NLP with Reddit

Deepthi Vaddi

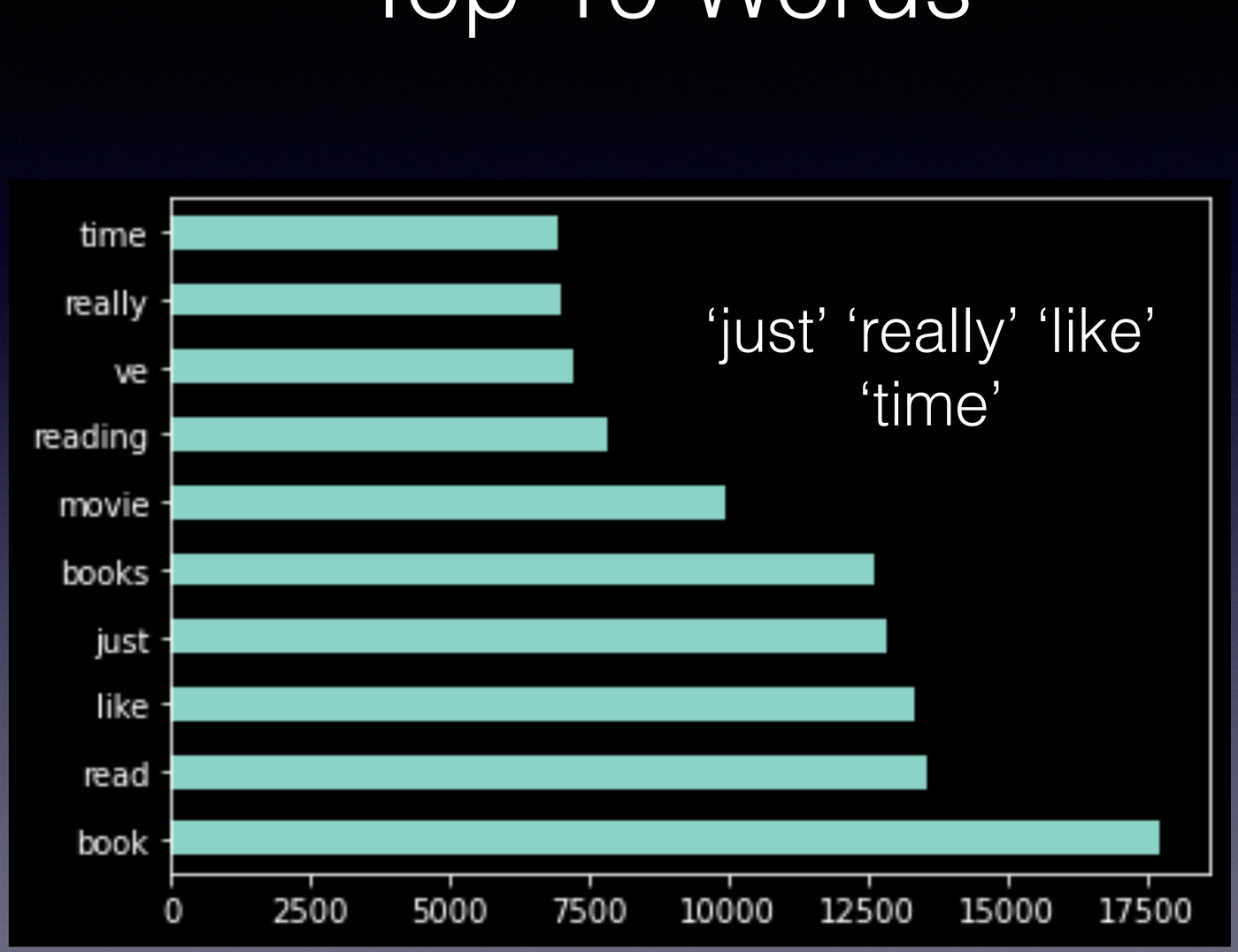
Collect Data

- Books : 26019 documents
- Movies : 16688 documents
- *day_window = 7, n = 520*

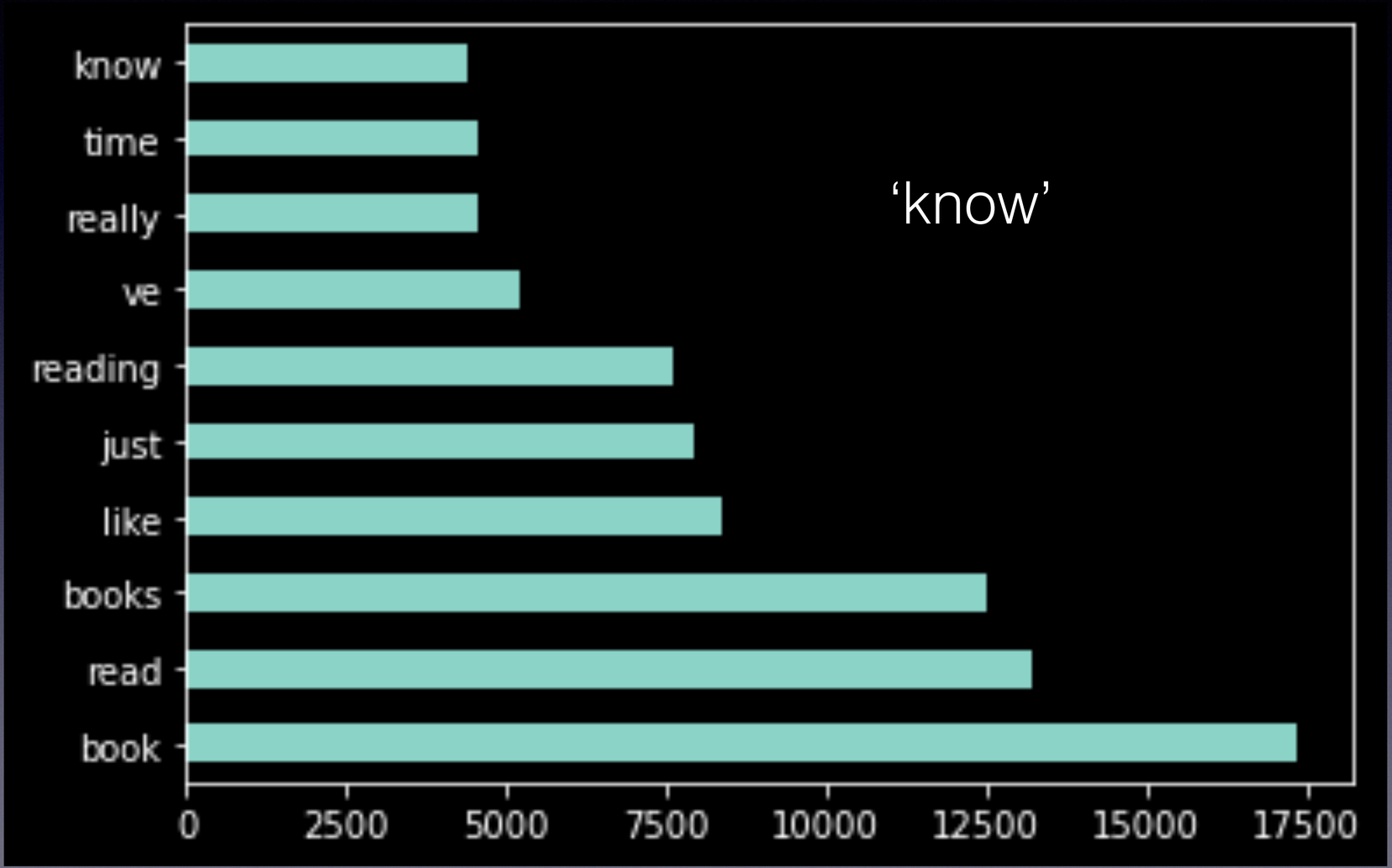
Data Preparation

- Concatenated the 2 datasets
- Deleted the observations with null, [deleted],[removed]
- Map movies:1 , books : 0
- 23,923 documents

Top 10 Words

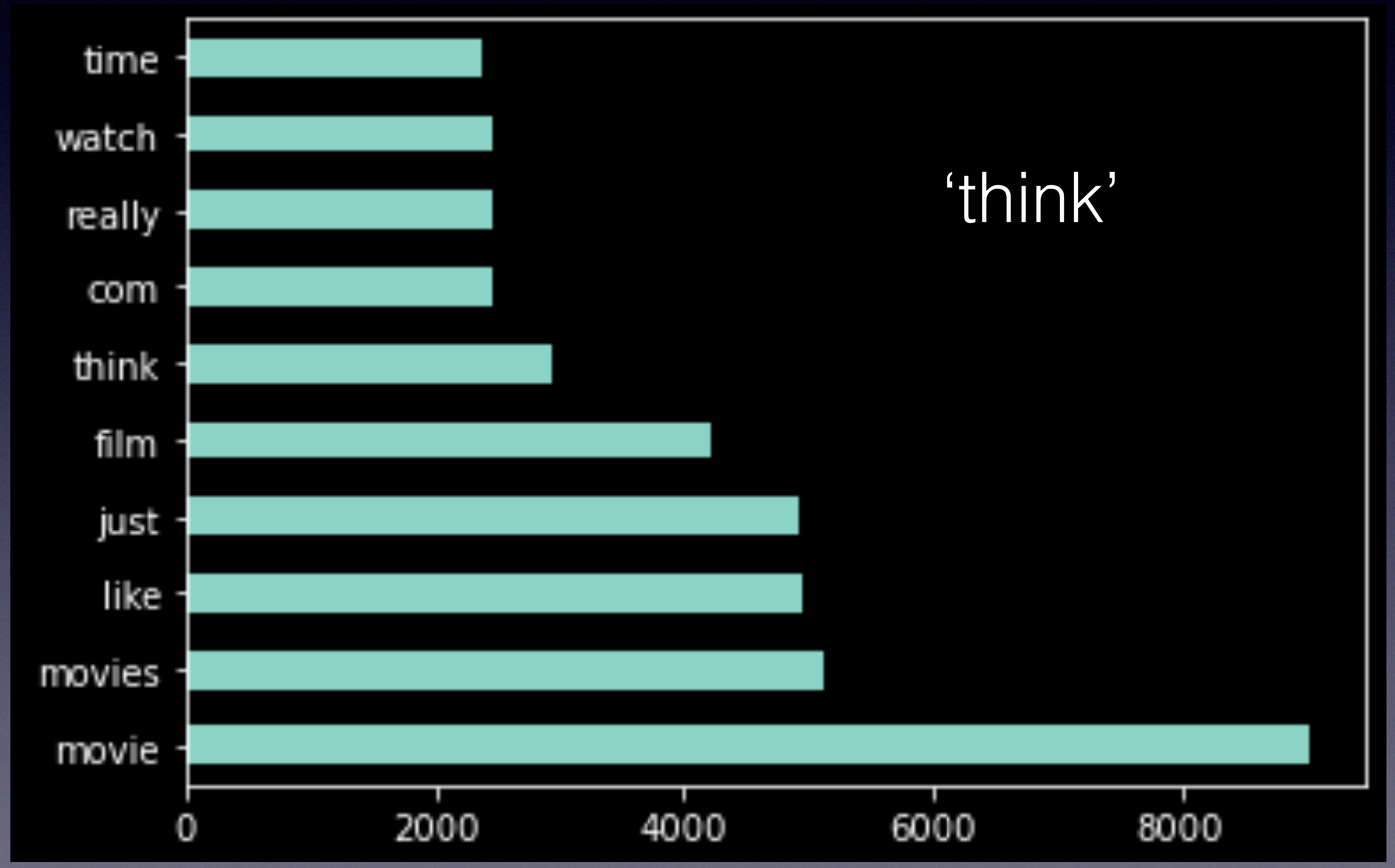

time
really
ve
reading
movie
books
just
like
read
book
0
2500
5000
7500
10000
12500
15000
17500
‘just’ ‘really’ ‘like’
‘time’

Books Top 10



‘know’

Movies Top 10



Baseline Measurement

- **Accuracy** will be our choice of measurement since it does not really matter if we are predicting books or movies better.
- Null model accuracy is **60.9%**

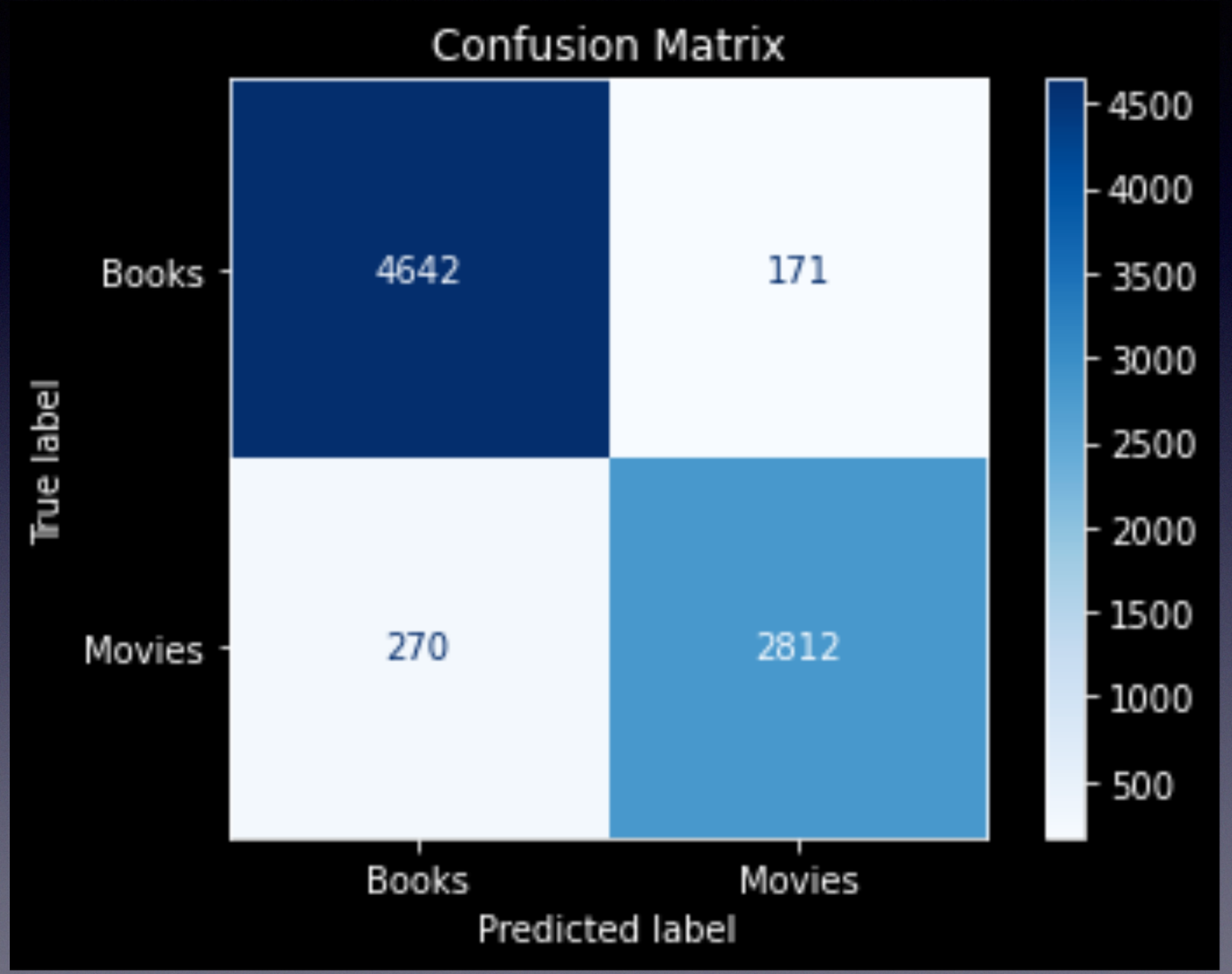
Model 1

- *Transformer* : Count Vectorizer
- *Model* : Multinomial Naive Bayes
- *Best Parameters* :
  - max_df = .9
  - min_df = 2
  - max_features = 5000
  - N gram = (1,1)

Naive Bayes with CV

Training Set Accuracy : **94.9**%

Test Set Accuracy:**94.4**%



Model 2

- *Transformer* : TFID vectorizer
- *Model* : Random Forest
- Train Accuracy : **99.8**%
- Test Accuracy: **93.4**%
- Overfit!


"The quest stands upon
edge of a knife. Stray
little and it will fail. But
hope remains, if friends stay
true."
J.R.R. Tolkien

GridSearch Random Forest

In preparation for the quest :

TFID stop words , max_df = .9

Results:

{'max_depth': 5,
 'max_features': 'auto',
 'min_samples_leaf': 2,
 'min_samples_split': 7,
 'n_estimators': 100}


Train Accuracy: **61.9%**
Test Accuracy: **61.7**%