PROJECT 3: WEB API'S & NLP

BY: KYLE REED

PROBLEM STATEMENT

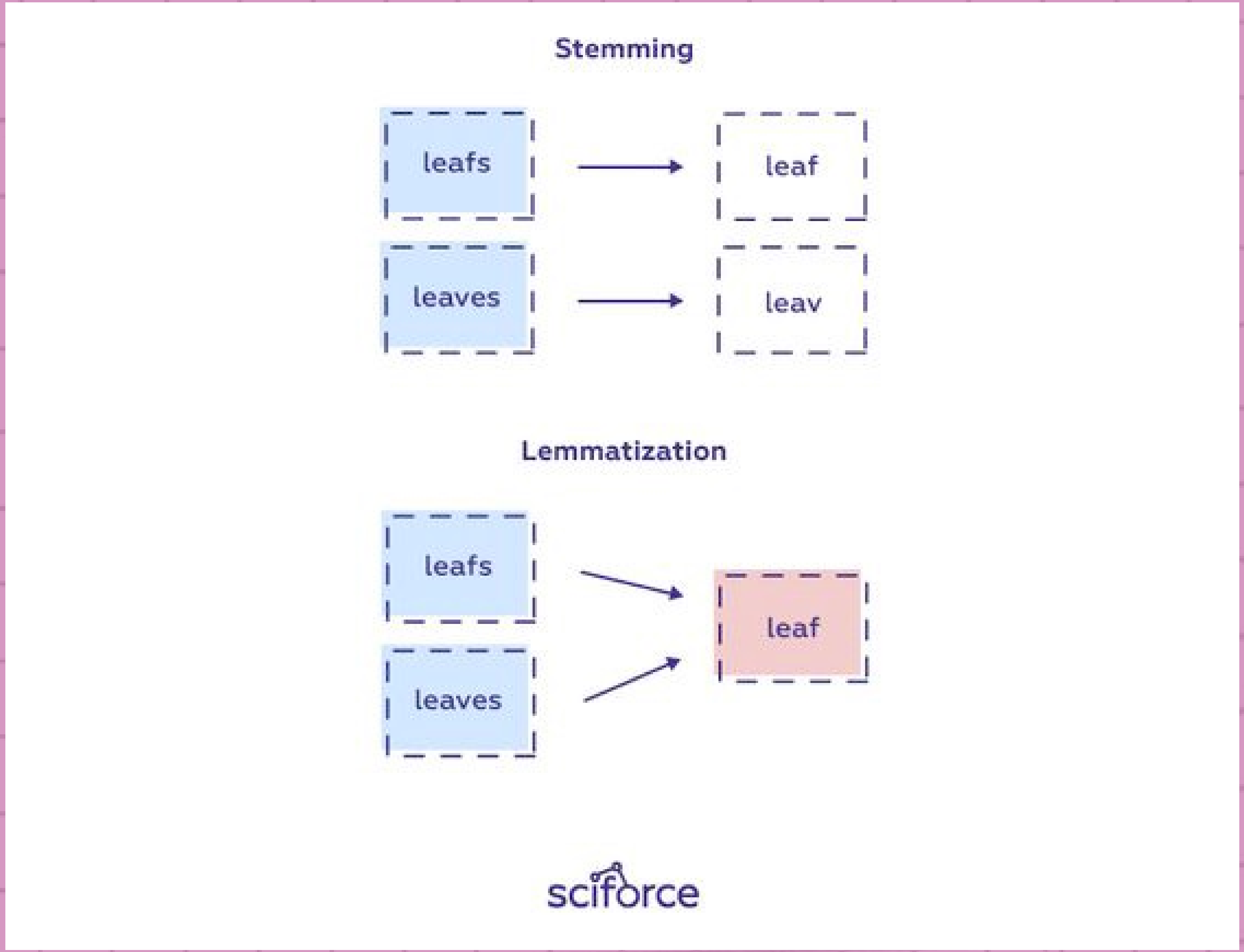
'I Love You Man' - DreamWorks

WHAT IS THE BEST POSSIBLE CLASSIFIER MODEL TO USE TO HELP PREDICT BETWEEN TWO SUBREDDITS GIVEN A RANDOM SUBMISSION POST?

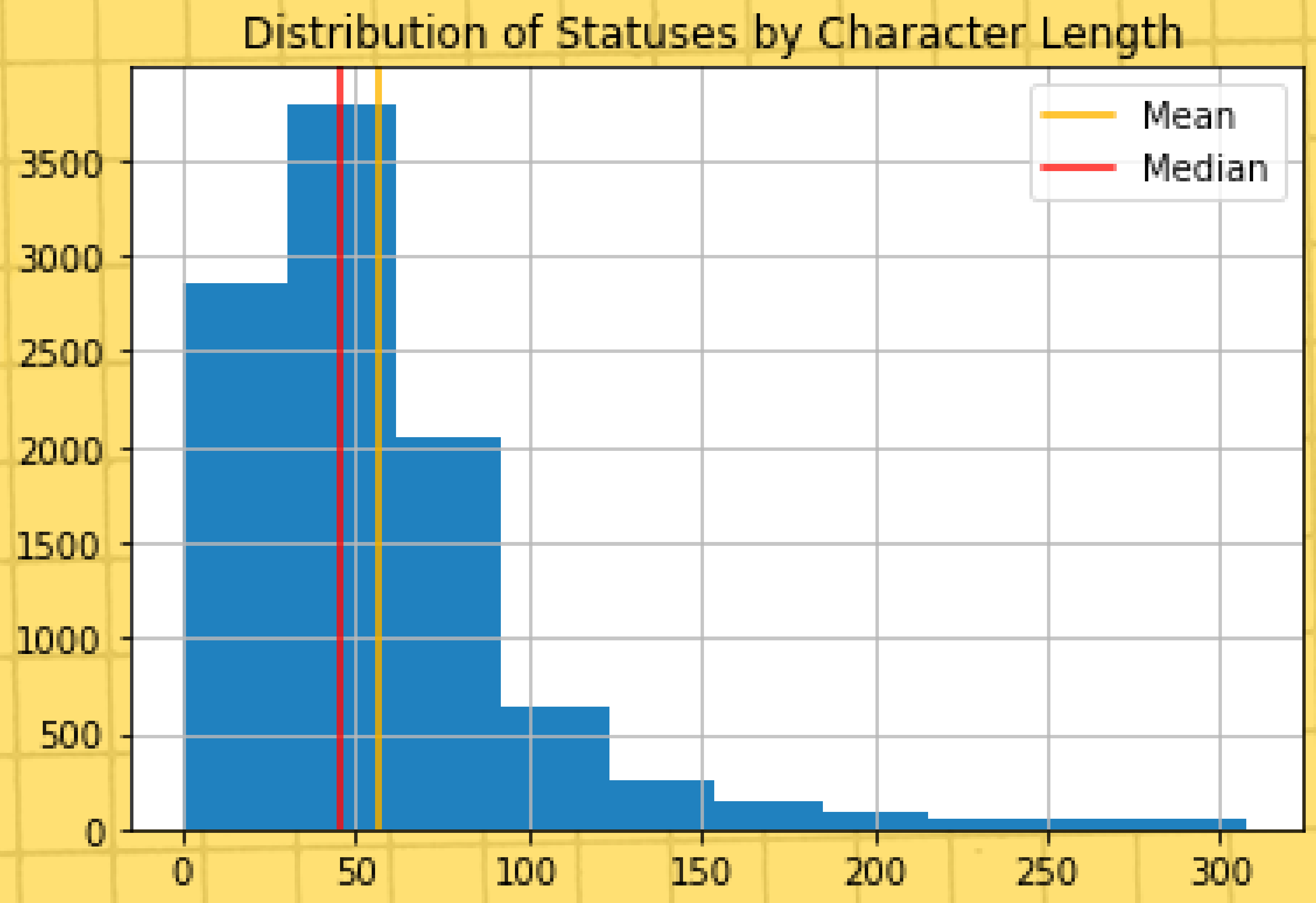
APPROACH

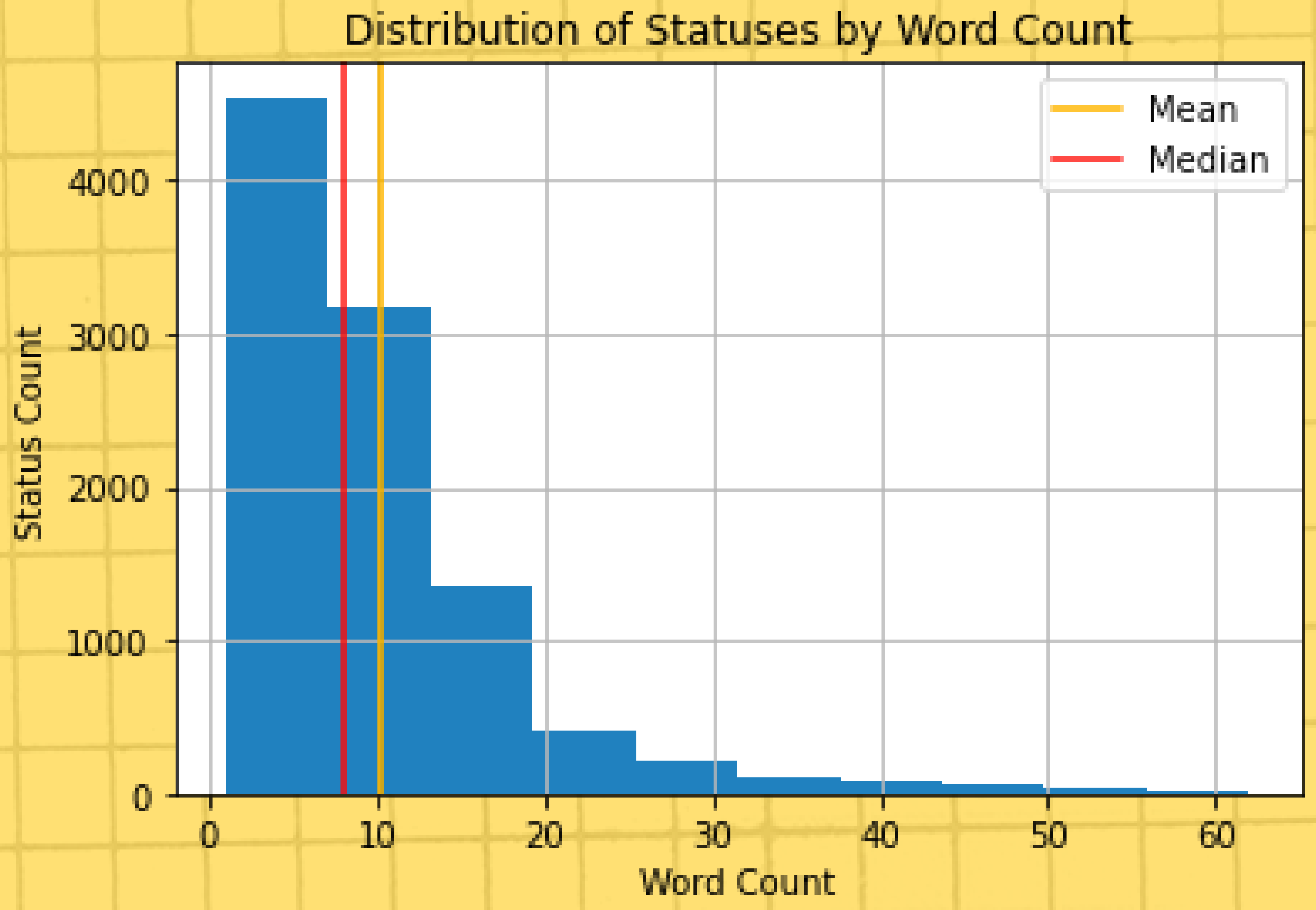
- WEBSCRAPING FUNCTION
- RETRIEVE,READ,CLEAN DATA
- EDA
- MODELING/EVALUATIONS

CLEANING

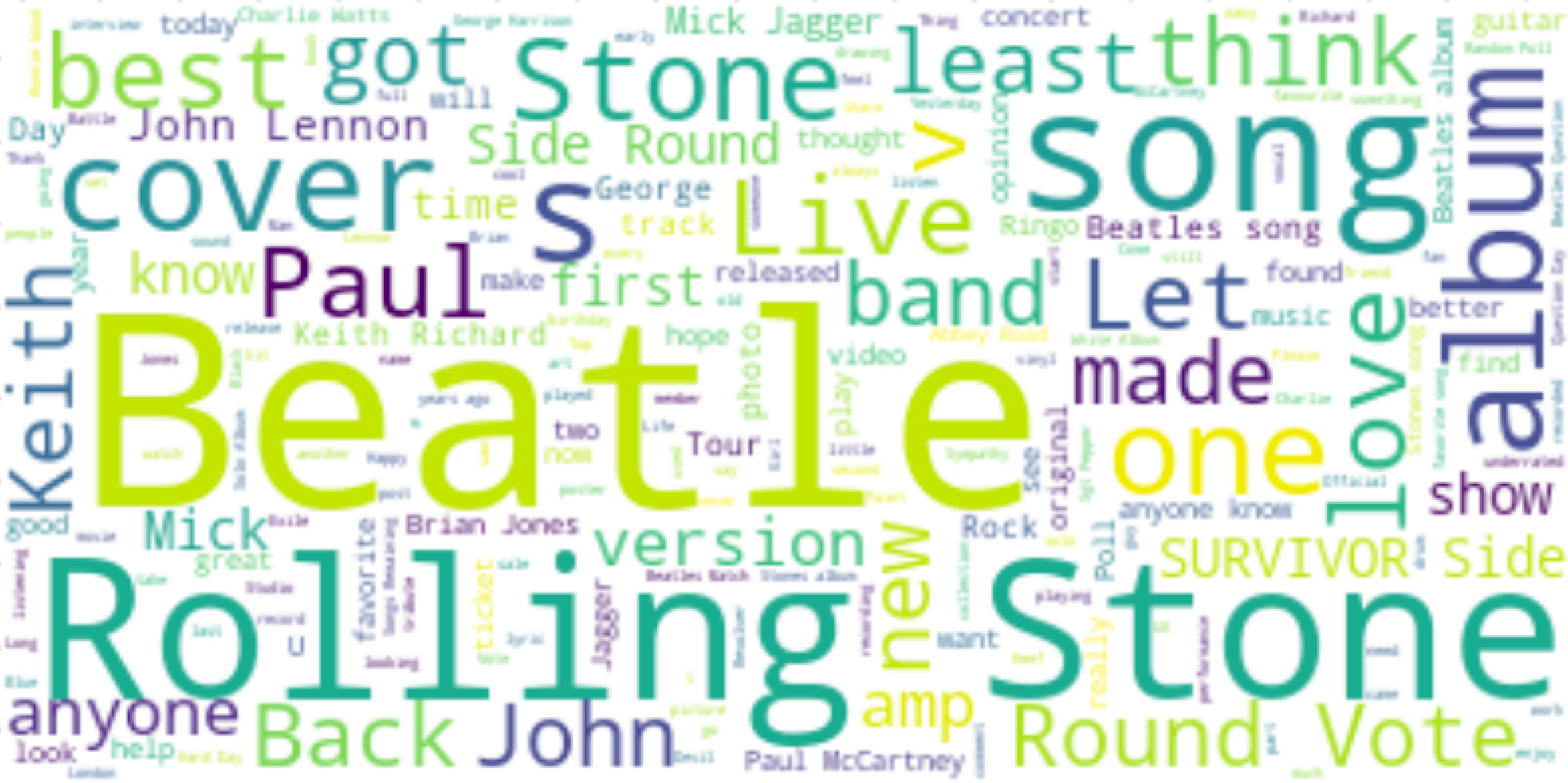

Stemming
leafs
leaf
leaves
leav
Lemmatization
leafs
leaves
leaf
sciforce

EDA


Distribution of Statuses by Character Length
Mean
Median
0
500
1000
1500
2000
2500
3000
3500
0
50
100
150
200
250
300


Distribution of Statuses by Word Count
Mean
Median
Status Count
0
1000
2000
3000
4000
0
10
20
30
40
50
60
Word Count

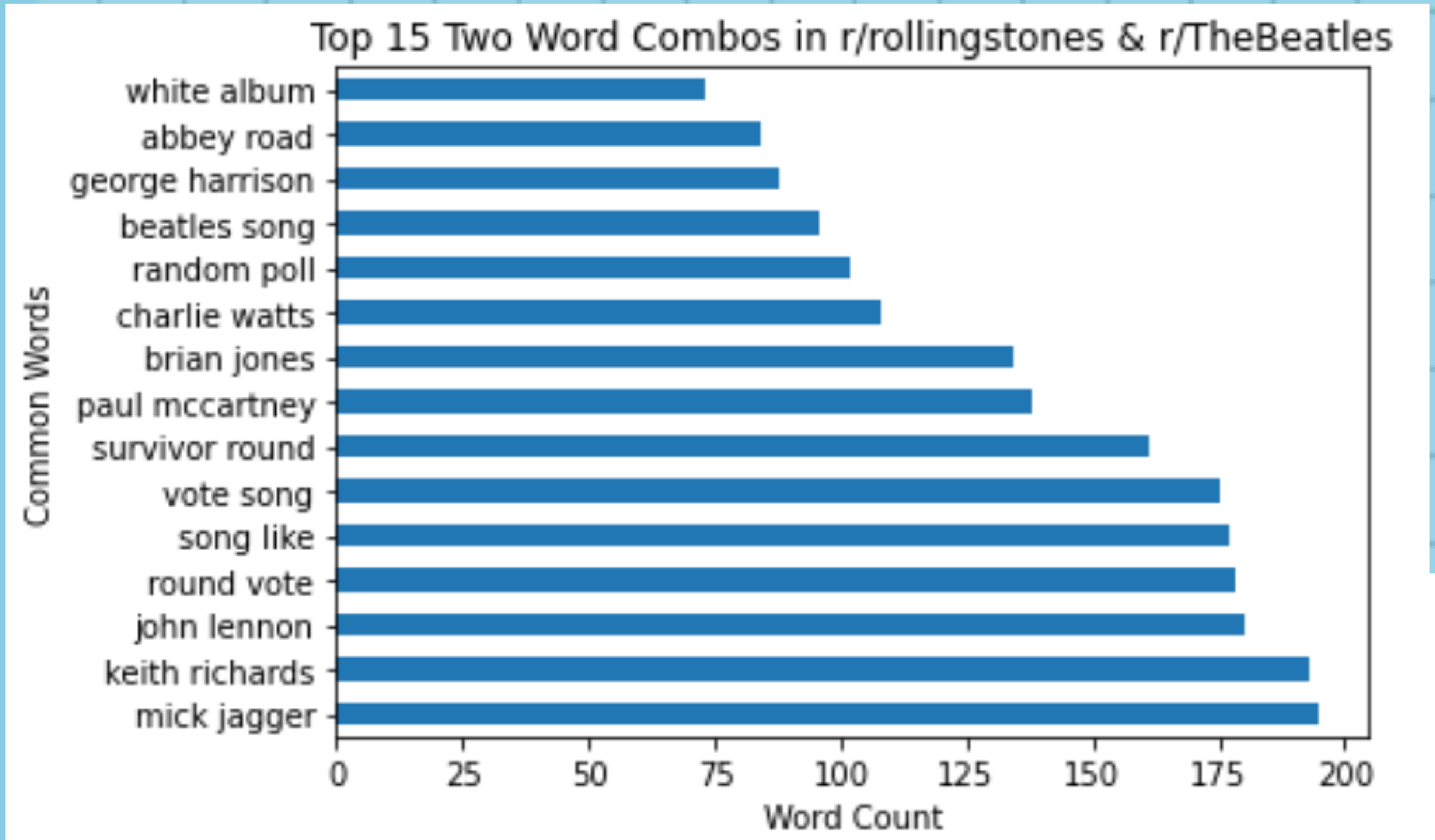
EDA



COMMON WORDS

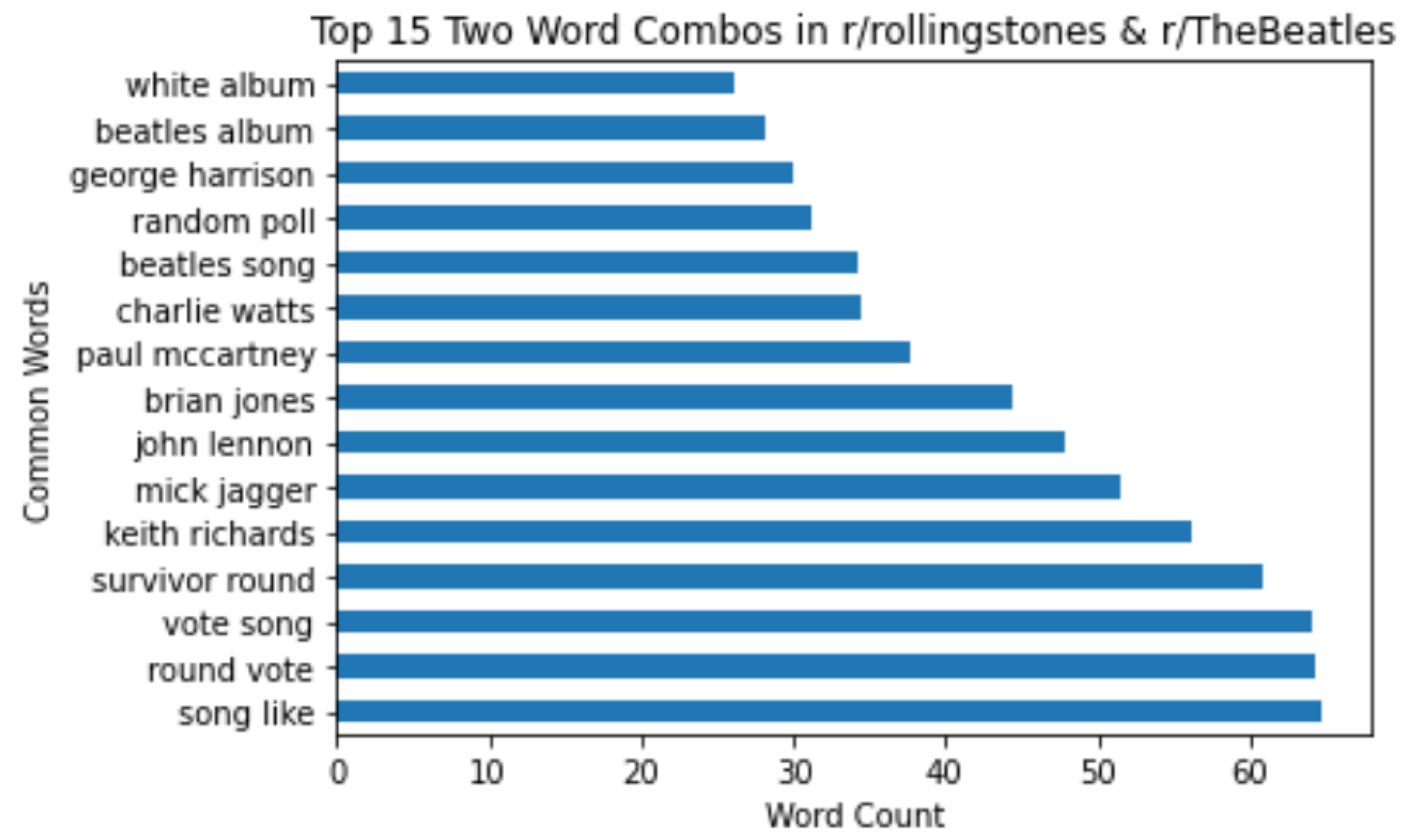
1. BEATLE
2. ROLLING STONE
3. SONG
4. ALBUM
5. COVER

EDA



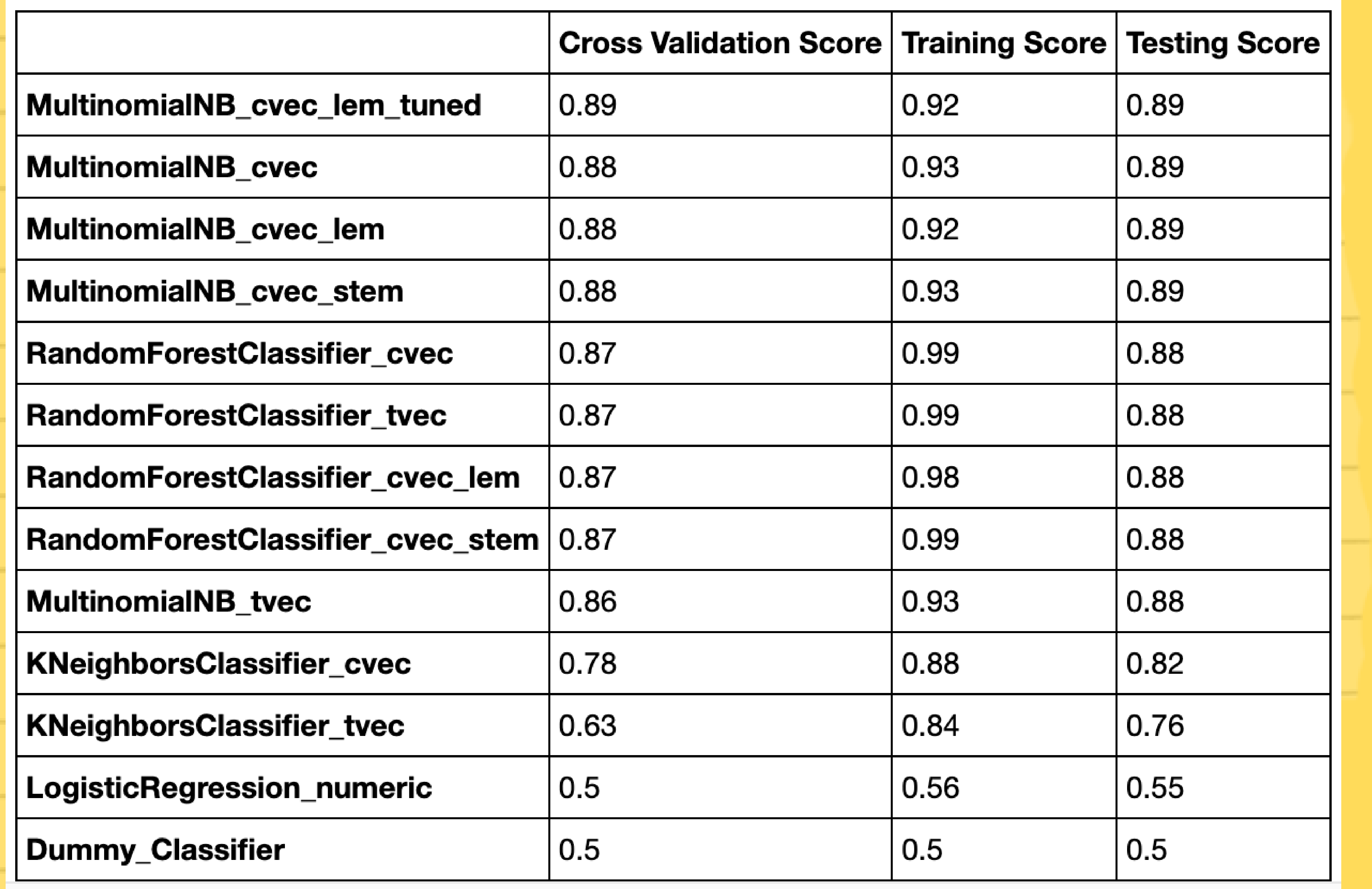
CountVectorizer

TFIDFVectorizer



MODELING SCORES

Photo by: Joel Kleon - Rolling Stone



	Cross Validation Score	Training Score	Testing Score
MultinomialNB_cvec_lem_tuned	0.89	0.92	0.89
MultinomialNB_cvec	0.88	0.93	0.89
MultinomialNB_cvec_lem	0.88	0.92	0.89
MultinomialNB_cvec_stem	0.88	0.93	0.89
RandomForestClassifier_cvec	0.87	0.99	0.88
RandomForestClassifier_tvec	0.87	0.99	0.88
RandomForestClassifier_cvec_lem	0.87	0.98	0.88
RandomForestClassifier_cvec_stem	0.87	0.99	0.88
MultinomialNB_tvec	0.86	0.93	0.88
KNeighborsClassifier_cvec	0.78	0.88	0.82
KNeighborsClassifier_tvec	0.63	0.84	0.76
LogisticRegression_numeric	0.5	0.56	0.55
Dummy_Classifier	0.5	0.5	0.5

EVALUATIONS

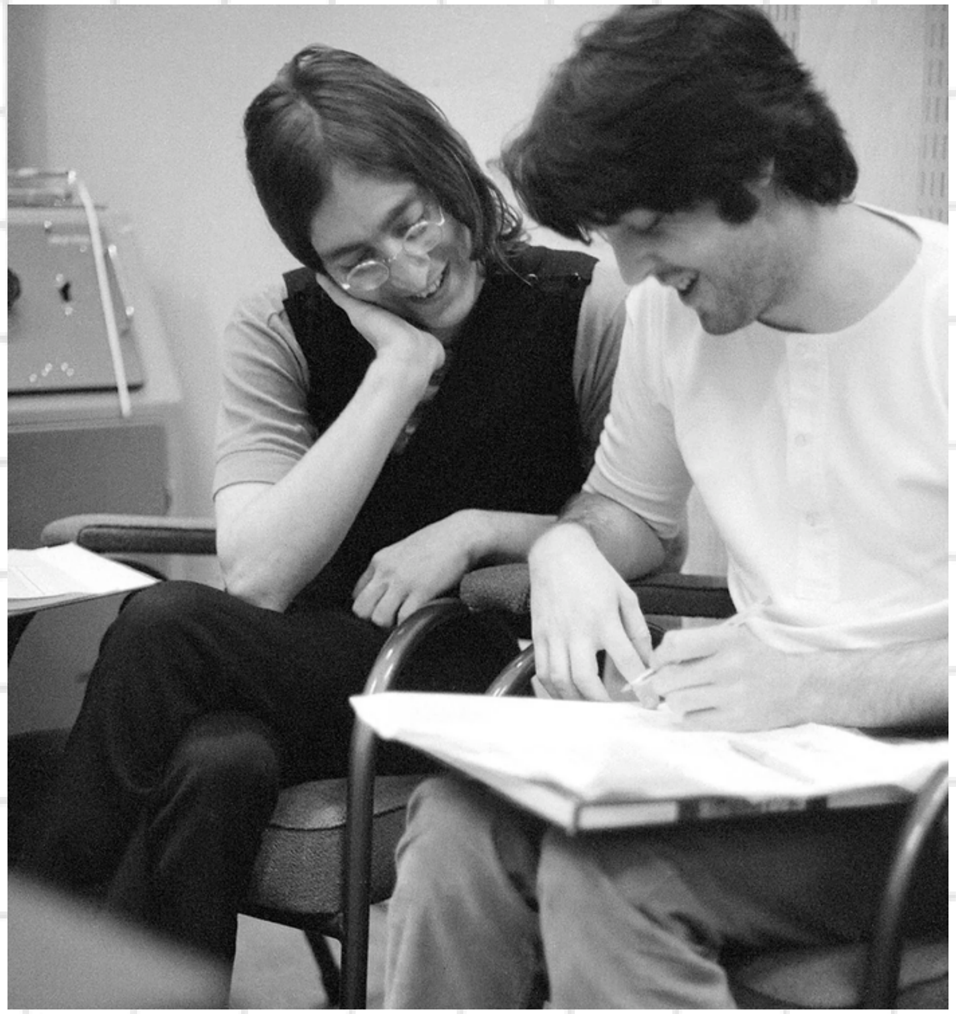
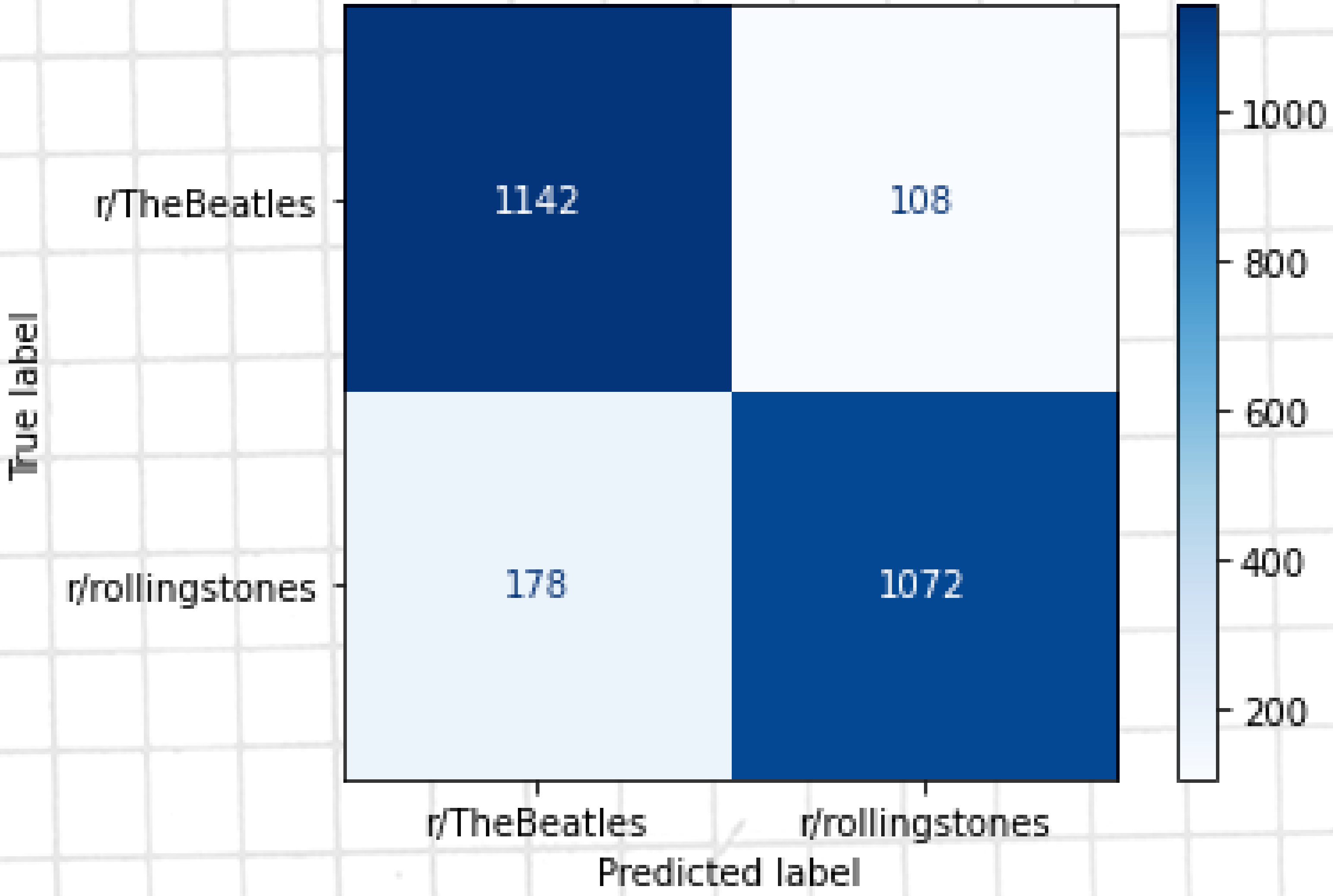
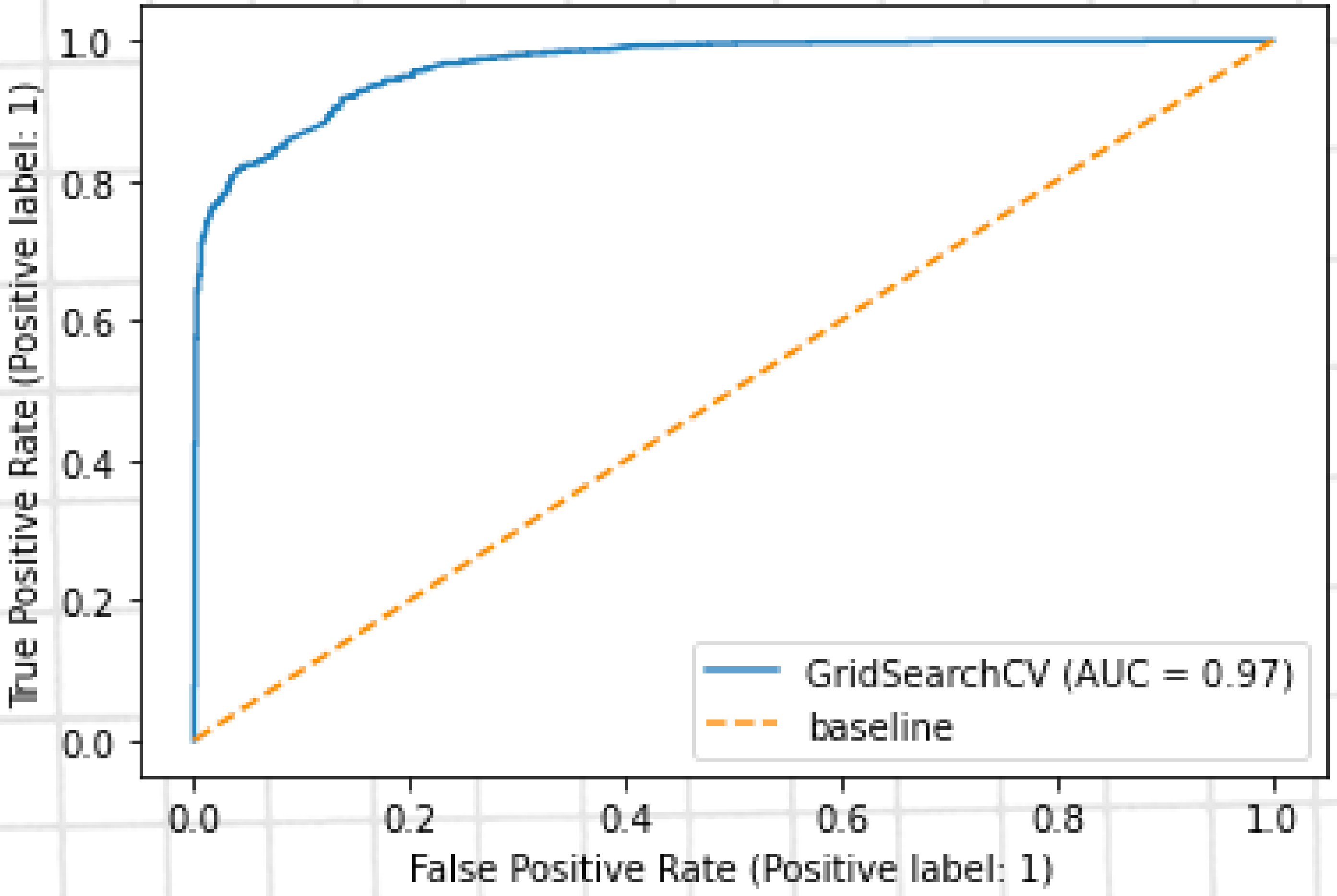


Photo by: Linda McCartney



	r/TheBeatles	r/rollingstones
r/TheBeatles	1142	108
r/rollingstones	178	1072



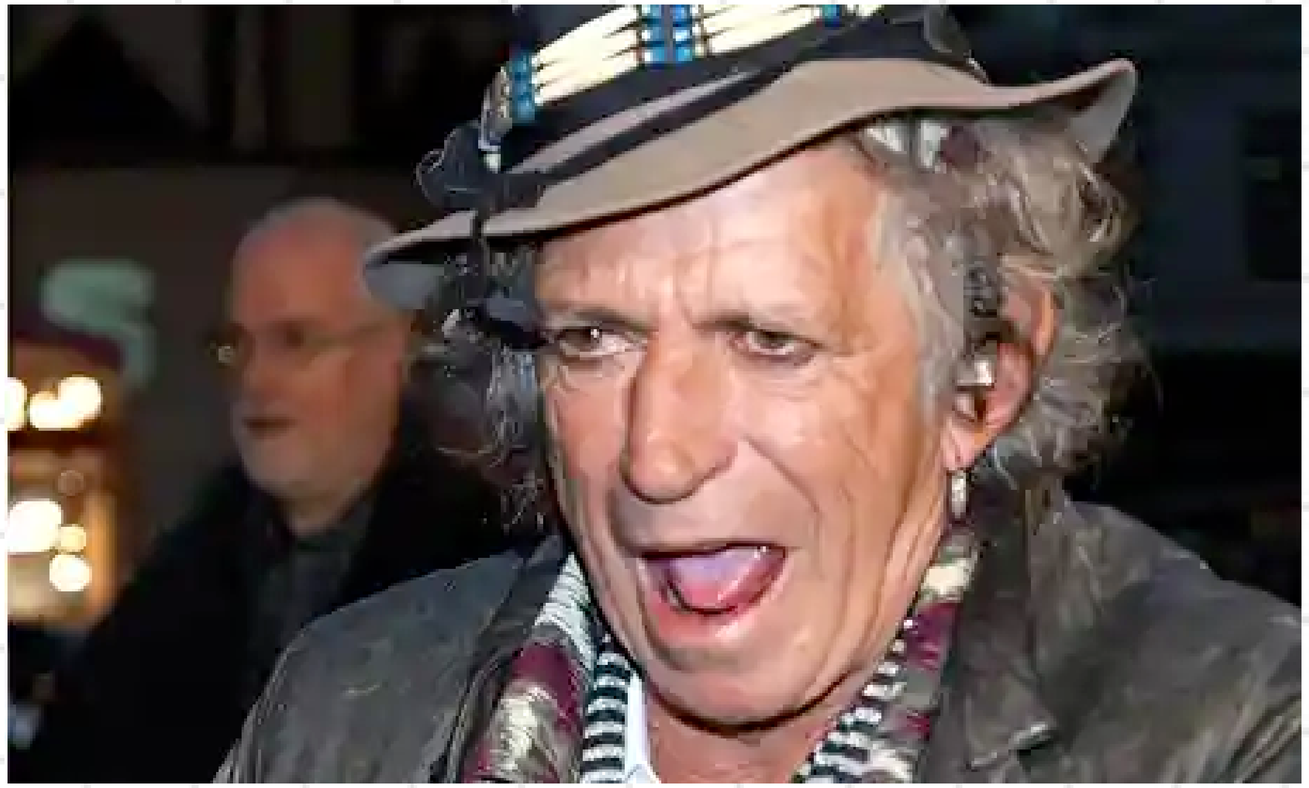


Photo by: Max Nash/AFP/Getty Images

SUMMARY

1. PRIMARY FEATURE USED: TITLE COLUMN
2. BEST PERFORMING CLASSIFIER MODEL: MULTINOMIAL NAIVE BAYES
3. TRANSFORMER OF CHOICE: COUNT-VECTORIZER
4. CV = .89, TRAIN = .92, TEST = .89
5. ROC AUC = 0.97

AND THAT'S A WRAP.

QUESTIONS?