Information Retrieval Final Project Report

IR Final Project (2025–2026)

Students Information

Roi Bublil 322352659, roibub@post.bgu.ac.il

Hadar Shir 314624842, hadaraf@post.bgu.ac.il

Github repo link

https://github.com/ROIBUB/ir-final-project-2026.git

The repository includes the full source code of the project, as well as a README file describing the project structure and execution instructions.

link to the Google Storage Bucket

https://storage.googleapis.com/roi-ir-bucket-1919

List of all index files

Have been saved and attach to the project on git in directory files and the file name index_files.txt

```
1.91 MiB     gs://roi-ir-bucket-1919/postings_gcp/9_012.bin
1.91 MiB     gs://roi-ir-bucket-1919/postings_gcp/9_013.bin
1.91 MiB     gs://roi-ir-bucket-1919/postings_gcp/9_014.bin
1.91 MiB     gs://roi-ir-bucket-1919/postings_gcp/9_015.bin
1.91 MiB     gs://roi-ir-bucket-1919/postings_gcp/9_016.bin
1.91 MiB     gs://roi-ir-bucket-1919/postings_gcp/9_017.bin
1.91 MiB     gs://roi-ir-bucket-1919/postings_gcp/9_018.bin
1.91 MiB     gs://roi-ir-bucket-1919/postings_gcp/9_019.bin
1.91 MiB     gs://roi-ir-bucket-1919/postings_gcp/9_020.bin
1.91 MiB     gs://roi-ir-bucket-1919/postings_gcp/9_021.bin
1.26 MiB     gs://roi-ir-bucket-1919/postings_gcp/9_022.bin
100.51 KiB   gs://roi-ir-bucket-1919/postings_gcp/9_posting_locs.pickle
18.45 MiB    gs://roi-ir-bucket-1919/postings_gcp/index.pkl
5.92 GiB     gs://roi-ir-bucket-1919/postings_gcp/
20.45 GiB    total
```

Experiments

We implemented and evaluated a retrieval system composed of multiple scoring components applied to the document body.
The index was constructed over the full English Wikipedia corpus using an inverted index representation. To improve efficiency and reduce index size, terms with very low document frequency ($df \leq 50$) were filtered out during index construction.
The primary scoring component is based on TF-IDF, where term frequencies are weighted by inverse document frequency. Document scores are normalized using a cosine-like normalization scheme to mitigate the effect of document length differences.
In addition, a BM25-style scoring component was implemented to model term frequency

saturation effects. This component uses standard BM25 parameters (k1 and b), with an average-length approximation rather than explicit per-document length normalization.
The final relevance score of each document is computed using a fixed linear combination of the TF-IDF-based score and the BM25-style score, assigning equal weight to both components.
To incorporate global document importance, PageRank values computed over the Wikipedia link graph were integrated into the scoring function as a multiplicative boost. The PageRank signal is logarithmically scaled and assigned a small weight, ensuring that it acts as a secondary signal and does not dominate textual relevance.
All experiments were conducted using a single fixed retrieval configuration, which was applied consistently across all queries.

Evaluation Methodology

Evaluation was performed using the official metrics and evaluation functions provided by the course staff. The evaluation was conducted using the supplied query set (queries_train.json), where each query is associated with a list of relevant document identifiers.
For internal analysis, the query set was randomly shuffled using a fixed random seed and split into training (70%) and validation (30%) subsets.
This split was used solely for diagnostic purposes and did not affect the retrieval configuration or ranking logic. For each query, the search engine returned up to 100 ranked documents.
System performance was evaluated using the following metrics, exactly as defined in the course materials:

- Precision@5, measuring the relevance of the top-ranked results
- F1@30, capturing a balance between precision and recall at a deeper cutoff
- results_quality, defined as the harmonic mean of Precision@5 and F1@30

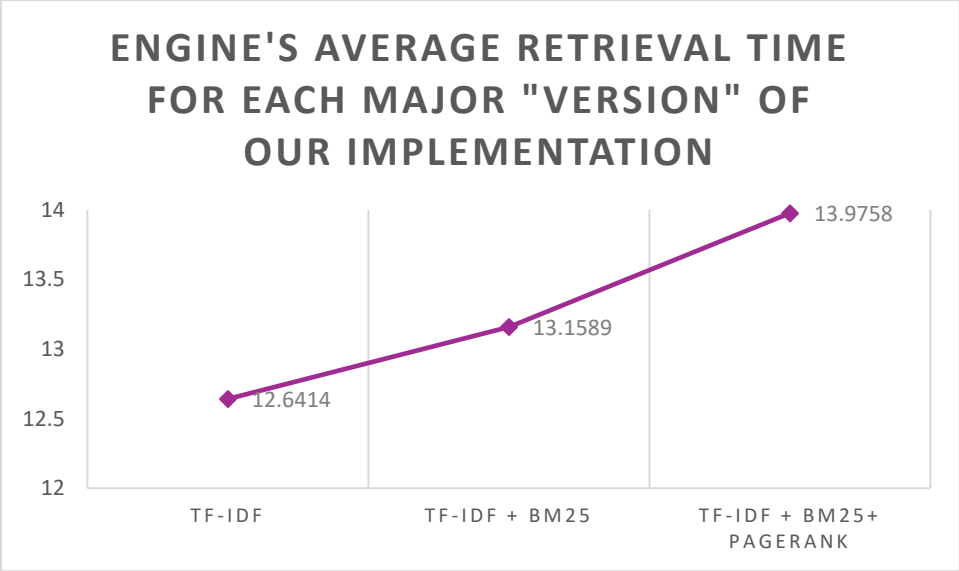
In addition to aggregate scores, per-query diagnostics were examined, including overlap between retrieved and relevant documents within the top 40 results, in order to better understand query-level performance behavior

Key Findings and Observations

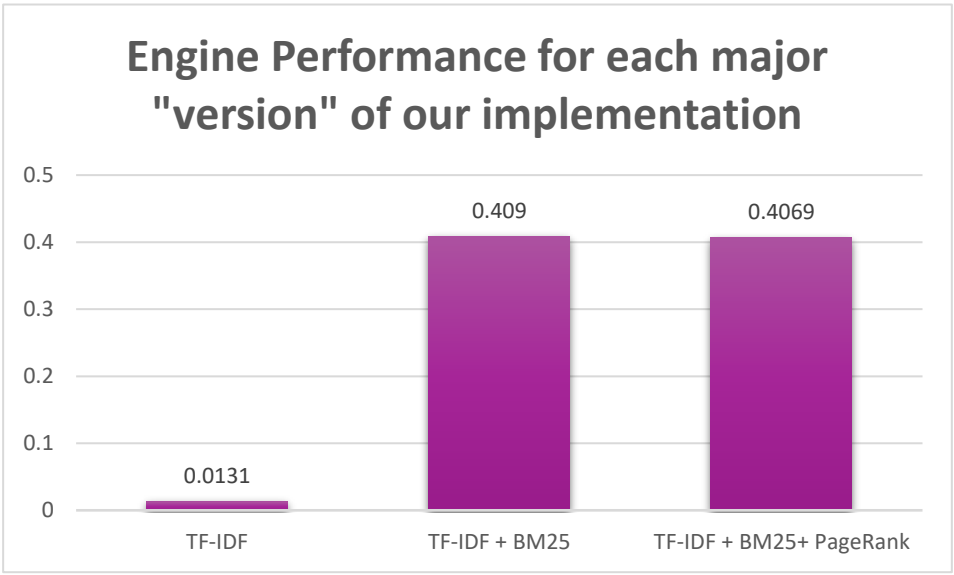
The evaluation results indicate consistent retrieval performance across the training and validation subsets.
The average results_quality score was approximately 0.418 on the training set and 0.407 on the validation set, suggesting stable behavior and limited overfitting.Qualitative inspection of individual queries revealed that the system performs particularly well on focused,

specific queries that contain distinctive terms.
In such cases, the combination of term-based scoring and global PageRank information effectively highlights relevant documents.
In contrast, more general or ambiguous queries tend to yield lower performance. These queries often correspond to broad topics with many relevant documents or terms that appear across diverse contexts, making it more difficult for term-based scoring methods to clearly distinguish highly relevant results.
Overall, the observed behavior aligns with known characteristics of term-based retrieval models, where precision is typically higher for narrow information needs and lower for broad or underspecified queries.

Graphs:



This graph shows the retrieval quality of our search engine for each major version.
Our basic TF-IDF model performs poorly, while adding BM25 significantly improves the results. Adding PageRank does not lead to a noticeable improvement in retrieval quality.



This graph shows the average retrieval time of our search engine for each version.
The TF-IDF version is the fastest, while adding BM25 and PageRank increases the response time. This shows that improving retrieval quality comes at the cost of higher query time.

Qualitative Evaluation

Good-performing query:

The query "Dana International Eurovision" demonstrates good qualitative performance. Manual inspection of the top 10 results returned by our search engine shows that most retrieved documents are clearly relevant and focus on Dana International, the song "Diva", and Israel's participation in the Eurovision Song Contest.

The retrieved results address key concepts directly related to the query, including Dana International's Eurovision appearances and general articles about the competition, indicating that the system successfully captures the thematic context of the query.

As a qualitative reinforcement, the results were compared to those returned by Wikipedia's native search interface, revealing that six out of the top 10 retrieved documents are shared between the two systems.
This overlap provides additional qualitative evidence that the system ranks relevant documents in a manner comparable to a well-established search engine.

Poor-performing query:

The query "Ben Gurion" illustrates a weaker qualitative performance due to ambiguity in user intent.
Manual inspection of the top 10 results returned by our search engine shows that, while many retrieved documents are semantically related to the query terms, several of them correspond primarily to the entity "Ben Gurion Airport" rather than to David Ben-Gurion as a historical figure.
Although the top-ranked result correctly identifies "David Ben-Gurion", many subsequent results focus on transportation- and infrastructure-related topics, such as "Ben Gurion Airport", "Tel Aviv", and "El Al".

A comparison with Wikipedia's native search interface further highlights this issue: Wikipedia's results emphasize biographical and historical entries related to David Ben-Gurion, whereas our system places greater emphasis on location-based and transportation-related entities.
This behavior reflects a limitation of term-based retrieval methods when handling short and ambiguous queries with multiple plausible interpretations and no explicit disambiguation signals.