# Information Retrieval

Aditya Viraj Rao

Kakunuri Shashank Reddy

Haricharan Korrapati

May 15, 2024

# Introduction to Prior IR System

- Depended on TF-IDF method for document indexing.
- Utilized cosine similarity for judging query relevance with documents.
- Effectiveness measured on the Cranfield dataset using precision-recall and ranking-based metrics.

# Shortcomings in Prior IR System

**d1**: Herbivores are typically plant eaters and not meat eaters
**d2**: Carnivores are typically meat eaters and not plant eaters
**d3**: Deer eat grass and leaves
**query : plant eaters**

**Issues:**
**Bag of Words**
**Sparsity**
**Semantic Relatedness**

# Evaluation Metrics for Vector Space Model

MAP @ 10: 0.61, nDCG @ 10: 0.46

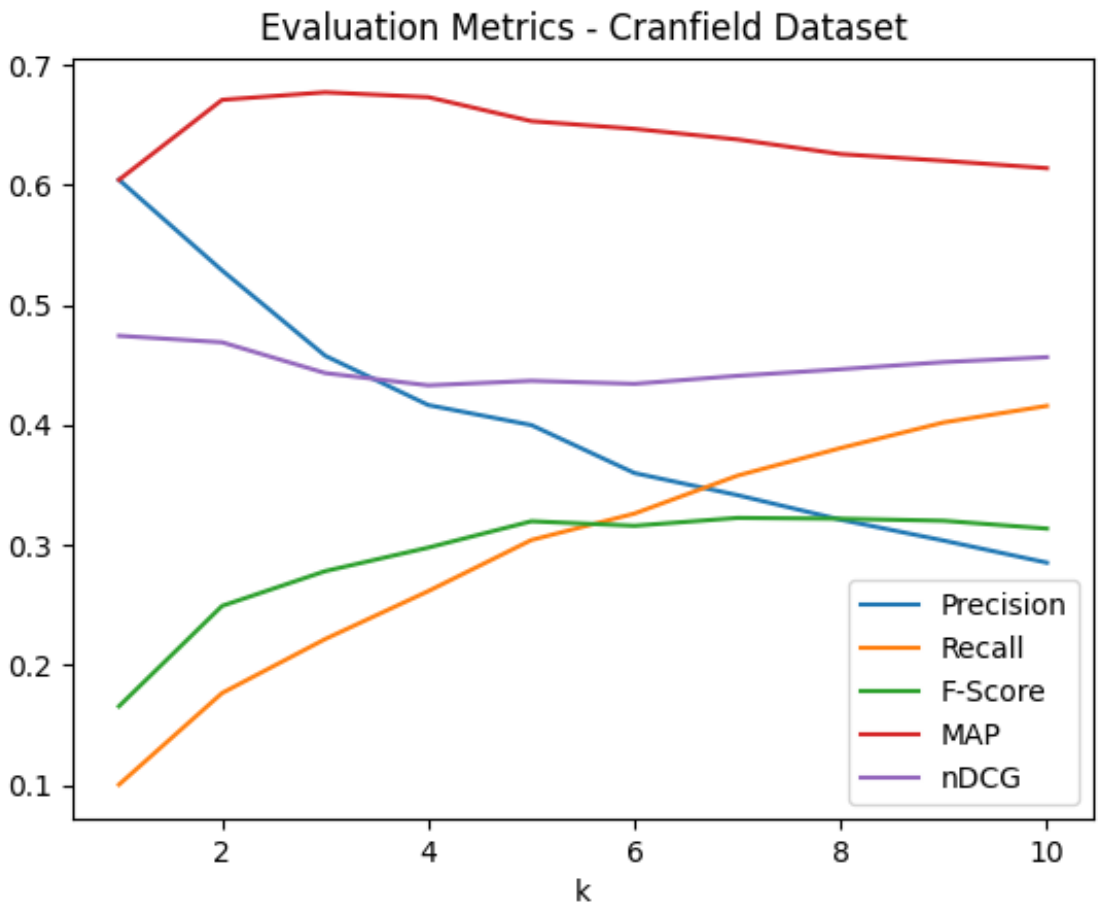


Figure: Evaluation Metrics for Vector Space Model on the Cranfield Dataset

# Challenges

- Sometimes, no relevant documents are found in the top 10 results
- The current model produces high-dimensional sparse vectors
- Synonyms used in queries and documents reduce recall
- The bag-of-words representation decreases precision. Retrieving irrelevant documents alongside relevant ones impacts the overall performance of the IR system.

# Latent Semantic Analysis

- LSA, or Latent Semantic Analysis, uncovers hidden semantic relationships between words and documents.
- It utilizes Singular Value Decomposition (SVD) to represent words and documents in a lower-dimensional space, where each dimension corresponds to a latent concept or topic.
- LSA operates on the principle that words appearing in similar contexts often possess similar meanings.
- Boosts search accuracy by identifying documents semantically related to user queries, even if they lack the exact query terms.
- Supports document classification by categorizing documents into different groups based on their latent semantic features.

# Determining Optimal 'c' Value

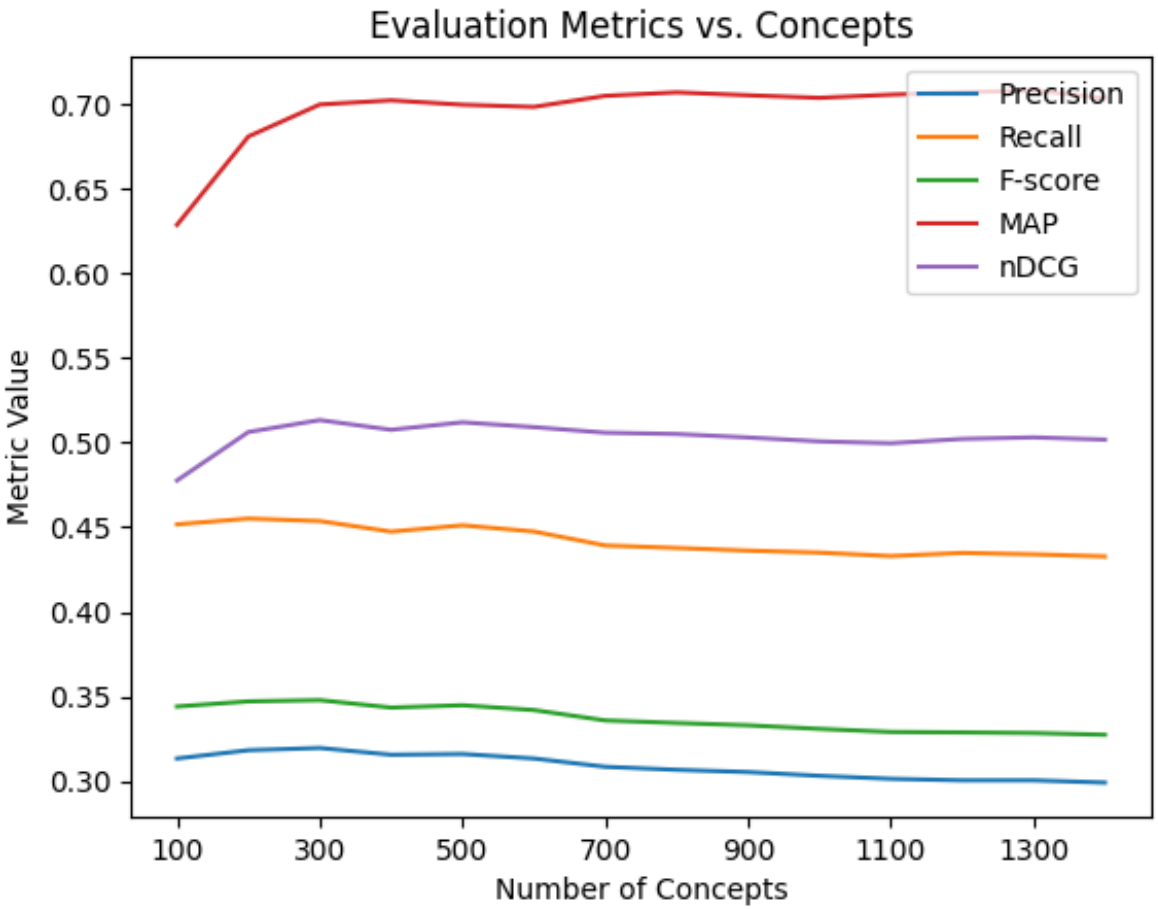


Figure: Performance Metrics of LSA on Cranfield Dataset with Varying Concepts

# LSA with Ngram Indexing

- Three different index structures were created, each utilizing a different type of n-gram: unigrams, bigrams, both unigrams and bigrams.
- LSA was applied to these indexes to extract underlying semantic information and improve retrieval performance.
- The experimental results indicated that the hybrid indexing method using both unigrams and bigrams, combined with LSA, outperformed the other indexing methods.

# Unigram and Bigram indexing with LSA

MAP @ 10: 0.70, nDCG @ 10: 0.51

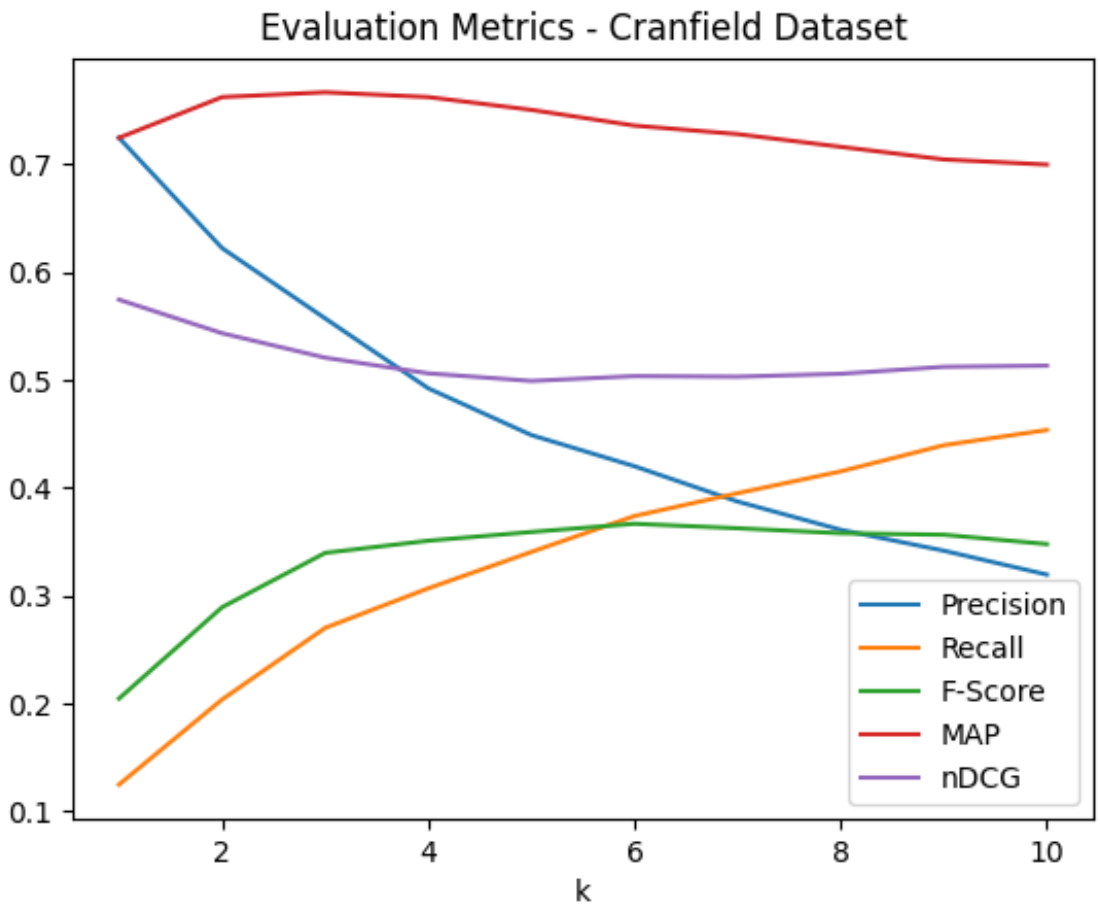


Figure: Unigram and Bigram indexing with LSA

# Only Unigram indexing with LSA

MAP @ 10: 0.69, nDCG @ 10: 0.50

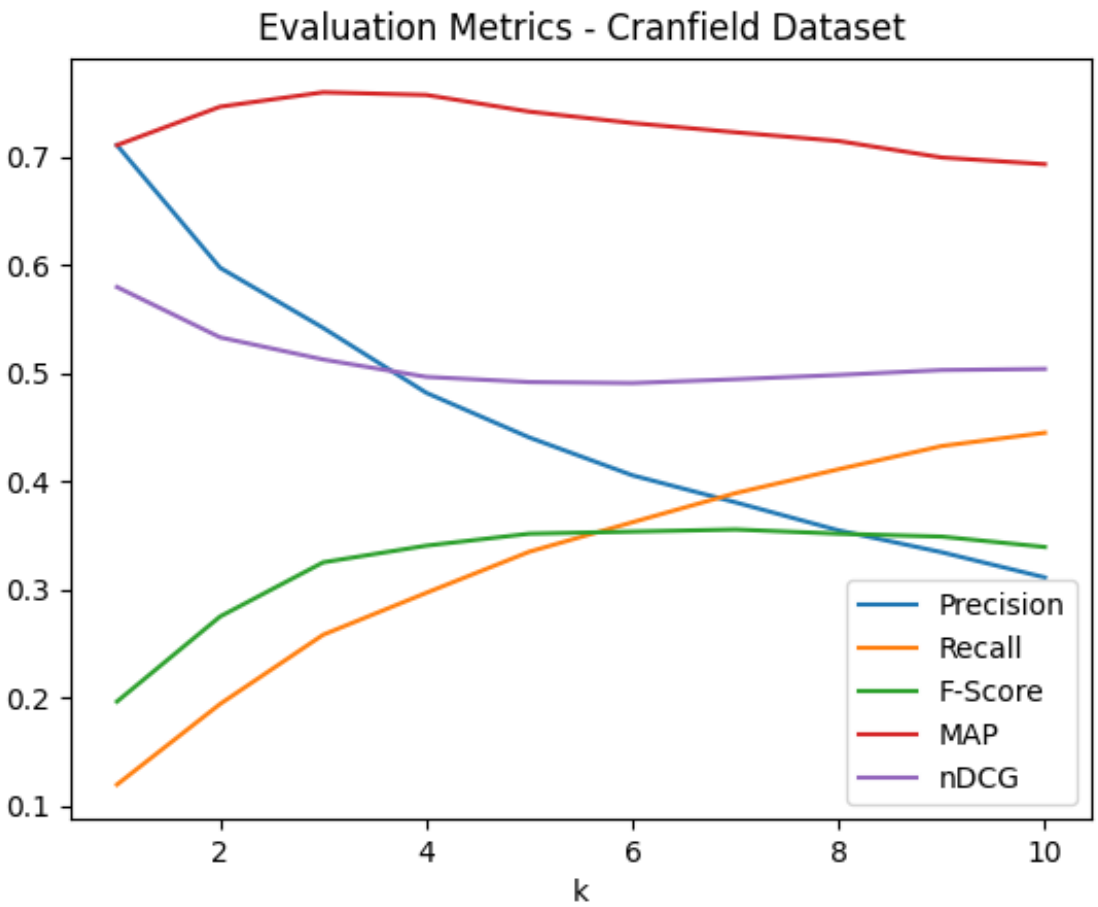


Figure: Only Unigram indexing with LSA

# Only Bigram indexing with LSA

MAP @ 10: 0.59, nDCG @ 10: 0.41

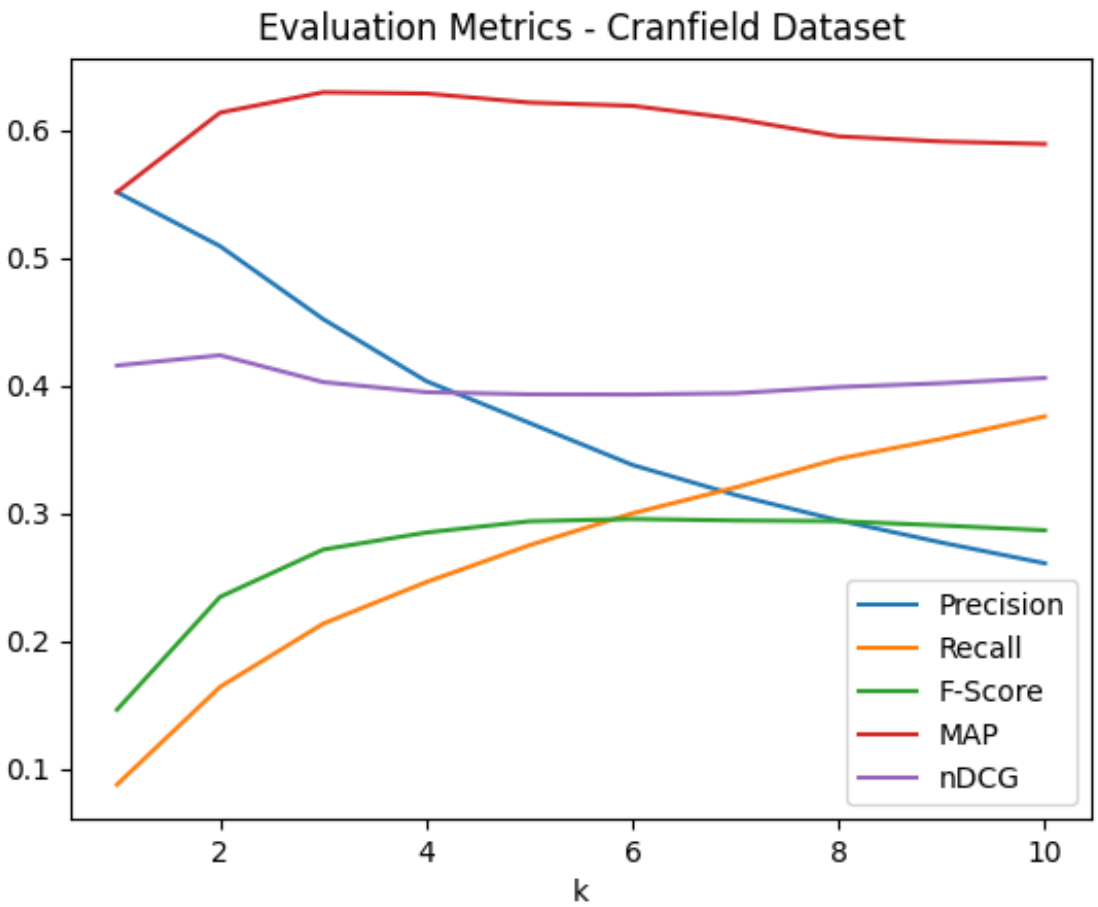


Figure: Only Bigram indexing with LSA

# Query Expansion

- The objective of query expansion is to enhance system performance by supplementing the original user query with additional terms.
- Its aim is to improve both recall and precision by retrieving more relevant documents while excluding irrelevant ones.
- It offers a more comprehensive view of the document collection by including related documents not explicitly mentioned in the original query.
- We utilized sysnet[0], which represents the most frequently used synset, for query expansion.

# Evaluation Metrics for Query Expansion Model

MAP @ 10: 0.68, nDCG @ 10: 0.50

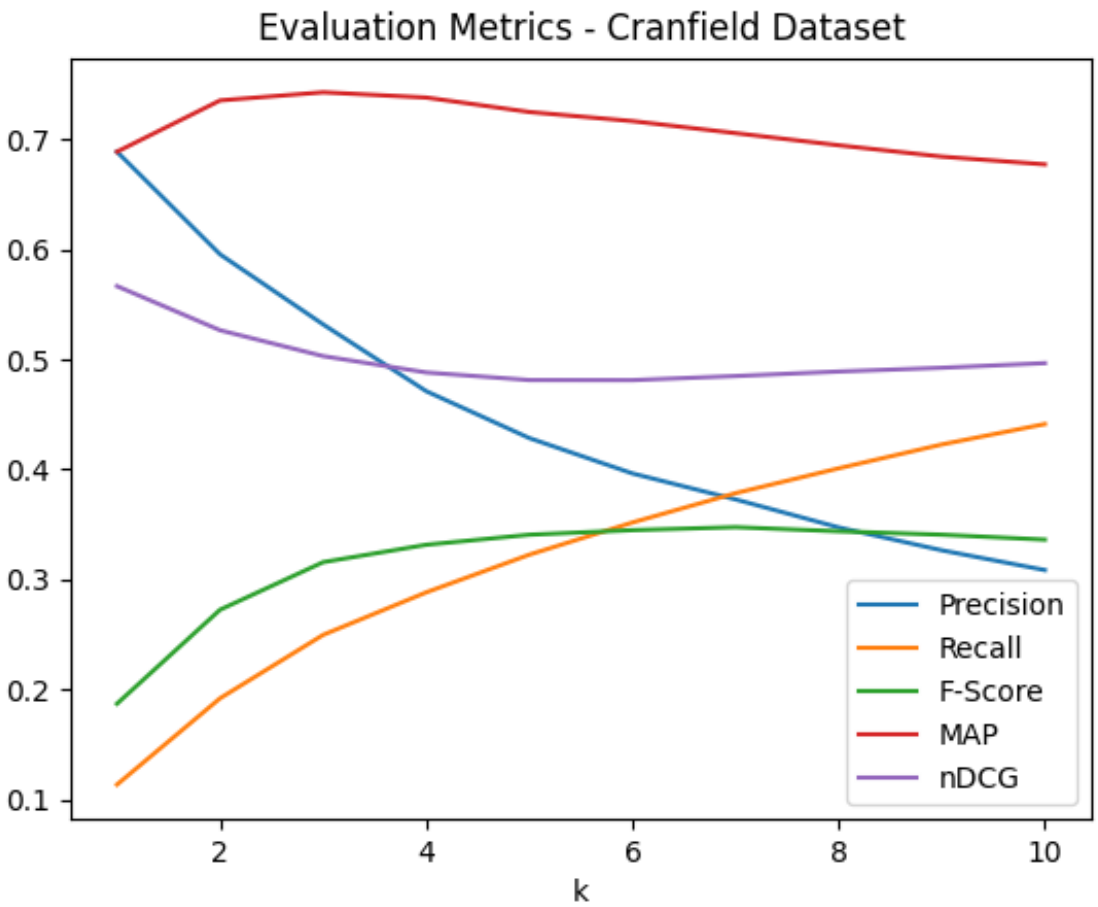


Figure: LSA with Query Expansion

# Spell Check

- Spell check is crucial in information retrieval tasks as it addresses spelling errors and enhances result accuracy.
- Utilizing the Pyspellchecker library, we performed spell check on queries to rectify any spelling errors and refine result precision.
- Although we intended to employ the spell check code from Assignment-1, it proved computationally expensive.
- Calculating the edit distance for each word with every other word in each query is not feasible in information retrieval systems, as it significantly increases processing time.

# Spell Check

- However, the impact on our system's performance before and after spell check was not substantial.
- The original queries may have contained few spelling errors, or any errors present may have been minor enough to not significantly affect system performance.

# Spell Check

MAP @ 10: 0.68, nDCG @ 10: 0.50

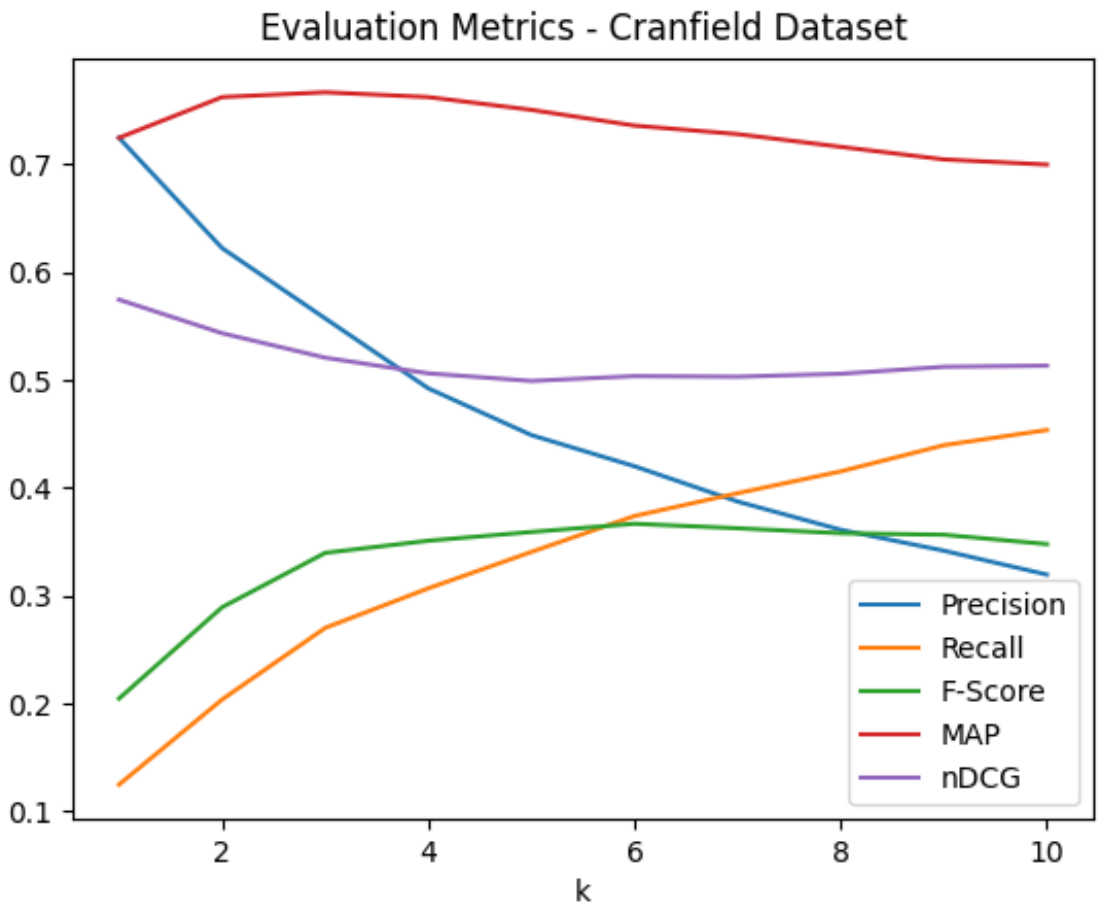


Figure: LSA with Spell Check

# Glove Embeddings

- This process, also known as word vectorization, maps words into a vector space using language models, where each word is represented by a vector of real numbers and of fixed dimension (hyper-parameter)
- Higher cosine similarity implies more related words
- Smaller number of dimensions compared to the vocabulary size reducing the sparsity
- Different types - glove, BERT etc. Created using Neural networks.
- Uses the pre-trained glove embedding model of 50 dimensions on a vocabulary of 6 Billion words.
- Representing the documents and the queries as the average of vectors of each constituent words.
- Still uses the bag of words approach so order of words is not captured.

# Improvements

- Train the Glove word embeddings on the Cranfield dataset
- We can create word embeddings that capture the semantic relationships between words in the documents that are specific to cranfield
- We can capture the word embedding representations for the deep aeronautical technical terms that are not present in the pre-trained model
- Compare sentence vs query instead of document vs query

# Conclusions

- Incorporating both the title, author and body of the documents produced the best outcomes for the Information Retrieval (IR) system.
- Implementing LSA with 300 components yielded optimal results for the IR system.
- Spell check didn't significantly enhance the results and considerably increased processing time compared to the model without spell check.
- Utilizing more concepts in LSA resulted in higher Mean Average Precision (MAP) values, indicating improved performance.
- Employing a hybrid model, which combines unigrams and bigrams in the vector space model, outperformed using only unigrams or bigrams.
- With the adopted strategies, the IR system achieved notable improvements, including a normalized Discounted Cumulative Gain (nDCG) of 0.513 and a maximum MAP of 0.70, marking a significant improvement over the baseline.