CS6370
NLP
Course Project

Improving the Simple VSM Search Engine

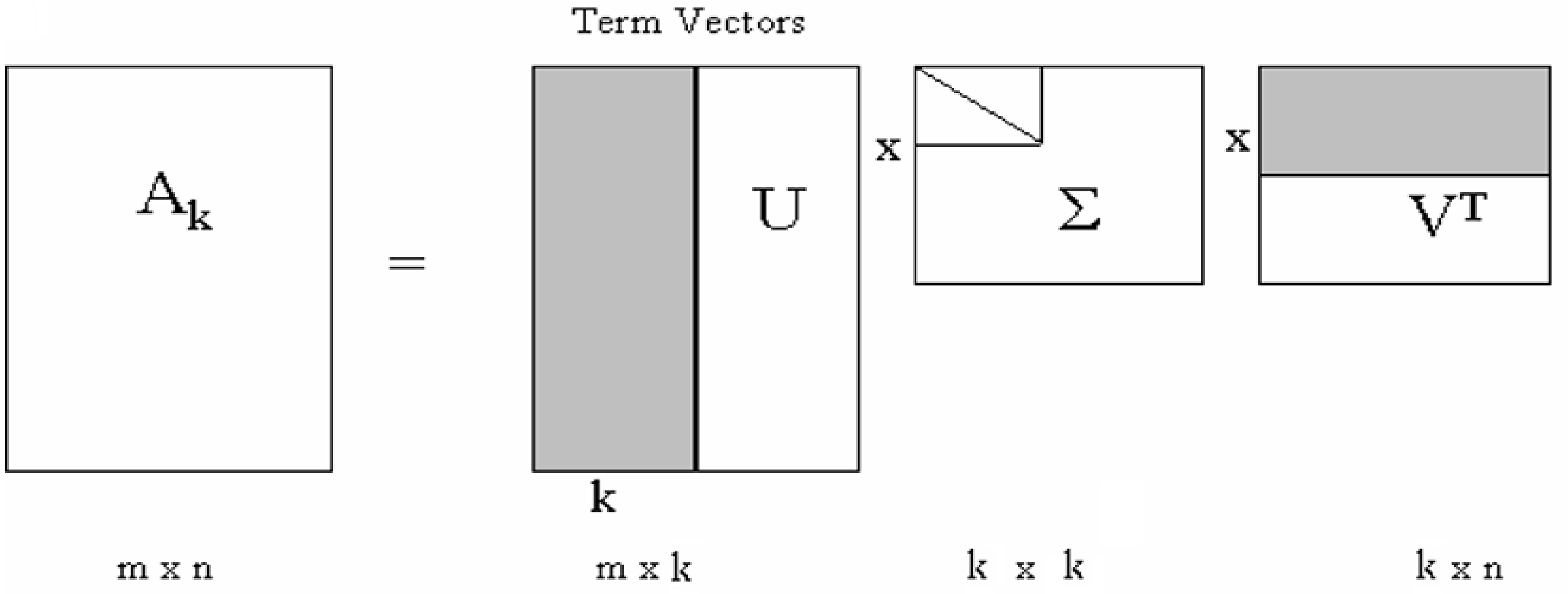
- Naveen Vakada (CS20S012)
- Arup Das (CS20S016)

Limitations of the current search engine

1. Problem of Synonymy: Vocabulary mismatch
2. Problem of Polysemy
3. Ignores co-occurrence relation between terms in similar documents.
4. The title of the document was not used.
5. Ignores word order.
6. Spelling errors negatively impact relevance results.

Solutions

- LSA



A – Tf idf matrix of the corpus
M – vocabulary size
N- Number of documents
U – Term vector representation (Eigen vectors of A'A)
V – Document vector representation (Eigen vectors of AA')
Σ - Singular values matrix
K- dimenationality of the latent space
q' - Query in TFIDF vector representation
$\hat{q}$ - Query in latent dimensions.

$$\hat{q} = q' U_k \Sigma_k^{-1}$$

Advantages:

- Dual mode for comparison
- Solves polysemy partially.
- Solves synonymy.
- Addresses co-occurrence problem.

Disadvantages:

- Boolean logic like statements are not interpreted correctly.
- Adding new document requires to perform SVD decomposition again.
- Word order is not considered.

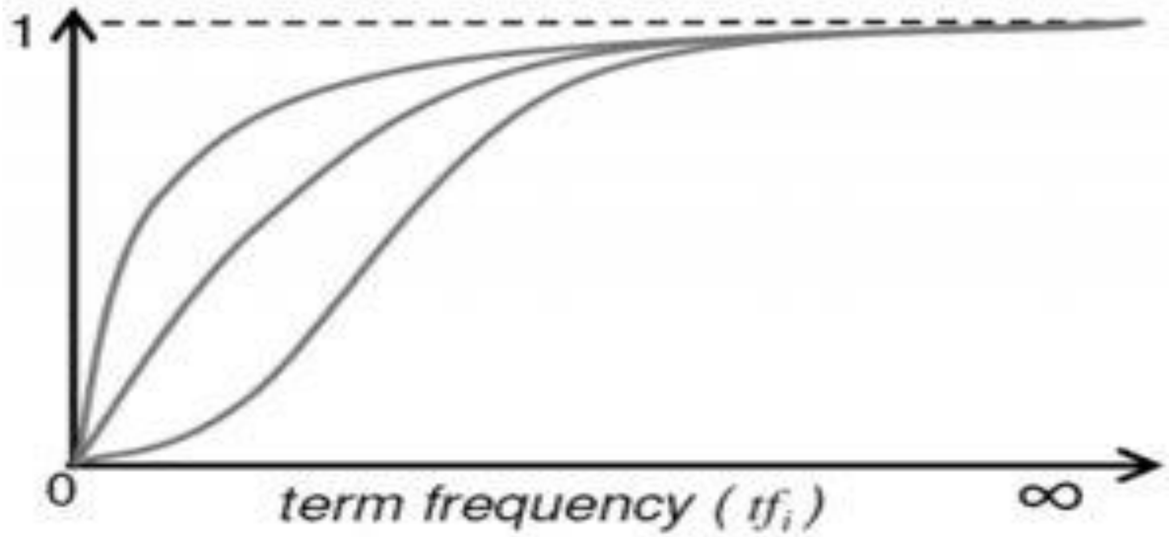
- BM25 – Eliteness (describes the aboutness of a document with respect to its terms.)

$$\Sigma_{t \in q} IDF(t) * \frac{(k_1 + 1) * tf(t, d)}{k_1[(1 - b) + \frac{b*dl}{avdl}] + tf(t, d)} * \frac{(k_2 + 1) * tf(t, q)}{k_2 + tf(t, q)}$$

$$\text{where, } IDF(t) = log\left(\frac{N}{df(t)}\right)$$

1. N is the total number of documents in the corpus.
2. df(t) is the number of documents containing the term t.
3. dl is document length measured in terms of the number of words in the document.
4. avdl is the average length of the document in the collection.
5. tf(t,d) is the frequency of term t in document d.
6. tf(t,q) is the frequency of term t in query q.
7. K_1 is a non-negative tuning parameter that scales the document term frequency.
8. k_2 is a non-negative tuning parameter that scales the query term frequency.
9. b is a non-negative tuning parameter that scales the document length.

- Problem with traditional TF IDF: Rewards term frequency and penalizes document frequency.



- Values: b = 0.75, k_1 in [1.2, 2].
- Advantages

1. Computationally efficient.
2. Penalizes term frequency in documents.

- Disadvantages

1. No guidelines to set the hyperparameters.
2. Penalizes very long documents.

Why are documents lengthy?

1. Verbosity & Scope Hypothesis

2. Soft normalization.

1. Word order – Accounting for the word sequences:

- Used n-gram approach to solve this problem.
- Higher values of n (3,4,..) are computationally expensive to experiment, but they can capture long term dependencies.

2. Query Expansion

- Two approaches can be used:
- Wordnet based
- LSA based approach.

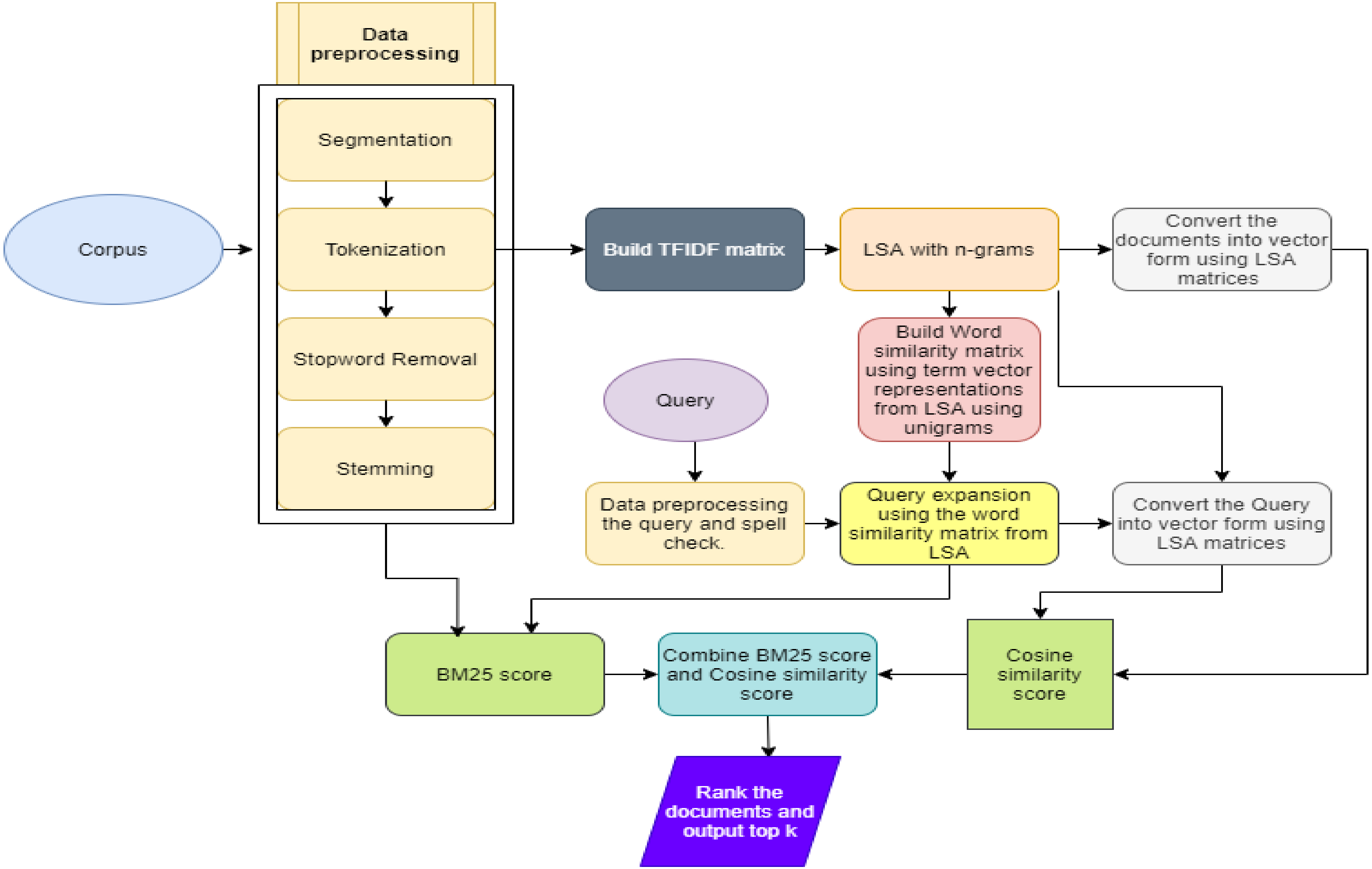
3. Title Inclusion

- Giving twice the weightage for the terms in the title of the documents improved the performance of the IRS.

4.Spell Check

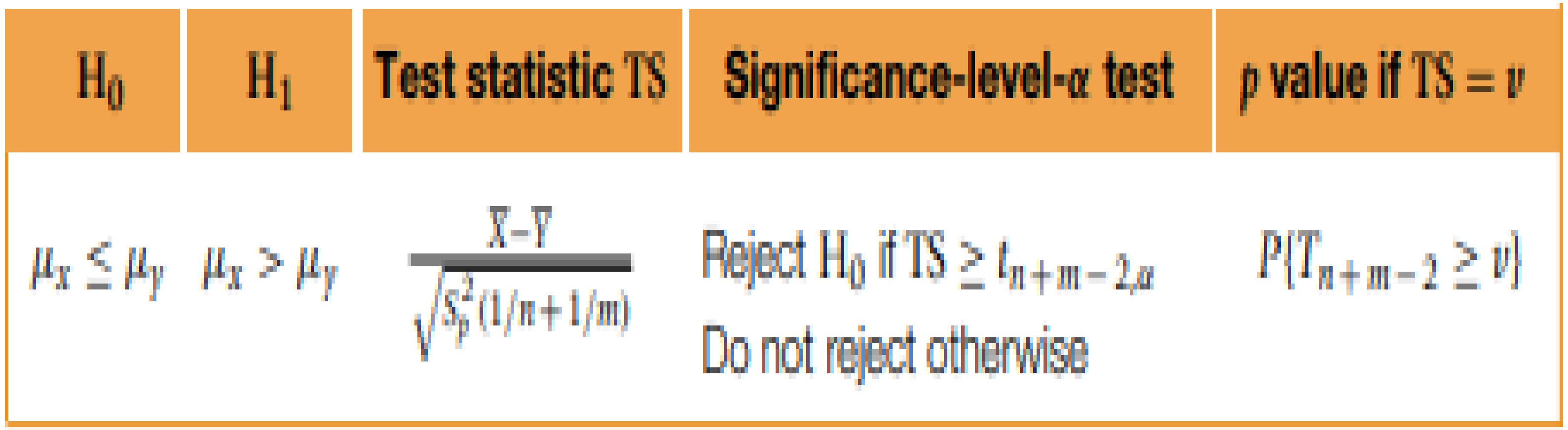
- A simplified version of the Spell Correction Program based on Noisy Channel model by Mark D. Kernighan inspired from Peter Norvig's Blog Post with 68% accuracy on the test set.

IRS Architecture


Data preprocessing
Segmentation
Tokenization
Stopword Removal
Stemming
Corpus
Build TFIDF matrix
LSA with n-grams
Convert the documents into vector form using LSA matrices
Build Word similarity matrix using term vector representations from LSA using unigrams
Query
Data preprocessing the query and spell check.
Query expansion using the word similarity matrix from LSA
Convert the Query into vector form using LSA matrices
BM25 score
Combine BM25 score and Cosine similarity score
Cosine similarity score
Rank the documents and output top k

Hypothesis Testing

$$S_p^2 = \frac{(n-1)S_x^2+(m-1)S_y^2}{n+m-2}$$



H_0	H_1	Test statistic TS	Significance-level-α test	p value if TS $= v$
$\mu_x \le \mu_y$	$\mu_x > \mu_y$	$\frac{\bar{X}-\bar{Y}}{\sqrt{S_p^2(1/n+1/m)}}$	Reject H_0 if TS $\ge t_{n+m-2,\alpha}$ Do not reject otherwise	$P\{T_{n+m-2} \ge v\}$

20 samples where each sample contains 10 queries.

k = 3

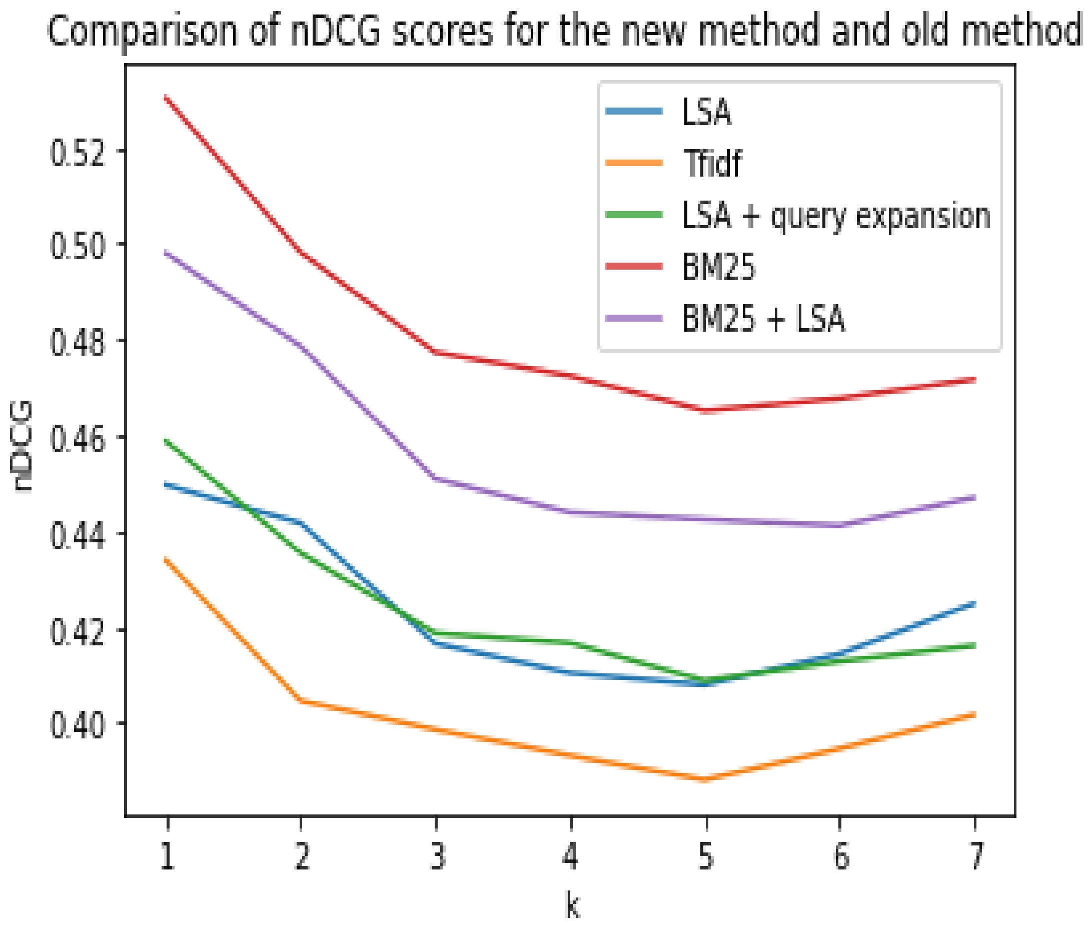
evaluation metrics used nDCG, Mean Precision, Mean Recall, and Mean F-Score.

Evaluation metric graphs were plotted for values of k from 1 to 7

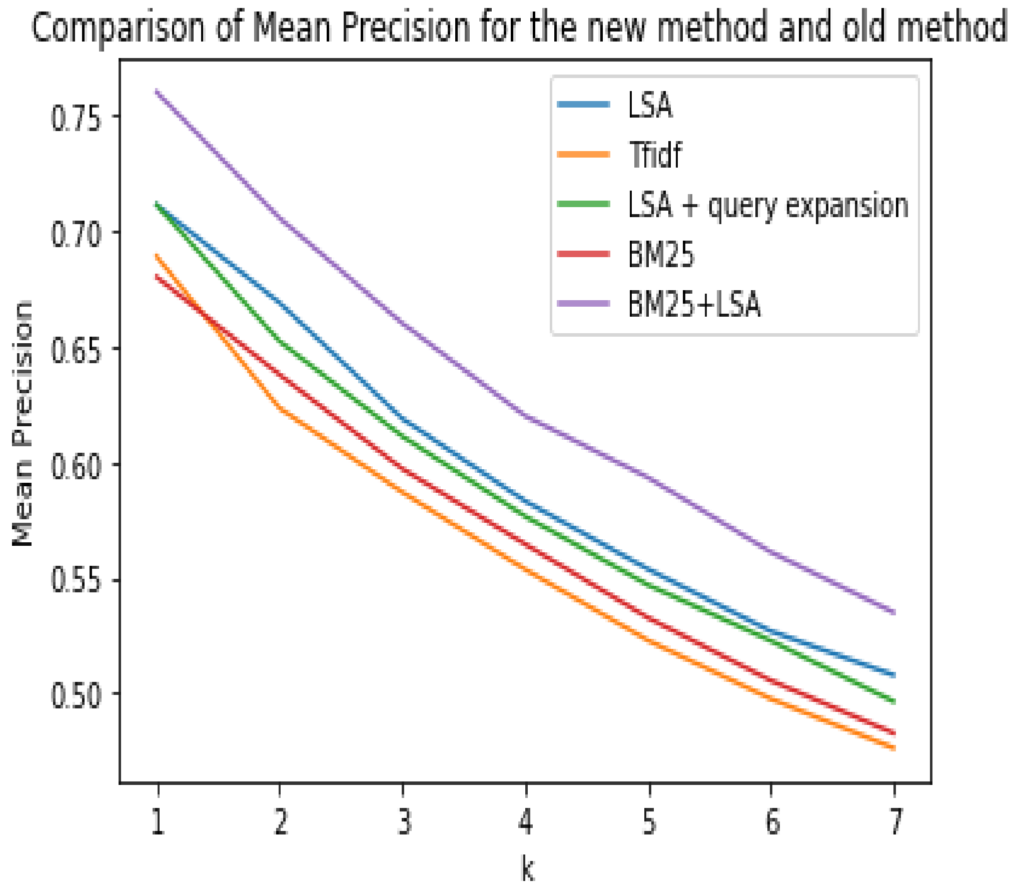
Null Hypothesis (H_0): Mean of the evaluation metric of a candidate hypothesis does not give a better score when compared to that of the current search engine.

Alternate Hypothesis (H_4): Mean of the evaluation metric of a candidate hypothesis gives a better score when compared to that of the current search engine.

Results

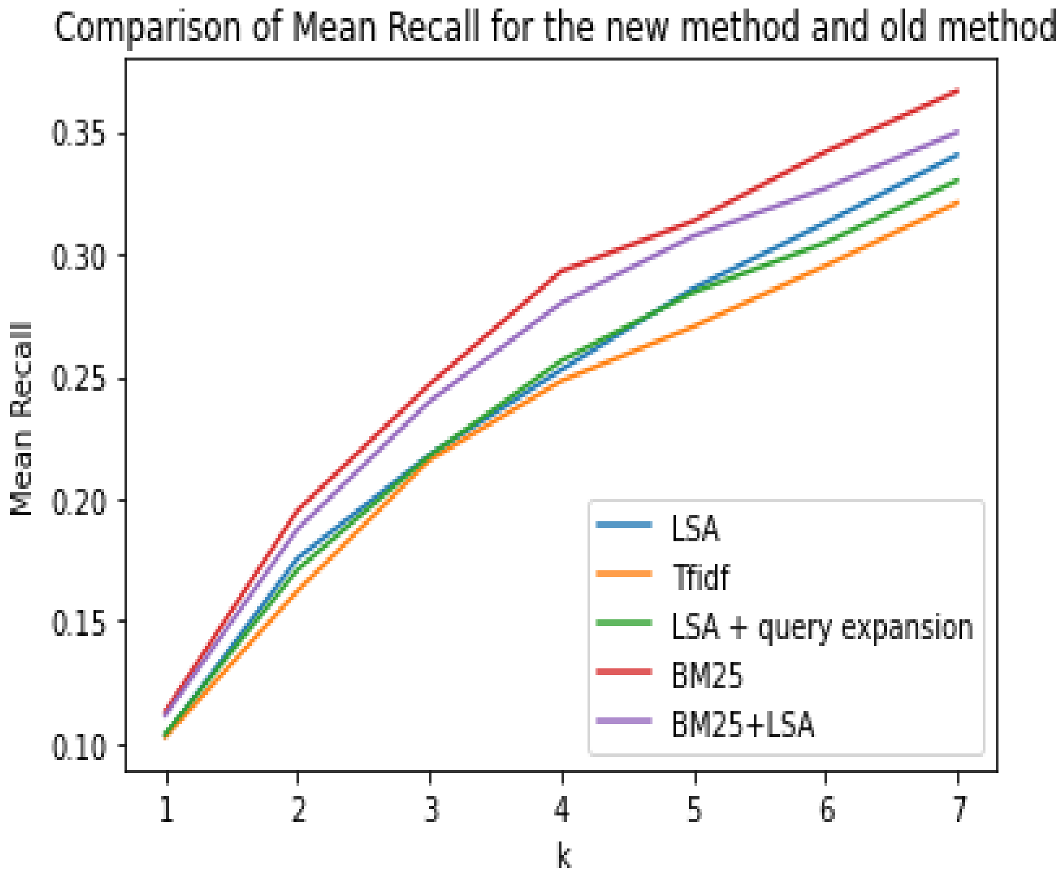


Comparison of nDCG scores for the new method and old method

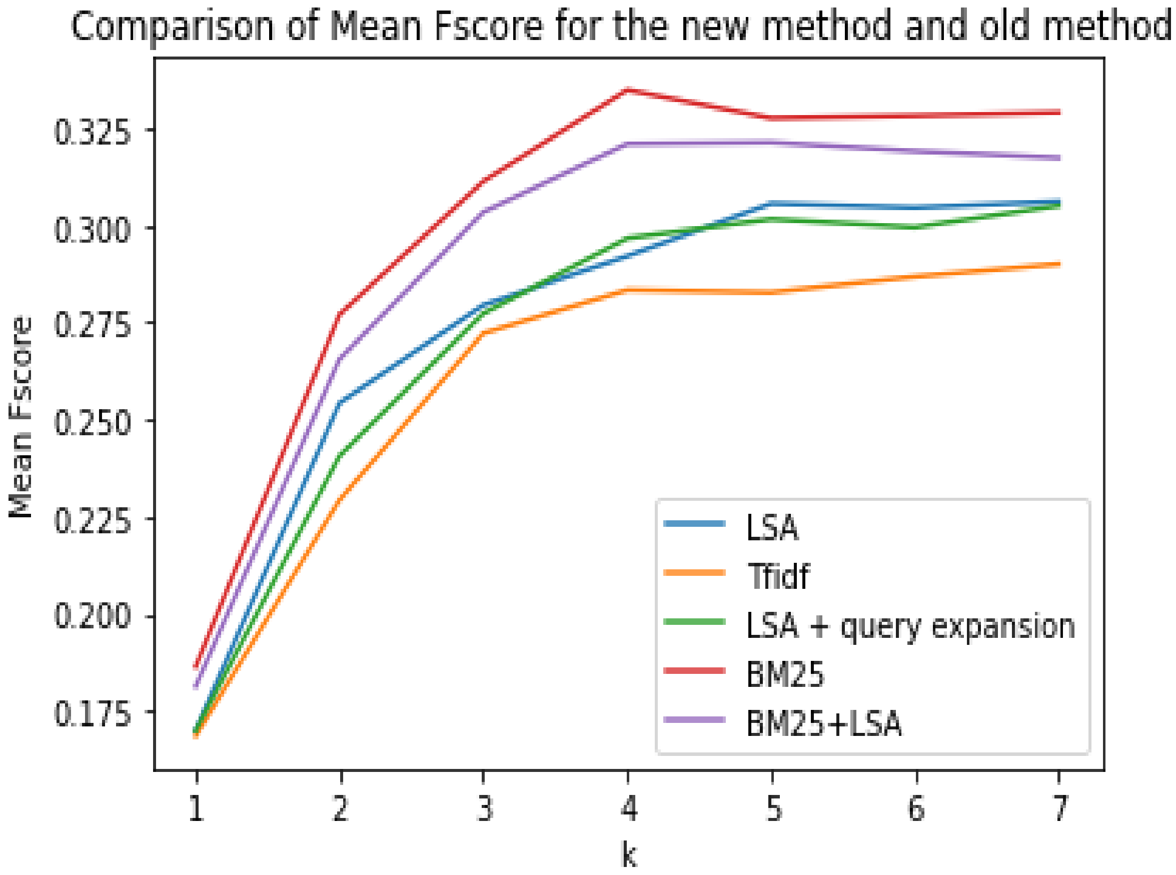


Comparison of Mean Precision for the new method and old method

Results



Comparison of Mean Recall for the new method and old method



Comparison of Mean Fscore for the new method and old method

Best Model

- **Convex combination procedure:**

α =a/(a + b) and β=1 -α

where a = nDCG(LSA) @K = 3 and b = nDCG(BM25)@K = 3

Final Score = α * CosSim(LSA) + β * BM25Score

$\alpha = 0.46$ and $\beta = 0.53$

- **Results from Hypothesis testing:**

The graphs obtained shows that BM25 is better in most of the metrics, from statistical viewpoint BM25+LSA outperforms other models when metrics like nDCG, mean Precision and mean F-Score are used.

Conclusion & Future enhancements

- Convex combination of BM25 and LSA along with the n-gram approach and query expansion is a better modeling choice than the current search engine.

- To further improve the retrieval results in our project, we can explore the possibility of using various neural models for information retrieval.

References

- Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R., 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), pp.391-407.
- Robertson et. al., The Probabilistic Relevance Framework: BM25 and Beyond, Foundations and Trends in Information Retrieval, 2009
- Pilato, Giovanni, et al. "A simple solution for improving the effectiveness of traditional information retrieval systems." WSEAS Transactions on Information Science & Applications 2 (2005): 189-194
- Peter Norvig's spell correction algorithm: http://norvig.com/spell-correct.html
- Sheldon M. Ross, Chapter 10 - Hypothesis Testing Concerning Two Populations, Introduction Statistics (Third edition), Academic Press, 2010, Pages 443 - 498, ISBN: 978-0-12-374388-6, DOI:: 10.1016/B978-0-12-374388-6.00001-6