IR – Model and Evaluation

NLP and Information Retrieval Class

20191181 Seunguk Yu

1. Entire Process
2. Boolean Model & Evaluation
3. Vector Model & Evaluation
4. Result & Conclusion

1. Entire Process

```
movie_title_df = pd.read_csv('./imsdb_5categories_title.csv')
movie_title_df.head(4)
```

	drama	romance	comedy	thriller	crime
0	12 and Holding	10 Things I Hate About You	10 Things I Hate About You	12 Monkeys	15 Minutes
1	12 Monkeys	17 Again	12	127 Hours	44 Inch Chest
2	12 Years a Slave	500 Days of Summer	17 Again	15 Minutes	A Few Good Men
3	127 Hours	Adjustment Bureau, The	30 Minutes or Less	2012	A Most Violent Year

```
movie_script_df = pd.read_csv('./imsdb_5categories_token.csv')
movie_script_df.head(4)
```

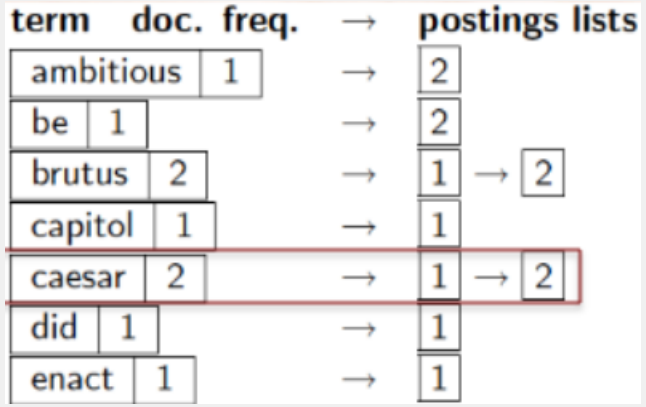
	drama	romance	comedy	thriller	crime
0	['12', 'holding', 'written', 'anthony', 's', '...	['ten', 'things', 'hate', 'written', 'karen', ...	['ten', 'things', 'hate', 'written', 'karen', ...	['twelve', 'monkeys', 'original', 'screenplay'...	['fade', 'words', 'czech', 'airline', 'panning...
1	['twelve', 'monkeys', 'original', 'screenplay'...	['17', 'again', 'written', 'jason', 'filardi',...	['cut', 'blacktitle', 'finexterior', 'la', 'da...	['127', 'hours', 'written', 'simon', 'beaufoy'...	['44', 'inch', 'chest', 'written', 'louis', 'm...
2	['12', 'years', 'slave', 'written', 'john', 'r...	['500', 'days', 'summer', 'written', 'scott', ...	['17', 'again', 'written', 'jason', 'filardi',...	['fade', 'words', 'czech', 'airline', 'panning...	['few', 'good', 'men', 'written', 'aaron', 'so...
3	['127', 'hours', 'written', 'simon', 'beaufoy'...	['adjustment', 'bureau', 'written', 'george', ...	['30', 'minutes', 'less', 'written', 'michael'...	['2012', 'written', 'roland', 'emmerich', 'har...	['violent', 'year', 'written', 'jc', 'chandor'...

1. Entire Process

Movie Script Tokens
:5 Categories
(Drama, Romance, Comedy, Thriller, Crime)

Indexing Terms

Postings Lists



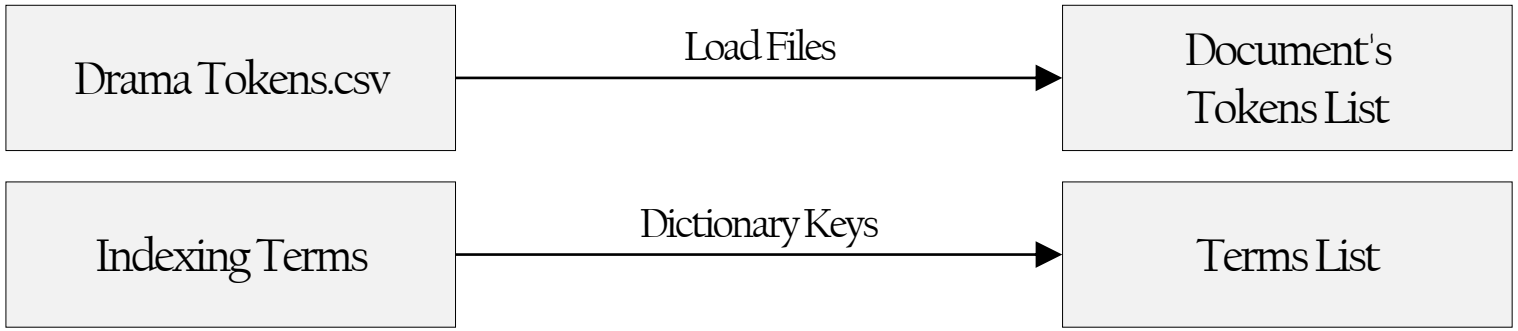
Boolean Model

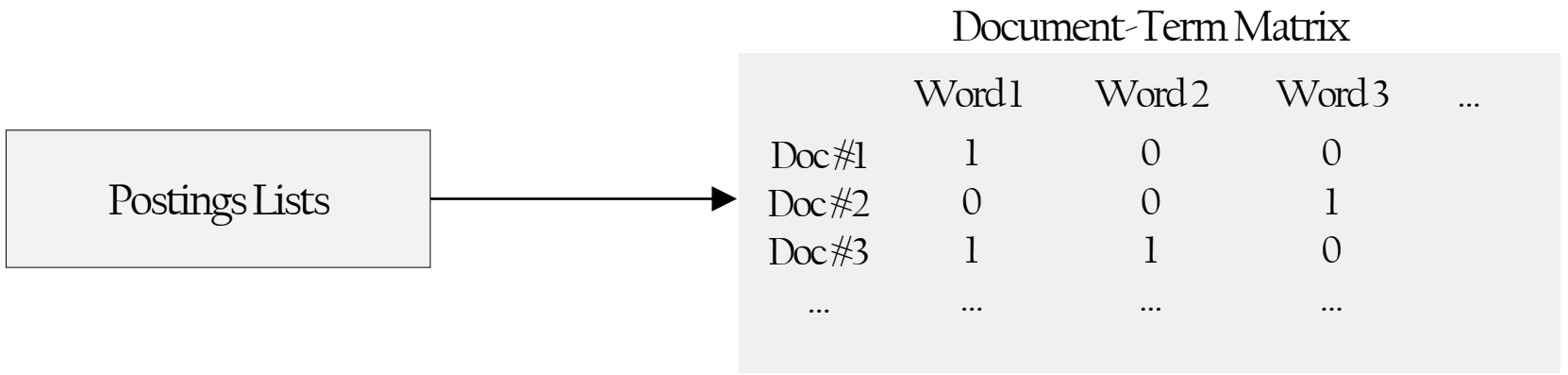
Vector Space Model

Precision / Recall
Evaluation

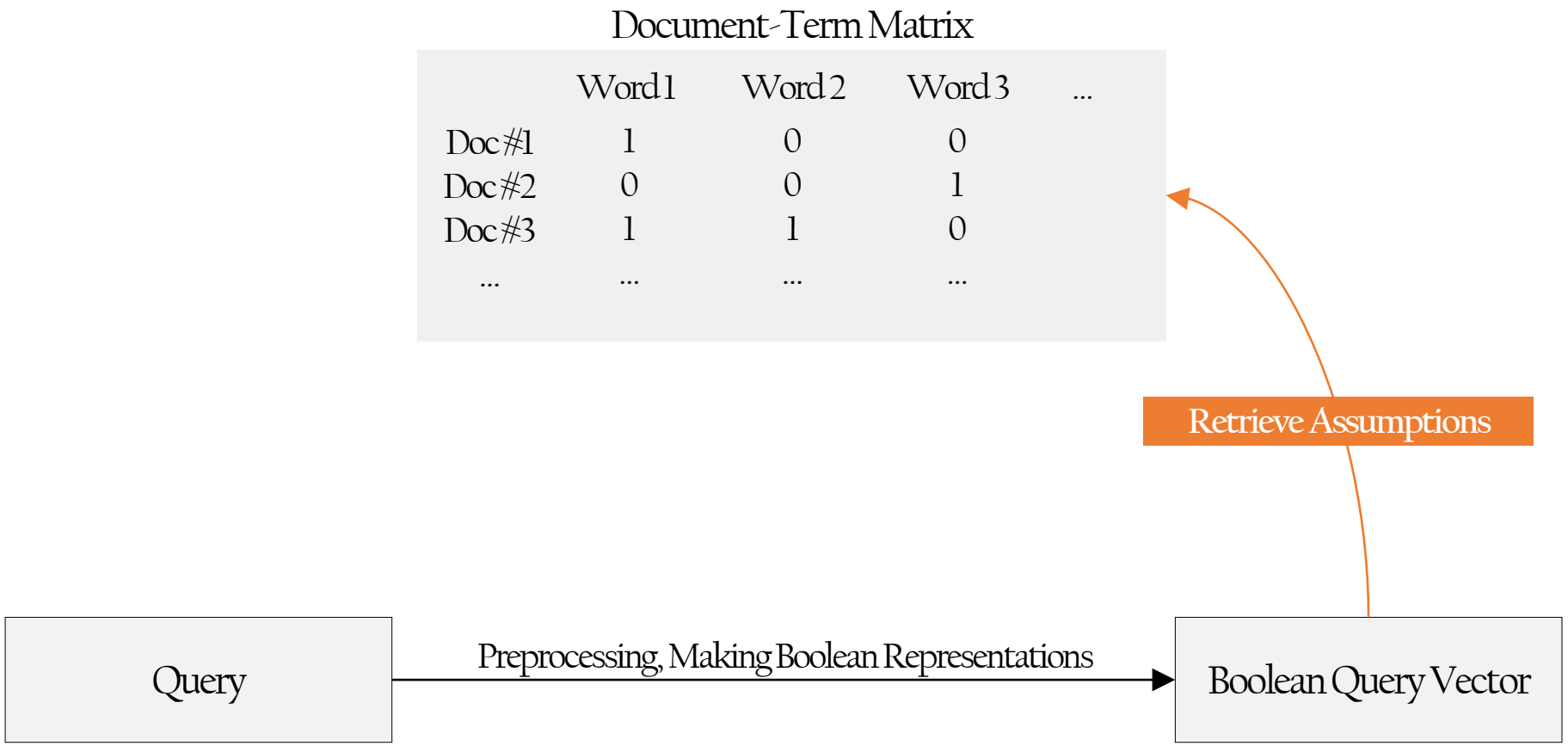
MAP / MAR
Evaluation, NDCG

2. Boolean Model & Evaluation





2. Boolean Model & Evaluation



2. Boolean Model & Evaluation

Retrieve Assumptions

1) When selecting common elements between query vector and all document vectors, the largest number of common elements was called N.

2) In the above boolean model,
only those with more than 0.95*N common elements
between the query vector and document vector were RETRIEVED.

3) If there are 0.9*N common elements between query vector and document vector, they are assumed to be RELEVANT.

2. Boolean Model & Evaluation

```
# 'Drama' category
query = "Please holding me tight.₩
         Probably he's saying forever.₩
         Discover my classroom and students.₩
         They're continued in morning."

matched_drama_movie_list = match_doc_and_query_byboolean('drama', drama_dtm, drama_rows, query, drama_terms)
```

```
The related movie to your query in genre 'drama' is '12 and Holding'
The related movie to your query in genre 'drama' is '187'
The related movie to your query in genre 'drama' is 'Apt Pupil'
The related movie to your query in genre 'drama' is 'Theory of Everything, The'
The related movie to your query in genre 'drama' is 'Twin Peaks'
```

```
# 'Romance' category
query = "I love ten things that seen in November against hot summer.₩
         They looks so beautiful, really joyful.₩
         Imagine what's coming.₩
         There's no time to block anything."

matched_romance_movie_list = match_doc_and_query_byboolean('romance', romance_dtm, romance_rows, query, romance_terms)
```

```
The related movie to your query in genre 'romance' is '10 Things I Hate About You'
The related movie to your query in genre 'romance' is 'Autumn in New York'
The related movie to your query in genre 'romance' is 'Heavenly Creatures'
The related movie to your query in genre 'romance' is 'Jane Eyre'
The related movie to your query in genre 'romance' is 'Marty'
The related movie to your query in genre 'romance' is 'Out of Sight'
The related movie to your query in genre 'romance' is 'Spanglish'
```

2. Boolean Model & Evaluation

```
# 'Comedy' category
query = "Boys and girls are different.₩
         They never work a lot.₩
         Amazingly, ladies catch sunglasses and laugh.₩
         There's some question about eating."

matched_comedy_movie_list = match_doc_and_query_byboolean('comedy', comedy_dtm, comedy_rows, query, comedy_terms)
```

```
The related movie to your query in genre 'comedy' is '10 Things I Hate About You'
The related movie to your query in genre 'comedy' is 'Change-Up, The'
The related movie to your query in genre 'comedy' is 'Man on the Moon'
The related movie to your query in genre 'comedy' is 'Twins'
```

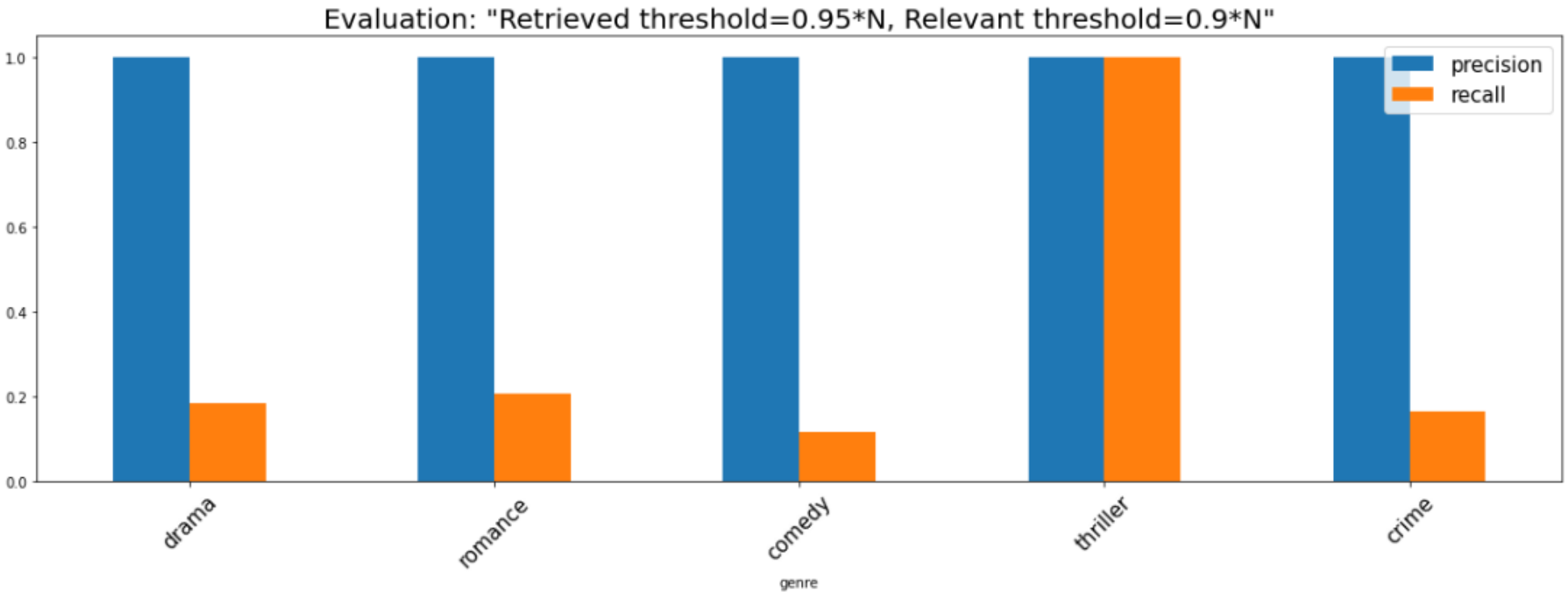
```
# 'Thriller' category
query = "Monkeys are inspired by strong people!₩
         Newspapers mean breath carefully.₩
         You'll play like lion.₩
         There are choices."

matched_thriller_movie_list = match_doc_and_query_byboolean('thriller', thriller_dtm, thriller_rows, query, thriller_terms)
```

```
The related movie to your query in genre 'thriller' is '12 Monkeys'
```

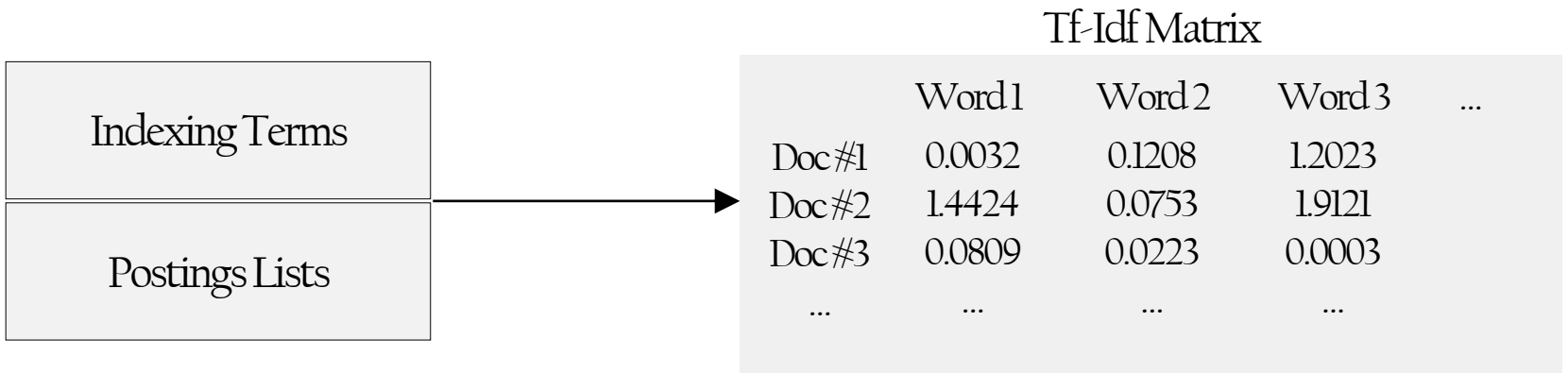
```
# 'Crime' category
query = "Money and bills are coming.₩
         They're not outdated...₩
         It's called America joke.₩
         Not a joke, made by dirty story"

matched_crime_movie_list = match_doc_and_query_byboolean('crime', crime_dtm, crime_rows, query, crime_terms)
```

```
The related movie to your query in genre 'crime' is '15 Minutes'
The related movie to your query in genre 'crime' is 'American Gangster'
```

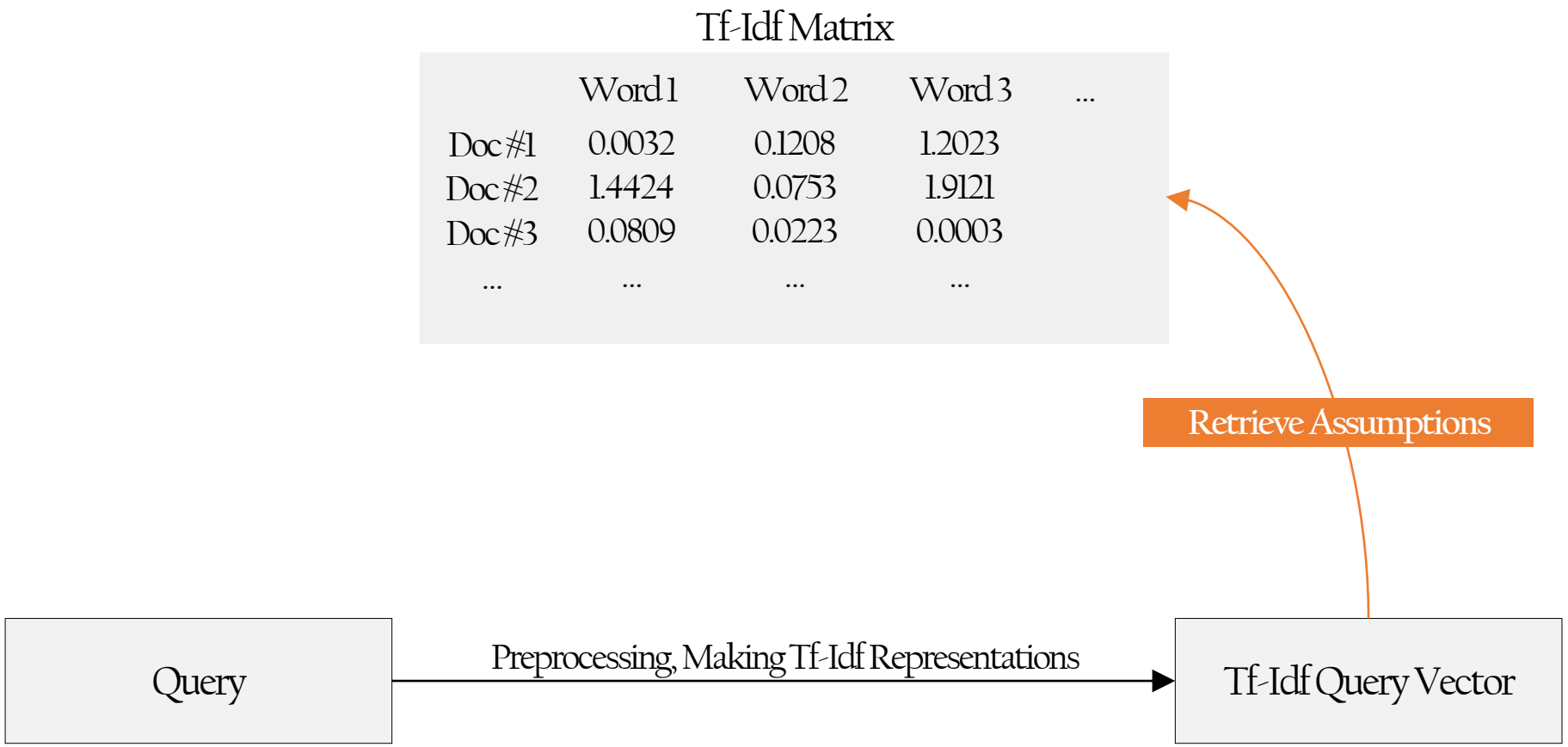
2. Boolean Model & Evaluation



3. Vector Model & Evaluation



3. Vector Model & Evaluation



3. Vector Model & Evaluation

Retrieve Assumptions

1) In the above vector space model,
only RETRIEVED the top 10 cos similarity movies
between query vector and document vector.

2) Among them, only 3-5 (for scoring MAP/MAR) or 5 (for scoring NDCG)
randomly selected movies are assumed to be RELEVANT.

3) (for scoring NDCG) Cos similarity is treated as a 'relevance' in each movie.

3. Vector Model & Evaluation

```
# 'Drama' category
query = "Please holding me tight.₩
         Probably he's saying forever.₩
         Discover my classroom and students.₩
         They're continued in morning."

matched_drama_movie_list = match_doc_and_query_byvector('drama', drama_dtm, query, drama_terms, drama_postings_lists)
```

```
TOP 10 movies to your query in genre 'drama' are...
 '12 and Holding                ' with score 0.04890273811399546
 'Apt Pupil                     ' with score 0.040419990012928744
 '187                           ' with score 0.0358933582165348
 'Almost Famous                 ' with score 0.02938131941458541
 'Ali                           ' with score 0.028607373702858586
 'American Splendor             ' with score 0.027571239565206948
 'American Gangster             ' with score 0.026189930550036617
 '50-50                         ' with score 0.026027418014808682
 'An Education                  ' with score 0.025129429693557772
 'After.Life                    ' with score 0.024825005481282768
```

```
# 'Romance' category
query = "I love ten things that seen in November against hot summer.₩
         They looks so beautiful, really joyful.₩
         Imagine what's coming.₩
         There's no time to block anything."

matched_romance_movie_list = match_doc_and_query_byvector('romance', romance_tim, query, romance_terms, romance_postings_lists)
```

```
TOP 10 movies to your query in genre 'romance' are...
 '500 Days of Summer            ' with score 0.07058284286157498
 'Artist, The                   ' with score 0.0034213392071774824
 'Brothers Bloom, The           ' with score 0.0033127129333557943
 '10 Things I Hate About You    ' with score 0.002410674987584413
 'Clueless                      ' with score 0.002354104552827887
 'Almost Famous                 ' with score 0.002254437882465098
 'Autumn in New York            ' with score 0.00196422785268089
 'American President, The       ' with score 0.001930453104120181
 'Bruce Almighty                ' with score 0.001365393788235247
 'Cooler, The                   ' with score 0.0011099663255404472
```

3. Vector Model & Evaluation

```
# 'Comedy' category
query = "Boys and girls are different.₩
         They never work a lot.₩
         Amazingly, ladies catch sunglasses and laugh.₩
         There's some question about eating."

matched_comedy_movie_list = match_doc_and_query_byvector('comedy', comedy_dtm, query, comedy_terms, comedy_postings_lists)
```

```
TOP 10 movies to your query in genre 'comedy' are...
 '10 Things I Hate About You    ' with score 0.03927447047913445
 'American Pie                  ' with score 0.03162680167058886
 'Drop Dead Gorgeous            ' with score 0.031191640447265058
 'Fatal Instinct                ' with score 0.03099965604755074
 'Devil Wears Prada, The        ' with score 0.02772157020454454
 'Entrapment                    ' with score 0.027134010113659946
 'Croods, The                   ' with score 0.026956611006231552
 'Back-up Plan, The             ' with score 0.02658907004905023
 '48 Hrs.                       ' with score 0.025446718861336155
 '500 Days of Summer            ' with score 0.02392951683692961
```

```
# 'Thriller' category
query = "Monkeys are inspired by strong people!₩
         Newspapers mean breath carefully.₩
         You'll play like lion.₩
         There are choices."

matched_thriller_movie_list = match_doc_and_query_byvector('thriller', thriller_dtm, query, thriller_terms, thriller_postings_lists)
```

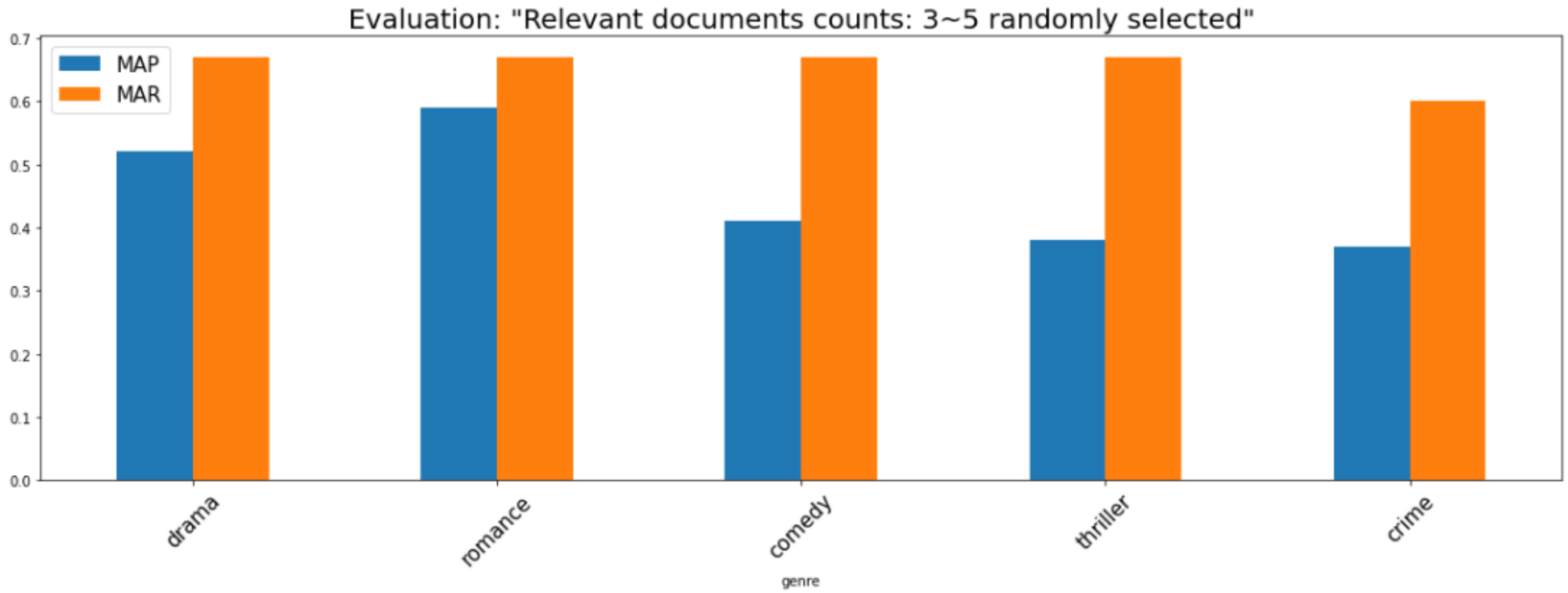
```
TOP 10 movies to your query in genre 'thriller' are...
 '12 Monkeys                    ' with score 0.03565008961896475
 'Adjustment Bureau, The        ' with score 0.01882440177767354
 'Basic                         ' with score 0.018593336125182082
 'Avventura, L' (The Adventure) ' with score 0.015792793091269873
 '8MM                           ' with score 0.014280168676499028
 'Black Dahlia, The             ' with score 0.014269050898844698
 'Batman 2                      ' with score 0.013735337514713534
 'Birds, The                    ' with score 0.012269021027917234
 '15 Minutes                    ' with score 0.011713840925709804
 'Air Force One                 ' with score 0.01131386922858502
```

3. Vector Model & Evaluation

```
# 'Crime' category
query = "Money and bills are coming.₩
         They're not outdated...₩
         It's called America joke.₩
         Not a joke, made by dirty story"

matched_crime_movie_list = match_doc_and_query_byvector('crime', crime_dtm, query, crime_terms, crime_postings_lists)
```

```
TOP 10 movies to your query in genre 'crime' are...
 '15 Minutes                    ' with score 0.031020131219329932
 'American Gangster             ' with score 0.02601192275342815
 'Bound                         ' with score 0.012319345252069147
 'Catch Me If You Can           ' with score 0.01226406675056572
 'Analyze That                  ' with score 0.011948710212007906
 'BlacKkKlansman                ' with score 0.011853095382854537
 'Batman                        ' with score 0.011555054998641213
 'Capote                        ' with score 0.011501321308417021
 'Black Rain                    ' with score 0.011288029388025853
 'Chinatown                     ' with score 0.011052800798142377
```

3. Vector Model & Evaluation



3. Vector Model & Evaluation

	doc#	relevance	DCG	IDCG	NDCG
0	2	0.00000	0.00000	0.00342	0.00000
1	12	0.00342	0.00342	0.00583	0.58662
2	35	0.00000	0.00342	0.00731	0.46785
3	0	0.00241	0.00462	0.00800	0.57750
4	45	0.00235	0.00563	0.00848	0.66392
5	6	0.00000	0.00563	0.00848	0.66392
6	14	0.00000	0.00563	0.00848	0.66392
7	7	0.00000	0.00563	0.00848	0.66392
8	36	0.00137	0.00606	0.00848	0.71462
9	48	0.00111	0.00639	0.00848	0.75354

'Drama' Category

	doc#	relevance	DCG	IDCG	NDCG
0	0	0.03927	0.03927	0.03927	1.00000
1	21	0.00000	0.03927	0.07046	0.55734
2	116	0.03119	0.05895	0.09002	0.65485
3	134	0.03100	0.07445	0.10332	0.72058
4	110	0.00000	0.07445	0.11428	0.65147
5	124	0.00000	0.07445	0.11428	0.65147
6	98	0.00000	0.07445	0.11428	0.65147
7	42	0.02659	0.08331	0.11428	0.72900
8	4	0.02545	0.09134	0.11428	0.79926
9	6	0.00000	0.09134	0.11428	0.79926

'Romance' Category

	doc#	relevance	DCG	IDCG	NDCG
0	0	0.03927	0.03927	0.03927	1.00000
1	21	0.00000	0.03927	0.07046	0.55734
2	116	0.03119	0.05895	0.09002	0.65485
3	134	0.03100	0.07445	0.10332	0.72058
4	110	0.00000	0.07445	0.11428	0.65147
5	124	0.00000	0.07445	0.11428	0.65147
6	98	0.00000	0.07445	0.11428	0.65147
7	42	0.02659	0.08331	0.11428	0.72900
8	4	0.02545	0.09134	0.11428	0.79926
9	6	0.00000	0.09134	0.11428	0.79926

'Comedy' Category

	doc#	relevance	DCG	IDCG	NDCG
0	0	0.00000	0.00000	0.01859	0.00000
1	9	0.00000	0.00000	0.03287	0.00000
2	39	0.01859	0.01173	0.04187	0.28015
3	32	0.00000	0.01173	0.04772	0.24581
4	5	0.01428	0.01788	0.05259	0.33999
5	46	0.01427	0.02340	0.05259	0.44495
6	42	0.00000	0.02340	0.05259	0.44495
7	45	0.00000	0.02340	0.05259	0.44495
8	2	0.01171	0.02709	0.05259	0.51512
9	11	0.01131	0.03049	0.05259	0.57977

'Thriller' Category

	doc#	relevance	DCG	IDCG	NDCG
0	0	0.03102	0.03102	0.03102	1.00000
1	8	0.02601	0.05703	0.05703	1.00000
2	39	0.01232	0.06480	0.06480	1.00000
3	45	0.00000	0.06480	0.07058	0.91811
4	12	0.00000	0.06480	0.07544	0.85896
5	28	0.00000	0.06480	0.07544	0.85896
6	21	0.01156	0.06892	0.07544	0.91357
7	43	0.00000	0.06892	0.07544	0.91357
8	27	0.01129	0.07248	0.07544	0.96076
9	50	0.00000	0.07248	0.07544	0.96076

'Crime' Category

4. Result & Conclusion

1) Through the two models (Boolean model, Vector space model),
it was able to find a document that matched to the query vector.

These two models can be basic information retrieval models.

2) Retrieved documents reflect the characteristics of the query
depending on the set of words they used.

Example: 'Romance' category in Boolean model

```
query = "I love ten things that seen in November against hot summer.₩
         They looks so beautiful, really joyful.₩
         Imagine what's coming.₩
         There's no time to block anything."
```

```
The related movie to your query in genre 'romance' is '10 Things I Hate About You'
The related movie to your query in genre 'romance' is 'Autumn in New York'
The related movie to your query in genre 'romance' is 'Heavenly Creatures'
The related movie to your query in genre 'romance' is 'Jane Eyre'
The related movie to your query in genre 'romance' is 'Marty'
The related movie to your query in genre 'romance' is 'Out of Sight'
The related movie to your query in genre 'romance' is 'Spanglish'
```

3) Evaluation didn't reflect the characteristics of the actual document
because 'relevance' criteria were set manually or randomly.

**For better evaluation, it is necessary to diversify the levels
for determining the 'relevance' criteria.**