Final Project – IMDb Reviews (Sentiment Analysis)

ANGELICA, ANEESH, BRYAN, DONALD, SORAIYA

What tools did we use?

Our steps

▶ Our approach to the sentiment analysis:

1) Clean and transform data

2) Tokenize words, remove stop words and punctuation

3) Create a bag of words

4) Build our model

5) Check its accuracy!

6) Visualize

Results – ML Algorithm

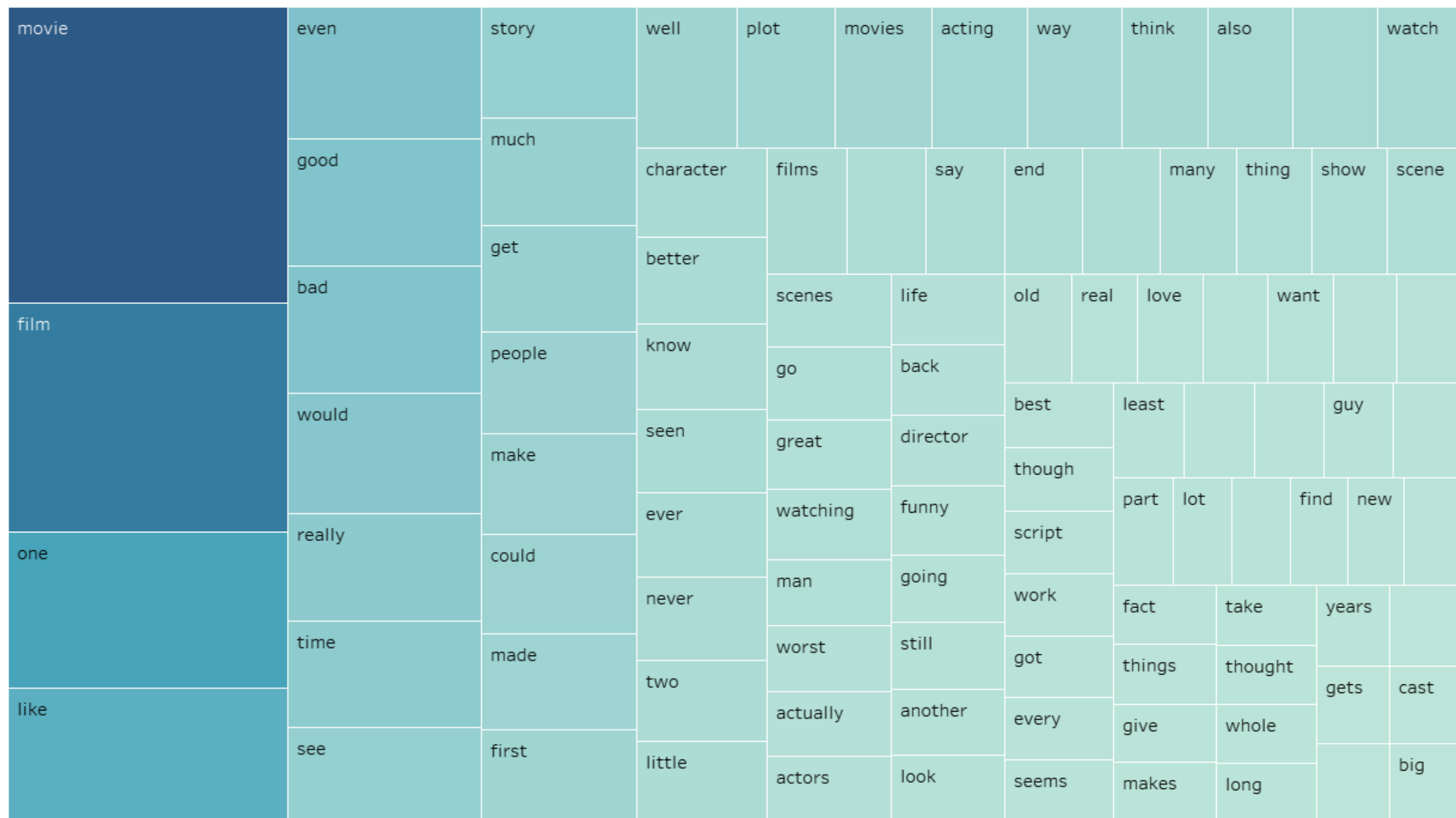▶ We tried, and ended up using, one model:

▶ Bag of Words model; Score: 0.998

In [36]:
```python
#Depicts term frequency vector for each review (bag of words)
vocab = CV.get_feature_names()
df_reviews = pd.DataFrame(cv_matrix, columns=vocab)
df_reviews.head(500)
```

Out[36]:

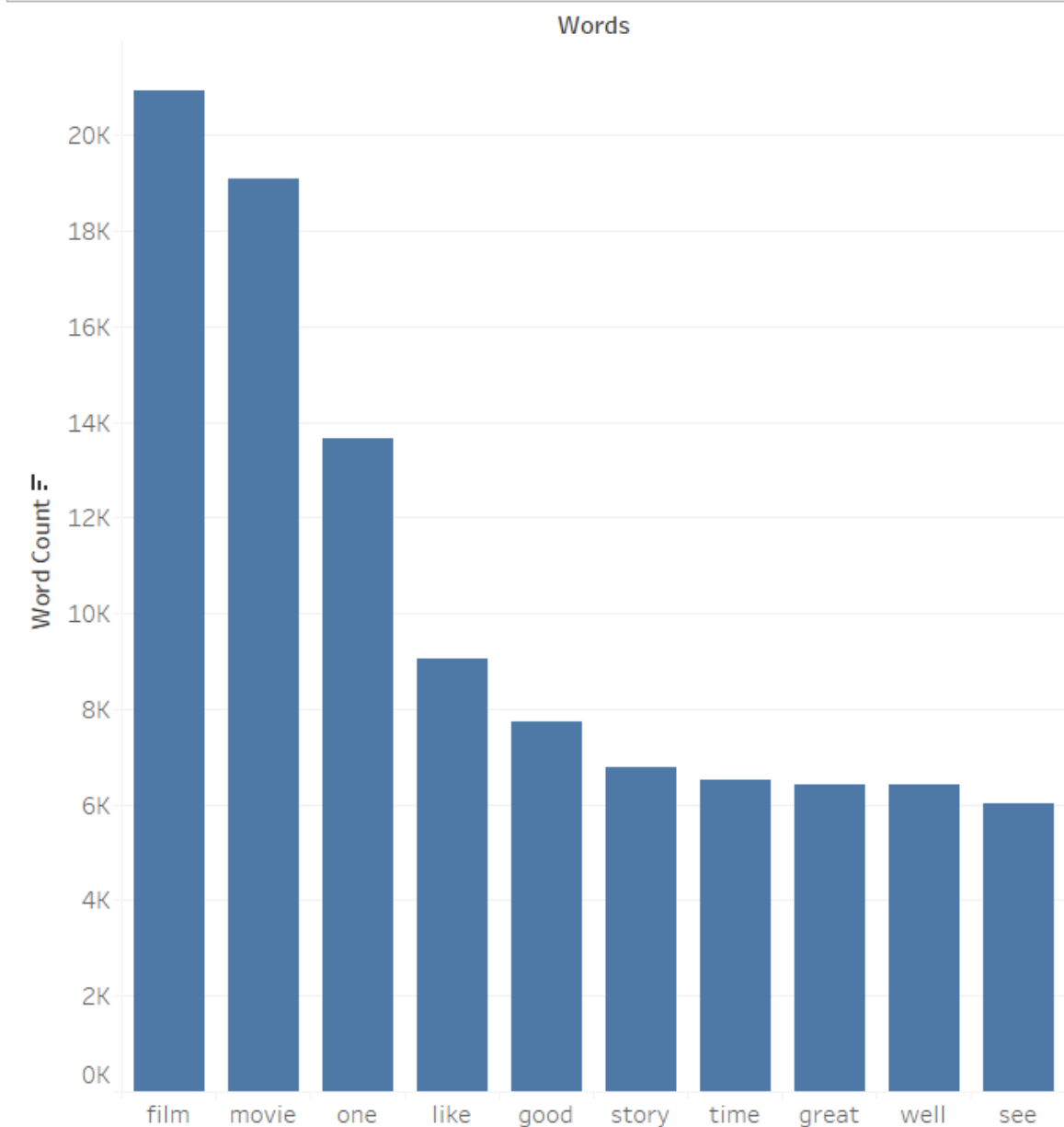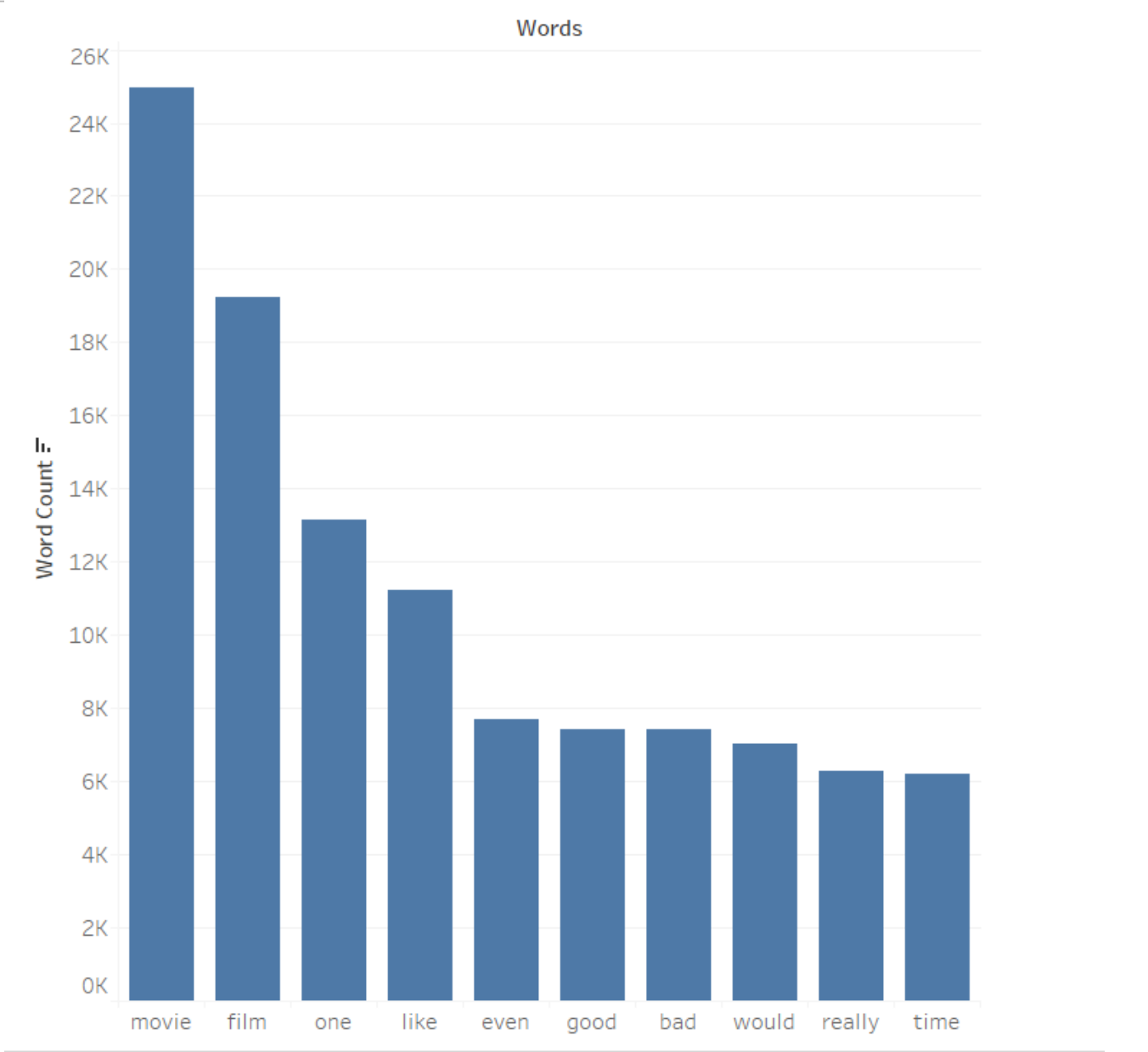|	aa	aaa	aaaaaaah	aaaaah	aaaahhhhhhh	aaaarrgh	aaah	aaargh	aaaugh	aaawwwwnnn	...	zyuranger	zz	zzzz	zzzz
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
495	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
496	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
497	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
498	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
499	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0

500 rows × 53052 columns

Top 100 Words (by Frequency)

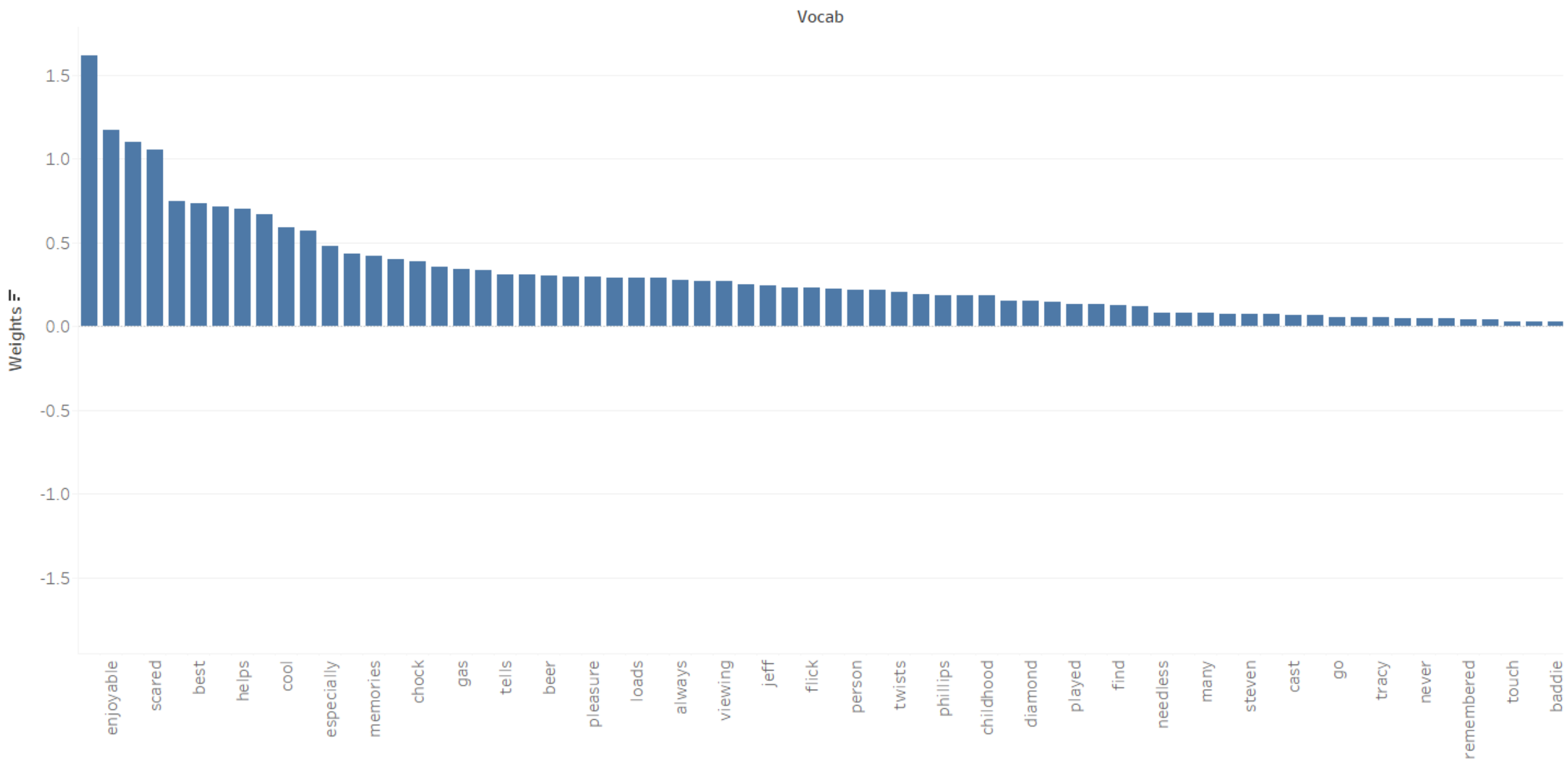# Top 10 Positive Review Words (by frequency)

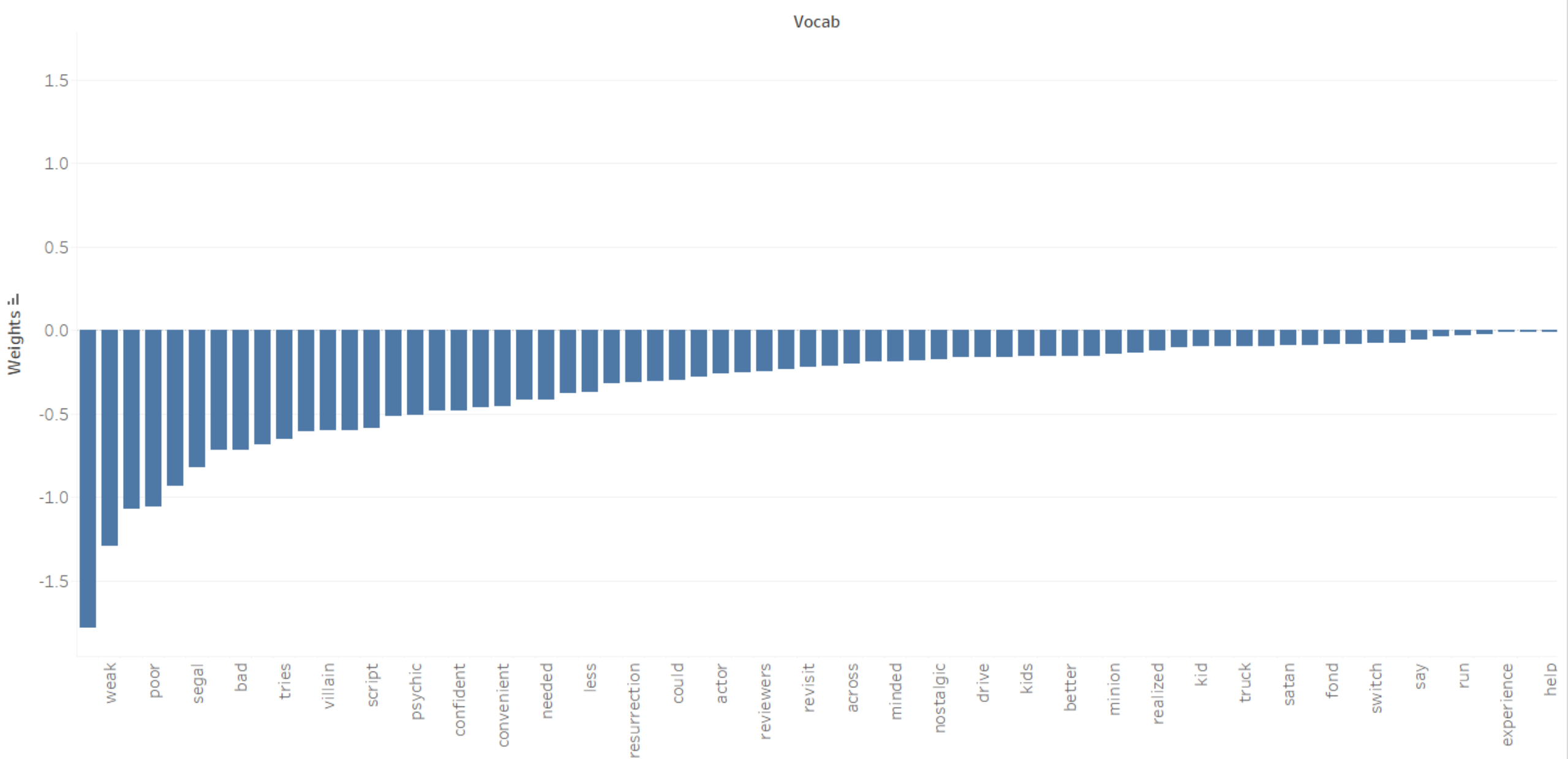# Top 10 Negative Review Words (by frequency)

Example

- *"childhood memory tells scary movie touch go whether revisit anyway remembered scary scene involving homeless person cool villain played jeff kober first power good movie sad say chock full cop clich poor script holes truck could drive along countless convenient twists help story run along lou diamond phillips confident bad ass cop sends baddie serial killer kober gas chamber find minion satan power resurrection along power possessing every weak minded person comes across mix poorly realized psychic helps case ahhh trash enjoyable especially fond memories scared kid scene homeless person still pretty good kind logic forget every scenario thrown good measure end cross steven segal action flick demonic flick earth thought good idea cast lou diamond phillips lead needless say convincing tries best never problem guy many reviewers seem tracy griffith psychic less said better kober pretty good killer always liked actor first power may doctor ordered hard day work brain switch needed beer likely enhance viewing experience definitely loads next time give movie spin good flick somewhat guilty pleasure nostalgic fans easily scared kids see around buddy boy"*

Words by Pos/Neg Weight

Words by Pos/Neg Weight

Vocab

Weights ▣

Example (cont.)

▶ Model correctly predicts a negative review!

▶ Probability of a positive review is 7.3%

Questions?