy
finally
Legend
Palpatine
Hope
Battle
find
Custom
poster
Rey

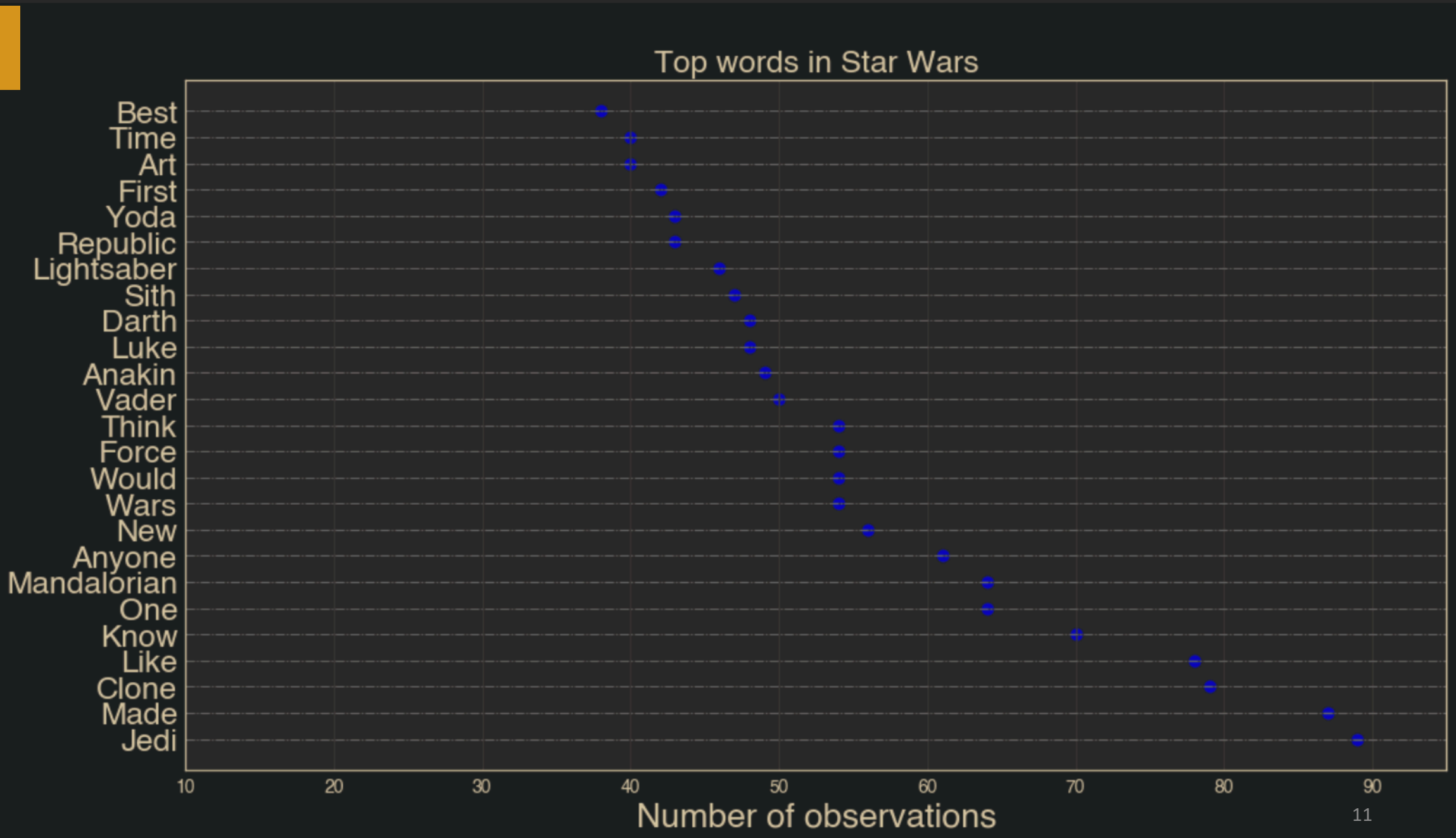

Star Trek
episode
Enterprise
DS9
TNG
Voyager
Picard
Season
Discovery
Klingon
question
think
Lower Decks
character
anyone
TOS
Federation
Starfleet
Captain
watch
series
ship
new
made
one
show
time
best
movie
will
really
Paramount
Data
much
scene
thought
love
know
see
universe
watched
Spoiler
favorite
watching
Vulcan
Borg
good
first
fan
start
thing
old
year
Sisko
never
amp
now
Romulan
make
Trek Discovery
look
Original
ST

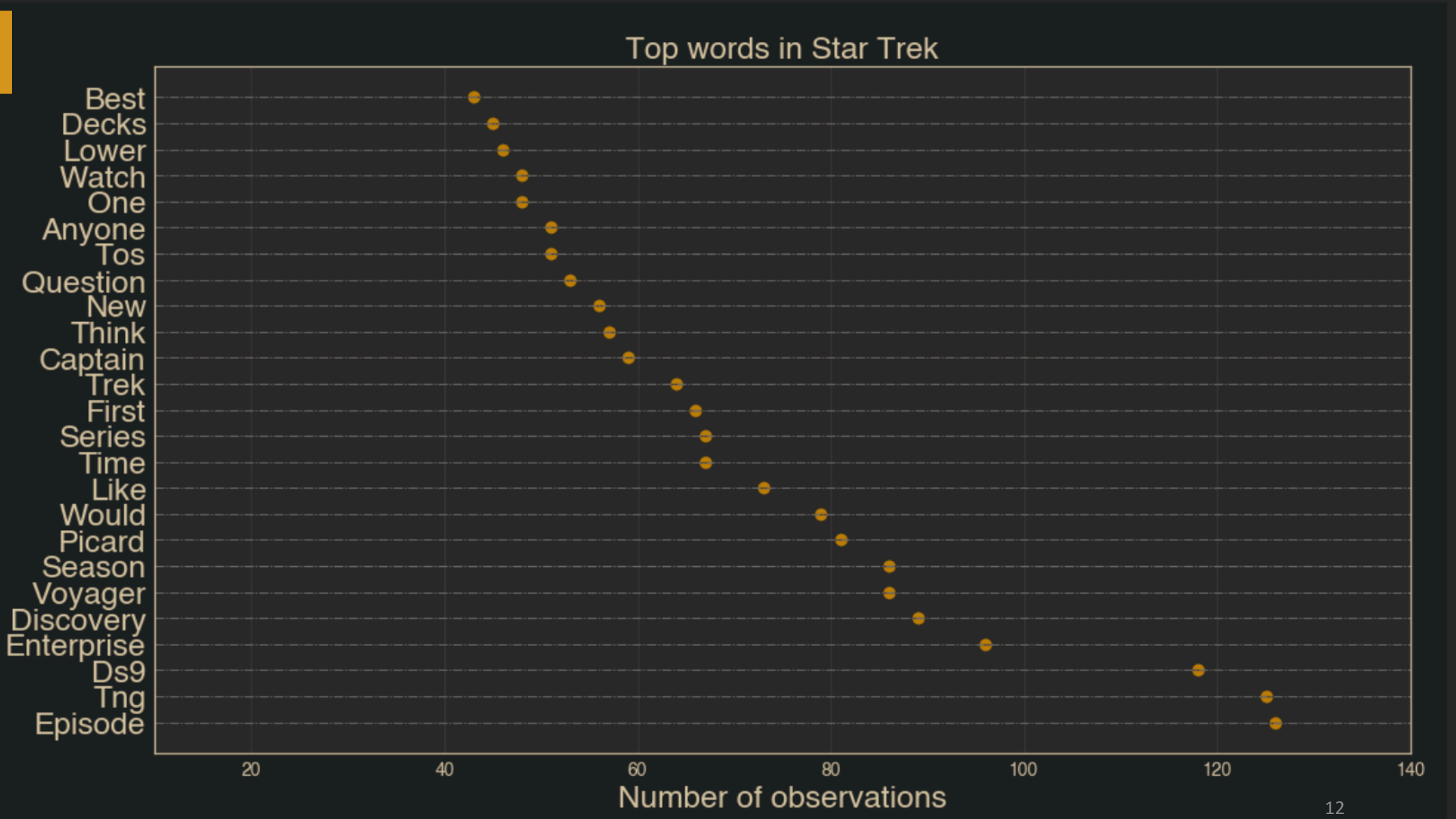


subreddit	count	mean	std	min	25%	50%	75%	max
StarWars	1989.0	54.332831	46.435670	1.0	24.0	41.0	68.0	298.0
startrek	1989.0	57.182001	40.685735	3.0	31.0	48.0	71.0	303.0



subreddit	count	mean	std	min	25%	50%	75%	max
StarWars	1989.0	9.918552	8.569247	1.0	4.0	7.0	13.0	61.0
startrek	1989.0	10.024635	7.397312	1.0	5.0	8.0	13.0	60.0


Top words in Star Wars
Best
Time
Art
First
Yoda
Republic
Lightsaber
Sith
Darth
Luke
Anakin
Vader
Think
Force
Would
Wars
New
Anyone
Mandalorian
One
Know
Like
Clone
Made
Jedi
10
20
30
40
50
60
70
80
90
Number of observations


Top words in Star Trek
Best
Decks
Lower
Watch
One
Anyone
Tos
Question
New
Think
Captain
Trek
First
Series
Time
Like
Would
Picard
Season
Voyager
Discovery
Enterprise
Ds9
Tng
Episode
20
40
60
80
100
120
140
Number of observations

Word overlap

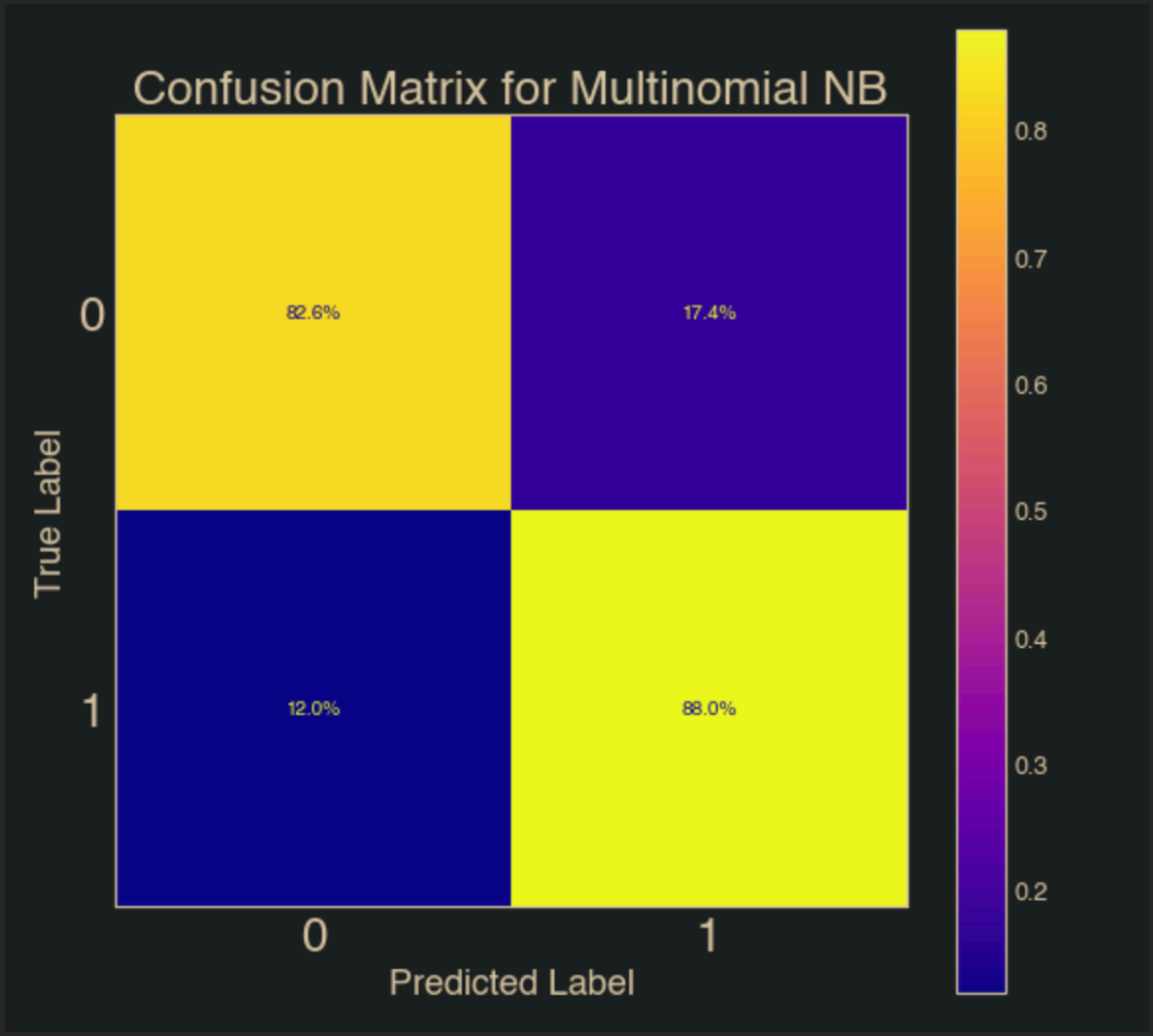
'like'
'one'
'anyone'
'new'
'would'

'think'
'first'
'time'
'best'

The Modeling process

Model	Training Score	Testing Score	Specificity	Sensitivity
Multinomial NB	93.6	85.3	82.6%	88.0%
Random Forest	98.5	81.4	75.6%	87.3%

The Modeling process



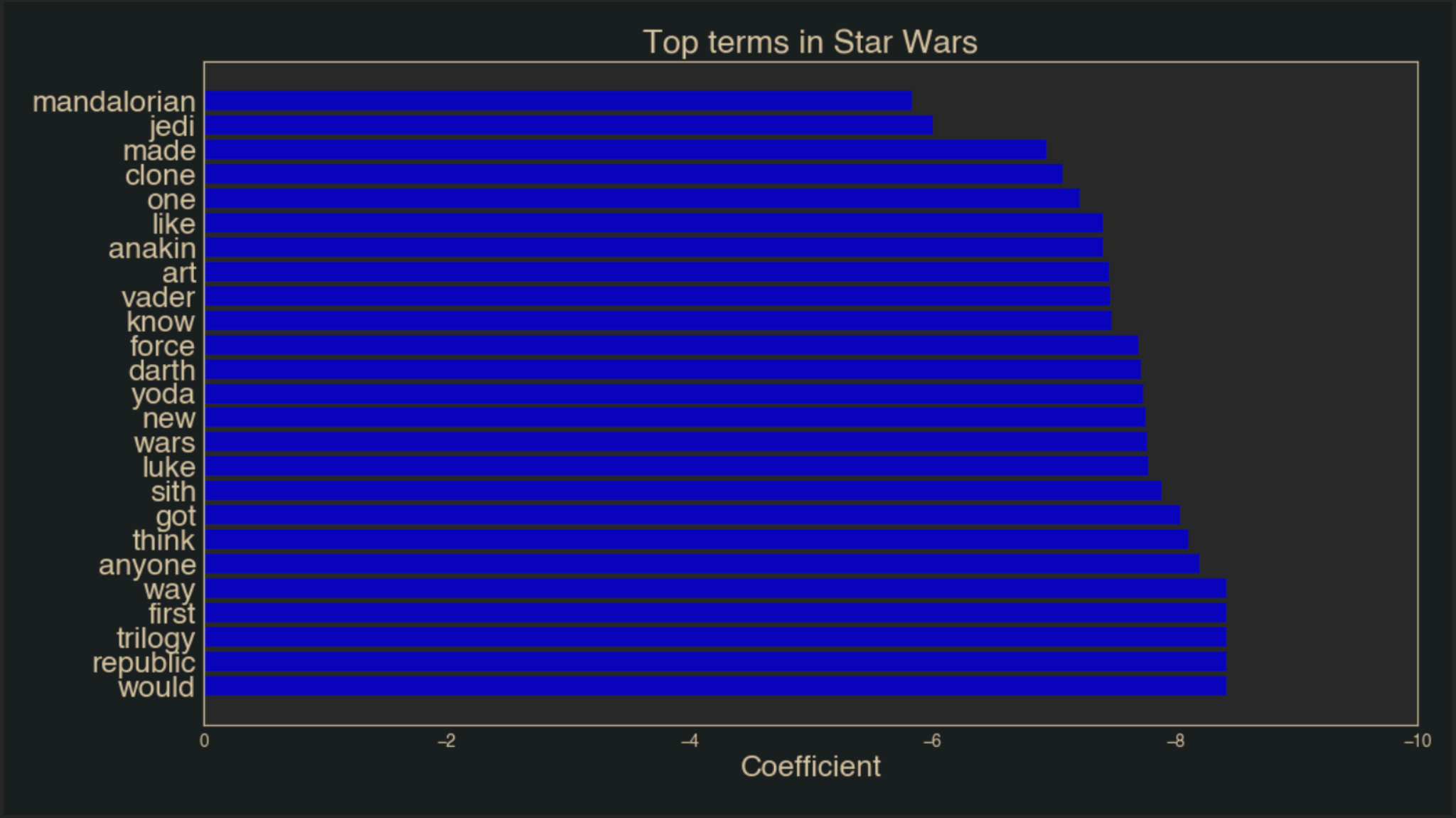
Specificity
(true negative rate):

82.6

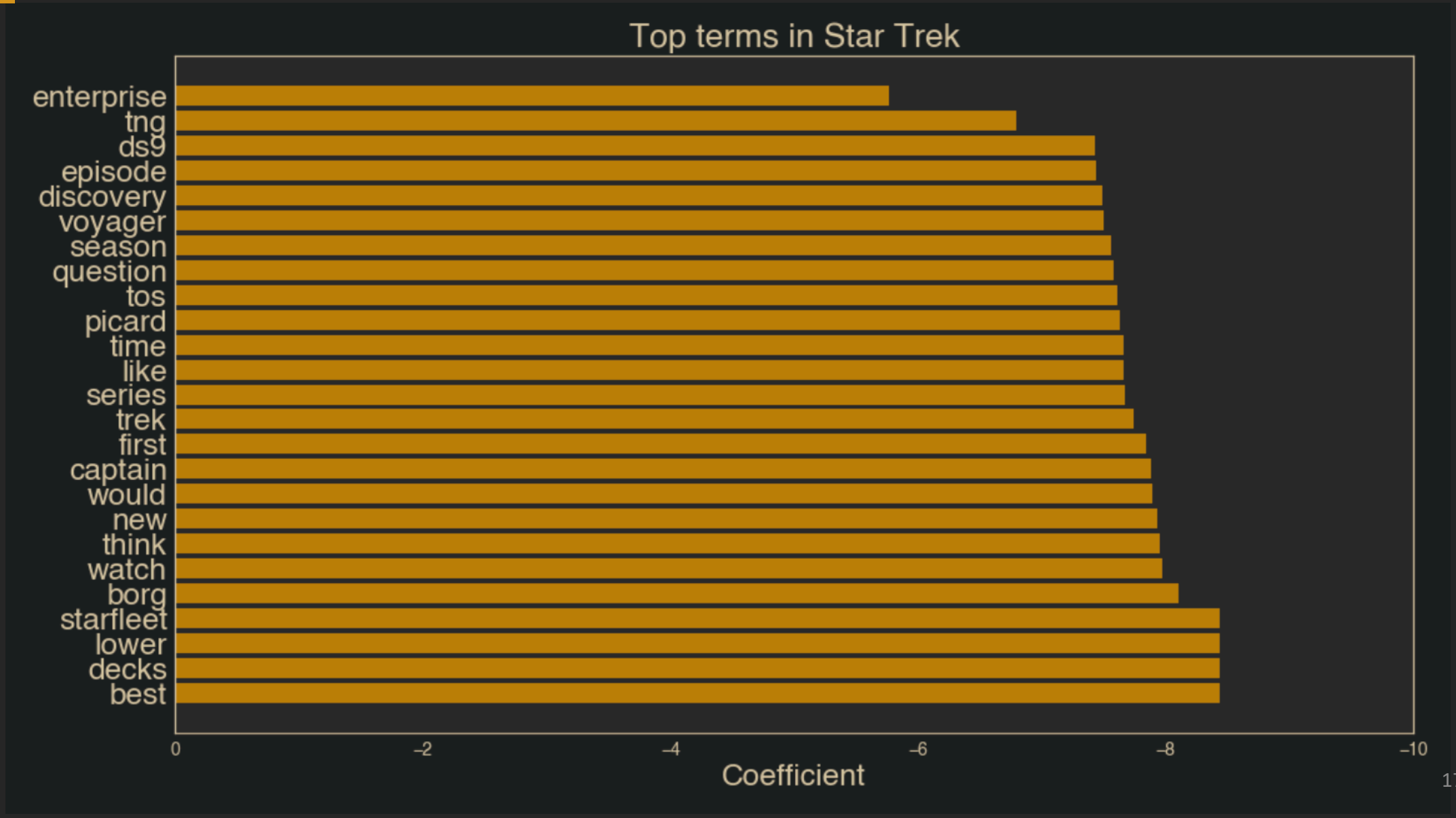
Sensitivity
(true positive rate):

88.0

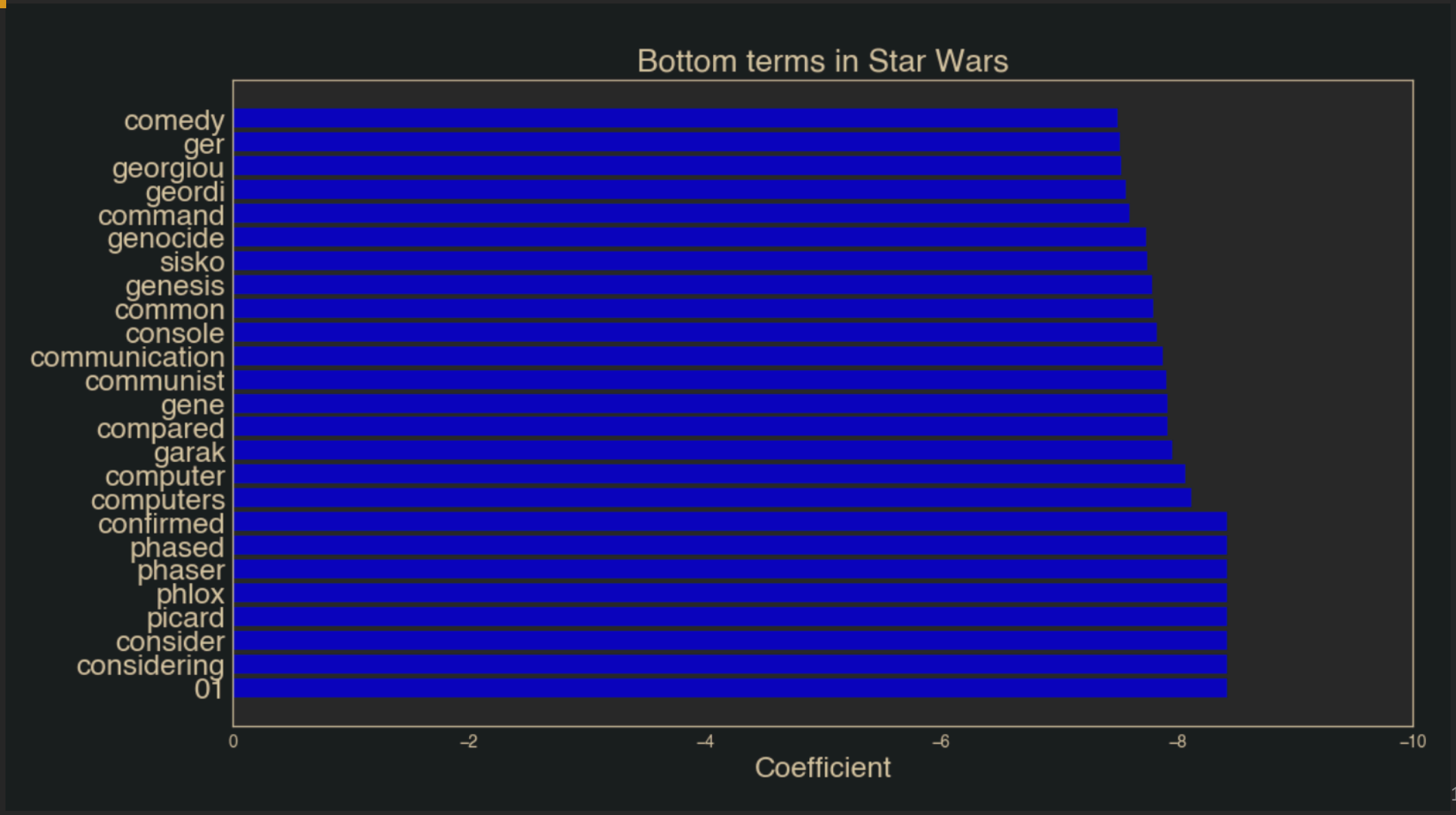
Conclusions and recommendations



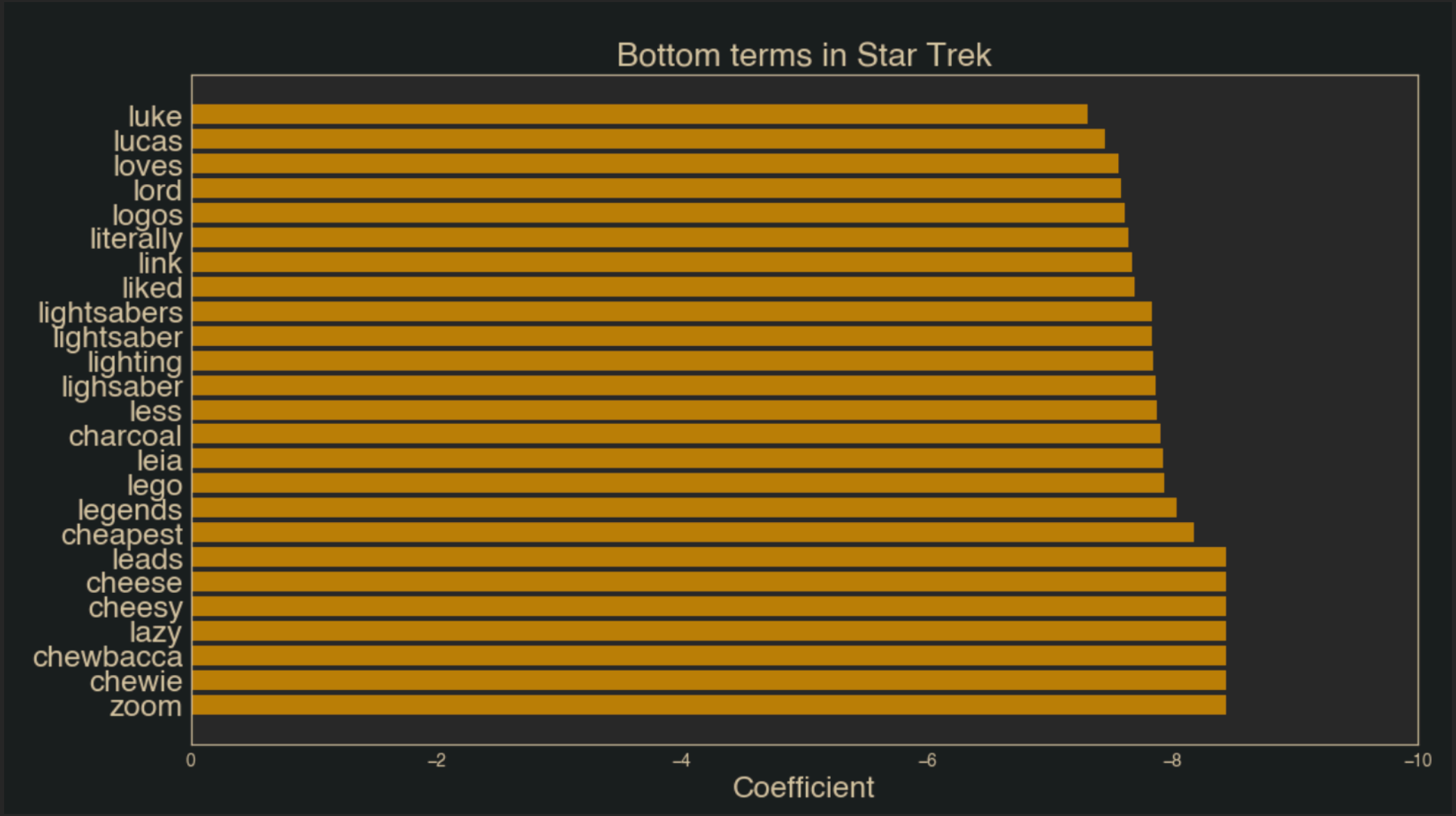
Conclusions and recommendations



Conclusions and recommendations



Conclusions and recommendations



Conclusions and recommendations

- Unique terms help identify the subreddit
- There could to be further tweaks to the modeling process
- Pipelines with grid search are recommended
- As of now, the project helped to learn difference between the two subreddits and help classify posts into correct categories.

Thank you

Q&A