Movie Genre Prediction - Milestone 1

Group 2: Charles, Giffin, Jon G, Canxiu

Problem Definition: Movie Genre prediction

The problem is supervised text classification, and our goal is to investigate which supervised machine learning methods are best suited to solve it.

Given a new movie information (overview, company, director, budget, ...), we want to assign it to one of top genres. The classifier makes the assumption that each new movie is assigned to at least one dominant genres. This is multi-label classification problem.

Input: Overview (Text) + other supporting features (Numerical)
Output: Genre

Initial Data Exploration

Link to dataset:
https://www.kaggle.com/datasets/juzershakir/tmdb-movies-dataset?select=tmdb_movies_data.csv

The dataset contains observations of 10k+ movies like title, popularity, budget, revenue, cast, director, tagline, keywords, **overview, genres,** release date, runtime etc.

- Clean and pre-processing ‘overview’
- Upto 5 Genres per movie and thousands of combination. Parse Genre column into five columns for further analysis
- Split the data 80/20

#	Column	Non-Null Count	Dtype
0	id	8693 non-null	int64
1	imdb_id	8684 non-null	object
2	popularity	8693 non-null	float64
3	budget	8693 non-null	int64
4	revenue	8693 non-null	int64
5	original_title	8693 non-null	object
6	cast	8642 non-null	object
7	homepage	2358 non-null	object
8	director	8658 non-null	object
9	tagline	6413 non-null	object
10	keywords	7525 non-null	object
11	overview	8689 non-null	object
12	runtime	8693 non-null	int64
13	genres	8675 non-null	object
14	production_companies	7872 non-null	object
15	release_date	8693 non-null	object
16	vote_count	8693 non-null	int64
17	vote_average	8693 non-null	float64
18	release_year	8693 non-null	int64
19	budget_adj	8693 non-null	float64
20	revenue_adj	8693 non-null	float64
21	genres_1	8675 non-null	object
22	genres_2	6795 non-null	object
23	genres_3	4044 non-null	object
24	genres_4	1564 non-null	object
25	genres_5	410 non-null	object

Text Cleaning

Movie: The Hunger Games: Mockingjay - Part 2

Overview:

With the nation of Panem in a full scale war, Katniss confronts President Snow in the final showdown. Teamed with a group of her closest friends Ã¢â‚¬â€œ including Gale, Finnick, and Peeta Ã¢â‚¬â€œ Katniss goes off on a mission with the unit from District 13 as they risk their lives to stage an assassination attempt on President Snow who has become increasingly obsessed with destroying her. The mortal traps, enemies, and moral choices that await Katniss will challenge her more than any arena she faced in The Hunger Games.

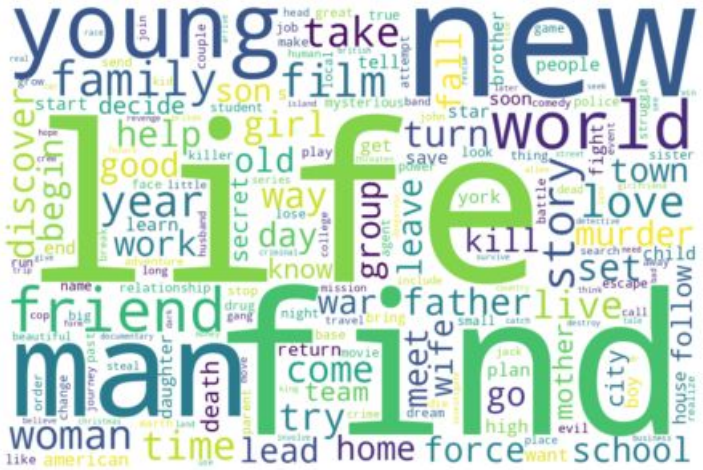
Data preprocessing steps by using spacy:

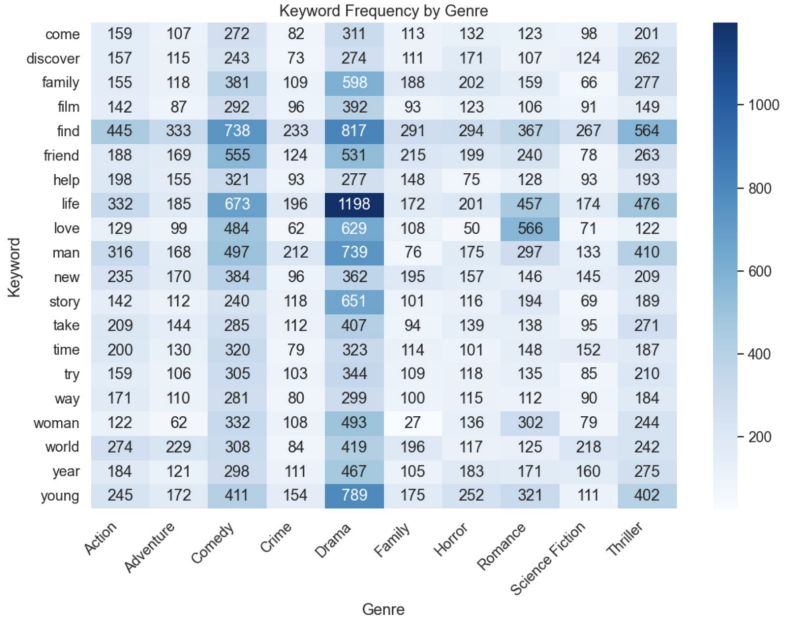
- Remove accent characters
- Expand contractions
- Remove special characters
- Remove stop words

NLP Pre-processing: Lemmatization and Part-of-speech (POS) tagging

New columns:

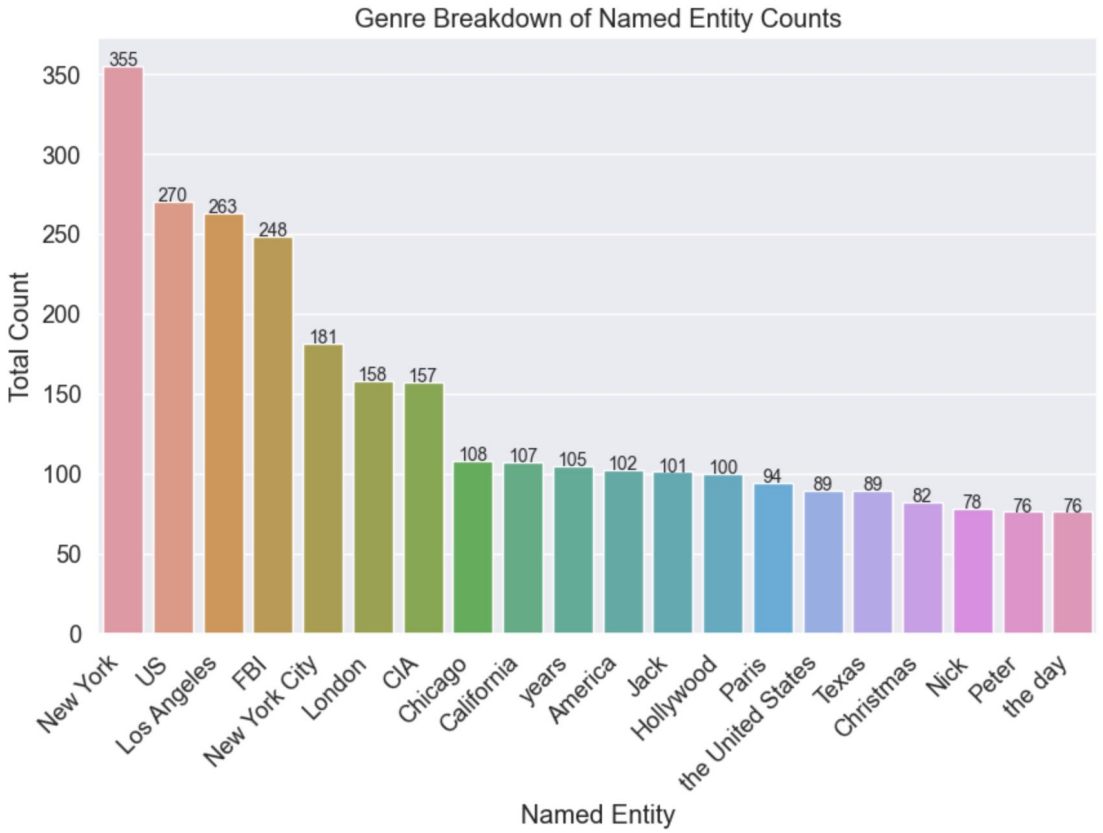
1. **Overview_lemma**; Overview_nouns; Overview_adjectives; Overview_verbs and Overview_nav
2. No_tokens
3. Overview_person; Overview_org; Overview_date; Overview_time; Overview_money and overview_gpe



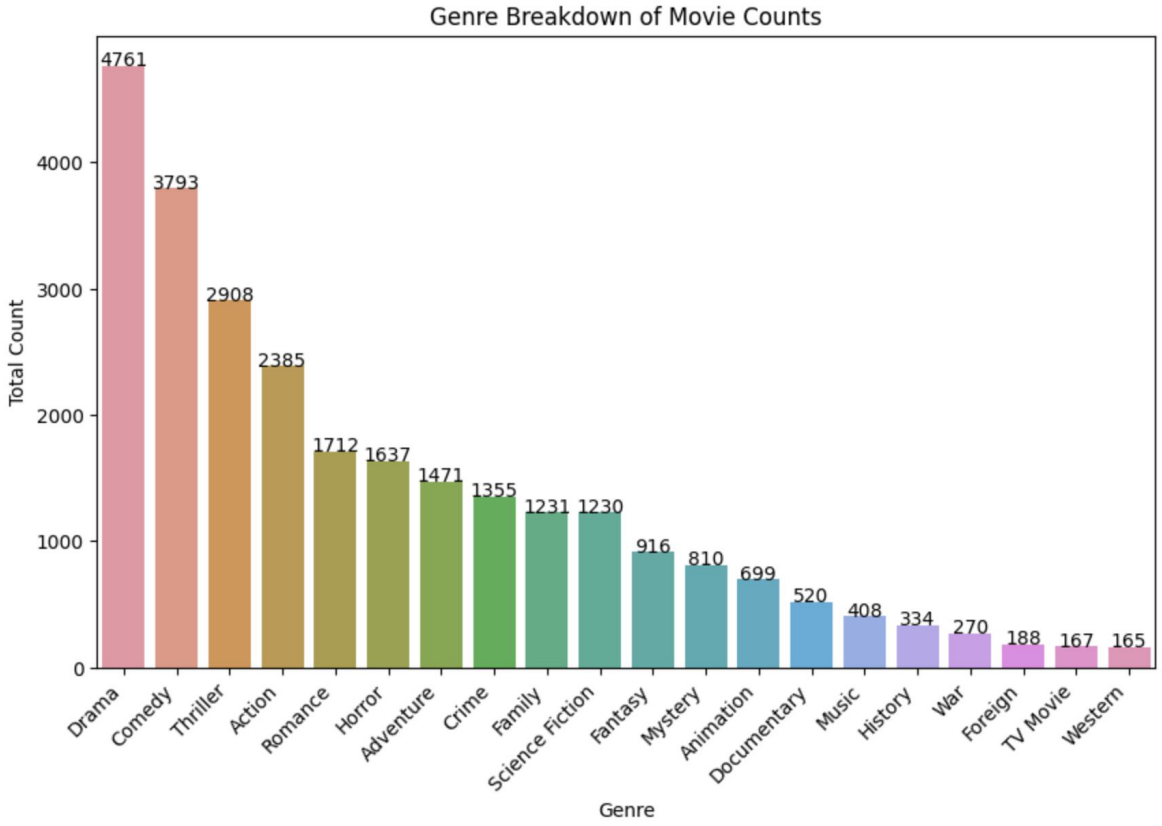


NLP Pre-processing - Named Entity Recognition

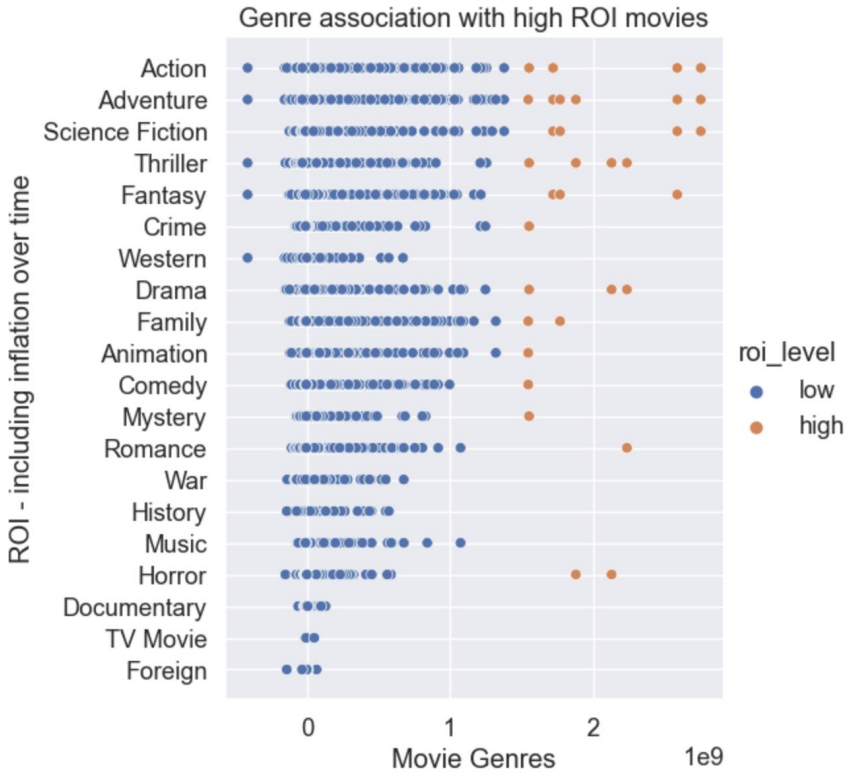
- Person
- Organization
- Date, Time
- Money
- Geopolitical entity



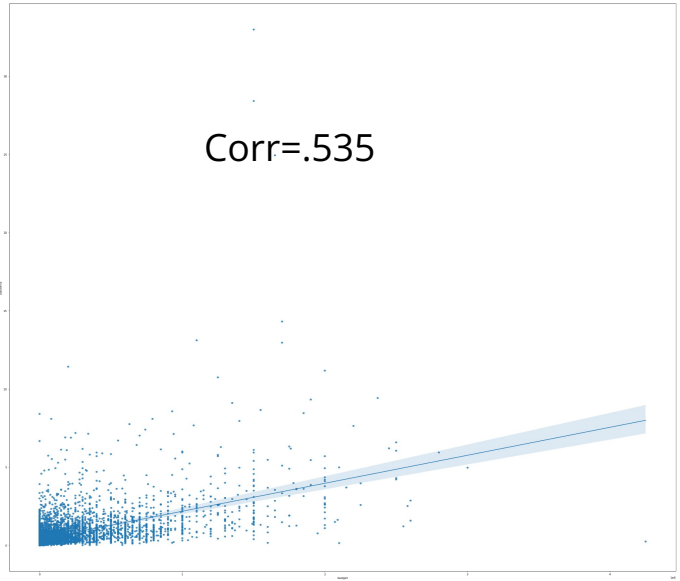
Define scope: Narrow down to top 10 genres

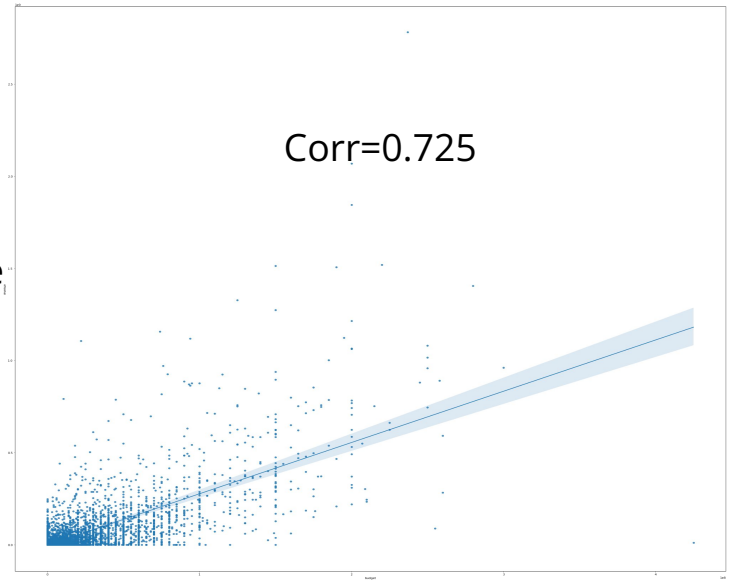


Data Exploration - ROI and Genre

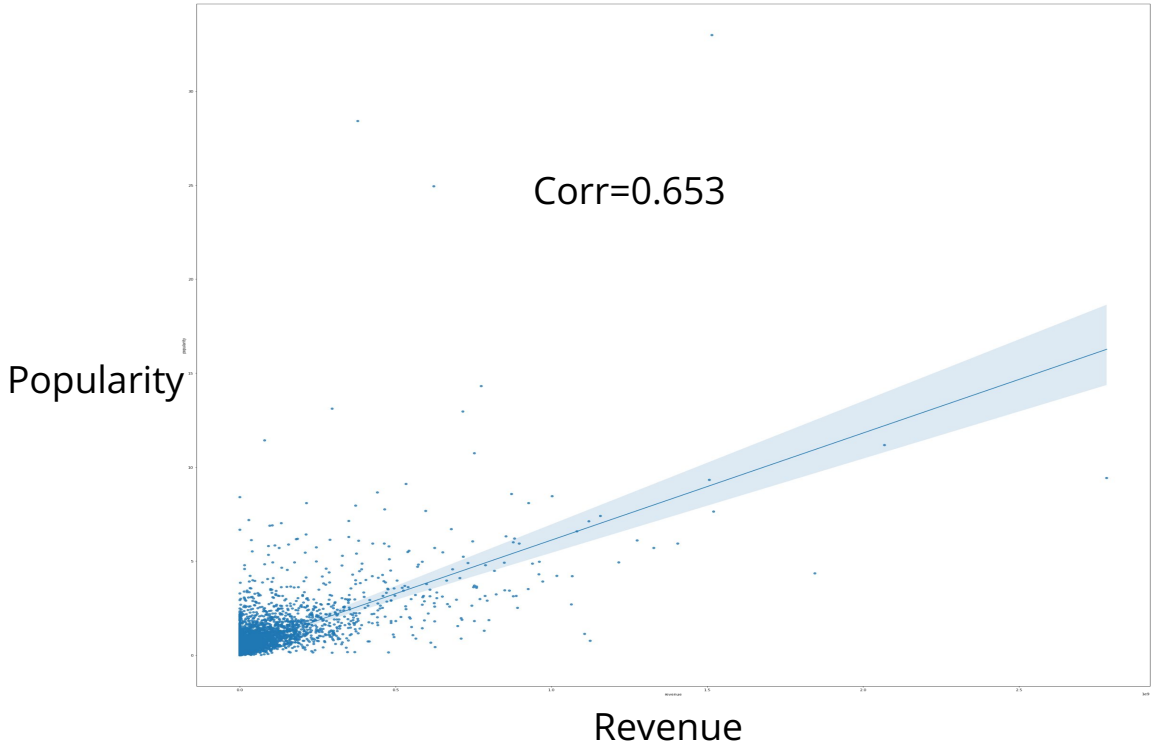


Data Exploration - Explore important features related to Genres





Data Exploration - Explore important features related to Genres



ANOVA For ‘Budget’ and Genre Columns

```
ANOVA results for genres_1:
                   sum_sq      df          F         PR(>F)
C(genres_1)  6.914923e+17    19.0  42.760898  1.423597e-152
Residual     7.366378e+18  8655.0        NaN            NaN
ANOVA results for genres_2:
                   sum_sq      df          F        PR(>F)
C(genres_2)  4.754949e+17    19.0  24.454908  1.678480e-83
Residual     6.933229e+18  6775.0        NaN           NaN
ANOVA results for genres_3:
                   sum_sq      df          F        PR(>F)
C(genres_3)  3.085076e+17    19.0  11.667588  1.947609e-35
Residual     5.600014e+18  4024.0        NaN           NaN
ANOVA results for genres_4:
                   sum_sq      df         F        PR(>F)
C(genres_4)  1.503698e+17    19.0  4.77677  5.693018e-11
Residual     2.558114e+18  1544.0      NaN           NaN
ANOVA results for genres_5:
                   sum_sq     df         F    PR(>F)
C(genres_5)  4.075705e+16   18.0  1.129875  0.320091
Residual     7.835680e+17  391.0       NaN       NaN
```

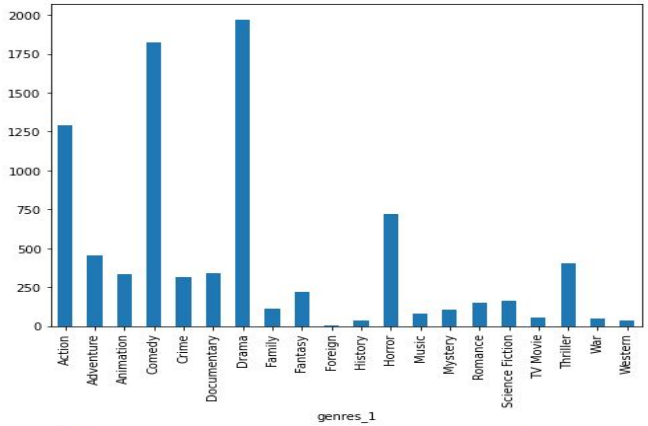
ANOVA For ‘Popularity’ and Genre Columns

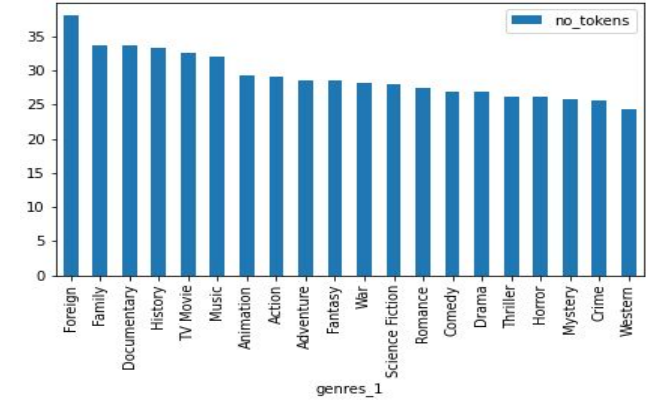
```
ANOVA results for genres_1:
                 sum_sq      df         F        PR(>F)
C(genres_1)   405.888959    19.0  21.47669  6.684103e-73
Residual     8609.012900  8655.0       NaN           NaN
ANOVA results for genres_2:
                 sum_sq      df          F        PR(>F)
C(genres_2)   239.451115    19.0  10.412953  2.981967e-31
Residual     8199.713423  6775.0        NaN           NaN
ANOVA results for genres_3:
                 sum_sq      df         F        PR(>F)
C(genres_3)   226.894664    19.0  7.226218  1.443606e-19
Residual     6649.937577  4024.0       NaN           NaN
ANOVA results for genres_4:
                 sum_sq      df         F    PR(>F)
C(genres_4)    99.488065    19.0  2.479652  0.000399
Residual     3260.423445  1544.0       NaN       NaN
ANOVA results for genres_5:
                sum_sq     df         F    PR(>F)
C(genres_5)   23.128089   18.0  1.988585  0.009611
Residual     252.638728  391.0       NaN       NaN
```

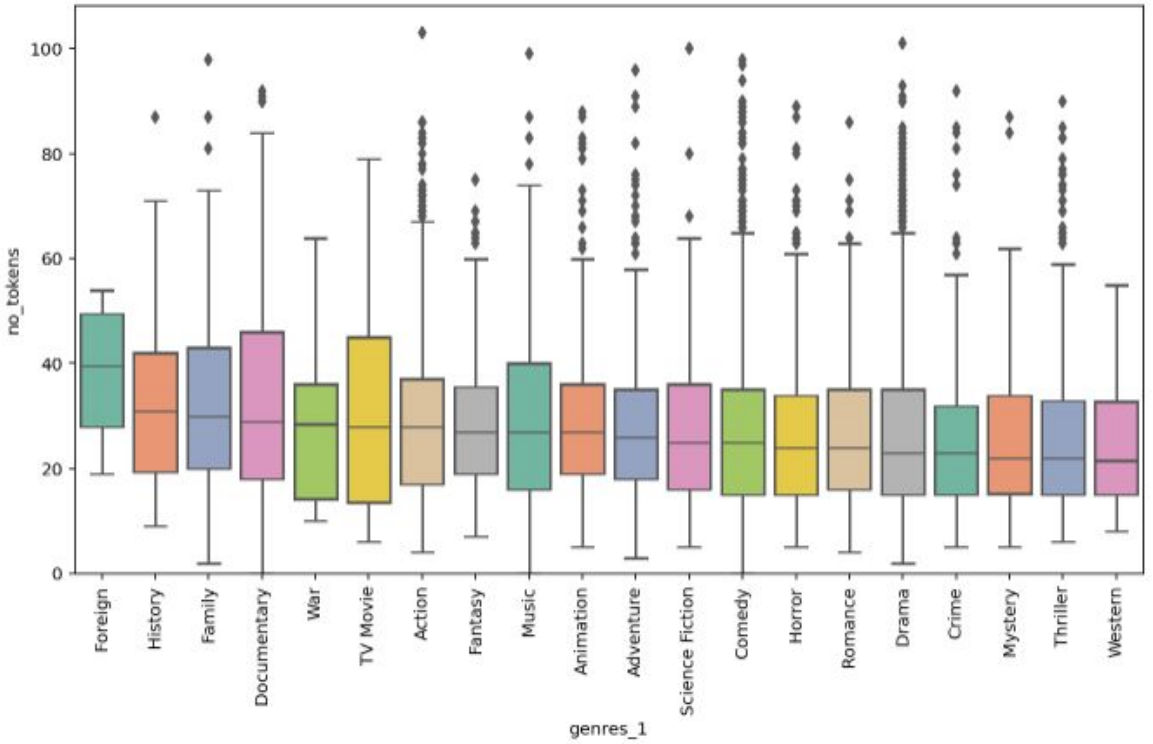
ANOVA For ‘Revenue’ and Genre Columns

```
ANOVA results for genres_1:
                   sum_sq      df           F        PR(>F)
C(genres_1)  6.152287e+18    19.0  24.944903  5.406090e-86
Residual     1.123487e+20  8655.0        NaN           NaN
ANOVA results for genres_2:
                   sum_sq      df           F        PR(>F)
C(genres_2)  4.403851e+18    19.0  14.500926  2.228284e-46
Residual     1.082911e+20  6775.0        NaN           NaN
ANOVA results for genres_3:
                   sum_sq      df          F        PR(>F)
C(genres_3)  3.341307e+18    19.0  7.760037  1.915735e-21
Residual     9.119206e+19  4024.0       NaN           NaN
ANOVA results for genres_4:
                   sum_sq      df          F        PR(>F)
C(genres_4)  1.984719e+18    19.0  4.068595  9.572616e-09
Residual     3.964133e+19  1544.0       NaN           NaN
ANOVA results for genres_5:
                   sum_sq     df          F    PR(>F)
C(genres_5)  8.039354e+17   18.0  2.284664  0.002159
Residual     7.643692e+18  391.0       NaN       NaN
```

Genre_1 tokenization based on Overview_lemma







Next steps

Solve Multi-label classification problem by using problem transformation methods to transform the multi-label problem into a set of binary classification problems, which can then be handled using single-class classifiers.

genres
Drama|Adventure|Science Fiction
Family|Animation|Adventure|Comedy
Comedy|Animation|Family
Action|Adventure|Crime
Science Fiction|Fantasy|Action|Adventure

Horror: 0.02%
Romance: 0.02%
Adventure: 99.96%
Documentary: 0.0%

Feature engineering for consideration:

1. Counting methods (Bag of Words, Bag-of-ngram and TF-IDF)
2. Word Embedding model (Word2Vec, GloVe and fasttext)
3. Language model (BERT)
4. Topic model (LDA/LSI)
5. Document Clustering with Similarity Features