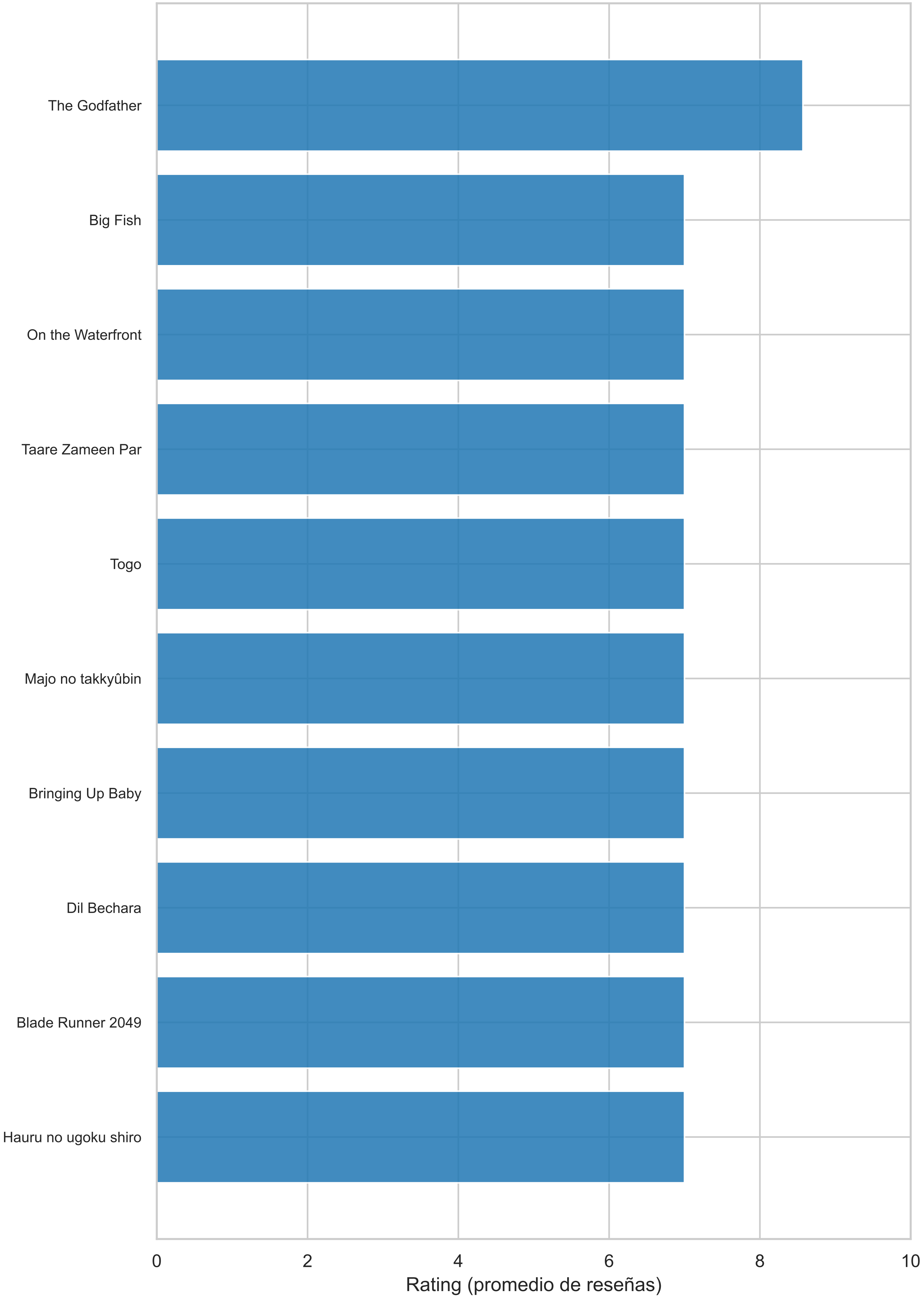
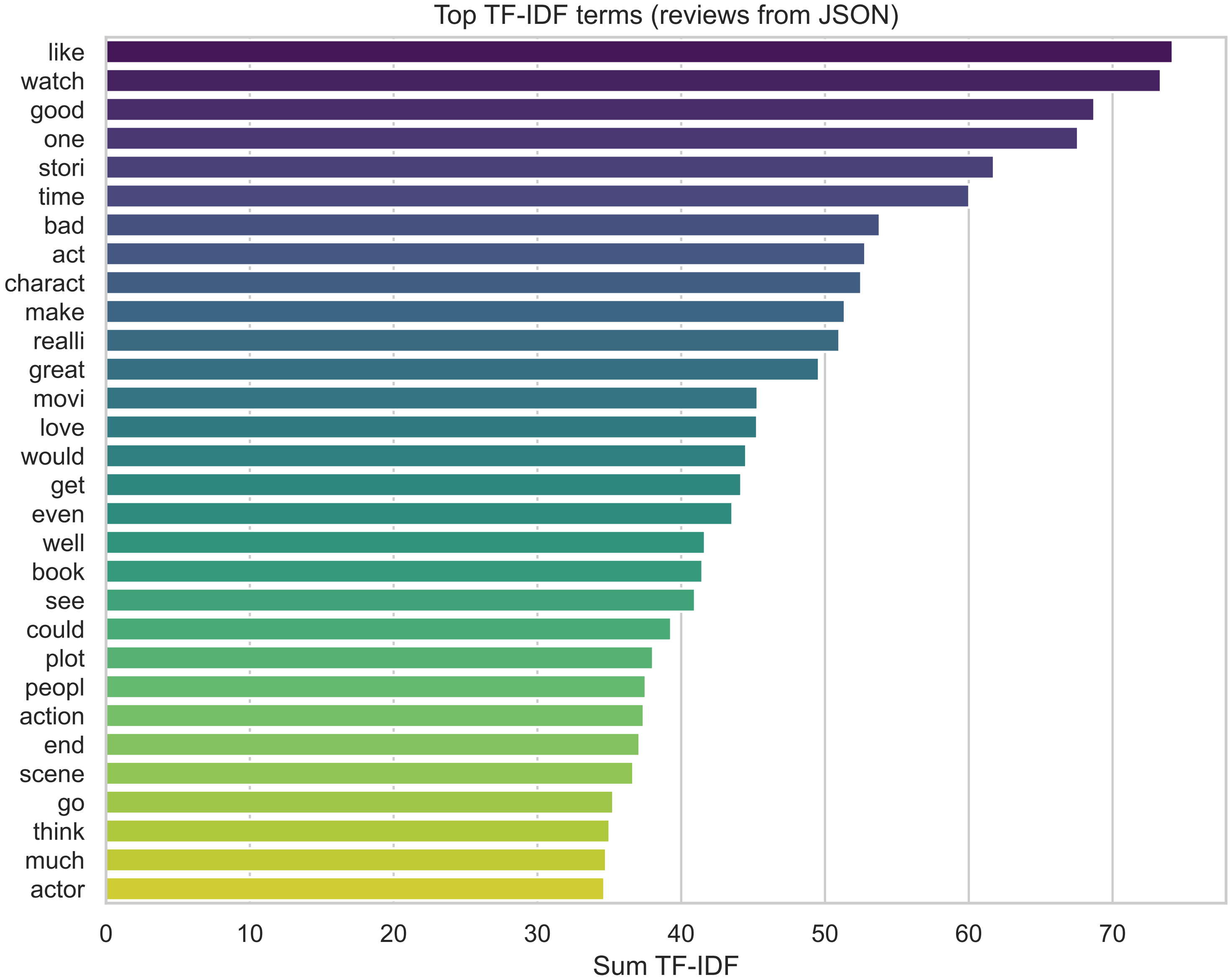
Top 10 películas (rating por promedio de reseñas)

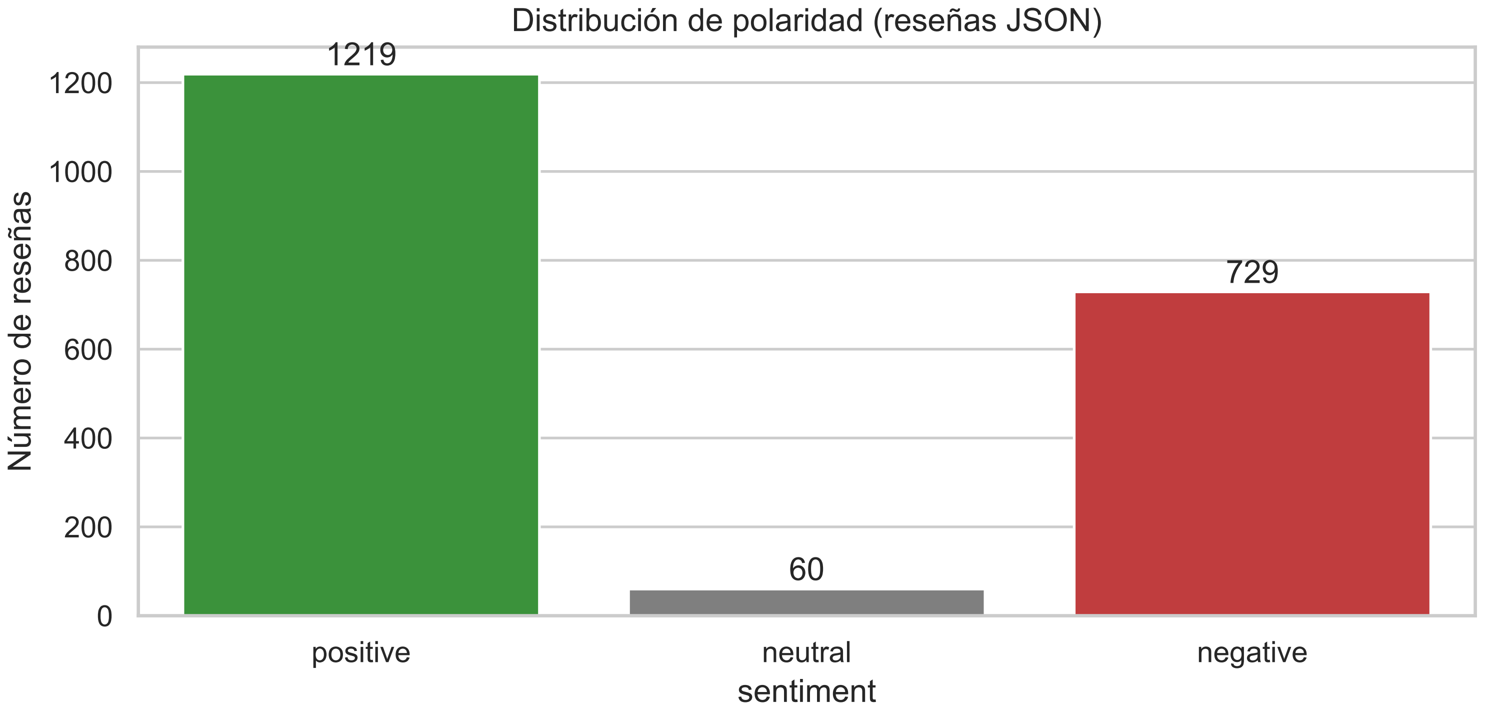


IMDB - Model Results

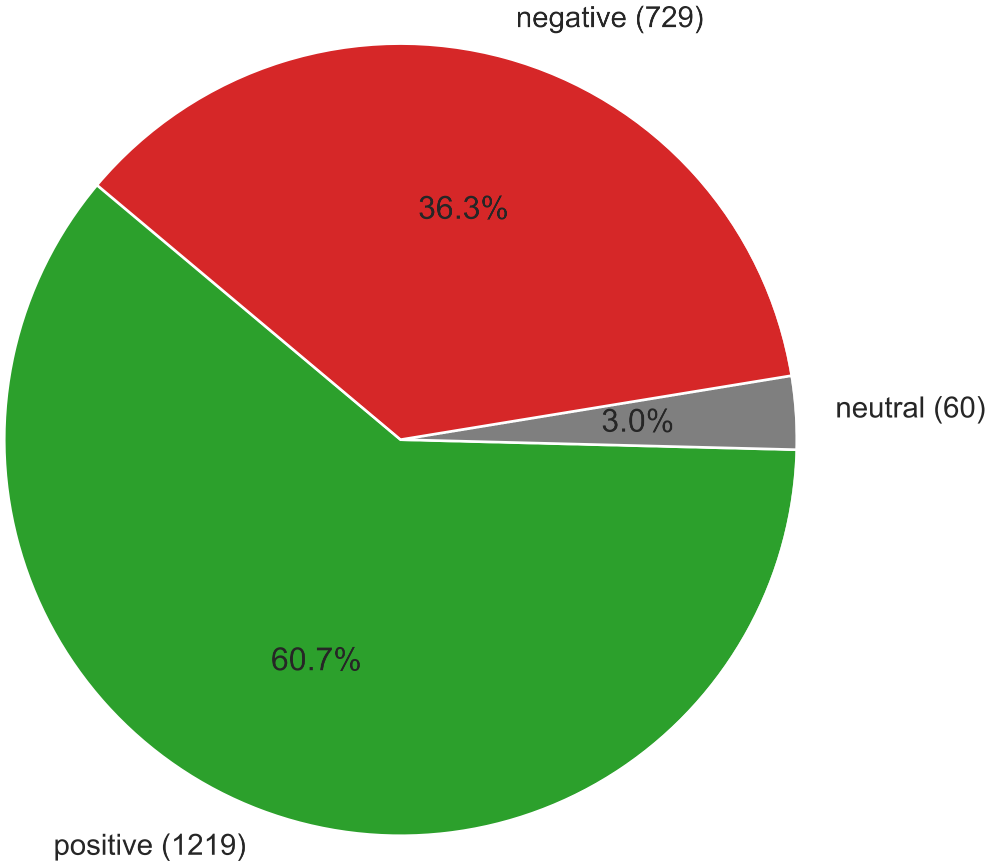
Total reviews in DB: 5458
Reviews used for modeling (first 4000): 3997
Reviews used for PDF visuals/evaluation (JSON): 2008

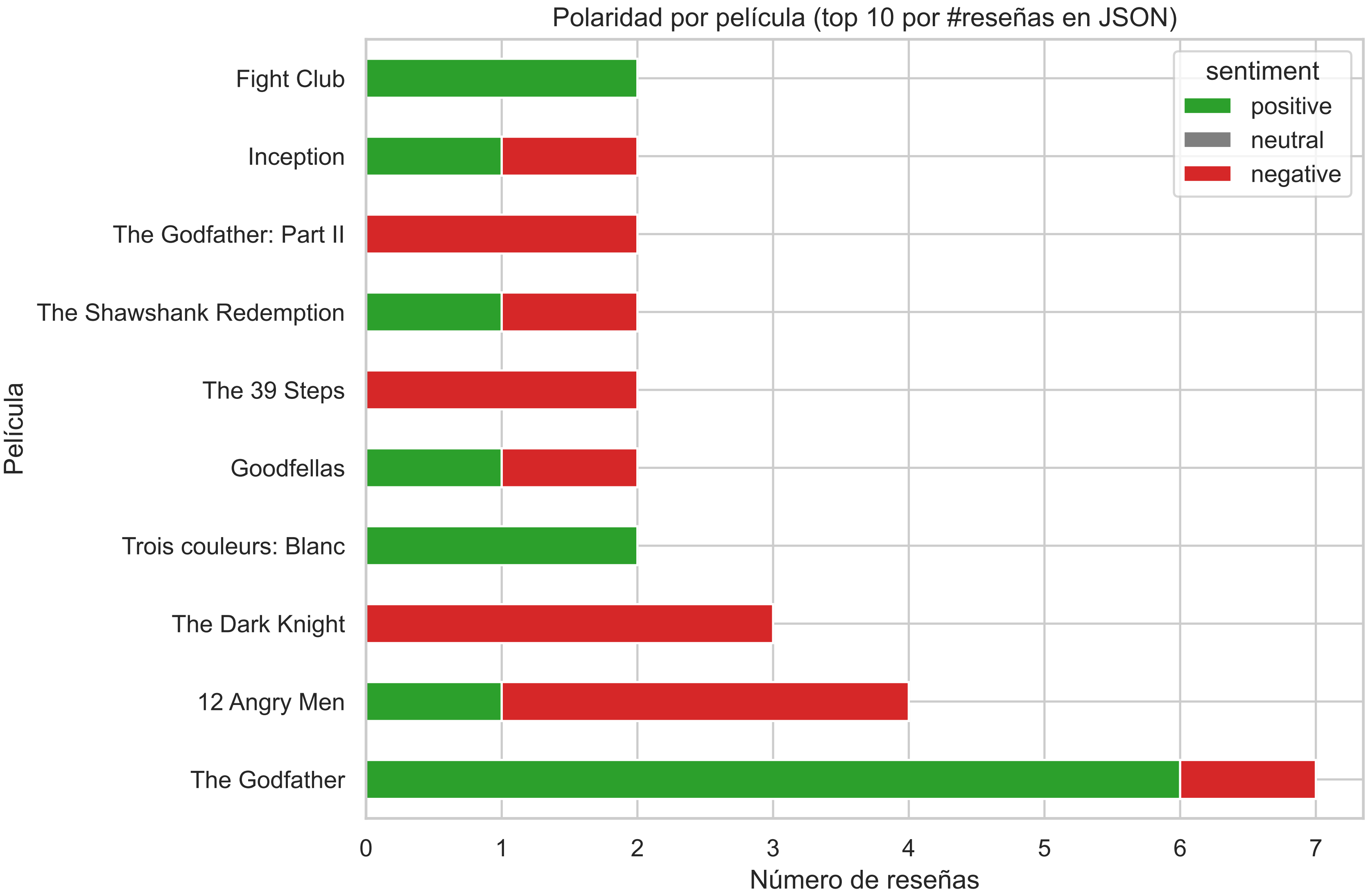
Models trained: Naive Bayes, Linear SVM, Decision Tree

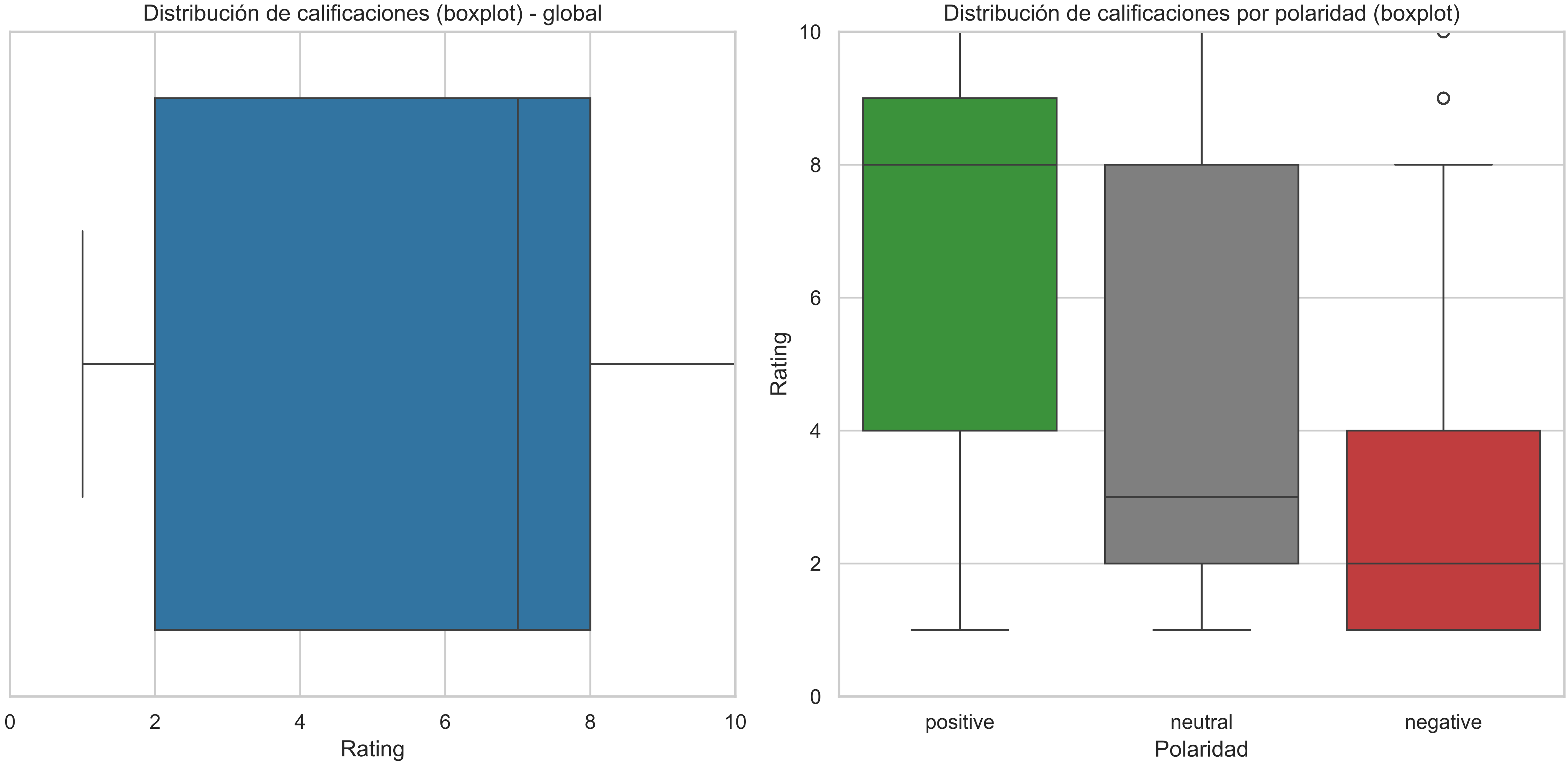

Top TF-IDF terms (reviews from JSON)
like
watch
good
one
stori
time
bad
act
charact
make
realli
great
movi
love
would
get
even
well
book
see
could
plot
peopl
action
end
scene
go
think
much
actor
0
10
20
30
40
50
60
70
Sum TF-IDF


Distribución de polaridad (reseñas JSON)
Número de reseñas
0
200
400
600
800
1000
1200
1219
60
729
positive
neutral
negative
sentiment

Porcentaje de polaridad (reseñas JSON)

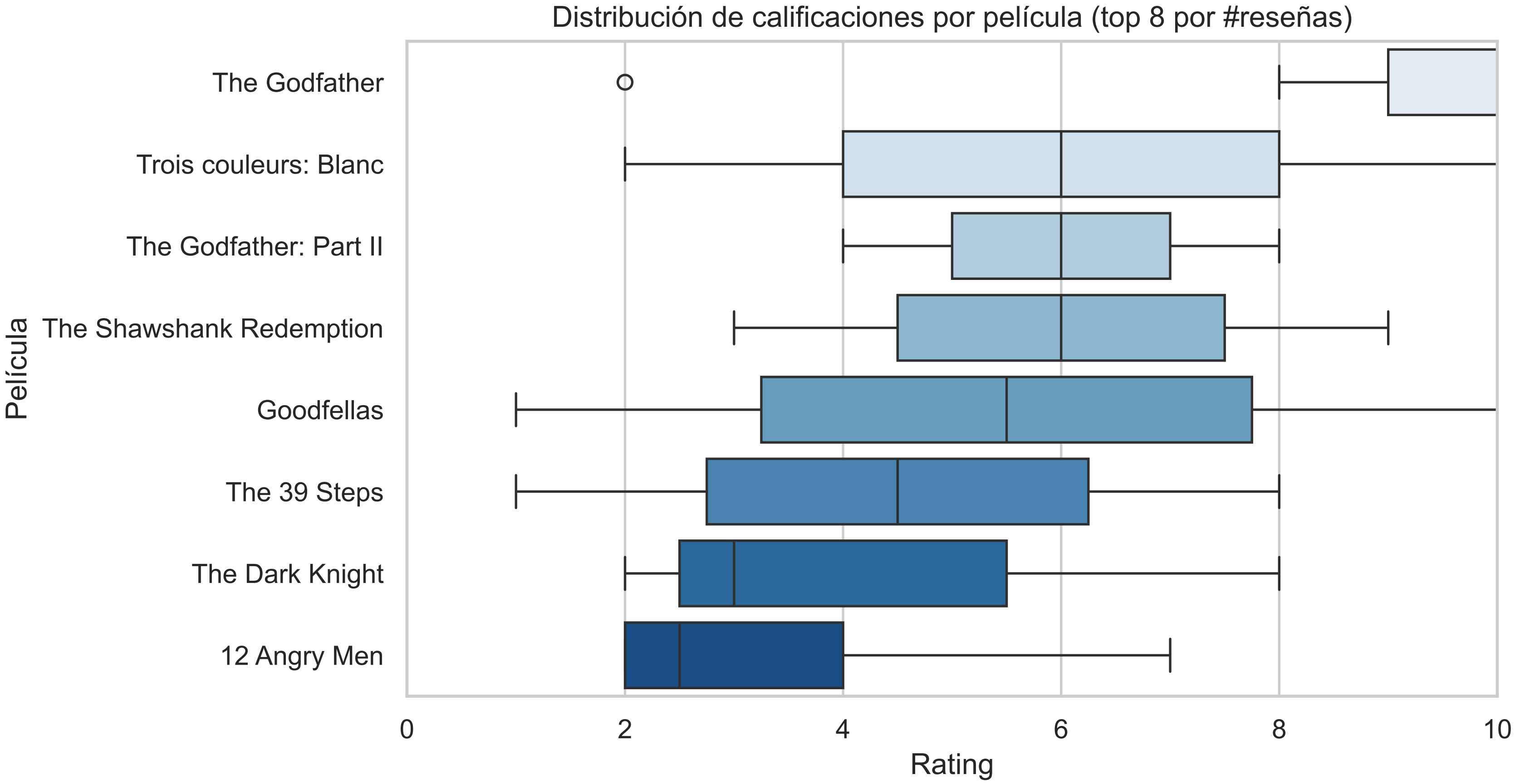



Polaridad por película (top 10 por #reseñas en JSON)
sentiment
positive
neutral
negative
Fight Club
Inception
The Godfather: Part II
The Shawshank Redemption
The 39 Steps
Goodfellas
Trois couleurs: Blanc
The Dark Knight
12 Angry Men
The Godfather
Película
0
1
2
3
4
5
6
7
Número de reseñas

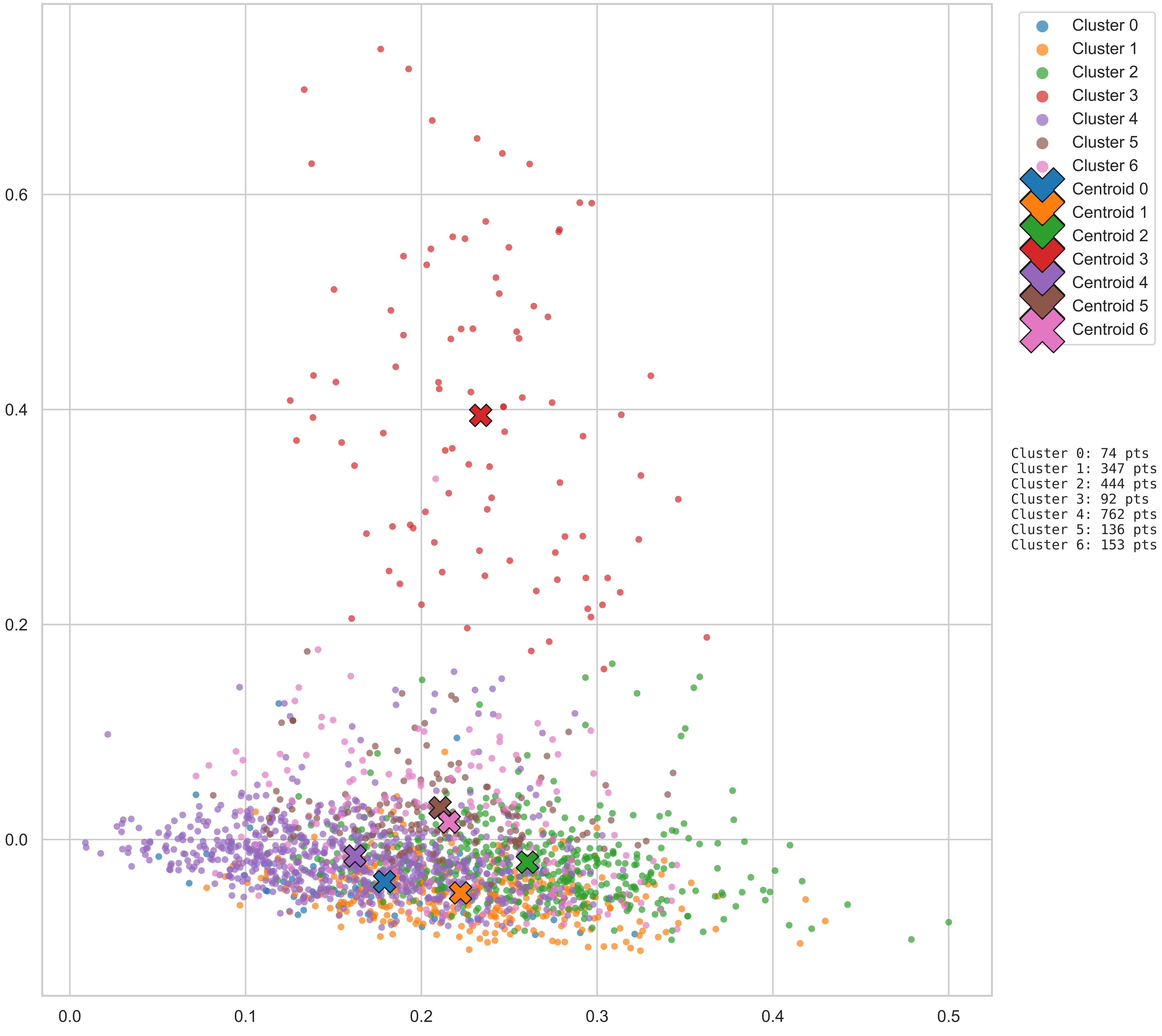


Distribución de calificaciones (boxplot) - global

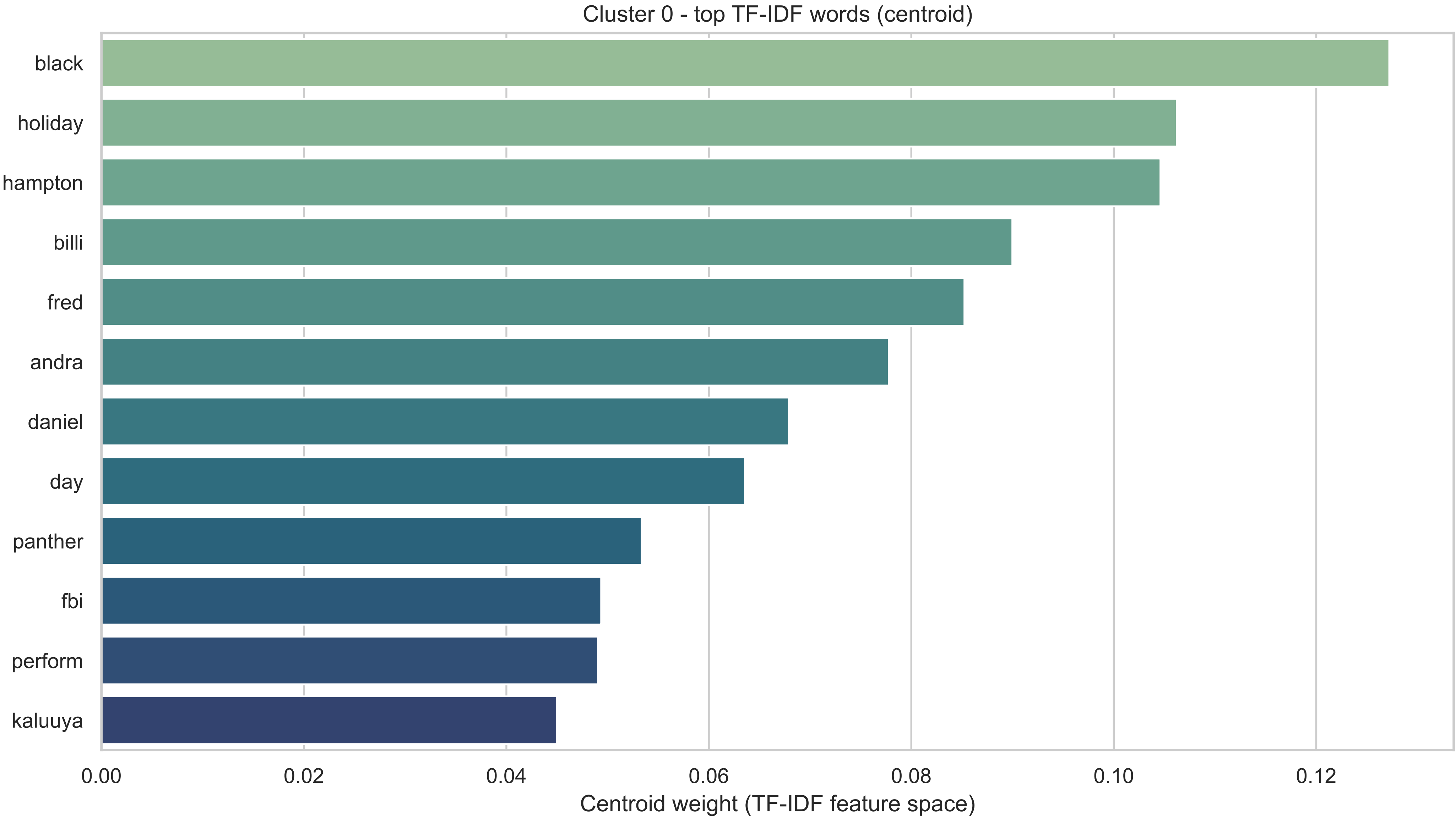
Distribución de calificaciones por polaridad (boxplot)

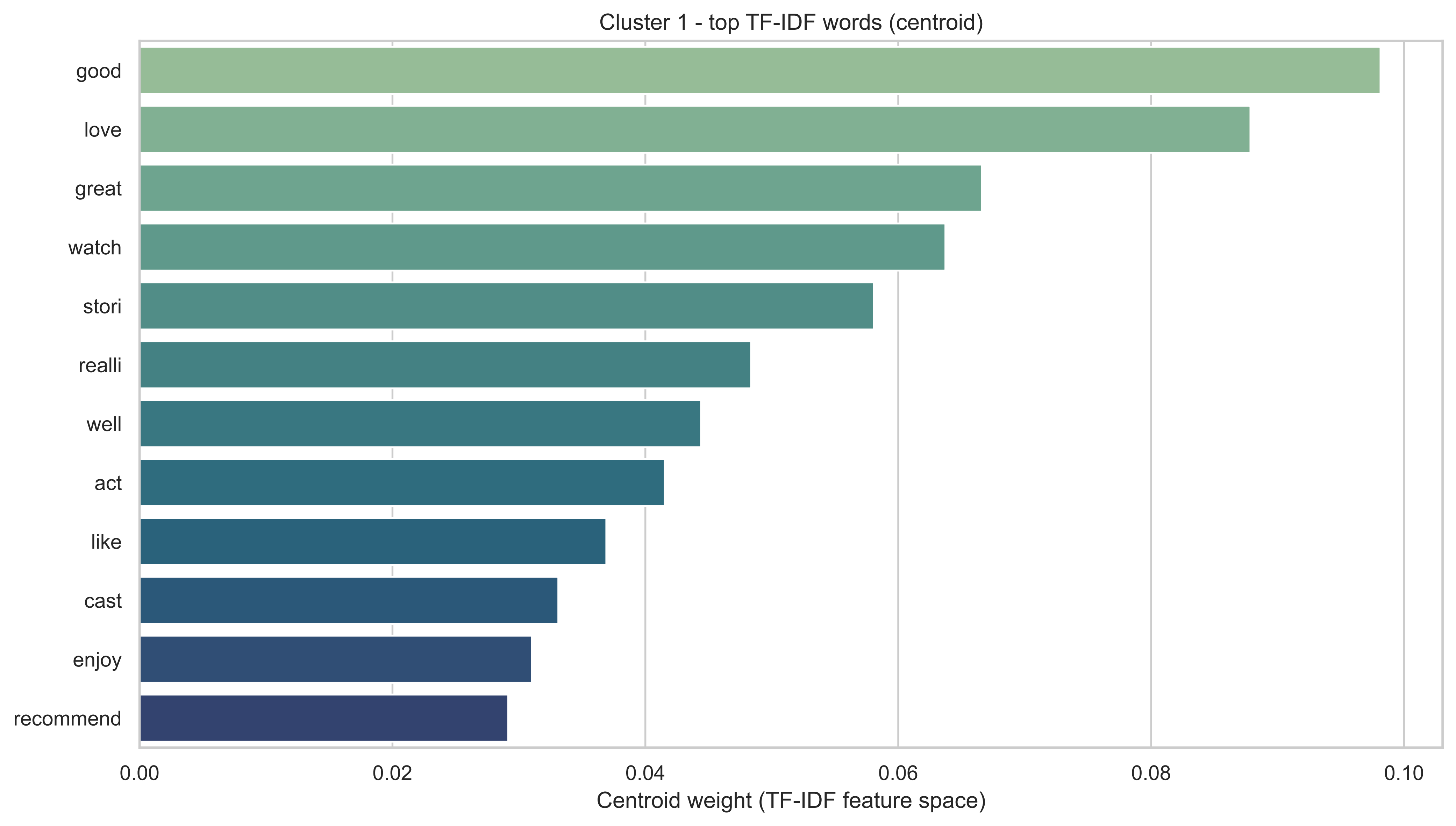

Distribución de calificaciones por película (top 8 por #reseñas)
Película
The Godfather
Trois couleurs: Blanc
The Godfather: Part II
The Shawshank Redemption
Goodfellas
The 39 Steps
The Dark Knight
12 Angry Men
0
2
4
6
8
10
Rating

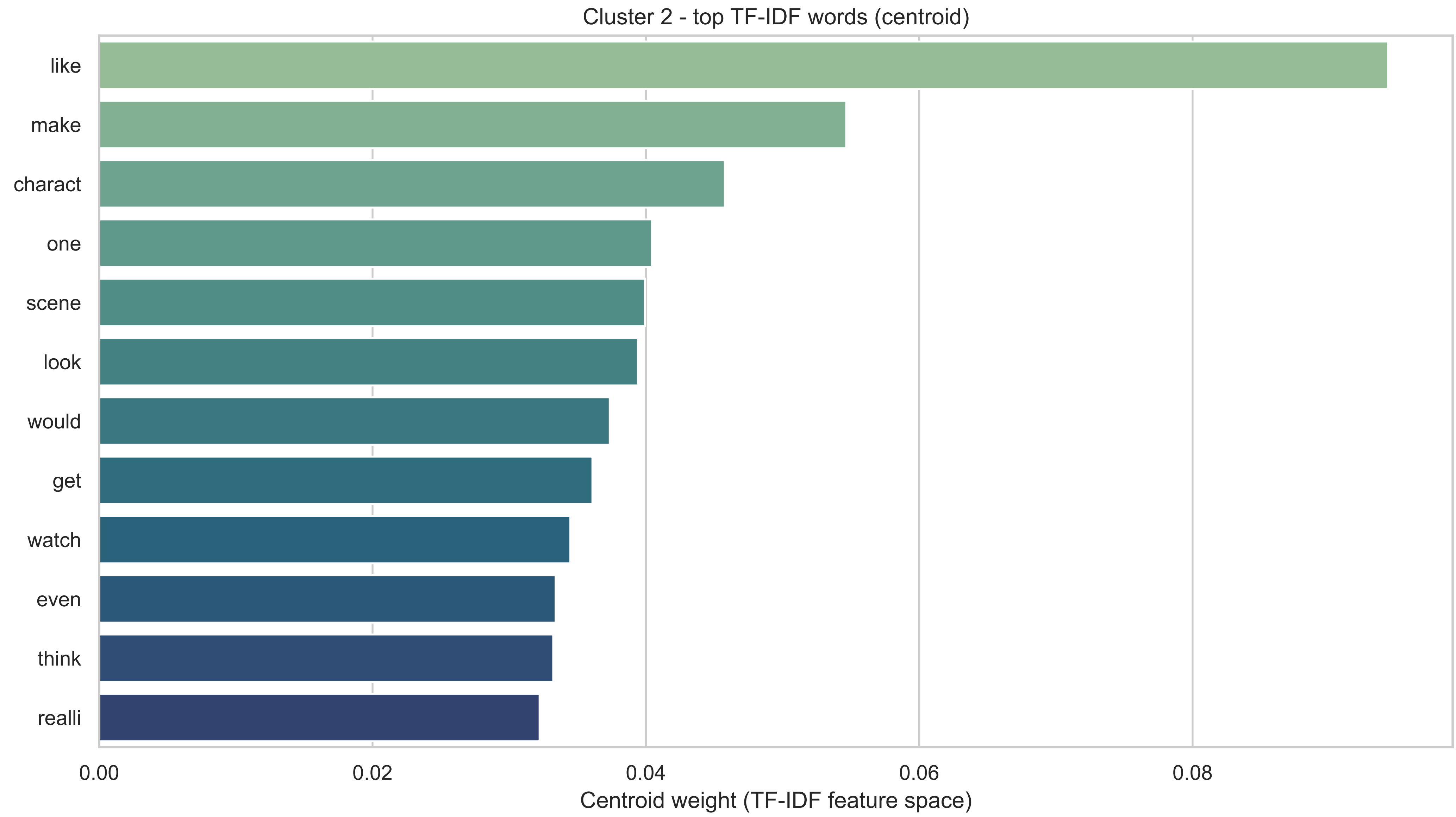
KMeans clusters (TruncatedSVD 2D) - JSON reviews

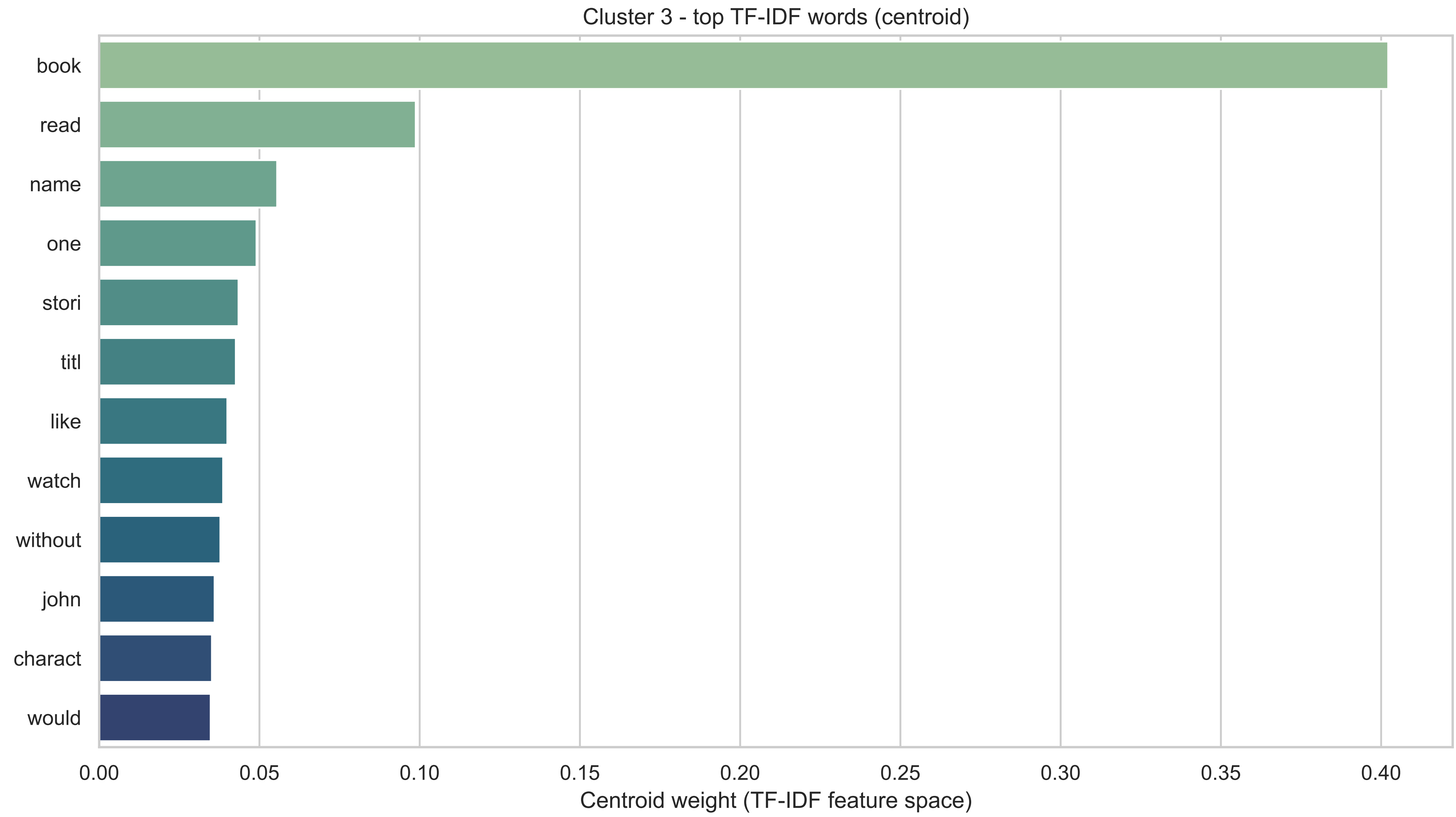


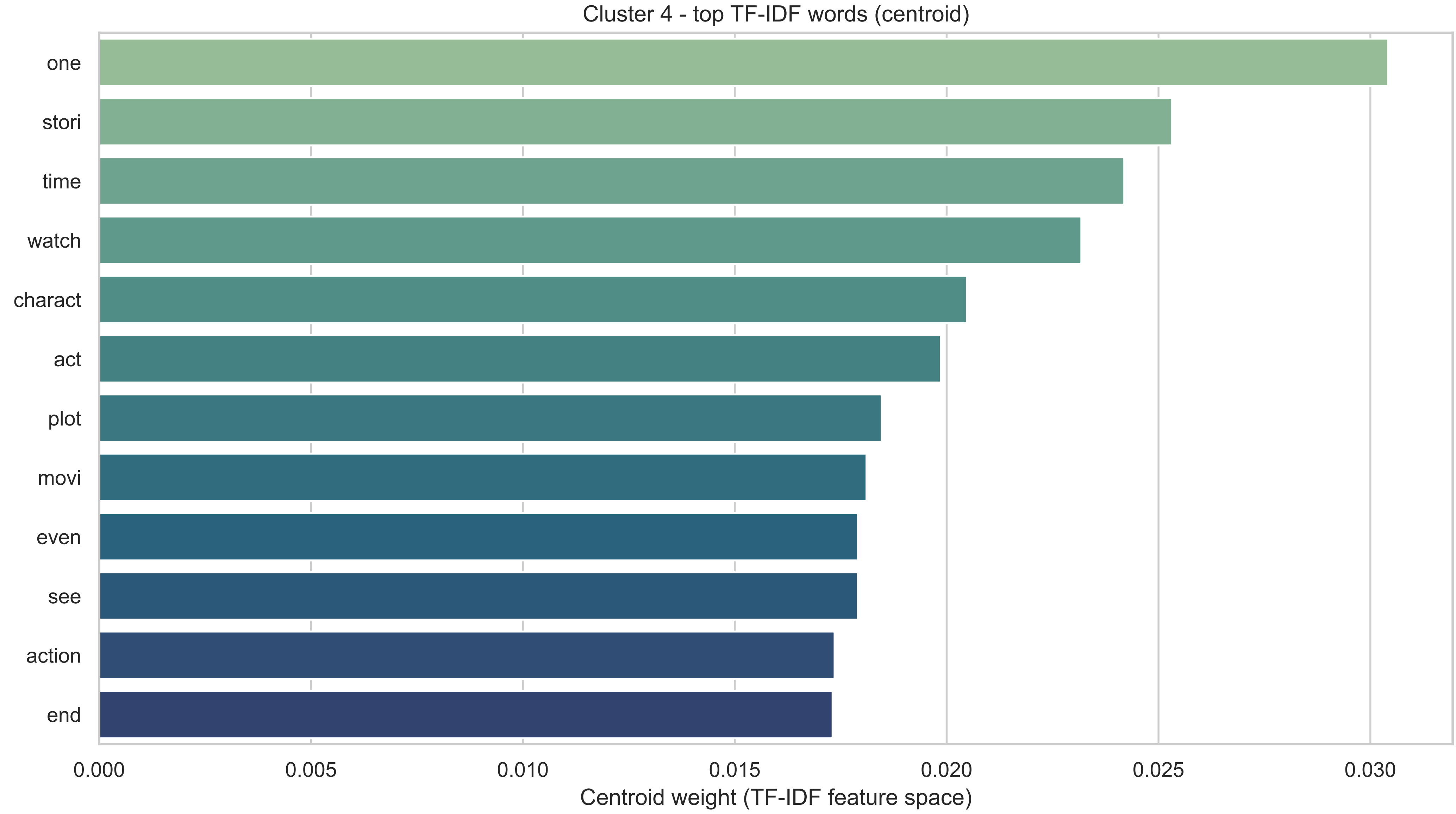
Cluster 0: 74 pts
Cluster 1: 347 pts
Cluster 2: 444 pts
Cluster 3: 92 pts
Cluster 4: 762 pts
Cluster 5: 136 pts
Cluster 6: 153 pts

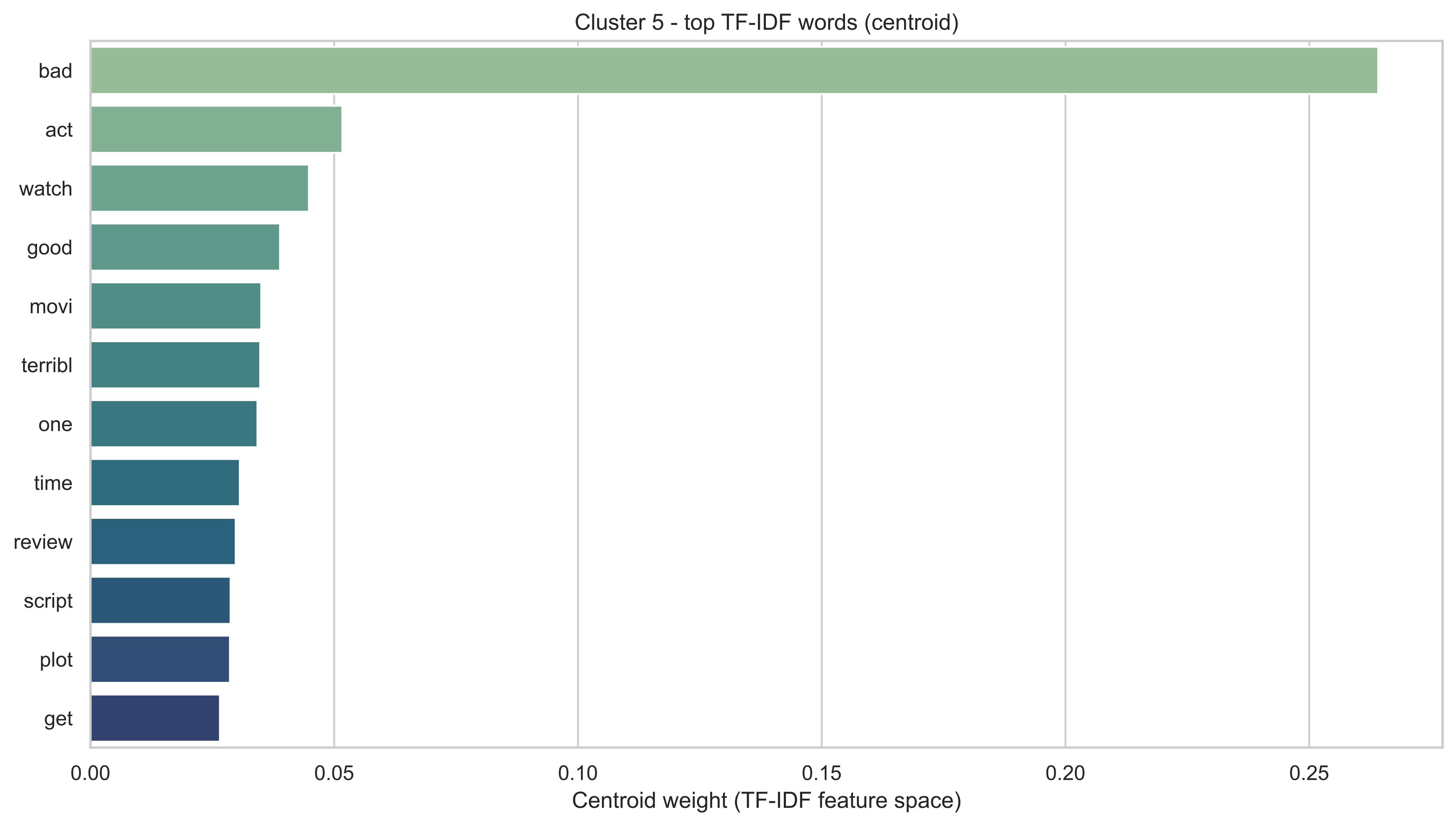

Cluster 0 - top TF-IDF words (centroid)
black
holiday
hampton
billi
fred
andra
daniel
day
panther
fbi
perform
kaluuya
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Centroid weight (TF-IDF feature space)

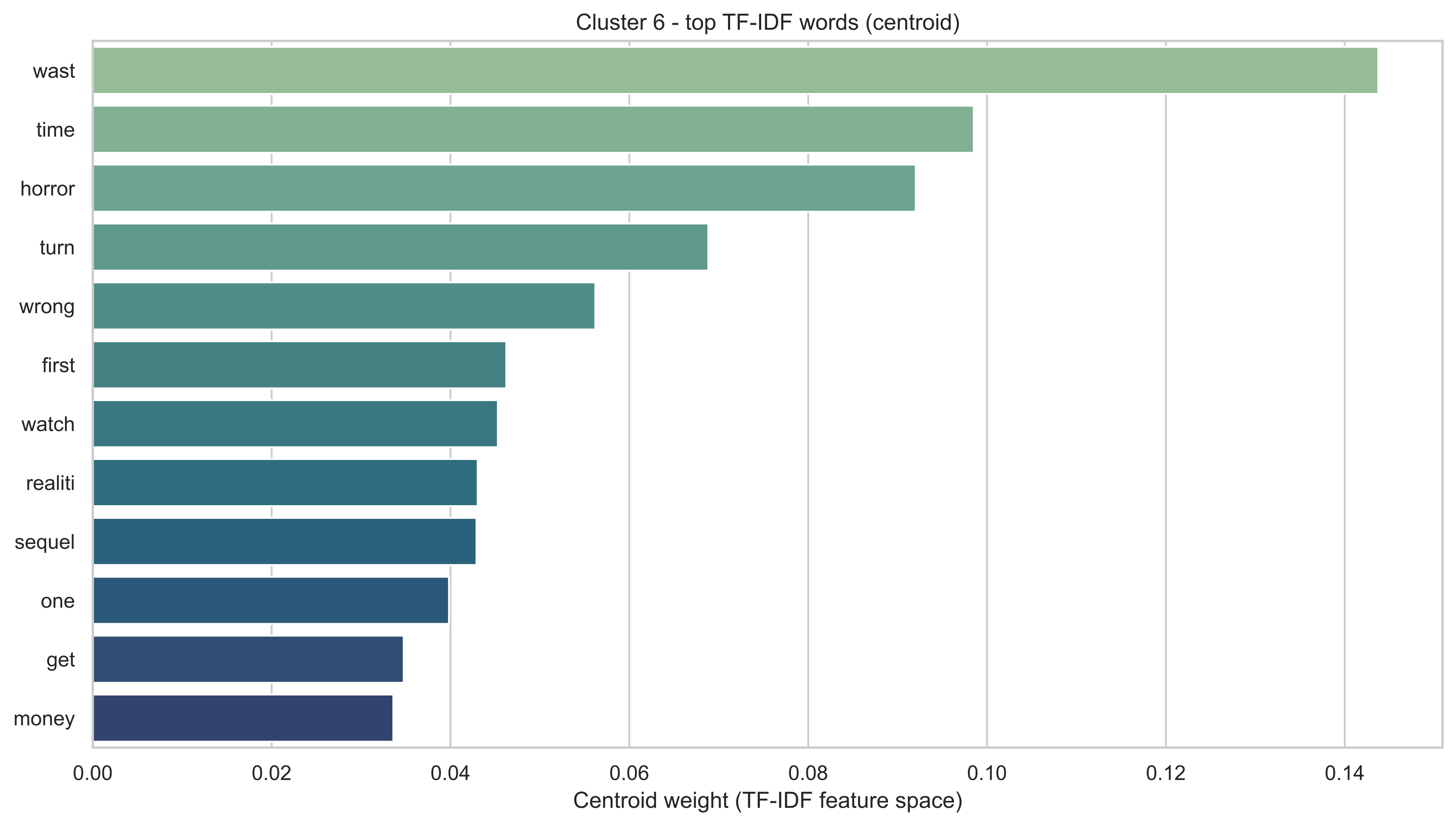

Cluster 1 - top TF-IDF words (centroid)
good
love
great
watch
stori
realli
well
act
like
cast
enjoy
recommend
0.00
0.02
0.04
0.06
0.08
0.10
Centroid weight (TF-IDF feature space)

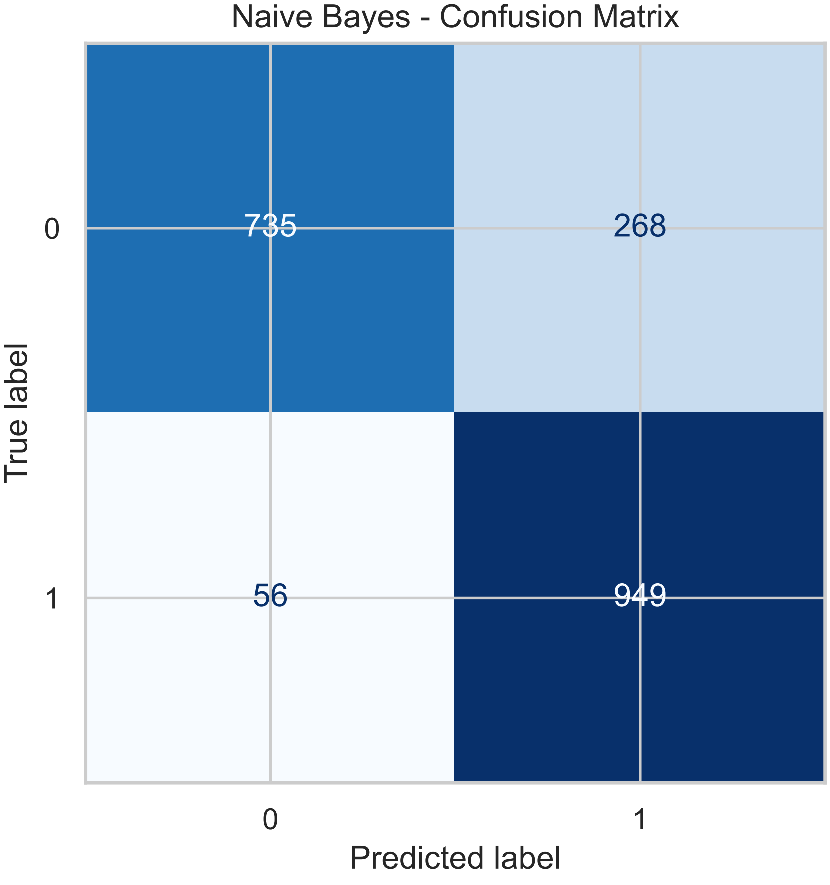

Cluster 2 - top TF-IDF words (centroid)
like
make
charact
one
scene
look
would
get
watch
even
think
realli
0.00
0.02
0.04
0.06
0.08
Centroid weight (TF-IDF feature space)

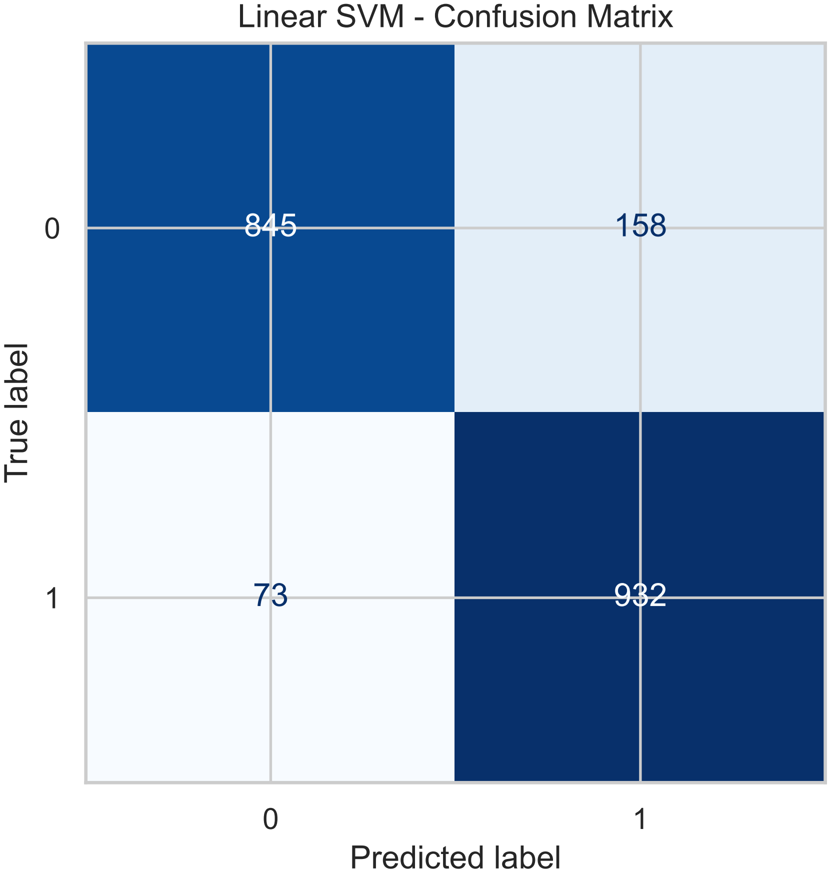

Cluster 3 - top TF-IDF words (centroid)
book
read
name
one
stori
titl
like
watch
without
john
charact
would
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Centroid weight (TF-IDF feature space)

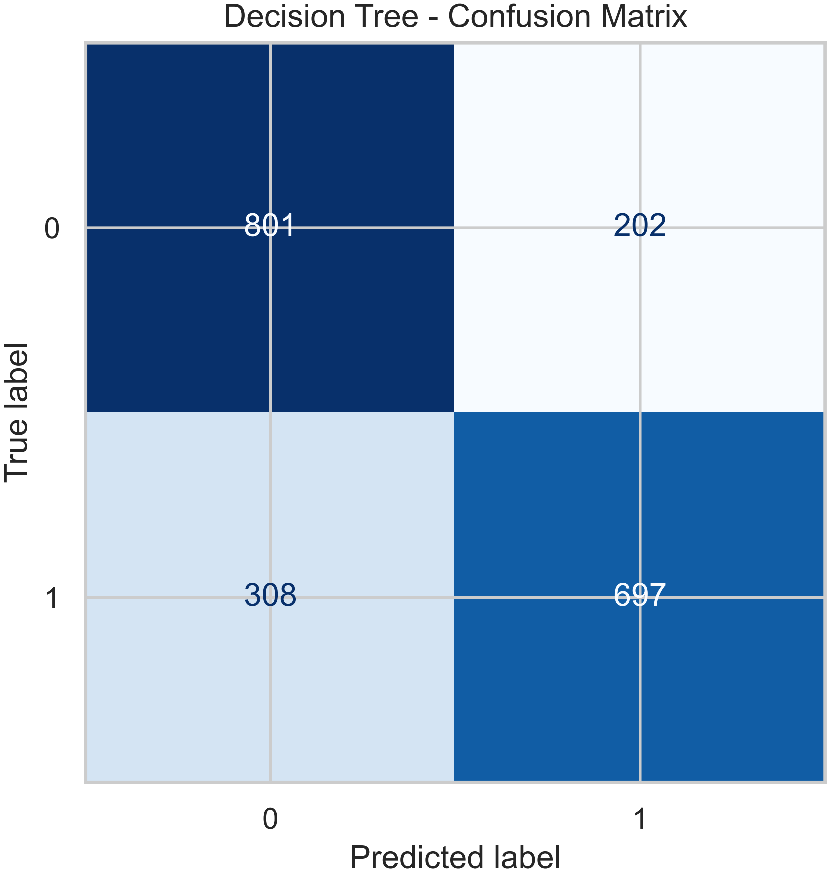

Cluster 4 - top TF-IDF words (centroid)
one
stori
time
watch
charact
act
plot
movi
even
see
action
end
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Centroid weight (TF-IDF feature space)

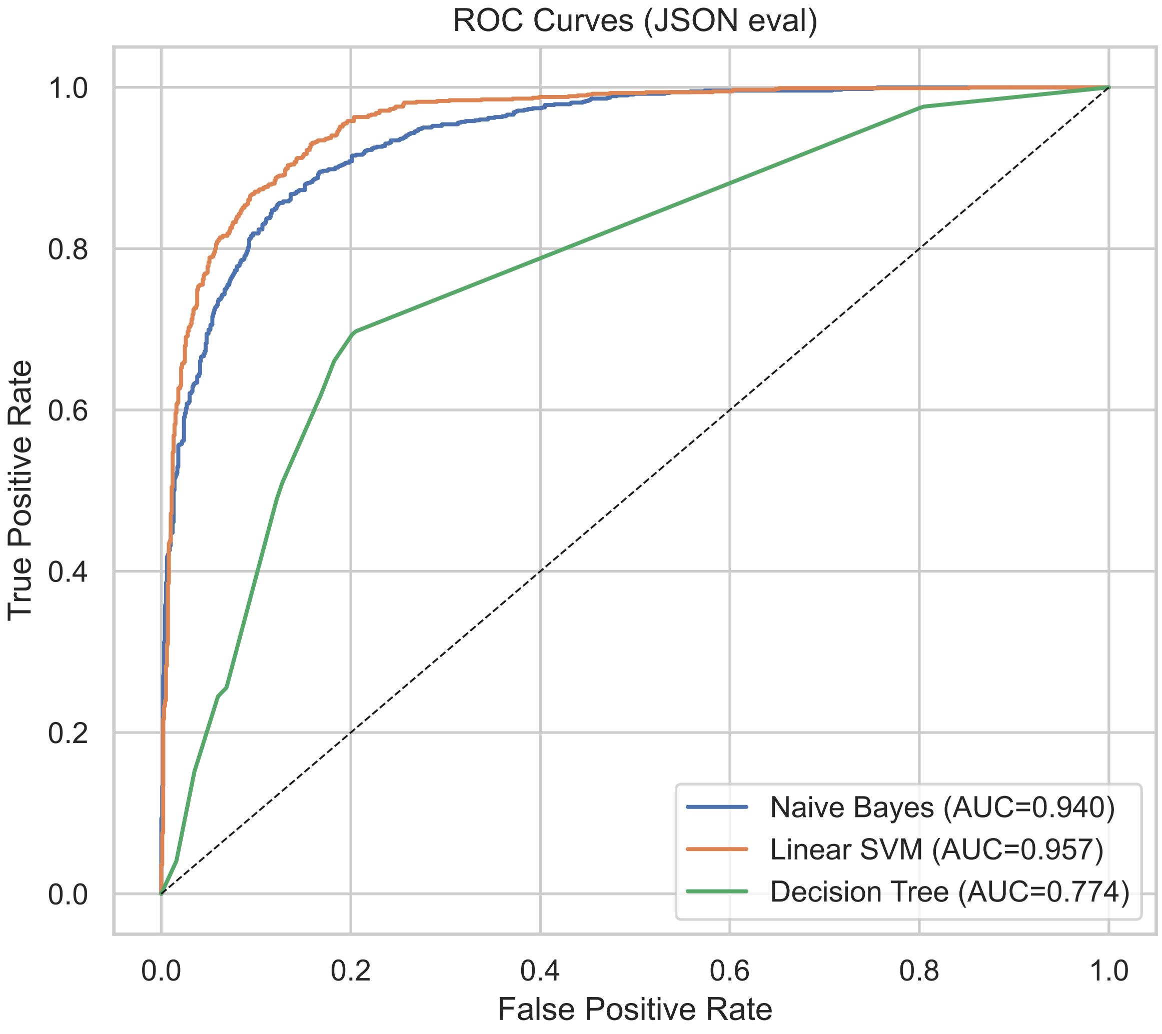

Cluster 5 - top TF-IDF words (centroid)
bad
act
watch
good
movi
terribl
one
time
review
script
plot
get
0.00
0.05
0.10
0.15
0.20
0.25
Centroid weight (TF-IDF feature space)

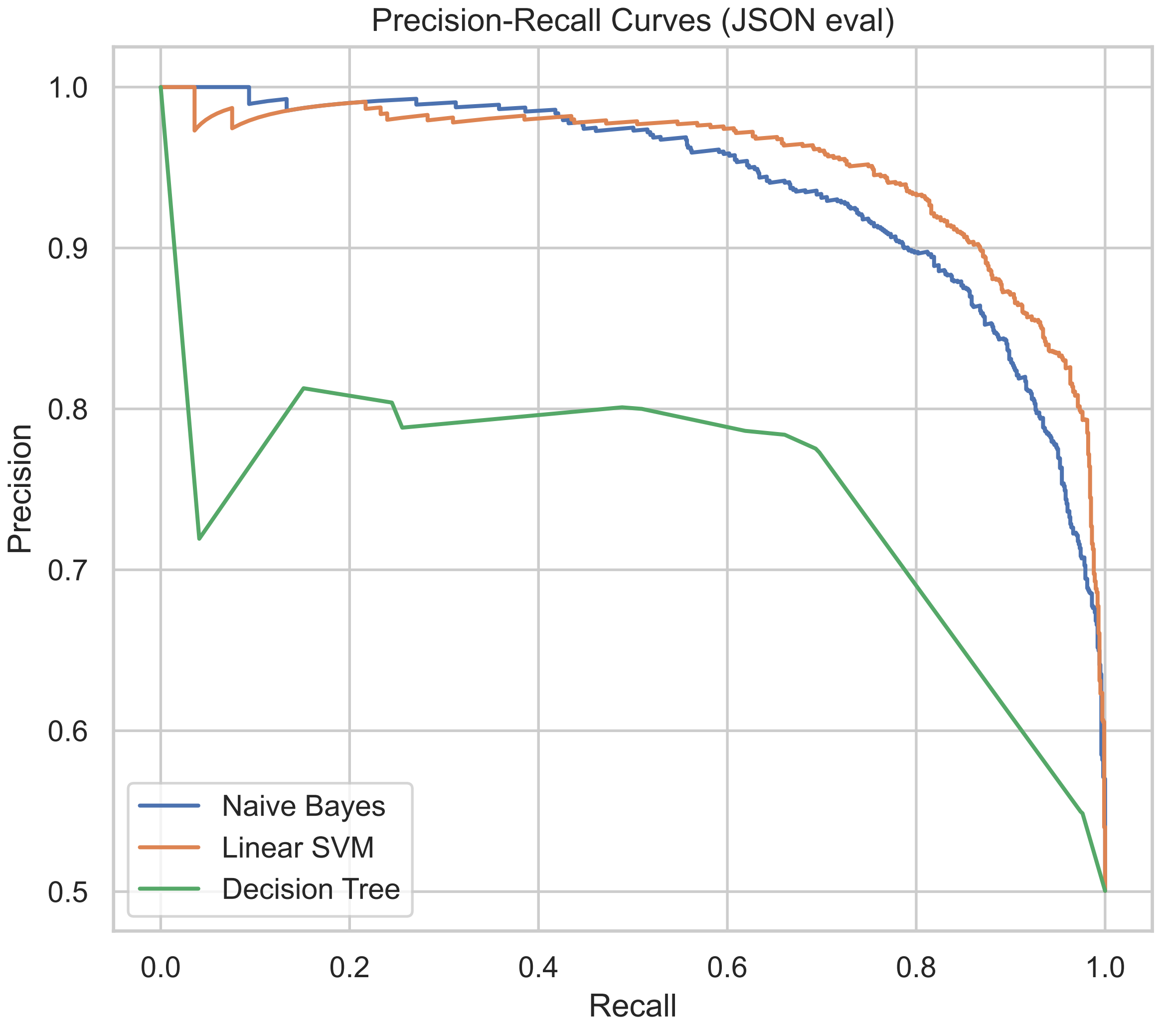

Cluster 6 - top TF-IDF words (centroid)
wast
time
horror
turn
wrong
first
watch
realiti
sequel
one
get
money
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Centroid weight (TF-IDF feature space)


Naive Bayes - Confusion Matrix
True label
0
1
735
268
56
949
0
1
Predicted label

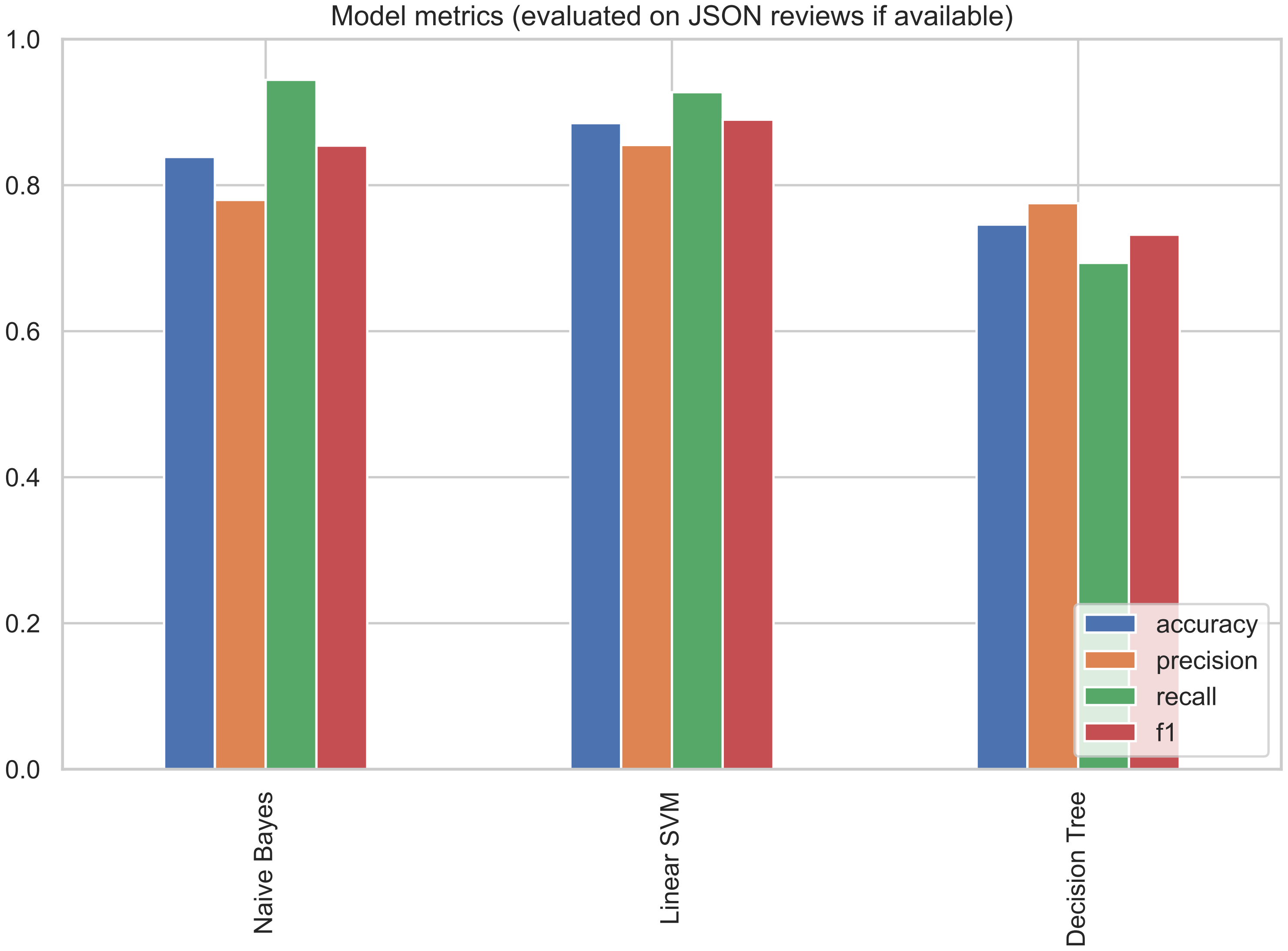

Linear SVM - Confusion Matrix
845
158
73
932
0
1
True label
0
1
Predicted label


Decision Tree - Confusion Matrix
801
202
308
697
0
1
True label
0
1
Predicted label

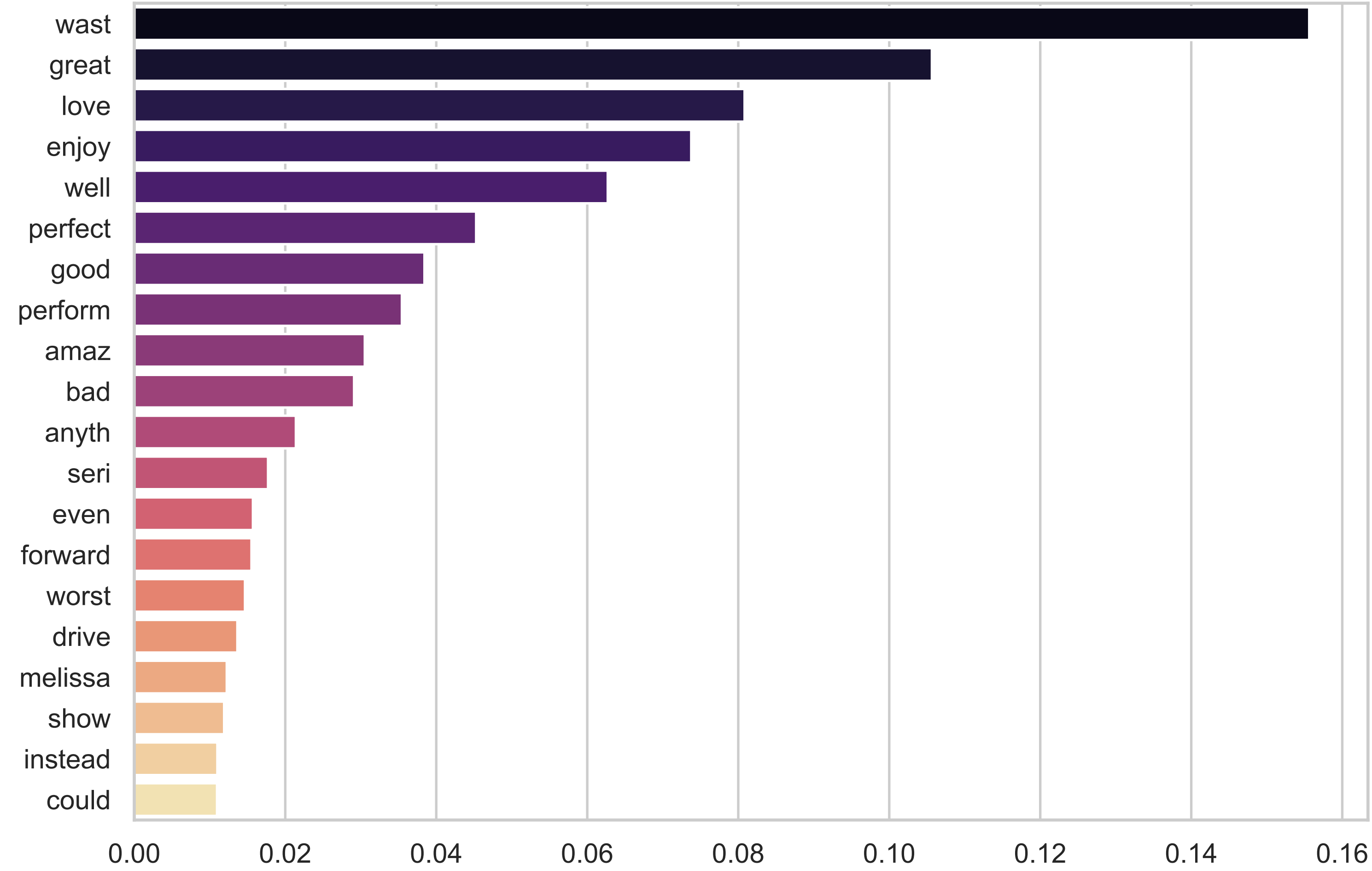

ROC Curves (JSON eval)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Naive Bayes (AUC=0.940)
Linear SVM (AUC=0.957)
Decision Tree (AUC=0.774)

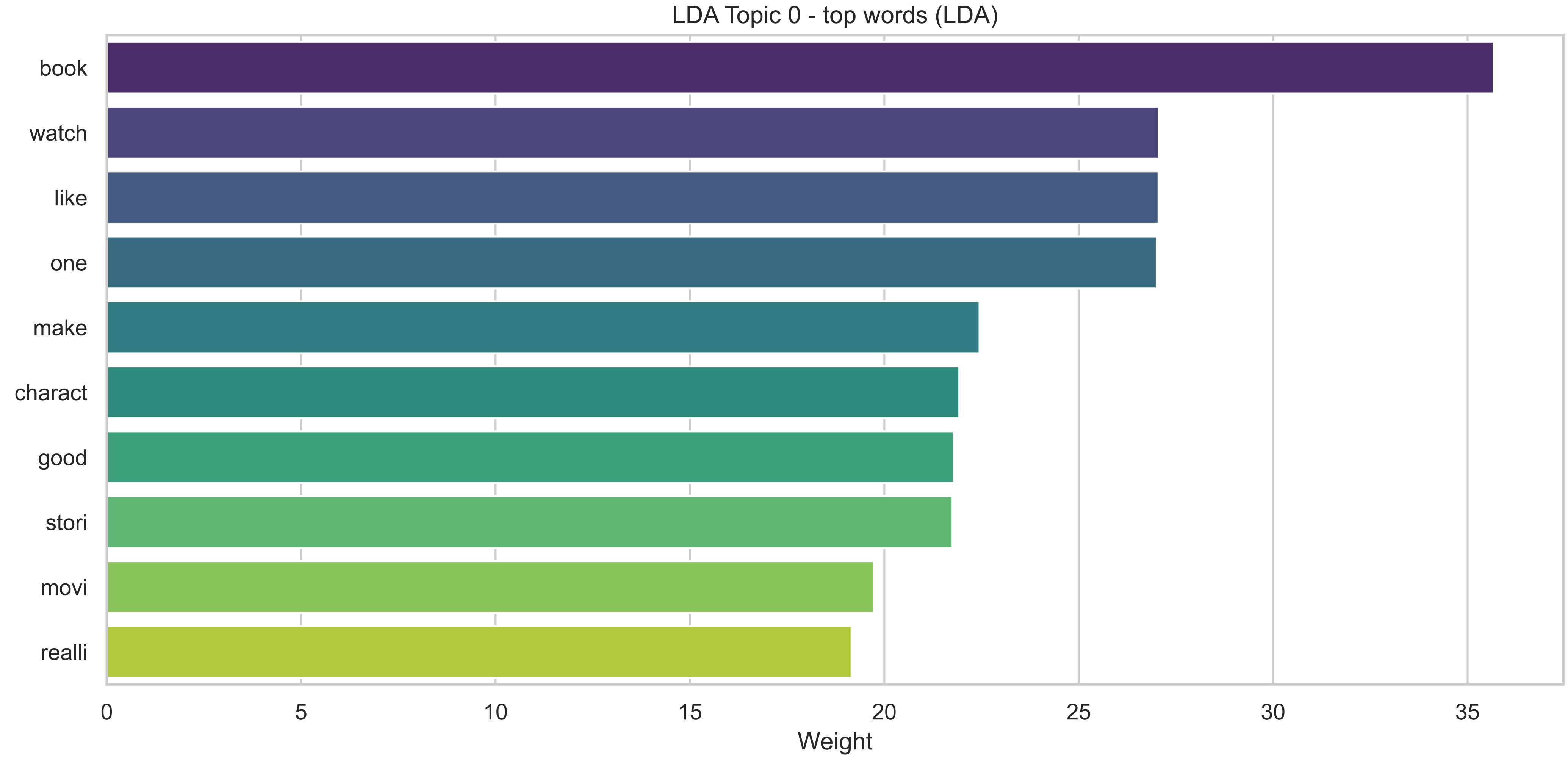


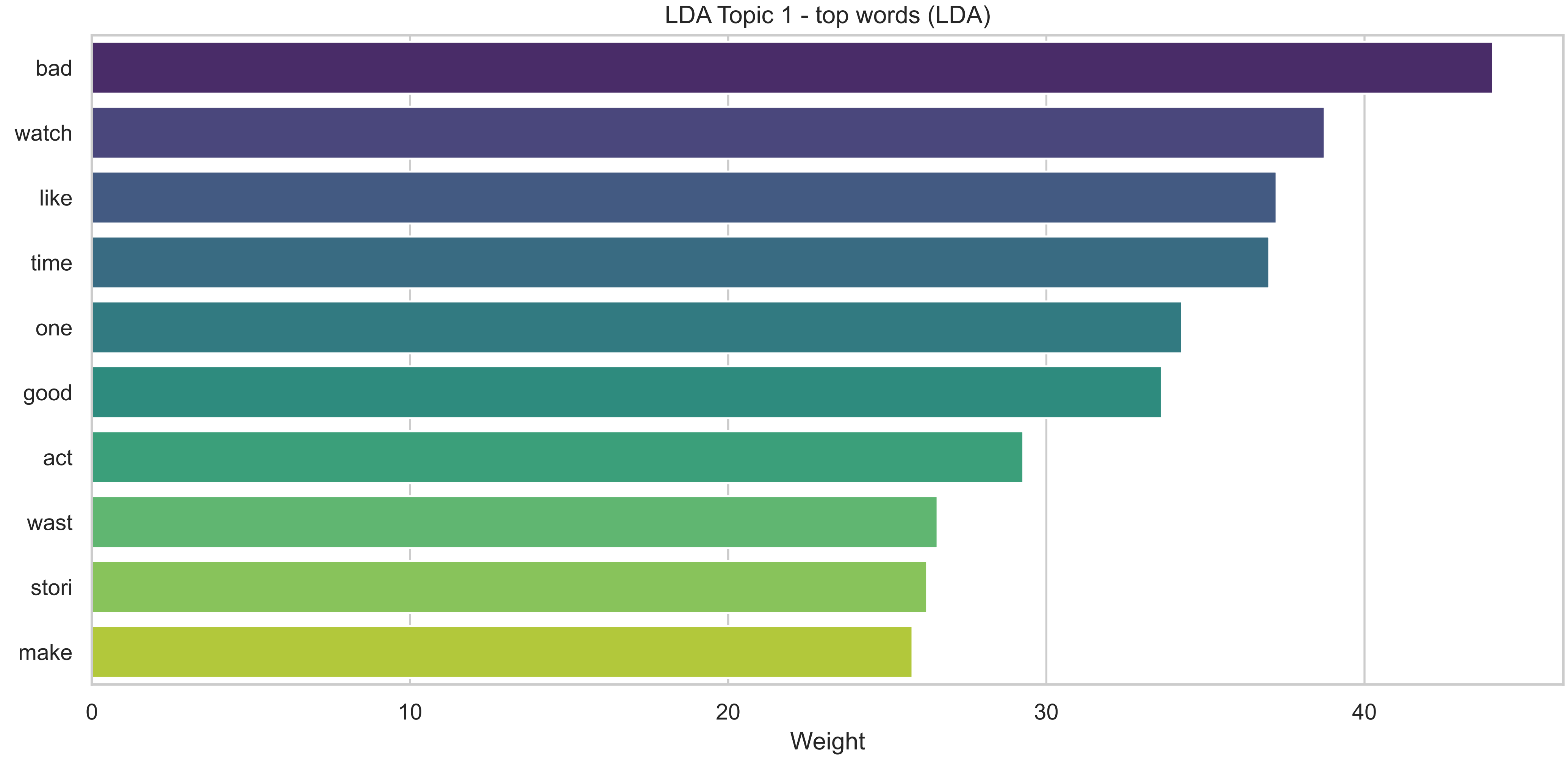
Precision-Recall Curves (JSON eval)
Precision
1.0
0.9
0.8
0.7
0.6
0.5
Recall
0.0
0.2
0.4
0.6
0.8
1.0
Naive Bayes
Linear SVM
Decision Tree

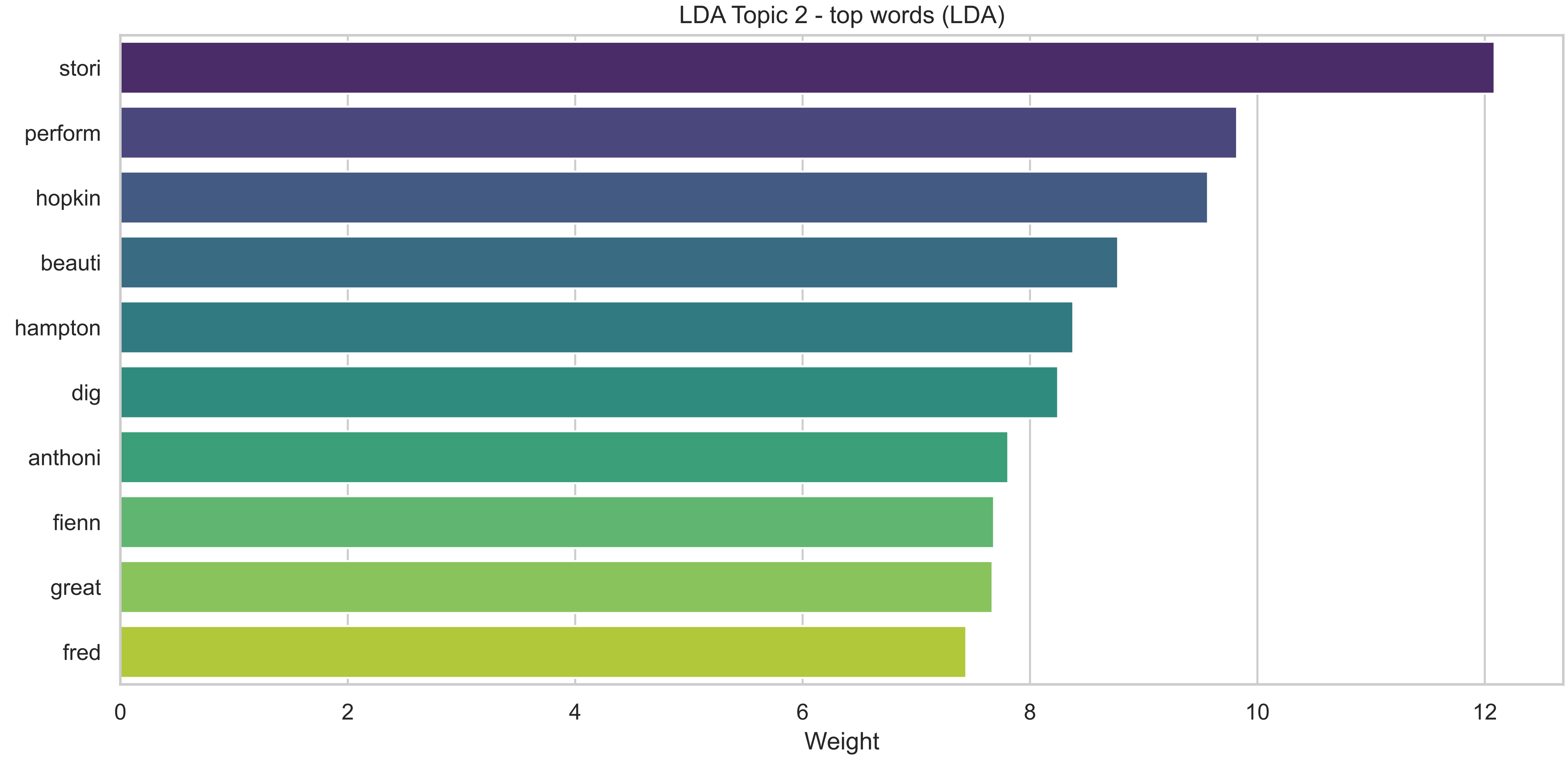

Model metrics (evaluated on JSON reviews if available)
1.0
0.8
0.6
0.4
0.2
0.0
Naive Bayes
Linear SVM
Decision Tree
accuracy
precision
recall
f1

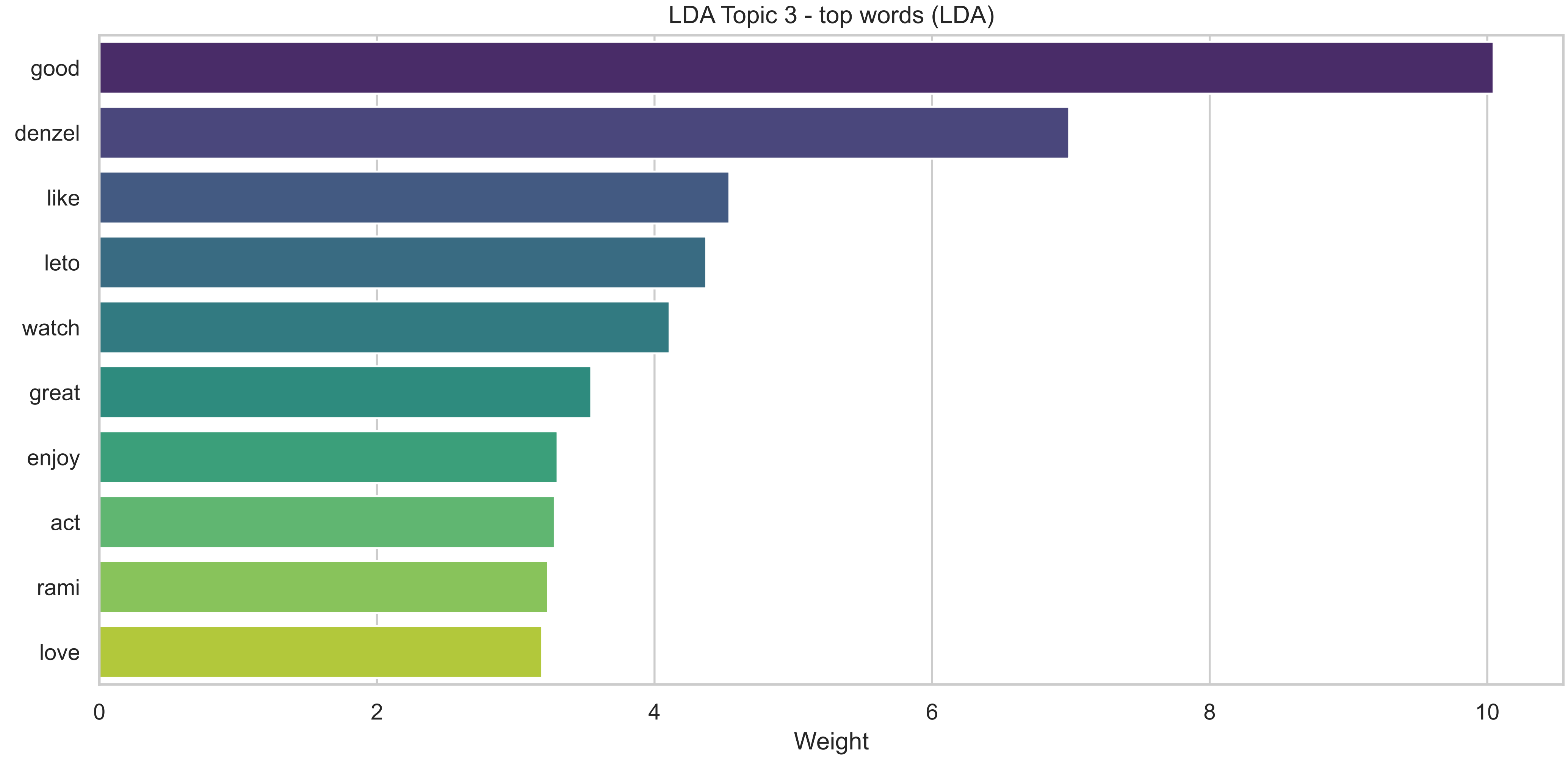
Decision Tree - top feature importances (training)

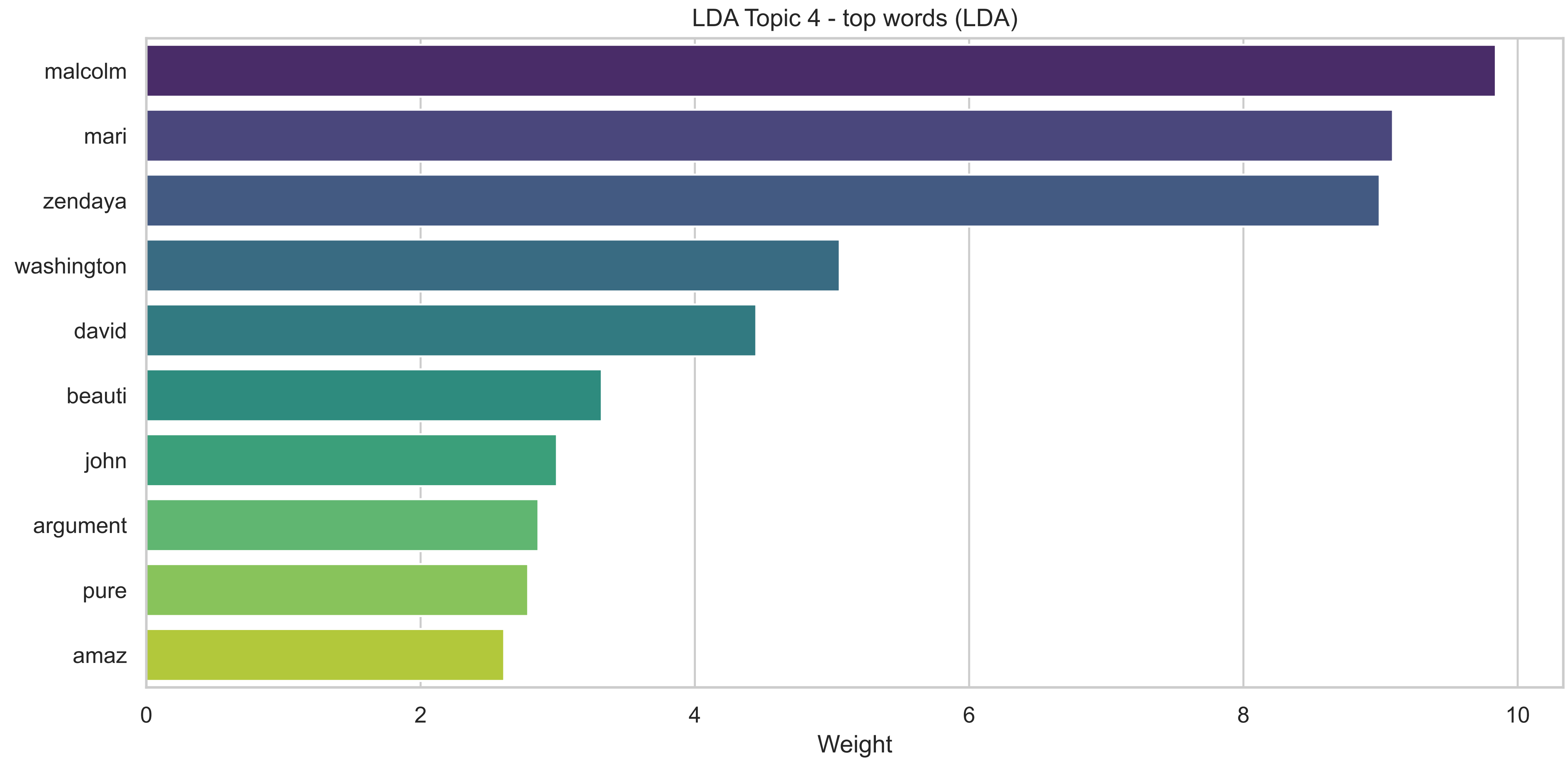



LDA Topic 0 - top words (LDA)
book
watch
like
one
make
charact
good
stori
movi
realli
0
5
10
15
20
25
30
35
Weight


LDA Topic 1 - top words (LDA)
bad
watch
like
time
one
good
act
wast
stori
make
0
10
20
30
40
Weight


LDA Topic 2 - top words (LDA)
stori
perform
hopkin
beauti
hampton
dig
anthoni
fienn
great
fred
0
2
4
6
8
10
12
Weight


LDA Topic 3 - top words (LDA)
good
denzel
like
leto
watch
great
enjoy
act
rami
love
0
2
4
6
8
10
Weight


LDA Topic 4 - top words (LDA)
malcolm
mari
zendaya
washington
david
beauti
john
argument
pure
amaz
0
2
4
6
8
10
Weight

```
Classification reports (evaluated on JSON reviews if available)

Model: Naive Bayes
              precision    recall  f1-score   support

         NEG       0.93      0.73      0.82      1003
         POS       0.78      0.94      0.85      1005

    accuracy                           0.84      2008
   macro avg       0.85      0.84      0.84      2008
weighted avg       0.85      0.84      0.84      2008


Model: Linear SVM
              precision    recall  f1-score   support

         NEG       0.92      0.84      0.88      1003
         POS       0.86      0.93      0.89      1005

    accuracy                           0.88      2008
   macro avg       0.89      0.88      0.88      2008
weighted avg       0.89      0.88      0.88      2008


Model: Decision Tree
              precision    recall  f1-score   support

         NEG       0.72      0.80      0.76      1003
         POS       0.78      0.69      0.73      1005

    accuracy                           0.75      2008
   macro avg       0.75      0.75      0.75      2008
weighted avg       0.75      0.75      0.75      2008
```