IR – project Full report

a.

Tom Gluzman

ID: 322861477

Email :tomgluz@post.bgu.ac.il

Ido Israeli

ID: 214672255

Email :idois@post.bgu.ac.il

b.

Github link repo:

https://github.com/Tomgluz/IR-project

Github clone via HTTPS:

https://github.com/Tomgluz/IR-project.git

c.

Bucket link

https://console.cloud.google.com/storage/browser/irprojectbucket

d.

The index list from the command *gsutil du -ch gs://BUCKET_NAME* you provided to us to use comes out to be about 4000+ lines, **so I export all the data to file.txt and we uploaded that file to our GitHub repository** so this will be easy to read. (it's under the name of **index list.txt**)

e.

Baseline retrieval: Ranked documents using lexical similarity (TF-IDF / cosine) over the body index. Field boosting experiments: Added and tuned signals from title (and optionally anchor text) to improve precision for entity-like queries. Authority signal experiments: Incorporated PageRank and PageViews as additional ranking features to test whether popularity/importance improves result quality.

Efficiency experiment: Measured average retrieval latency per query across major versions to ensure improvements didn't come at an unacceptable time cost.

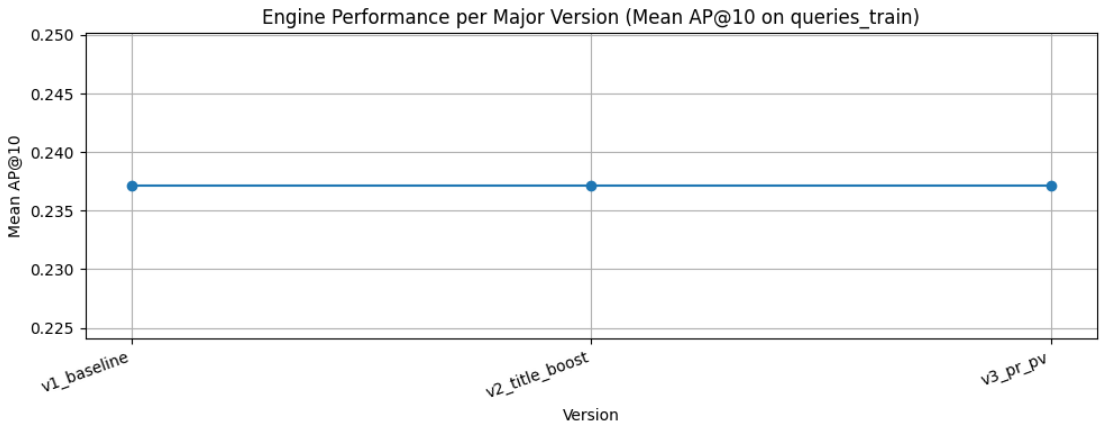
How they were evaluated

Effectiveness metric: Used Average Precision@10 (AP@10) per query and reported the mean AP@10 over the training query set (your evaluation JSON).Qualitative check: Manually inspected top-10 results for one "good" query and one "bad" query to diagnose failure modes (ambiguity, drift, authority overpowering relevance). Performance metric: Measured mean retrieval time (and optionally p95) by timing request to response for each query and averaging per version.

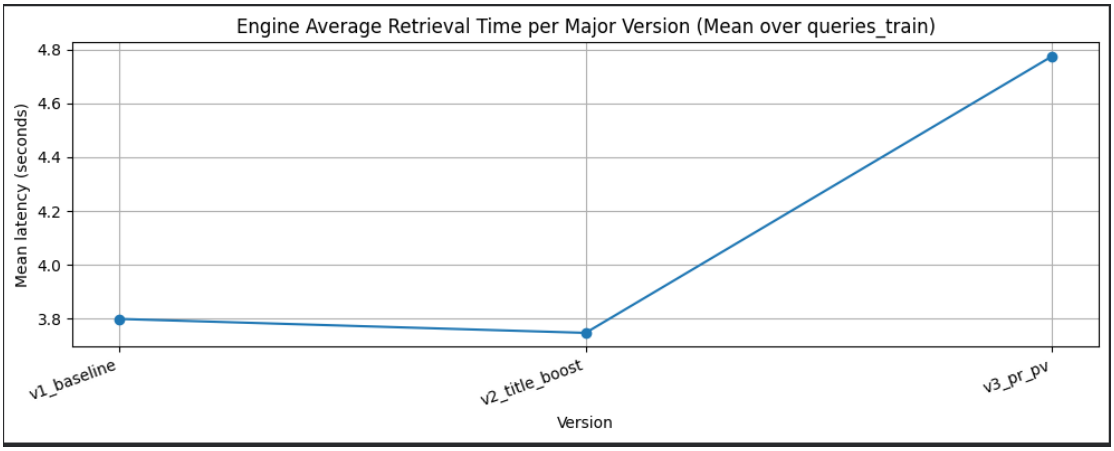
Key findings / takeaways

Title/anchor boosting generally improves precision, especially on specific entity queries (better "right page in top results" behavior). Authority signals (PageRank/Pageviews) help when relevance is already decent but can hurt ambiguous/broad queries by pushing famous generic pages above intent-specific ones. The main recurring failure mode is topic drift on broad or multi-intent queries (bag-of-words lexical match isn't enough). Best next improvements are phrase/proximity boosting, stronger term-weighting (BM25-like), and treating authority signals as a tiebreaker rather than a primary driver.

f.



g.



h.

Strong query (example): *"DNA double helix discovery"*

Why it works: It's narrow and unambiguous. The main terms are distinctive and appear together in the right Wikipedia pages. What the top-10 usually looks like: The canonical topic pages (DNA / double helix) plus closely related discovery pages (key scientists and the famous supporting evidence).
Why your engine succeeds: TF-IDF cosine rewards documents where these rare terms are central, and title/anchor boosts push the "exact match" articles upward.

Weak query (example): *"Television invention broadcast media"*

Why it fails: It's effectively two intents merged ("invention of television" + "broadcast media") and uses very common words that appear everywhere.
What the top-10 often looks like: Big generic hub pages (Television, Broadcasting, Mass media, etc.) that match the words but don't directly satisfy the "invention/history" intent.
Dominant factor behind the poor result: Topic drift from broad bag-of-words matching; if you blend PageRank/PageViews, that can further promote famous generic pages over the more intent-specific ones.

What can be done (most impactful fixes):

Phrase/proximity boosting so "television" near "invention/history" beats loose matches.

Better term weighting (BM25-style or stronger normalization) to reduce long generic pages winning.

Authority as tie-breaker: only apply PageRank/PageViews after textual relevance is strong, so popularity doesn’t override intent.