IR PROJECT
MOVIE SEARCH ENGINE

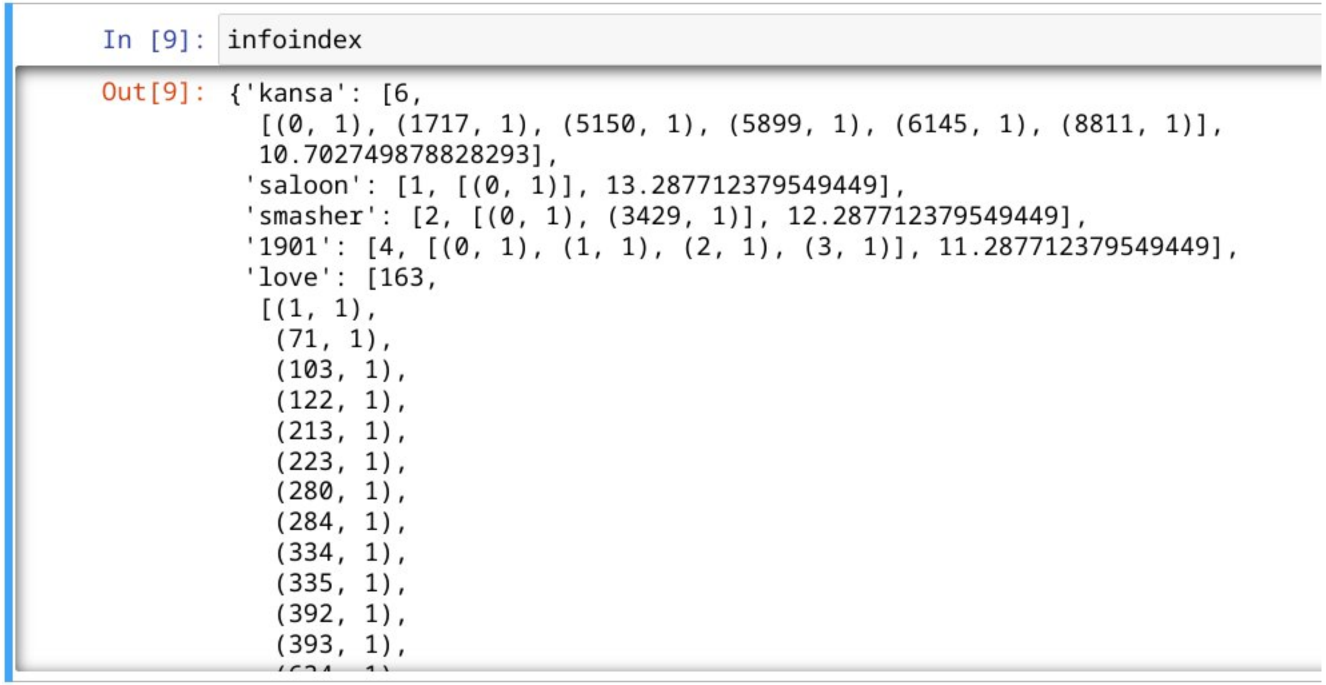
MS DHEERAJ

About the project:

- Users can search the movie based on the plot,Director name,cast,year of release
- Components we included:
- Indexing
- Searching
- Refining searches based on filters
- Assessment components
- Capturing feedback

INDEXING:

- Created two inverted indexes for plot and genre Director cast year of release combined
- Dataset contains 10k documents
- We dumped the inverted indexex in the pickel file
- We appended IDF Values to posting list for offline mode
- We tokenized the documents and stemmed them using nltk packages



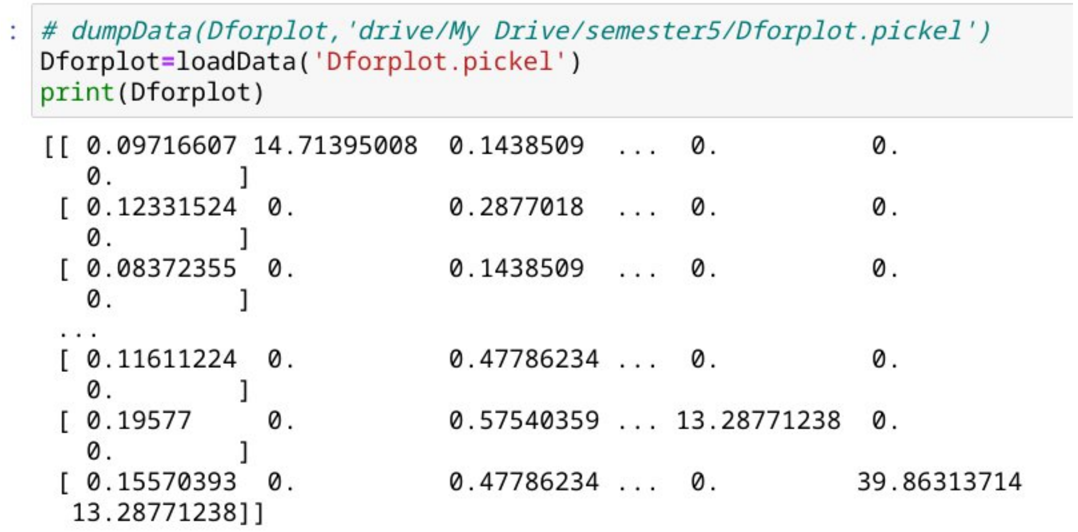
```
In [9]: infoindex

Out[9]: {'kansa': [6,
          [(0, 1), (1717, 1), (5150, 1), (5899, 1), (6145, 1), (8811, 1)],
          10.702749878828293],
         'saloon': [1, [(0, 1)], 13.287712379549449],
         'smasher': [2, [(0, 1), (3429, 1)], 12.287712379549449],
         '1901': [4, [(0, 1), (1, 1), (2, 1), (3, 1)], 11.287712379549449],
         'love': [163,
          [(1, 1),
           (71, 1),
           (103, 1),
           (122, 1),
           (213, 1),
           (223, 1),
           (280, 1),
           (284, 1),
           (334, 1),
           (335, 1),
           (392, 1),
           (393, 1),
```

Searching:

- Calculated TF-idf ranking for all documents in the corpus
- Offline stored the tf-idf high dimentional vectors inorder to save time as corpus won’t change
- The cosine similarity is found between the query and each document is stored and used for ranking the documents.

tfidf vectors for all documents

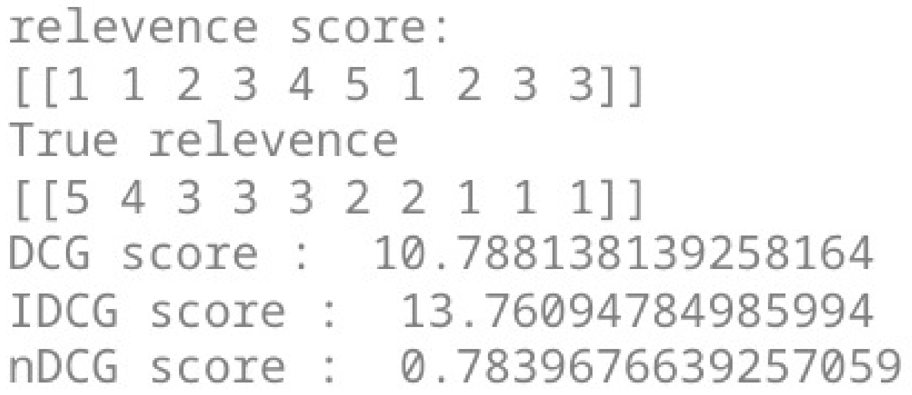


```
# dumpData(Dforplot,'drive/My Drive/semester5/Dforplot.pickel')
Dforplot=loadData('Dforplot.pickel')
print(Dforplot)
```

```
[[ 0.09716607 14.71395008  0.1438509  ...  0.          0.
   0.        ]
 [ 0.12331524  0.           0.2877018  ...  0.          0.
   0.        ]
 [ 0.08372355  0.           0.1438509  ...  0.          0.
   0.        ]
 ...
 [ 0.11611224  0.           0.47786234 ...  0.          0.
   0.        ]
 [ 0.19577     0.           0.57540359 ... 13.28771238  0.
   0.        ]
 [ 0.15570393  0.           0.47786234 ...  0.         39.86313714
  13.28771238]]
```

ASSESMENT COMPONENTS:Evaluation

- Relevance feedback is collected from the user
- Based on the feedback we used the normalized discounted cumulative gain metric for IR evaluation



```
relevence score:
[[1 1 2 3 4 5 1 2 3 3]]
True relevence
[[5 4 3 3 3 2 2 1 1 1]]
DCG score :  10.788138139258164
IDCG score :  13.76094784985994
nDCG score :  0.7839676639257059
```

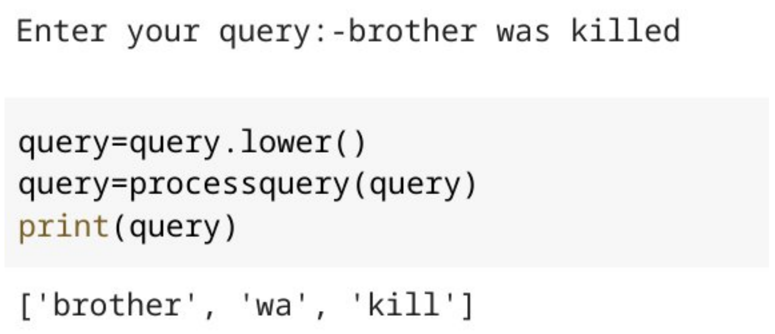
Capturing Feedback

- We captured the feedback from user
- Applied rocchio-algorithm for query expansion based on relevance feedback from the users
- Calculated the centroid of known relevance set and centroid of known non-relevence set
- Took the parameters as follows
- Alpha=2
- Beta=0.9, gamma=0.5

Refining searches based on specific filters

- As second inverted index was created specifically to search based on cast or based on year of release or Director of the movie ,we can filter and search the corpus by Director name or year of release
- we are doing filtering by Genre where we are retrieving all
- the documents which are released in the same year

QUERY: ‘brother was killed’

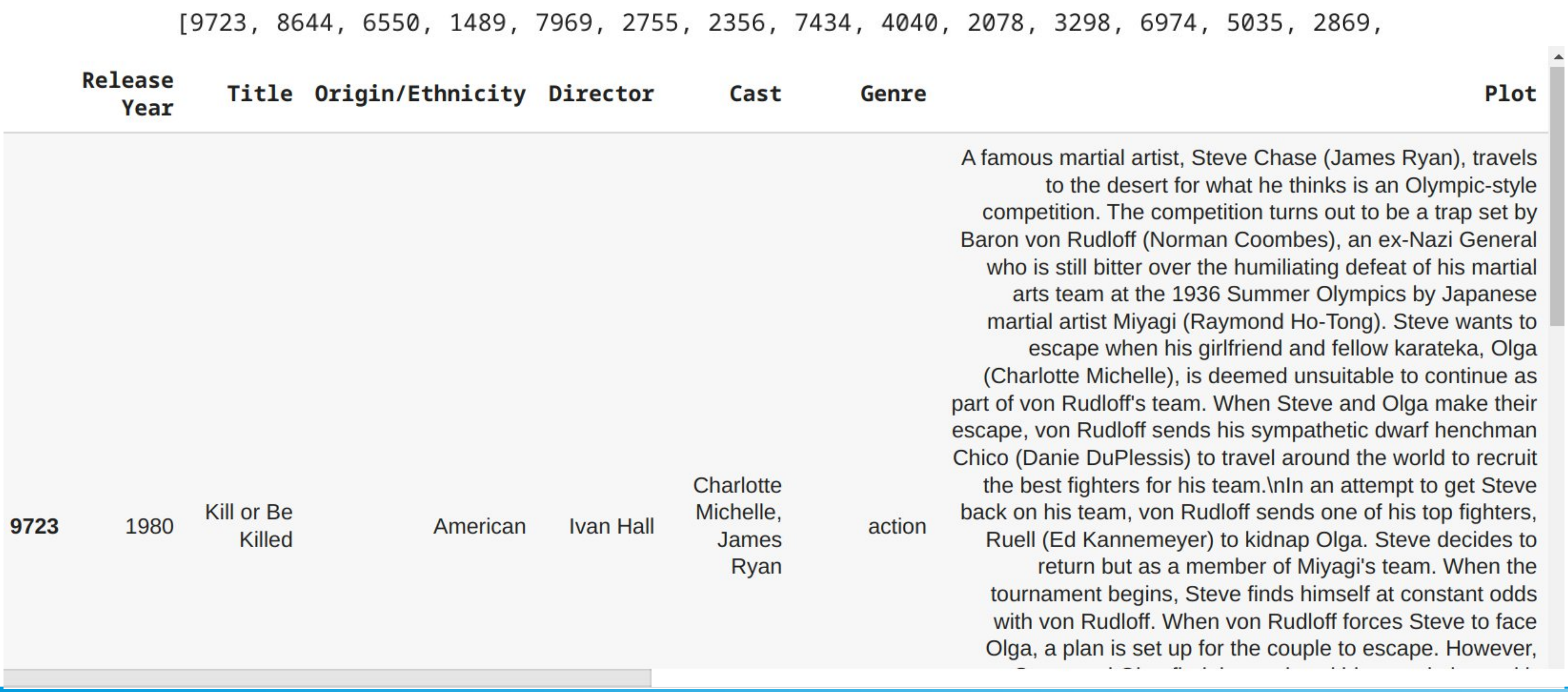


```
Enter your query:-brother was killed
```

```
query=query.lower()
query=processquery(query)
print(query)
```

```
['brother', 'wa', 'kill']
```

retrieved document ids are:



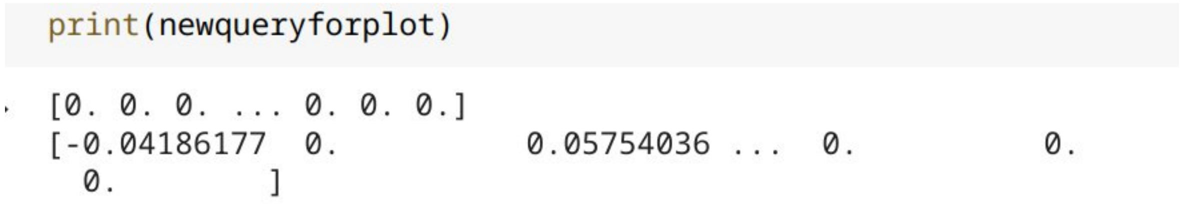
```
[9723, 8644, 6550, 1489, 7969, 2755, 2356, 7434, 4040, 2078, 3298, 6974, 5035, 2869,
```

	Release Year	Title	Origin/Ethnicity	Director	Cast	Genre	Plot
9723	1980	Kill or Be Killed	American	Ivan Hall	Charlotte Michelle, James Ryan	action	A famous martial artist, Steve Chase (James Ryan), travels to the desert for what he thinks is an Olympic-style competition. The competition turns out to be a trap set by Baron von Rudloff (Norman Coombes), an ex-Nazi General who is still bitter over the humiliating defeat of his martial arts team at the 1936 Summer Olympics by Japanese martial artist Miyagi (Raymond Ho-Tong). Steve wants to escape when his girlfriend and fellow karateka, Olga (Charlotte Michelle), is deemed unsuitable to continue as part of von Rudloff's team. When Steve and Olga make their escape, von Rudloff sends his sympathetic dwarf henchman Chico (Danie DuPlessis) to travel around the world to recruit the best fighters for his team.\nIn an attempt to get Steve back on his team, von Rudloff sends one of his top fighters, Ruell (Ed Kannemeyer) to kidnap Olga. Steve decides to return but as a member of Miyagi's team. When the tournament begins, Steve finds himself at constant odds with von Rudloff. When von Rudloff forces Steve to face Olga, a plan is set up for the couple to escape. However,

feedback from user is

[1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

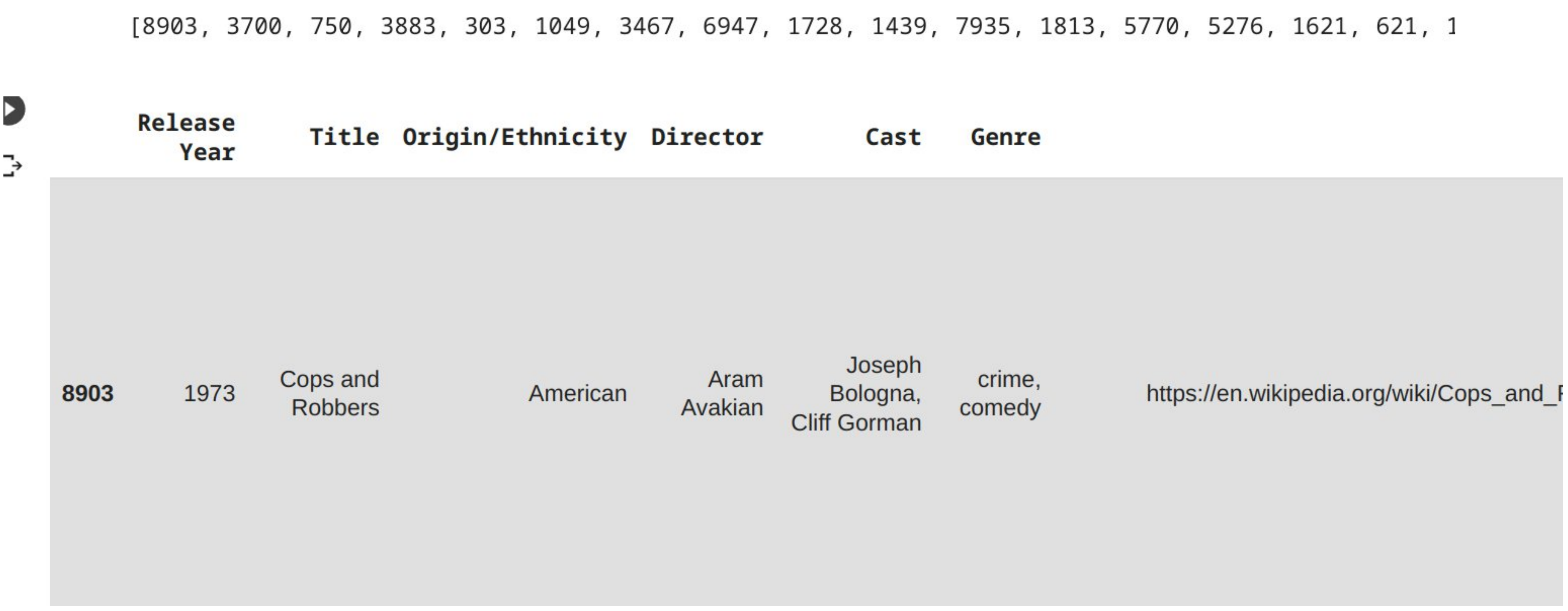
by rochhio algorithm New query vector is



```
print(newqueryforplot)
```

```
[0. 0. 0. ... 0. 0. 0.]
[-0.04186177  0.          0.05754036 ...  0.          0.
  0.        ]
```

new ranking after relevence feedback



```
[8903, 3700, 750, 3883, 303, 1049, 3467, 6947, 1728, 1439, 7935, 1813, 5770, 5276, 1621, 621, 1
```

	Release Year	Title	Origin/Ethnicity	Director	Cast	Genre	
8903	1973	Cops and Robbers	American	Aram Avakian	Joseph Bologna, Cliff Gorman	crime, comedy	https://en.wikipedia.org/wiki/Cops_and_