LSH Correctness:

We first test on only fake data and see the models performance:

```
PS E:\University\Term 6\Modern Information Retreival\Project\MIR-imdb> python -u "e:\University\Term 6\Modern Inf
\MIR-imdb\Logic\core\LSH.py"
found 9 similar pairs
detected pairs: {(0, 1), (10, 11), (12, 13), (18, 19), (2, 3), (6, 7), (8, 9), (14, 15), (16, 17)}
your final score in near duplicate detection: 1.0
```

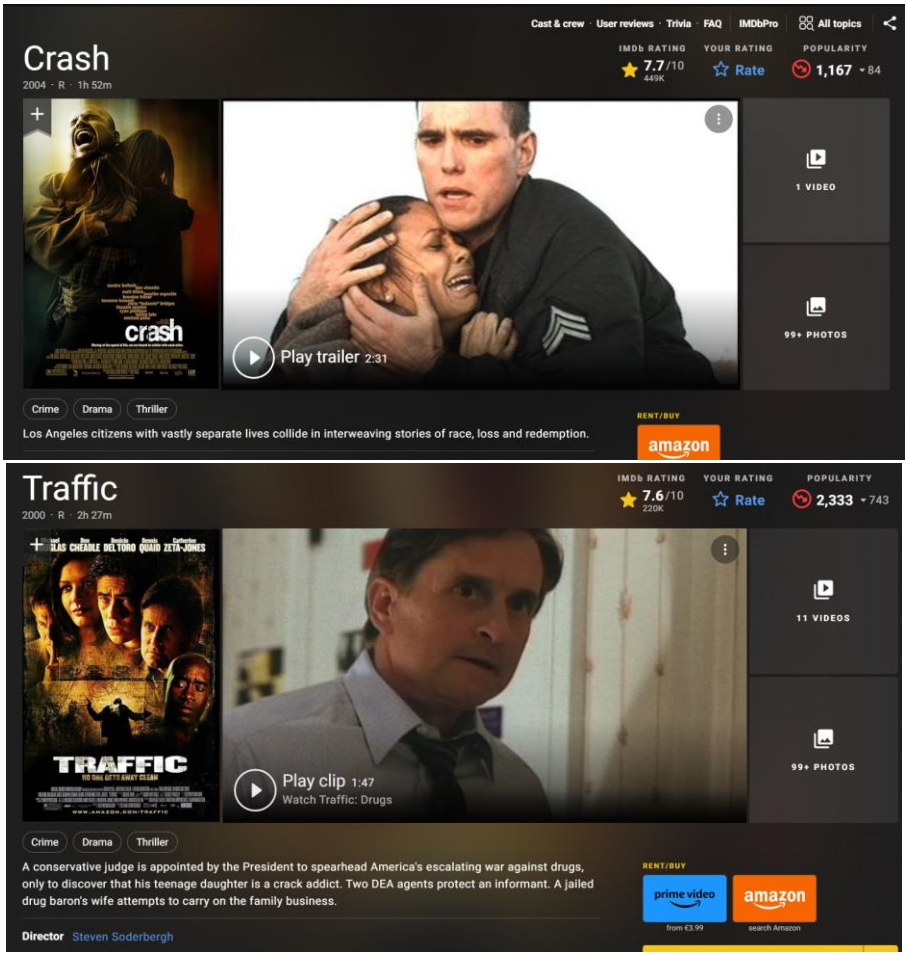
As you can see model detected 9 out of 10 similar movies correctly. Also the score is 1

Now let's test on all the movies in corpus:

```
found 16 similar pairs
detected pairs: {(2096, 2095), (820, 1653), (2080, 2079), (840, 1044), (1017, 1325), (1864, 2041), (1864, 1286), (2091, 2092), (2085, 2086)
, (2077, 2078), (1716, 510), (2093, 2094), (2041, 1286), (2083, 2084), (474, 1622), (2089, 2090)}
your final score in near duplicate detection: 0.9857142857142858
```

As you see the score is still very high. And some extra pairs have been detected from our data as well.

Example of a pair of movies detected as similar:



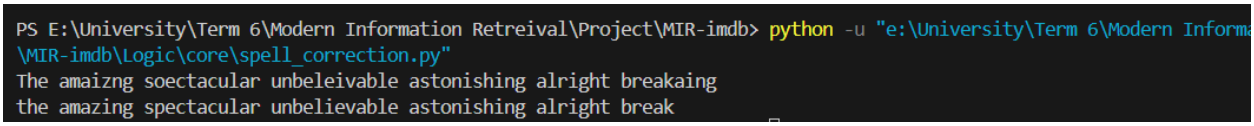
Index correctness:

We perform several test to check the correctness of index and compare it’s performance against the brute force method.

```
PS E:\University\Term 6\Modern Information Retreival\Project\MIR-imdb> python -u "e:\University\Term 6\Modern Information Retreival\Project
\MIR-imdb\Logic\core\indexer\index.py"
Add is correct
Remove is correct
documents load correctness: True
stars load correctness: True
genres load correctness: True
summaries load correctness: True
### stars index evaulation:
Brute force time:  0.002978324890136718
Implemented time:  0.0
Indexing is correct
Indexing is good
True
Brute force time:  0.0020003318786621094
Implemented time:  0.0
Indexing is correct
Indexing is good
True
### genres index evaulation:
Brute force time:  0.0
Implemented time:  0.0
Indexing is correct
Indexing is good
True
Brute force time:  0.0
Implemented time:  0.0
Indexing is correct
Indexing is good
True
### summaries index evaulation:
Brute force time:  0.0009989738464355469
Implemented time:  0.0
Indexing is correct
Indexing is good
True
Brute force time:  0.00099945068359375
Implemented time:  0.0
Indexing is correct
Indexing is good
True
```

Spell correction:

Here’s an example of spell corrections functionality:



Evaluation

We run evaluation for 3 queries: (dune, spider-man spider-verse, matrix)

The metrics are as described below:

```
\MIR-imdb\Logic\core\utility\evaluation.py"Retreival\Project\MIR-imdb>
name = test
precision = 0.4666666666666666
recall = 0.4666666666666666
f1 = 0.4666666666666666
map = 0.6433333333333332
ndcg = 263.6333072229436
mrr = 0.8333333333333334
wandb: Currently logged in as: s-emad-emamjomeh (emadej). Use `wandb login --relogin` to force relogin
wandb: Tracking run with wandb version 0.16.6
wandb: Run data is saved locally in E:\University\Term 6\Modern Information Retreival\Project\MIR-imdb\wandb\run-20240410_185605-99c0n4x
wandb: Run `wandb offline` to turn off syncing.
wandb: Syncing run neat-plasma-5
wandb:  View project at https://wandb.ai/emadej/mir-project
wandb:  View run at https://wandb.ai/emadej/mir-project/runs/99c0n4x1
wandb: \ 0.007 MB of 0.007 MB uploaded
wandb: Run history:
wandb:                              F1 Score _
wandb:                 Mean Average Precision _
wandb:                   Mean Reciprocal Rank _
wandb: Normalized Discounted Cumulative Gain _
wandb:                             Precision _
wandb:                               Recall _
wandb:
wandb: Run summary:
wandb:                              F1 Score 0.46667
wandb:                 Mean Average Precision 0.64333
wandb:                   Mean Reciprocal Rank 0.83333
wandb: Normalized Discounted Cumulative Gain 263.63331
wandb:                             Precision 0.46667
wandb:                               Recall 0.46667
wandb:
wandb:  View run neat-plasma-5 at: https://wandb.ai/emadej/mir-project/runs/99c0n4x1
wandb:  View project at: https://wandb.ai/emadej/mir-project
wandb: Synced 4 W&B file(s), 0 media file(s), 0 artifact file(s) and 0 other file(s)
wandb: Find logs at: .\wandb\run-20240410_185605-99c0n4x1\logs
PS E:\University\Term 6\Modern Information Retreival\Project\MIR-imdb>
```

Wandb Page: