INTRODUCTION TO MOVIE RECOMMENDATION SYSTEM

PROBLEM

- Everyone's preference of movie is different
- Movies are hit driven – a handful of movie can drive financial success

THE CHALLENGE

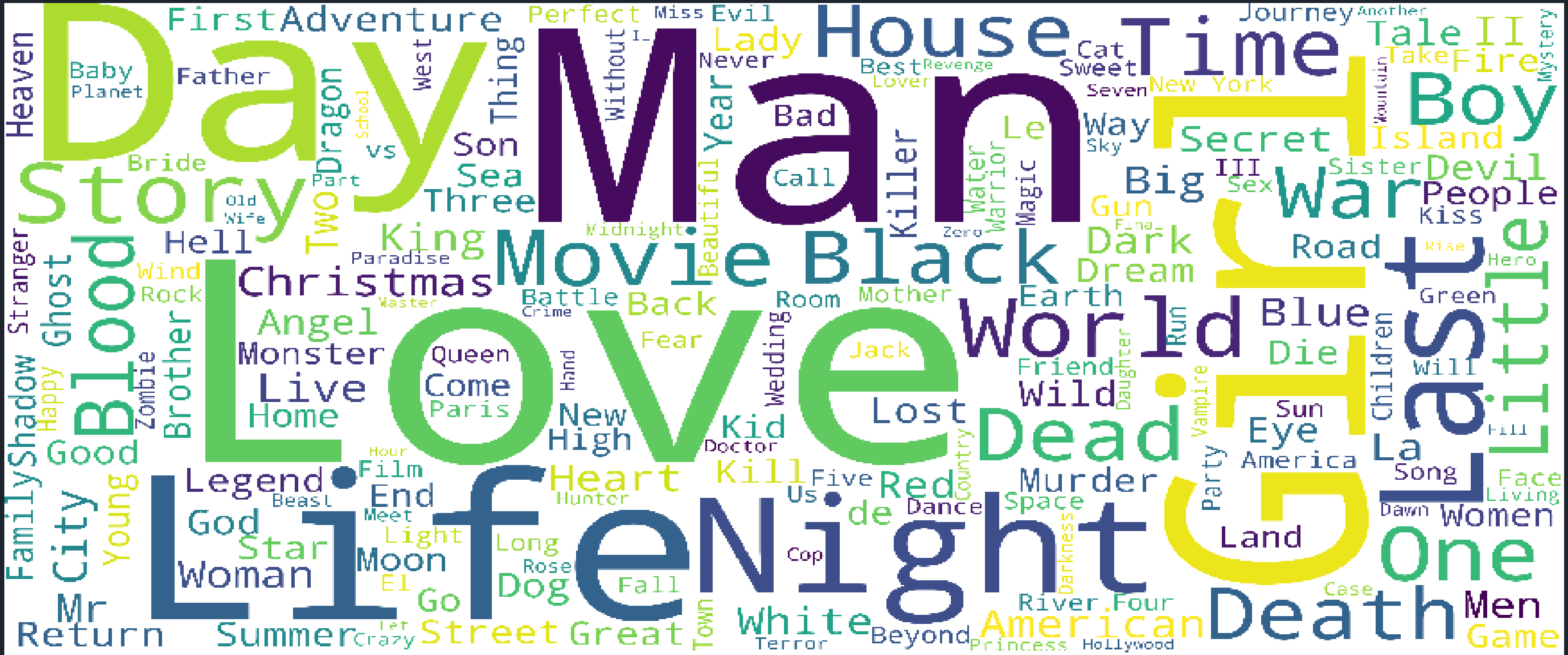
- Apply a recommendation system to market the right movie to the right audience
- Improve Effectiveness of advertisements
- Sometimes improve visibility for less popular movies

DATA

- I used the Movielens Data hosted on Kaggle.
- Containing 26 million ratings from 270,000 users for all 45,000 movies.
- Ratings are on a scale of 1-5 and have been obtained from the official GroupLens website.
- Source: https://www.kaggle.com/rounakbanik/the-movies-dataset

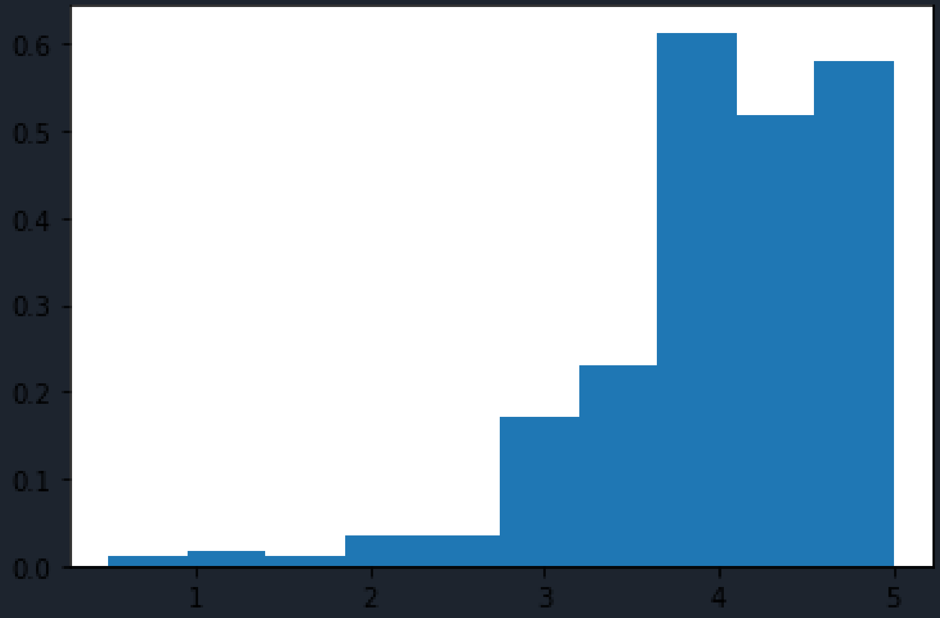
EXPLORATORY ANALYSIS

- Sample Word Cloud- Title



EXPLORATORY ANALYSIS

- Ratings



TYPES OF RECOMMENDATION SYSTEM

- Content Base

  Finding and ranking similarity score

- Collaborative

  Using Surprise package to predict a single user's rating

- Hybrid

CONTENT BASE FILTERING

- 1st attempt using Taglines and Overview and applying TF-IDF Vectorizor

In [40]:
```
j = get_recommendations('Toy Story').head(10).index

metadata_df_sub.loc[j,['title','tagline','overview']]
```

Out[40]:

	title	tagline	overview
3024	Toy Story 2	The toys are back!	Andy heads off to Cowboy Camp, leaving his toy...
10389	The 40 Year Old Virgin	The longer you wait, the harder it gets	Andy Stitzer has a pleasant life with a nice a...
8410	The Champ	NaN	Dink Purcell loves his alcoholic father, ex-he...
3084	Man on the Moon	Hello, my name is Andy and this is my movie.	A film about the life and career of the eccent...
11715	Factory Girl	When Andy met Edie, life imitated art.	In the mid-1960s, wealthy debutant Edie Sedgwi...
6504	What's Up, Tiger Lily?	WOODY ALLEN STRIKES BACK!	In comic Woody Allen's film debut, he took the...
1092	Rebel Without a Cause	The bad boy from a good family.	After moving to a new town, troublemaking teen...
11508	For Your Consideration	NaN	Three actors learn that their respective perfo...
5859	Class of 1984	Class of 1984. Is this the future?	Andy is a new teacher at a inner city high sch...
1952	Condorman	An action adventure romantic comedy spy story.	Comic artist and writer Woody performs a simpl...

CONTENT BASE FILTERING

- 2nd Attempt using Meta Data (Cast, Genre, Keywords, etc)

```
get_recommendations('The Godfather', cosine_sim=cosine_sim2)
```

```
3]: 1199                 The Godfather: Part II
    1934                The Godfather: Part III
    10261                            The Outfit
    11733            The Consequences of Love
    5309                            The Gambler
    3327                  ...And Justice for All
    4602                        The Cotton Club
    9517                         The Black Lapp
    1614                         The Rainmaker
    12221                          10th & Wolf
    1648                       Ill Gotten Gains
    3487          Jails, Hospitals & Hip-Hop
    6744                                  Ruby
    7772                              Mitchell
    8001       The Night of the Following Day
    5793                      True Confessions
    1430                         Donnie Brasco
    549                          Trial by Jury
    2112                      The Paradine Case
    10729                   House of Strangers
    13361                   Manhattan Melodrama
    276                     Murder in the First
    9874                       Amongst Friends
    1186                        Apocalypse Now
    3640                               Serpico
    4080                                Pixote
    7052                 Shoot the Piano Player
    8698                       Branded to Kill
    11287                                G-Men
    13597                             Il Divo
```

COLLABORATIVE FILTERING

- Using Surprise Package
- Using other users rating to predict the rating of target user

HYBRID RECOMMENDER

- Content-base filtering first then apply collaborative filtering

```
hybrid(1,"The Godfather",cosine_sim2)
```

2]:

	title	vote_count	vote_average	id	est
1186	Apocalypse Now	2112.0	8.0	28	4.613409
7052	Shoot the Piano Player	69.0	7.2	1818	4.352920
3640	Serpico	429.0	7.5	9040	4.330476
1199	The Godfather: Part II	3418.0	8.3	240	4.316102
4012	Gardens of Stone	25.0	5.5	28368	4.285536
1430	Donnie Brasco	1175.0	7.4	9366	4.125541
11733	The Consequences of Love	125.0	7.6	24653	4.119607
1614	The Rainmaker	239.0	6.7	11975	4.048667
3327	...And Justice for All	118.0	7.1	17443	4.004266
4080	Pixote	24.0	8.4	42148	3.994565

MODEL COMPARISON

- Precision at K

A method to see at threshold rating K, arbitrarily set, that how many items the system rated at least K or higher for a user

SVD

```
[86]:  for trainset, testset in kf.split(data):
           algo.fit(trainset)
           predictions = algo.test(testset)
           precisions, recalls = precision_recall_at_k(predictions, k=15, threshold=3.5)

           # Precision and recall can then be averaged over all users
           print(sum(prec for prec in precisions.values()) / len(precisions))
           print(sum(rec for rec in recalls.values()) / len(recalls))

0.6773641924688965
0.5724082177629027
0.6781626724448315
0.5809743942524471
0.6856334618028078
0.5842087998220512
0.6876841223307756
0.5858252813718233
0.6831897090921978
0.5819061289397528
```

MODEL COMPARISON

- Precision At K for Baseline

```
In [88]:  for trainset, testset in kf.split(data):
              algo_norm.fit(trainset)
              predictions = algo_norm.test(testset)
              precisions, recalls = precision_recall_at_k(predictions, k=15, threshold=3.5)

              # Precision and recall can then be averaged over all users
              print(sum(prec for prec in precisions.values()) / len(precisions))
              print(sum(rec for rec in recalls.values()) / len(recalls))

          0.5821217419887966
          0.4348215174573881
          0.5853237597005768
          0.4336387928090972
          0.5765403600913539
          0.4320687982419843
          0.583347459044794
          0.42839280690906817
          0.5825129631943087
          0.43548330660210877
```

FUTURE CONSIDERATION

- Pre-grouping users with similar rating patterns for Collaborative filtering
- Better computer hardware to use more of the data
- Tree based model?