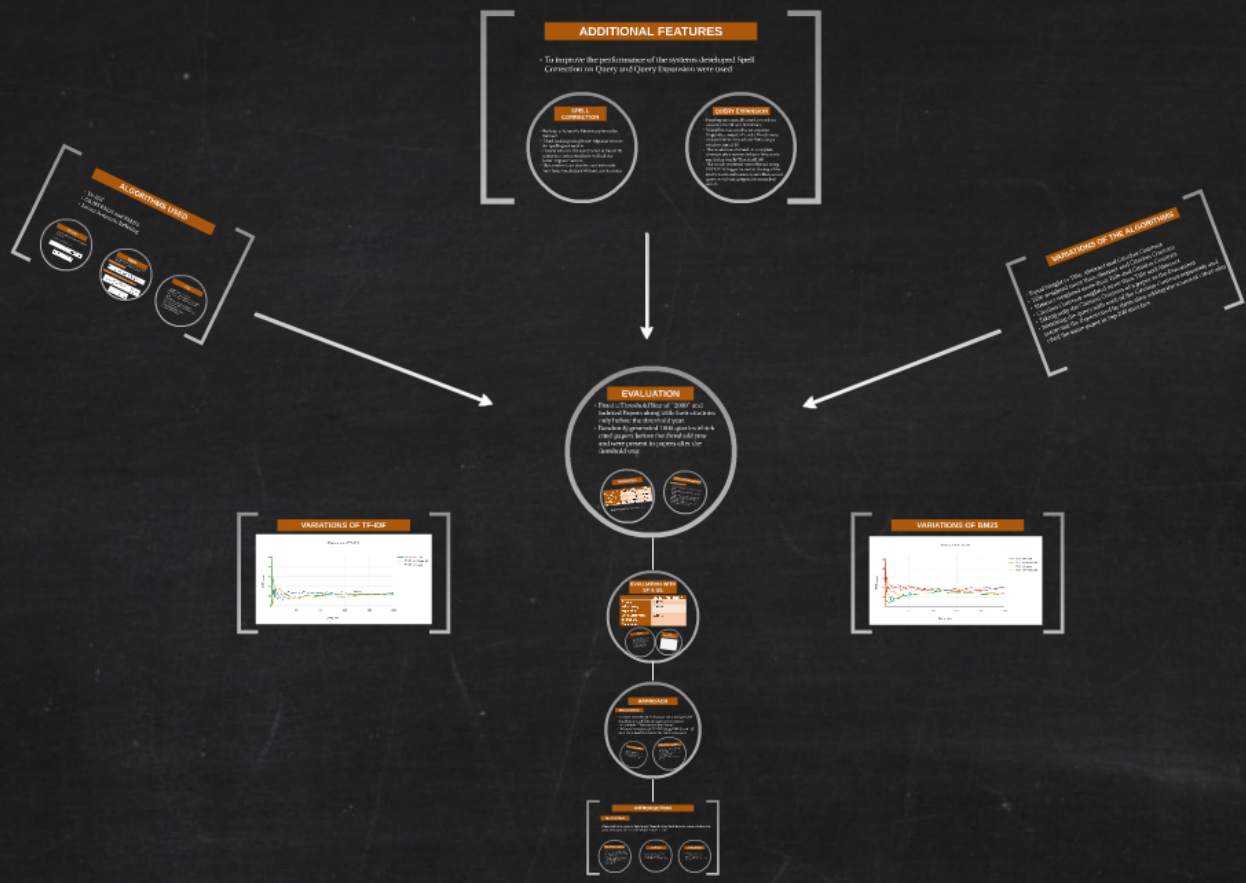


Citation Recommendation Based On Citation Context

Mentor: Mayank Singh

Based on Jim Harvey's speech structures

INTRODUCTION

MOTIVATION

When you write papers, how many times do you want to make some citations at a place but you are not sure which papers to cite?

RELATED WORKS

- He, Qi, et al. "Context-aware citation recommendation." Proceedings of the 19th international conference on World wide web. ACM, 2010.
- He, Qi, et al. "Citation recommendation without author supervision." Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011.

DATASET

- No. Of Scholarly Articles = 1,750,915
- No. Of Citations = 26,043,795
- Average Citations Per Document = 14.87
- Maximum Citations in a Document = 11520

CHALLENGES

- Volume of articles is huge.
- Different Authors may cite the same paper in different ways.
- Hard to get the Information Need from the Query.

RELATED WORKS

- He, Qi, et al. "Context-aware citation recommendation." Proceedings of the 19th international conference on World wide web. ACM, 2010.
- He, Qi, et al. "Citation recommendation without author supervision." Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011.

DATASET

- No. Of Scholarly Articles = 1,750,915
- No. Of Citations = 26,043,795
- Average Citations Per Document = 14.87
- Maximum Citations in a Document = 11520

CHALLENGES

- Volume of articles is huge.
- Different Authors may cite the same paper in different ways.
- Hard to get the Information Need from the Query.

APPROACH

PROCESSING

- Citation Contexts of each paper were merged with the Abstract and Title to capture the essence completely - "Wisdom-of-the-Group"
- Different Variations of TF-IDF, Okapi BM25 and LSI were then used to retrieve the relevant papers.

CORPUS PRUNING

- The given raw data was structured
- Stray Character Removal
- Example: Words having '- ' in between due split of lines were joined

PREPROCESSING

- Tokenized the total dataset using NLTK Punkt Tokenizer
- Removal of Non-Essential Characters , eg.,Punctuation Marks
- Stop Word Removal
- Stemming of the Tokens using Porter Stemmer 2

CORPUS PRUNING

- The given raw data was structured
- Stray Character Removal
- Example: Words having '- ' in between due split of lines were joined

PREPROCESSING

- Tokenized the total dataset using NLTK Punkt Tokenizer
- Removal of Non-Essential Characters , eg.,Punctuation Marks
- Stop Word Removal
- Stemming of the Tokens using Porter Stemmer 2

ALGORITHMS USED

- TF-IDF
- OKAPI BM25 and BM25L
- Latent Semantic Indexing

TF-IDF

- Term Frequency-Inverse Document Frequency
- ltc.lnc scheme is used

$$w_{t,d} = (1+\log(tf_{t,d})) \times \log(N/df_t)$$

$$w_{t,q} = (1+\log(tf_{t,q}))$$

BM25

- Non-Binary Model based on the fact that relevance of term frequency saturates

Standard Okapi BM25

$$RSV_d = \sum_{t \in q} \left[\log \frac{N}{df_t}\right] \cdot \frac{(k_1+1)tf_{td}}{k_1((1-b)+b \times (L_d/L_{ave})) + tf_{td}}$$

BM25 FOR LONG DOCUMENTS (Lv Y, Zhai C.)

$$f'(q,D) = \begin{cases} \frac{(k_1+1)\cdot[c'(q,D)+\delta]}{k_1+[c'(q,D)+\delta]} & \text{if } c'(q,D) > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$c'(q,D) = \frac{c(q,D)}{1-b+b\frac{|D|}{avdl}}$$

LSI

- TF-IDF Matrix is formed
- SVD on the above Matrix is computed
- The value of **k** used for truncation is 2000.
- The Final matrix is normalized
- Two variations of the Truncated SVD have been used :
  1. Randomized Algorithm
  2. ARPACK Algorithm

TF-IDF

- Term Frequency-Inverse Document Frequency
- ltc.lnc scheme is used

$$w_{t,d} = (1 + \log(\mathrm{tf}_{t,d})) \times \log(N / \mathrm{df}_t)$$

$$w_{t,q} = (1 + \log(\mathrm{tf}_{t,q}))$$

BM25

- Non-Binary Model based on the fact that relevance of term frequency saturates

Standard Okapi BM25

$$RSV_d = \sum_{t \in q} \left[\log \frac{N}{\mathrm{df}_t} \right] \cdot \frac{(k_1 + 1)\mathrm{tf}_{td}}{k_1((1 - b) + b \times (L_d / L_{\mathrm{ave}})) + \mathrm{tf}_{td}}$$

BM25 FOR LONG DOCUMENTS (Lv Y, Zhai C.)

$$f'(q, D) = \begin{cases} \frac{(k_1+1)\cdot[c'(q,D)+\delta]}{k_1+[c'(q,D)+\delta]} & \text{if } c'(q, D) > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$c'(q, D) = \frac{c(q, D)}{1 - b + b\frac{|D|}{avdl}}$$

LSI

- TF-IDF Matrix is formed
- SVD on the above Matrix is computed
- The value of **k** used for truncation is 2000.
- The Final matrix is normalized
- Two variations of the Truncated SVD have been used :
  1. Randomized Algorithm
  2. ARPACK Algorithm

VARIATIONS OF THE ALGORITHMS

- Equal Weight to Title, Abstract and Citation Contexts
- Title weighted more than Abstract and Citation Contexts
- Abstract weighted more than Title and Citation Contexts
- Citation Contexts weighted more than Title and Abstract
- Taking only the Citation Contexts of a paper as the Document
- Matching the query with each of the Citation Contexts separately and retrieving the Papers cited by them after adding the scores of citers who cited the same paper in top 200 matches

ADDITIONAL FEATURES

- To improve the performance of the systems developed Spell Correction on Query and Query Expansion were used

SPELL CORRECTION

- Built up a Scientific Dictionary from the Dataset
- Word hashing using letter-trigrams vectors for spelling correction
- Closest word to the query word is found by using the cosine similarity with all the letter-trigram vectors.
- This method can also be used to handle very large vocabulary without any learning

QUERY EXPANSION

- Feeding semantically similar words to improve Recall and Relevance
- Word2Vec was used to reconstruct linguistic context of words. Words were mapped to vectors of size 500 using a window size of 30
- The model was trained on complete datatset after removal of non frequently occurring words(Threshold 10).
- The words retrieved were filtered using NLTK POS Tagger to match the tag of the query word and a score lesser than actual query word was assigned to extended words

SPELL CORRECTION

- Built up a Scientific Dictionary from the Dataset
- Word hashing using letter-trigrams vectors for spelling correction
- Closest word to the query word is found by using the cosine similarity with all the letter-trigram vectors.
- This method can also be used to handle very large vocabulary without any learning

QUERY EXPANSION

- Feeding semantically similar words to improve Recall and Relevance
- Word2Vec was used to reconstruct linguistic context of words. Words were mapped to vectors of size 500 using a window size of 30
- The model was trained on complete datatset after removal of non frequently occurring words(Threshold 10).
- The words retrieved were filtered using NLTK POS Tagger to match the tag of the query word and a score lesser than actual query word was assigned to extended words

EVALUATION

- Fixed a Threshold Year of "2000" and indexed Papers along with their citations only before the threshold year.
- Randomly generated 1000 queries which cited papers before the threshold year and were present in papers after the threshold year

RESULTS(I)

	TF-IDF	BM25	BM25L
Title Priority	0.1308	0.1971	0.1516
Equal Weight	0.1191	0.1407	0.1437
Abstract Priority	0.1156	0.1310	0.1356
Citation Priority	0.1108	0.1272	0.1393
Citation Only	0.0280	0.0228	0.0227
Citation One By One	0.0224	0.0144	0.0229

MRR FOR LSA(Title Priority) = 0.078

EVALUATION METRIC

MEAN RECIPROCAL RANK

- Cited Document is known for each test query
- Reciprocal of rank of Cited Document is noted
- Average taken over 1000 queries is the metric of evaluation of the Retrieval Algorithms = MRR
- Reciprocal of MRR is the expected position for the document

EVALUATION METRIC

MEAN RECIPROCAL RANK

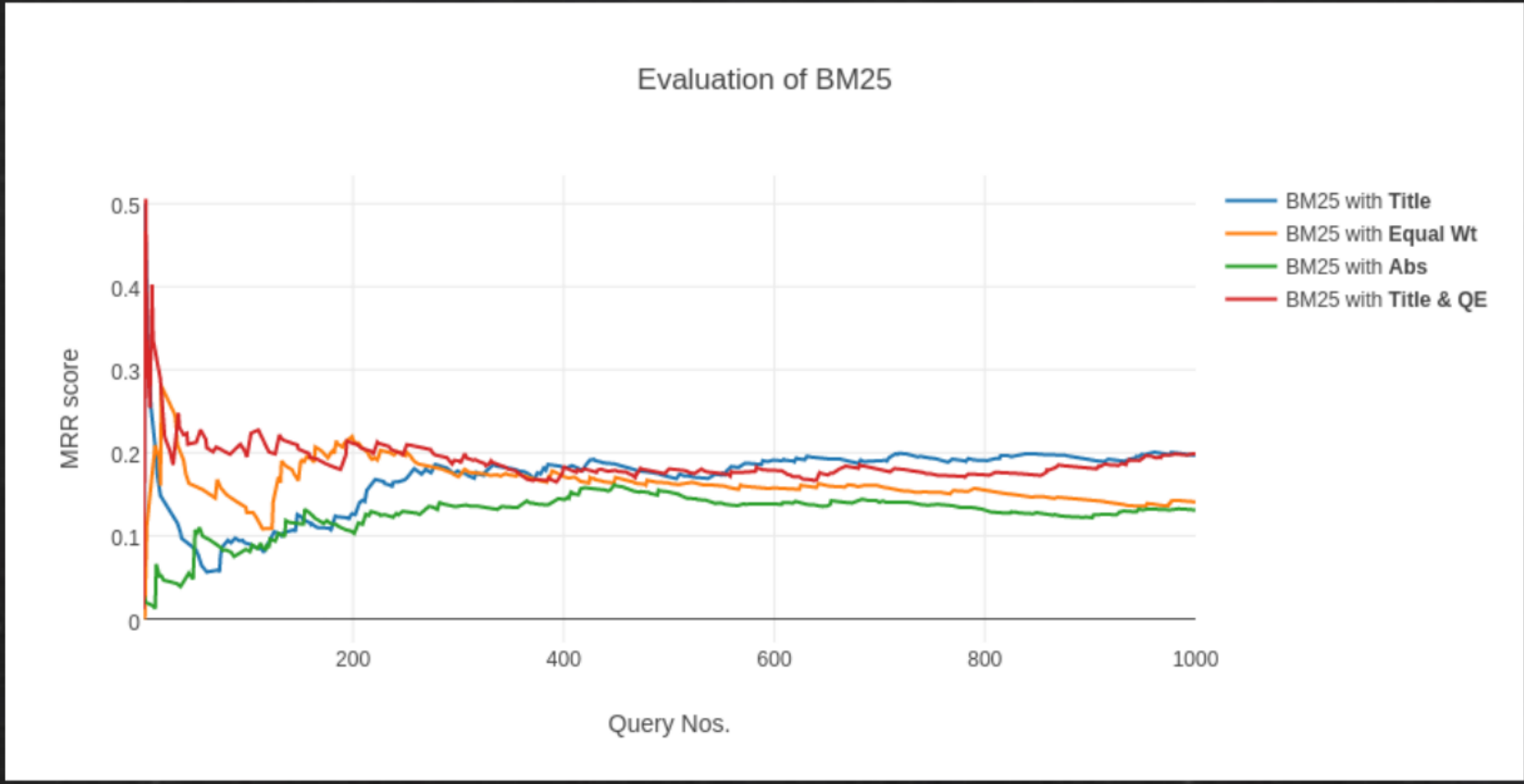
- Cited Document is known for each test query
- Reciprocal of rank of Cited Document is noted
- Average taken over 1000 queries is the metric of evaluation of the Retrieval Algorithms = MRR
- Reciprocal of MRR is the expected position for the document

RESULTS(I)

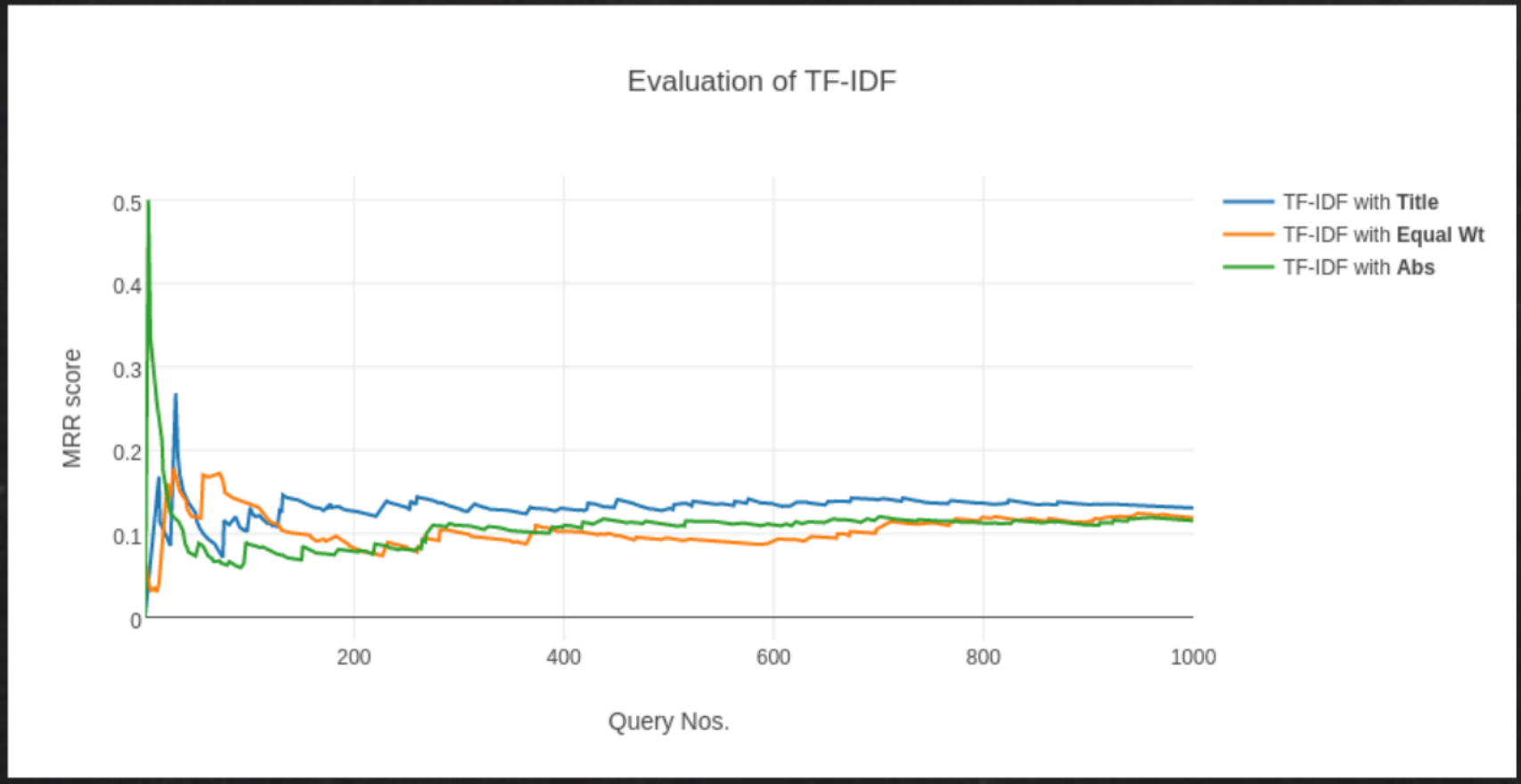
	TF-IDF	BM25	BM25L
Title Priority	0.1308	0.1971	0.1516
Equal Weight	0.1191	0.1407	0.1437
Abstract Priority	0.1156	0.1310	0.1356
Citation Priority	0.1108	0.1272	0.1393
Citation Only	0.0280	0.0228	0.0227
Citation One By One	0.0224	0.0144	0.0229

MRR FOR LSA(Title Priority) = 0.078

VARIATIONS OF BM25



VARIATIONS OF TF-IDF



EVALUATION WITH SP & QE

	BM25 (Title Priority)
Original	0.1971
With Query Expansion	0.1988
Spell Correction and Query Expansion	0.1850

TEAM

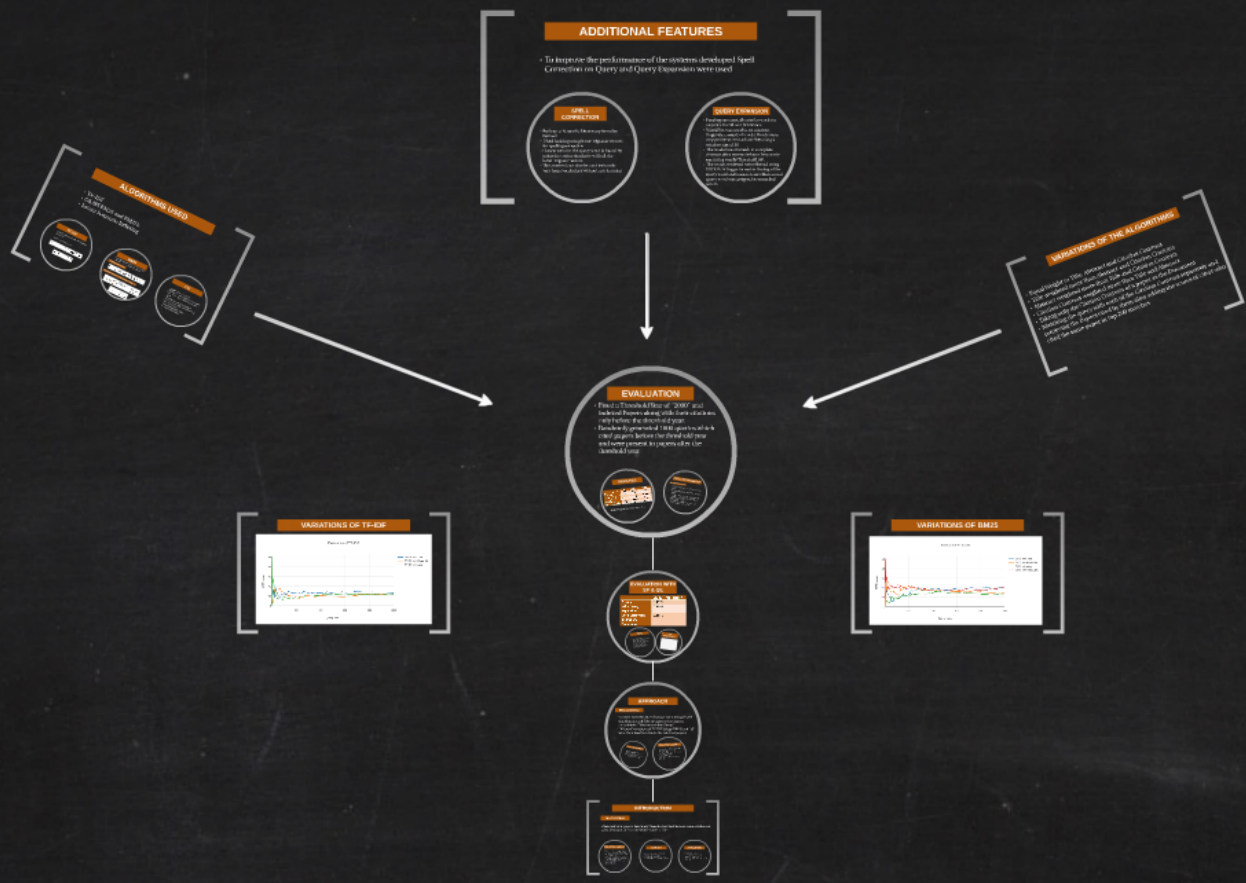
- Aishik Chakraborty (13CS30041)
- Ashish Sharma (13CS30043)
- Chinmaya Pancholi (13CS30010)
- Jatin Arora (13CS10057)
- Jeenu Grover (13CS30042)
- Prabhat Agarwal (13CS10060)
- Sumit Agarwal (13CS10061)

ANY QUESTIONS??

TEAM

- Aishik Chakraborty (13CS30041)
- Ashish Sharma (13CS30043)
- Chinmaya Pancholi (13CS30010)
- Jatin Arora (13CS10057)
- Jeenu Grover (13CS30042)
- Prabhat Agarwal (13CS10060)
- Sumit Agarwal (13CS10061)

ANY QUESTIONS??



Citation Recommendation Based On Citation Context

Mentor: Mayank Singh

Based on Jim Harvey's speech structures