Information retrieval

PROJECT- 2024

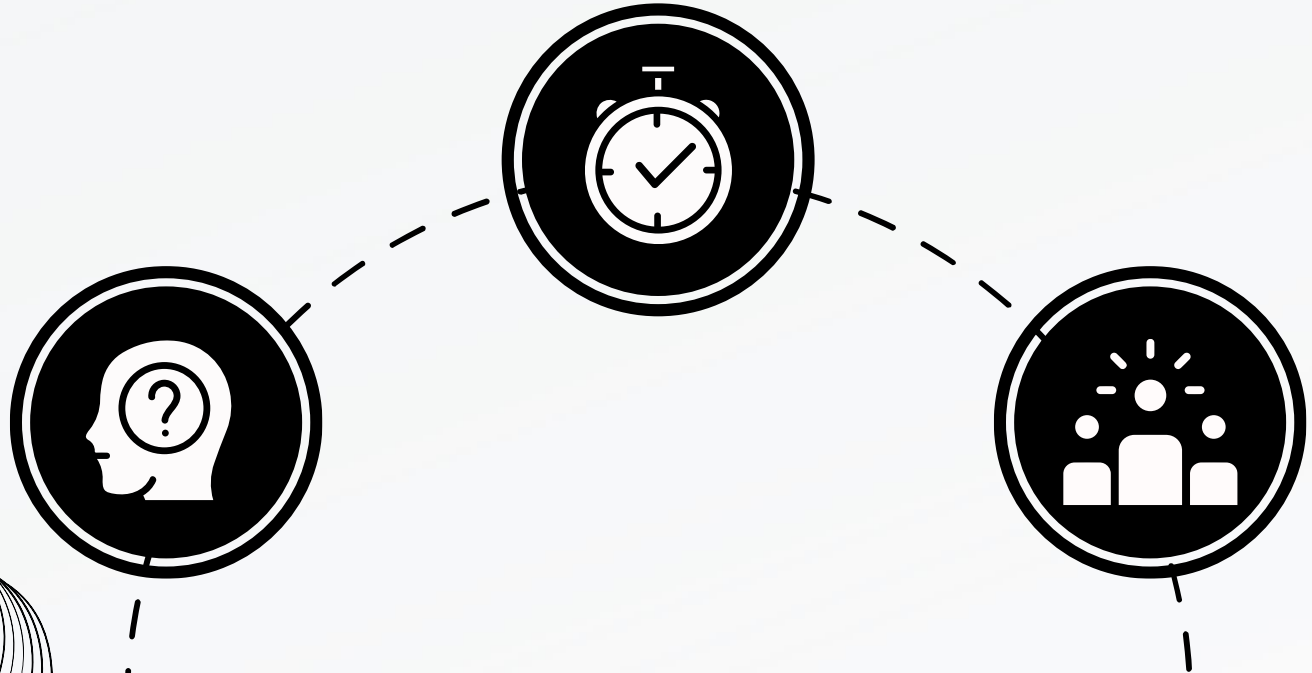
Ariel Siman Tov & Tal Klein

GOALS OF THE PROJECT

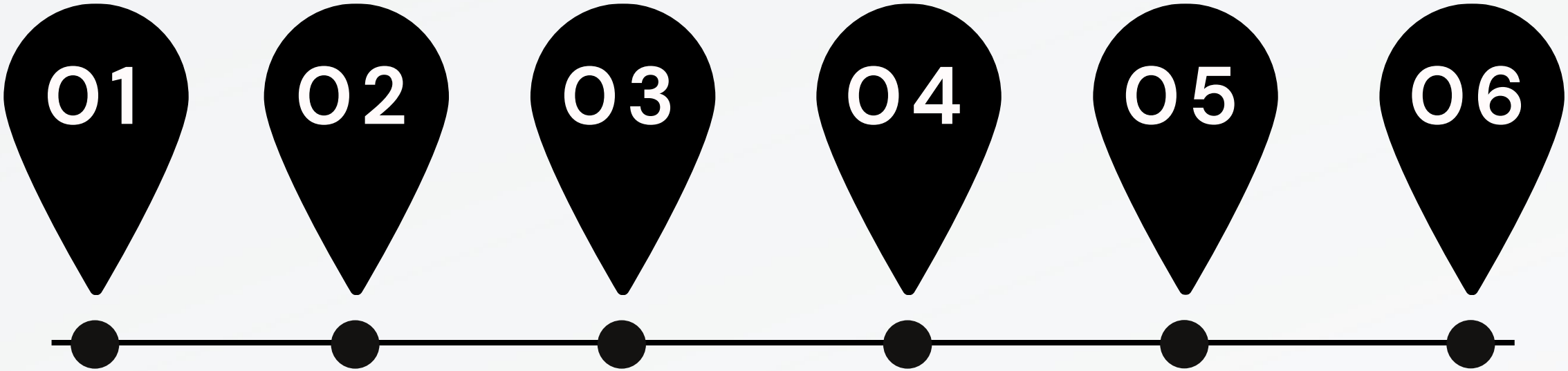
Building a search engine for English Wikipedia

Efficiency of minimum average retrieval time

Results quality by high harmonic mean of Precision@5 & F1@30



KEY EXPERIMENTS STEPS



Basic Model

Cosine similarity + Stemming

Ranking Method

Change the ranking method to BM25

AVG_Time Improvement

Use treads, Use Heap data structure, Use “Impact Ordering” by setting thresholds

NLP Model

Try to use “query expansion” methods by using Word2Vec

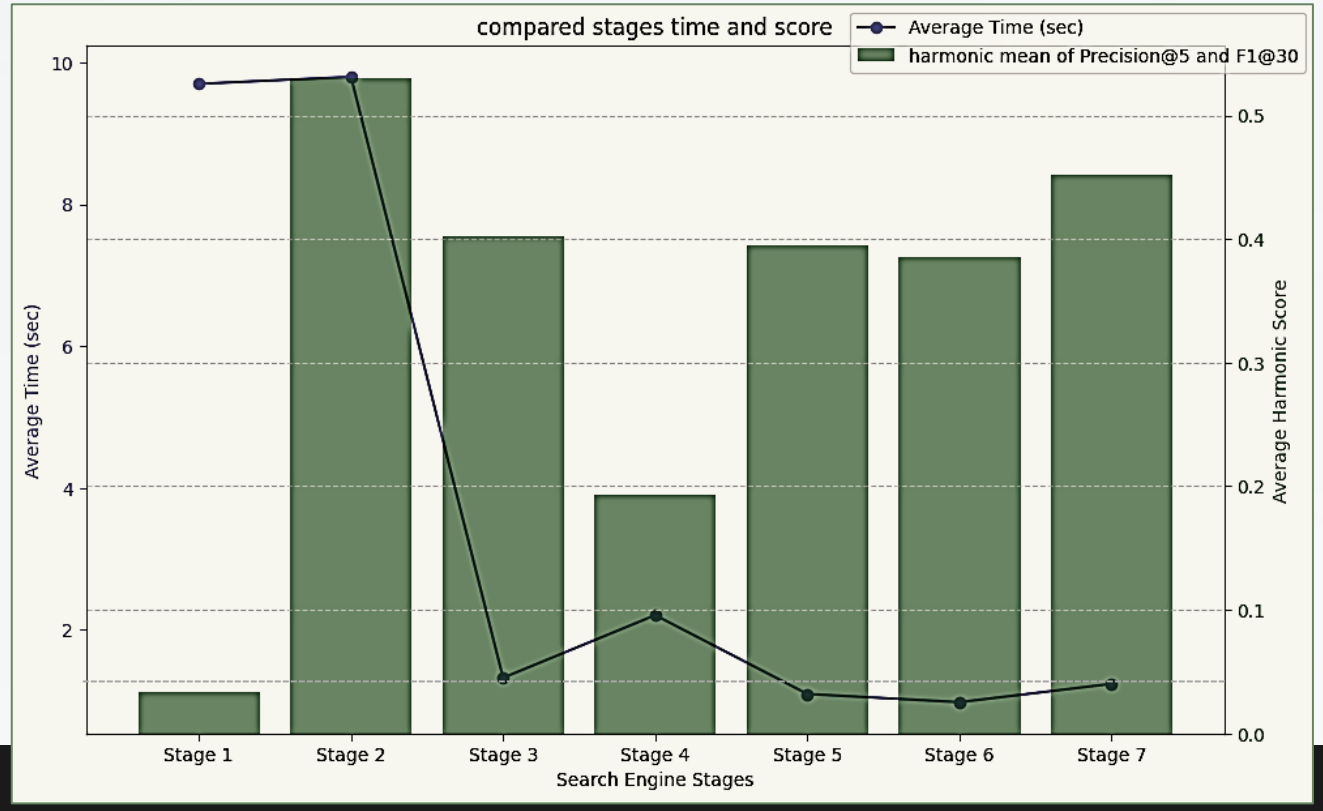
Normalization Methods

Try lemmatization as an alternative to stemming. Stay with No stemming and no lemmatization

Performance Rate Improvement

Adding Pageview, Improve our Tokenization, increase the Dictionary size threshold

ENGINE PREFORMENCE AND RETREVAL TIME BY MAIN STEPS



FINAL RESULTS:

AVG_Time: 1.238 | AVG_Hermonic: 0.4525 | AVG_Precision@10: 0.63

TAKEWAYS AND KEY FINDINGS

The choice of a ranking method significantly impacts our engine's performance.

Adopting the BM25 ranking method consistently improves our results

Optimization strategies involving multi-threading, suitable data structure (like heap) and ordering/sorting the candidate, contributed to enhanced efficiency.

NLP models, such as Word2Vec and Doc2Vec, has a big potential to improve performance, especially for case of static corpus.

Removing stemming or lemmatization in certain cases may resulted better performance.

In our understanding for two reasons: Time complexity and Evaluate method which lay mostly on precision

QUALITATIVE EVALUATION EXAMPELS

BAD QUERY EXAMPLE

"When did World War II end?"

Time – 3.595 | Hermonic_Score - 0.154

Our engine faced challenges because in our tokenization process, the term "II" was omitted from the query, leading to inaccuracies as the engine referred to World War in a general context, not specifically to World War II.

GOOD QUERY EXAMPLE

"Computer"

Time – 0.3243 | Hermonic_Score - 0.754

This query performs well because computer is a concise and frequently used term and a widespread word across various contexts, making it easier to match within documents. Moreover, this word has general relevance and Non-Hidden-Semantic meaning which helps the retrieval.

THANK'S FOR YOUR LISTENING

QUSTIONS?