

## Retrieval Evaluation Report of Multiple Experiments

### Loading Queries, Qrels, Runs for Analysis

See Details ▼

#### Overview

This section allows you to load and select the necessary data files for your retrieval evaluation.

#### How it works

The system loads query files, qrel (relevance judgment) files, and retrieval run files from predefined directories. You can select multiple run files for comparison.

#### How to use

Use this section to set up your evaluation environment. Ensure you have the correct files selected to get accurate and meaningful results in the subsequent analyses.

Select Queries
Select Qrels
Select Retrieval Runs

queries.xml
qrels.csv
results\_bm25\_hi... ×
results\_tf\_idf\_b... ×
results\_bm25\_b... ×
...

Begin the Experimental Evaluation!

### Retrieval Performance - Overall Retrieval Characteristics

See Analysis Details and Interpretations ^

#### Overview

This section provides an overview of the retrieval performance for each selected run.

#### How it works

The results are calculated using standard information retrieval metrics such as the number of queries, retrieved documents, relevant documents, and relevant retrieved documents.

#### How to use

Use these overall metrics to get a quick comparison between different retrieval runs. They can help identify which runs are performing better in terms of retrieving relevant documents.

Relevance judgements are binary, so **relevance threshold is set to 1**.

### Overall Measures Combined

	results_bm25_hist_qe	results_tf_idf_baseline	results_bm25_baseline
Total Queries	10,000	10,000	10,000
Relevant Documents	10,000	10,000	10,000
Relevant Retrieved Documents	6,646	7,124	7,121

### Precision Measures Combined

	results_bm25_hist_qe	results_tf_idf_baseline	results_bm25_baseline
P@5	0.0552	0.0592	0.0591
P@10	0.0276	0.0296	0.0296
P@25	0.011	0.0118	0.0118
P@50	0.0055	0.0059	0.0059
P@100	0.0028	0.003	0.003
Rprec	0.2126	0.2367	0.2365

## Recall Measures Combined

	results_bm25_hist_qe	results_tf_idf_baseline	results_bm25_baseline
Recall@50	0.276	0.2958	0.2957
Recall@1000	0.276	0.2958	0.2957

## Retrieval Performance - Experimental Evaluation

See Analysis Details and Interpretations



## Overview

This section provides a detailed evaluation of retrieval performance using various commonly used IR metrics.

## How it works

The evaluation is performed using selected IR measures (e.g., MAP, nDCG) and includes statistical significance testing against a chosen baseline.

## How to use

Use these results to compare different retrieval runs in depth. Pay attention to statistically significant differences to identify meaningful improvements or degradations in performance.

Relevance judgements are binary, so **relevance threshold is set to 1**.

Correction Value (alpha)



Select a correction method:

Bonferroni



Select a baseline run file:

results\_bm25\_hist\_qe.csv



Statistical significance is tested against the selected baseline ([results\\_bm25\\_hist\\_qe.csv](#)) using a paired two-sided t-test at a significance level (**0.05**). Multiple testing correction is performed using the (**Bonferroni**) method.

	AP@100	P@10	nDCG@10	R@50	RR@1000
results_bm25_hist_qe (Baseline)	0.237	0.028	0.247	0.276	0.237
results_tf_idf_baseline	<u>0.260</u>   0.001 0-001	<u>0.030</u>   0.001 0-001	<u>0.269</u>   0.001 0-001	<u>0.296</u>   0.001 0-001	<u>0.260</u>   0.001 0-001
results_bm25_baseline	0.260   0.001 0-001	0.030   0.001 0-001	0.269   0.001 0-001	0.296   0.001 0-001	0.260   0.001 0-001

Format is **Measure | p-value** corrected p-value. If the observed difference from the baseline is statistically significant, the background of the measure is green. The highest value per measure is underscored.

Select additional measures:

AP x nDCG x R x

Enter measure cutoff value:

x v

5

- +

	AP@5	nDCG@5	R@5
results_bm25_hist_qe (Baseline)	0.237	0.247	0.276
results_tf_idf_baseline	<u>0.260</u>   0.001 0.001	<u>0.269</u>   0.001 0.001	<u>0.296</u>   0.001 0.001
results_bm25_baseline	0.260   0.001 0.001	0.269   0.001 0.001	0.296   0.001 0.001

Format is **Measure** | **p-value** corrected p-value. If the observed difference from the baseline is statistically significant, the background of the measure is green. The highest value per measure is underscored.

## Retrieval Performance - Positional Distribution of Relevant and Unjudged Retrieved Documents

See Analysis Details and Interpretations ^

### Overview

This analysis shows how relevant and unjudged documents are distributed across different ranking positions.

### How it works

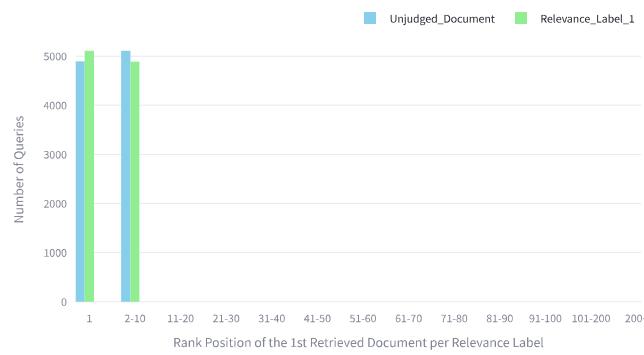
The graphs display the number of relevant, non-relevant, and unjudged documents at each ranking position for each retrieval run.

### How to use

Use these distributions to understand how well each system ranks relevant documents. A good system should have more relevant documents at higher ranking positions.

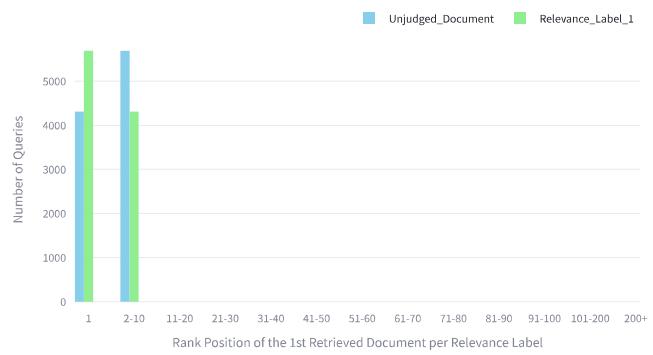
#### Experiment: results\_bm25\_hist\_qe

##### Distribution of Document Ranking Positions



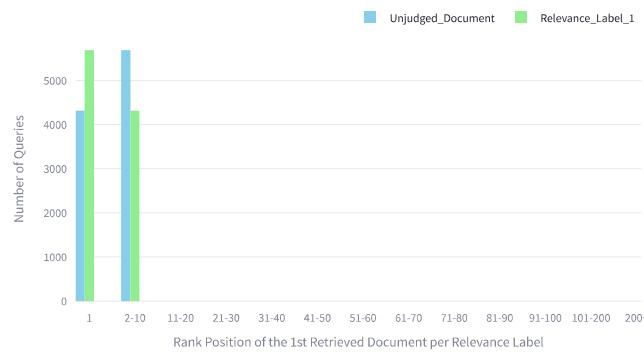
#### Experiment: results\_tf\_idf\_baseline

##### Distribution of Document Ranking Positions



#### Experiment: results\_bm25\_baseline

##### Distribution of Document Ranking Positions



## Retrieval Performance - Precision/Recall Curve

See Analysis Details and Interpretations ^

### Overview

This section presents precision-recall curves for each retrieval run.

### How it works

The curves show the trade-off between precision and recall at different cutoff points in the ranked list of retrieved documents.

### How to use

Use these curves to compare the overall performance of different runs. A curve that is higher and to the right indicates better performance, as it maintains higher precision at higher recall.

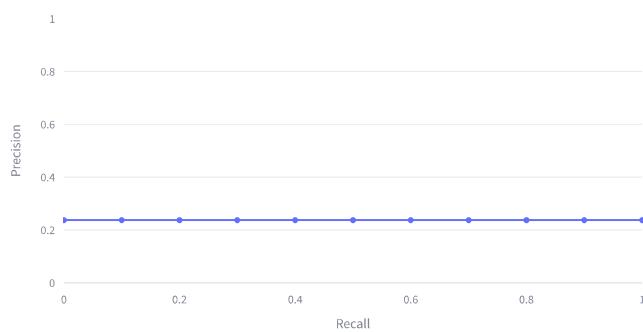
levels.

Relevance judgements are binary, so **relevance threshold is set to 1**.

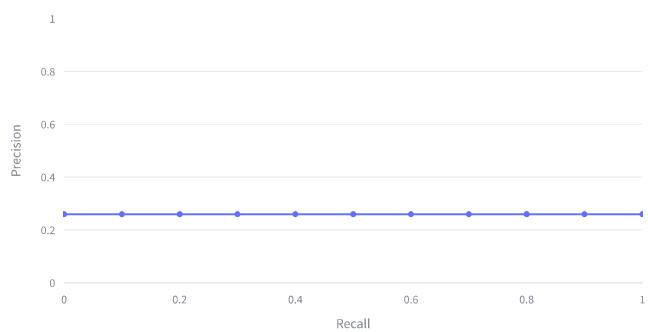
Experiment: **results\_bm25\_hist\_qe**

Experiment: **results\_tf\_idf\_baseline**

Precision-Recall Curve (Relevance\_threshold=1)

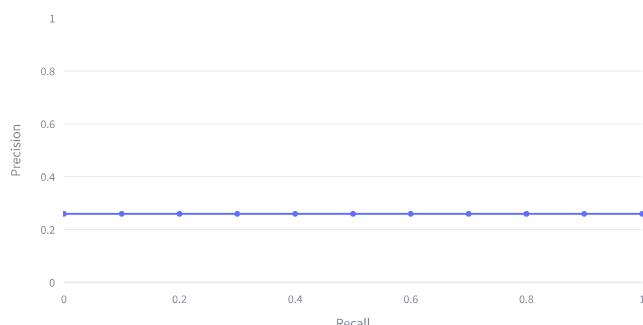


Precision-Recall Curve (Relevance\_threshold=1)



Experiment: **results\_bm25\_baseline**

Precision-Recall Curve (Relevance\_threshold=1)



## Retrieval Performance - Retrieved Document Intersection

See Analysis Details and Interpretations

### Overview

This analysis examines the overlap in retrieved documents between different retrieval runs.

### How it works

The intersection is calculated by comparing the top-ranked documents from each run against a selected baseline run.

### How to use

Use this analysis to understand how similar or different the retrieval results are across runs. High overlap might indicate similar approaches, while low overlap could suggest diversity in retrieval strategies.

This analysis calculates the intersection of retrieved documents between a selected baseline experiment and other retrieval experiments. It showcases the similarity between the document rankings across different retrieval approaches.

Select a baseline run file:

Top-ranked documents considered (ranking cutoff value):

results\_bm25\_hist\_qe.csv

5

**Interpretation:**

- **Intersected Documents:** The number of documents that appear in both the baseline and the compared experiments within the top 5 results.
- **Total Documents:** The total number of documents considered (number of queries  $\times$  cutoff value).
- **Intersection Percentage:** The percentage of documents that intersect, calculated as  $(\text{Intersected Documents} / \text{Total Documents}) \times 100$ .

**A higher intersection percentage indicates greater similarity with the baseline results.****Selected Baseline:** results\_bm25\_hist\_qe. All other experiments will be compared against this baseline.**Cutoff Value:** 5. The cutoff value determines how many top-ranked documents from each experiment are considered in the intersection analysis.

	Intersected Documents	Total Documents	Intersection Percentage
results_tf_idf_baseline.csv	29,464	49,525	59.49
results_bm25_baseline.csv	30,290	49,517	61.17

**Retrieval Performance - Documents Retrieved by All Systems**

See Analysis Details and Interpretations

**Overview**

This analysis identifies documents that are consistently retrieved across all selected retrieval runs.

**How it works**

The system finds documents that appear in the top-N results of all runs for each query, where N is a user-specified cutoff.

**Note!** Additional Analysis regarding the documents retrieved by several experiments and queries, can be conducted in [Query Collection-based Performance Report](#) page.

This analysis identifies documents that are retrieved by all selected retrieval systems within a specified cutoff rank. These documents represent a consensus among different retrieval approaches and may be particularly relevant or central to the queries.

Number of top-ranked documents considered:

1

Documents (IDs) retrieved by all systems:

10

**Cutoff Value:** 1. The analysis considers the top 1 ranked documents for each query across all experiments.

**Interpretation:**

- The query IDs listed above represent queries where at least one document was retrieved by all systems within the specified cutoff.
- For instance, if the cutoff value is 1, the analysis presents those queries for which all systems retrieve the same document in the 1st rank position.
- The documents listed are a sample of those retrieved by all systems, prioritized by frequency across queries. Sample size can be adjusted.
- These documents may be highly relevant to multiple queries or represent core content in your collection.
- Their consistent retrieval across all systems suggests they are important in the context of your retrieval task.

**Total Queries with documents retrieved by all systems: 7234**

Query IDs: test\_10, test\_100, test\_1000, test\_10001, test\_10002, test\_10003, test\_10004, test\_10005, test\_10007, test\_10008, test\_10009, test\_1001, test\_10012, test\_10015, test\_10016, test\_10017, test\_10018, test\_10019, test\_1002, test\_10020, test\_10021, test\_10024, test\_10025, test\_10026, test\_10027, test\_10029, test\_10030, test\_10031, test\_10032, test\_10035, test\_10036, test\_10039, test\_1004, test\_10041, test\_10042, test\_10043, test\_10045, test\_10047, test\_10048, test\_1005, test\_10052, test\_10054, test\_10055, test\_10056, test\_10057, test\_10058, test\_10059, test\_1006, test\_10061, test\_10063, test\_10064, test\_10065, test\_10066, test\_10067, test\_10069, test\_1007, test\_10072, test\_10073, test\_10074, test\_10075, test\_10076, test\_10079, test\_1008, test\_10081, test\_10083, test\_10084, test\_10085, test\_10086, test\_10087, test\_10088, test\_1009, test\_10091, test\_10092, test\_10093, test\_10094, test\_10095, test\_101, test\_10100, test\_10101, test\_10102, test\_10104, test\_10105, test\_10106, test\_10108, test\_10109, test\_1011, test\_10110, test\_10111, test\_10113, test\_10115, test\_10116, test\_10117, test\_10118, test\_10119, test\_1012, test\_10120, test\_10121, test\_10122, test\_10123, test\_10124, test\_10125, test\_10126, test\_10128, test\_10129, test\_1013, test\_10130, test\_10131, test\_10133, test\_10134, test\_10135, test\_10137, test\_10139, test\_10142, test\_10143, test\_10144, test\_10145, test\_10146, test\_10147, test\_10149, test\_1015, test\_10150, test\_10151, test\_10152, test\_10153, test\_10155, test\_10156, test\_10157, test\_10158, test\_10159, test\_1016, test\_10162, test\_10163, test\_10164, test\_10165, test\_10166, test\_10167, test\_10168, test\_10169, test\_1017, test\_10171, test\_10172, test\_10173, test\_10174, test\_10175, test\_10176, test\_10179, test\_1018, test\_10181, test\_10182, test\_10183, test\_10184, test\_10185, test\_10187, test\_10188, test\_10189, test\_1019, test\_10190, test\_10194, test\_10196, test\_10198, test\_10201, test\_10202, test\_10203, test\_10204, test\_10206, test\_10207, test\_10208, test\_10209, test\_1021, test\_10210, test\_10214, test\_10216, test\_10217, test\_10218, test\_10219, test\_1022, test\_10220, test\_10221, test\_10223, test\_10224, test\_10225, test\_10226, test\_10227, test\_10228, test\_10229, test\_1023, test\_10230, test\_10232, test\_10234, test\_10236, test\_10237, test\_10238, test\_10239, test\_1024, test\_10241, test\_10242, test\_10243, test\_10244, test\_10245, test\_10246, test\_10247, test\_10249, test\_1025, test\_10250, test\_10252, test\_10253, test\_10254, test\_10255, test\_10257, test\_10259, test\_1026, test\_10261, test\_10262, test\_10263, test\_10264, test\_10266, test\_10267, test\_10268, test\_1027, test\_10270, test\_10272, test\_10274, test\_10275, test\_10276, test\_10277, test\_10278, test\_10279, test\_10280, test\_10281, test\_10283, test\_10284, test\_10288, test\_1029, test\_10291, test\_10292, test\_10293, test\_10294, test\_10295, test\_10296, test\_10297, test\_103, test\_1030, test\_10300, test\_10302, test\_10303, test\_10304, test\_10305, test\_10306, test\_10307, test\_10309, test\_10310, test\_10311, test\_10312, test\_10313, test\_10314, test\_10315, test\_10316, test\_10317, test\_10318, test\_10319, test\_10320, test\_10321, test\_10323, test\_10324, test\_10326, test\_10327, test\_1033, test\_10330, test\_10331, test\_10332, test\_10333, test\_10334

**Potential Next Steps:**

- Analyze experiments with high intersection percentages to understand what makes them similar to the baseline.
- Investigate experiments with low intersection percentages to determine if they're introducing beneficial diversity or are potentially underperforming.
- Conduct a query-level analysis to identify which types of queries lead to high or low intersection across experiments.
- Consider combining systems with low intersection to potentially improve overall performance.











test\_18547, test\_18548, test\_18549, test\_1855, test\_18550, test\_18551, test\_18552, test\_18553, test\_18555, test\_18556, test\_18557, test\_18558, test\_1856, test\_18560, test\_18561, test\_18563, test\_18564, test\_18566, test\_18567, test\_18568, test\_18569, test\_1857, test\_18570, test\_18571, test\_18572, test\_18573, test\_18574, test\_18575, test\_18576, test\_18577, test\_18578, test\_18580, test\_18581, test\_18583, test\_18584, test\_18586, test\_18587, test\_18589, test\_1859, test\_18590, test\_18592, test\_18594, test\_18595, test\_18596, test\_18597, test\_18599, test\_186, test\_18601, test\_18602, test\_18603, test\_18604, test\_18605, test\_18607, test\_18608, test\_18609, test\_1861, test\_18610, test\_18612, test\_18613, test\_18614, test\_18615, test\_18616, test\_18617, test\_18618, test\_18619, test\_1862, test\_18620, test\_18621, test\_18622, test\_18624, test\_18625, test\_18626, test\_18627, test\_18628, test\_18629, test\_18630, test\_18631, test\_18632, test\_18633, test\_18634, test\_18635, test\_18636, test\_18637, test\_18638, test\_1864, test\_18640, test\_18642, test\_18643, test\_18644, test\_18645, test\_18646, test\_18648, test\_18652, test\_18653, test\_18655, test\_18656, test\_18657, test\_18658, test\_1866, test\_18661, test\_18663, test\_18664, test\_18665, test\_18668, test\_18669, test\_1867, test\_18670, test\_18671, test\_18672, test\_18673, test\_18674, test\_18676, test\_18677, test\_18678, test\_18679, test\_1868, test\_18680, test\_18681, test\_18682, test\_18683, test\_18684, test\_18685, test\_18686, test\_18689, test\_1869, test\_18690, test\_18693, test\_18694, test\_18695, test\_18696, test\_18697, test\_18699, test\_187, test\_1870, test\_18702, test\_18705, test\_18706, test\_18707, test\_18708, test\_18709, test\_1871, test\_18711, test\_18713, test\_18714, test\_18717, test\_18718, test\_1872, test\_18721, test\_18722, test\_18723, test\_18725, test\_18726, test\_18729, test\_18730, test\_18731, test\_18733, test\_18734, test\_18735, test\_18737, test\_18738, test\_18739, test\_18740, test\_18741, test\_18742, test\_18743, test\_18744, test\_18745, test\_18746, test\_18747, test\_18748, test\_1875, test\_18750, test\_18752, test\_18754, test\_18756, test\_18757, test\_18758, test\_18759, test\_18760, test\_18762, test\_18763, test\_18764, test\_18767, test\_18768, test\_18769, test\_18770, test\_18771, test\_18772, test\_18773, test\_18776, test\_18777, test\_1878, test\_18781, test\_18782, test\_18783, test\_18785, test\_18786, test\_18787, test\_18788, test\_18789, test\_1879, test\_18790, test\_18792, test\_18793, test\_18794, test\_18797, test\_18798, test\_18799, test\_188, test\_18802, test\_18804, test\_18806, test\_18807, test\_18809, test\_1881, test\_18811, test\_18813, test\_18814, test\_18815, test\_18816, test\_18817, test\_18818, test\_1882, test\_18820, test\_18821, test\_18822, test\_18823, test\_18824, test\_18825, test\_18826, test\_18827, test\_18830, test\_18831, test\_18832, test\_18833, test\_18834, test\_18836, test\_18837, test\_18838, test\_18840, test\_18842, test\_18843, test\_18844, test\_18845, test\_18846, test\_18847, test\_18848, test\_18849, test\_1885, test\_18850, test\_18851, test\_18852, test\_18855, test\_18856, test\_18857, test\_18859, test\_18860, test\_18861, test\_18862, test\_18864, test\_18866, test\_18867, test\_18868, test\_18869, test\_1887, test\_18871, test\_18872, test\_18873, test\_18875, test\_18878, test\_18879, test\_18881, test\_18883, test\_18884, test\_18886, test\_18887, test\_18889, test\_1889, test\_18890, test\_18891, test\_18892, test\_18893, test\_18894, test\_18894, test\_18897, test\_18898, test\_18899, test\_189, test\_1890, test\_1890, test\_18901, test\_18903, test\_18905, test\_18906, test\_18907, test\_18908, test\_1891, test\_18910, test\_18911, test\_18913, test\_18914, test\_18916, test\_18917, test\_18918, test\_18919, test\_1892, test\_18923, test\_18924, test\_18925, test\_18926, test\_18927, test\_18929, test\_1893, test\_18930, test\_18932, test\_18933, test\_18935, test\_18936, test\_18937, test\_18938, test\_18939, test\_1894, test\_18940, test\_18941, test\_18942, test\_18943, test\_18944, test\_18945, test\_18946, test\_18947, test\_18948, test\_18949, test\_1895, test\_18950, test\_18951, test\_18952, test\_18953, test\_18954, test\_18955, test\_18956, test\_18957, test\_18958, test\_18959, test\_18960, test\_18961, test\_18964, test\_18967, test\_18969, test\_1897, test\_18970, test\_18972, test\_18973, test\_18974, test\_18975, test\_18976, test\_18977, test\_18978, test\_1898, test\_18980, test\_18982, test\_18983, test\_18984, test\_18985, test\_18986, test\_18988, test\_18989, test\_1899, test\_18995, test\_18996, test\_18997, test\_18999

Sample of documents retrieved by all 3 systems:

District\_of\_Columbia\_18, Virgin\_Islands\_of\_the\_U, Virginia\_18190611\_40, West\_Virginia\_18900608\_, Illinois\_18690424\_73, New\_Hampshire\_18130112\_, New\_Hampshire\_18150207\_, Wyoming\_18840812\_1, Massachusetts\_18611010\_, New\_Hampshire\_18181215\_6

### Retrieval Performance - Personal Notes

Please add additional comments regarding this experiment.

To export the report as PDF press (⌘+P or Ctrl+P)