Name - Dhaval Jayantibhai Talaviya
Student ID - s223501691

Question 1

Alternative to Boolean Search:

Propose an alternative to Boolean search and briefly outline its advantages over the traditional Boolean approach.

An effective alternative to Boolean search is the Vector Space Model (VSM).

The Vector Space Model represents documents and queries as vectors in a multidimensional space. Each dimension corresponds to a separate term, and relevance is determined by calculating similarity (often using cosine similarity) between the query vector and document vectors.

Advantages:

Ranked Retrieval:
Unlike Boolean search, which returns either matching or non-matching documents, VSM ranks documents by relevance, presenting users with the most relevant results first.

Partial Matching:
Documents that partially match the query terms still receive a relevance score, reducing the likelihood of zero-result queries.

Term Weighting:
Incorporates term importance by weighting terms based on frequency and rarity, allowing more informative terms to influence relevance scoring significantly.

User-Friendly Queries:
Does not require users to understand complex Boolean logic, making it simpler and more intuitive.

Improved Recall and Precision:
Delivers a balance between precision and recall, offering users more nuanced and contextually relevant results than Boolean logic.

The correct answer is not displayed for Written Response type questions.

Question 2

Inverse Document Frequency Benefit:
Explain the advantage of incorporating inverse document frequency in estimating the weight of terms within a query.

Incorporating Inverse Document Frequency (IDF) when estimating the weight of terms within a query offers the following advantage:

Emphasizes Informative Terms:
IDF assigns higher weights to terms that occur rarely across the

document collection and lower weights to terms that are common. The reasoning is that rare terms are typically more informative and indicative of specific content, thus they better distinguish relevant from non-relevant documents.

Improved Ranking Accuracy:
By using IDF, the system differentiates between frequent, less informative terms (e.g., common words or stop words) and rare, highly informative terms. Thus, documents containing rare query terms like "arachnocentric" are more likely to be relevant to that query and receive higher ranking.

Avoids Dominance of Common Terms:
IDF helps prevent very common terms from overly influencing retrieval results, ensuring that the relevance score reflects the specificity and uniqueness of query terms.

In short, the key advantage of incorporating IDF is its ability to highlight and leverage the discriminative power of rare terms, significantly improving the retrieval performance of information retrieval systems.

The correct answer is not displayed for Written Response type questions.

Question 3

Euclidean Distance Weakness:
Identify and discuss one weakness associated with using the

Euclidean distance metric for ranking documents in information retrieval.

One significant weakness associated with using Euclidean distance for ranking documents in information retrieval is sensitivity to document length.

Euclidean distance measures the straight-line distance between document vectors. When used for ranking documents, it tends to favor shorter documents because shorter vectors inherently produce smaller distances compared to longer vectors. Even if a longer document contains equally relevant content, it may be penalized due to its length, potentially misrepresenting relevance.

This sensitivity can result in inaccurate rankings, with shorter documents unfairly ranked higher than longer, potentially more comprehensive and relevant documents.

In comparison, cosine similarity addresses this weakness by considering only the angle between document vectors, not their magnitude or length. Thus, cosine similarity is usually preferred over Euclidean distance in Information Retrieval systems.

In summary, the key weakness of Euclidean distance is its sensitivity to document length, making it less suitable than cosine similarity for effectively ranking documents in information retrieval scenarios.

The correct answer is not displayed for Written Response type questions.

Question 4

Boolean Query Optimization:

For the boolean query "machine AND learning," considering "machine" and "learning" each having N postings, determine the minimum number of comparisons required to construct the final posting list of documents containing both terms. Analyze this for (a) sorted and (b) unsorted postings. For the sorted case, categorize the complexity as {constant, linear, quadratic, cubic, factorial, exponential} in N and provide an explanation. In the unsorted scenario, select one of the complexities and justify your choice while adhering to the constraint of not sorting the postings.

1) Sorted Postings:

When postings are sorted, the intersection ("AND") can be efficiently computed by a linear merge algorithm. This involves iterating simultaneously through both postings lists and comparing document IDs pairwise from the front of each list.

Complexity:

Number of comparisons: At most $2N-1$ comparisons (approximately linear in N).

Complexity category: Linear complexity in N.

Each step advances the pointers in one or both postings lists, requiring at most one comparison per advancement. Hence, the complexity scales directly with the number of documents (N), which is linear complexity.

2) Unsorted Postings:

Without sorted postings and under the constraint of not sorting the postings, each entry in the first postings list must potentially be compared to every entry in the second postings list. This approach resembles a brute-force method.

Complexity:

Number of comparisons: Up to N×N=N^2 comparisons.

Complexity category: Quadratic complexity in N.

For each of the N documents in the first postings list, the worst-case scenario requires comparison with every document in the second postings list, resulting in N^2 comparisons.

The correct answer is not displayed for Written Response type questions.

Question 5

Limitations of Accuracy Metric:

Elaborate on why accuracy is considered a suboptimal metric for estimating the performance of a search engine.

In typical IR scenarios, the number of non-relevant documents vastly exceeds the number of relevant documents. Hence, if a search engine simply classifies every document as non-relevant, it will still achieve a very high accuracy due to this imbalance.

For instance, suppose we have 1,000 documents with only 10 relevant documents. Predicting all documents as non-relevant gives 990 correct predictions out of 1,000 (99% accuracy). Although seemingly high, this is a meaningless result because the search engine has failed to identify any useful document for the user's query.

Accuracy does not reflect the actual user experience or search quality, especially in retrieval tasks characterized by significant class imbalance. Thus, precision, recall, and F-measure are considered superior metrics, as they better reflect the ability of a system to retrieve useful, relevant information for users.

The correct answer is not displayed for Written Response type questions.

Question 6

Boolean Search Analysis:

Identify and discuss two weaknesses inherent in the Boolean search methodology.

Two inherent weaknesses in the Boolean search methodology are:

1. "Feast or Famine" Problem:

Boolean queries often result in either too few (sometimes zero) or excessively many results. Users may get overwhelmed by thousands of results using OR operators or receive no useful results with overly restrictive AND operators.

Users must precisely formulate queries, making it difficult to achieve a balanced result set without considerable expertise and effort.

2. Lack of Ranking and Partial Matching:

Boolean search treats documents as strictly relevant or irrelevant (exact match), without ranking documents according to their relevance. It doesn't support partial matches or differentiate relevance based on how well documents match the query.

Users receive no indication of which document is the most relevant among those retrieved, forcing them to manually inspect potentially many irrelevant documents.

The correct answer is not displayed for Written Response type questions.

Question 7

Jaccard Coefficient Limitations:
Highlight and elaborate on two weaknesses associated with using the Jaccard coefficient for ranking documents.

Two weaknesses associated with using the Jaccard coefficient for ranking documents are:

1. Ignores Term Frequency:
The Jaccard coefficient treats each term as a binary presence or absence (0 or 1) and does not consider how frequently a term appears in a document. Thus, it overlooks the crucial factor of term frequency, failing to differentiate between documents where a relevant term appears once versus multiple times.

This reduces ranking accuracy, as documents containing more occurrences of key terms (likely more relevant) receive no special advantage compared to those containing fewer occurrences.

2. Lack of Term Importance (IDF):
Jaccard also neglects the fact that some terms are inherently more important or discriminative than others. Terms rare in a collection typically offer greater discriminative power, but Jaccard coefficient weights all terms equally.

Consequently, documents with rare, highly informative terms aren't distinguished from documents that primarily contain common, less informative terms, diminishing the retrieval system's effectiveness.

The correct answer is not displayed for Written Response type questions.

Question 8

Logarithmic Smoothing with Term Frequency:
Outline the benefits of employing logarithmic smoothing in conjunction with term frequency within the context of search engine ranking.

1. Diminishing Returns on High Term Frequency:
Logarithmic smoothing ensures that relevance does not increase linearly with the frequency of terms. Instead, it assigns diminishing incremental weights to additional occurrences of the same term.

This avoids overly favoring documents merely because a term is repeated extensively, resulting in more balanced and accurate ranking.

2. Reduced Bias Toward Longer Documents:
Without logarithmic smoothing, long documents could artificially appear more relevant simply by frequently repeating certain terms. Logarithmic smoothing dampens this effect.

Document length alone does not disproportionately influence ranking, allowing both short and long documents an equal opportunity based on the true relevance of their content.

The correct answer is not displayed for Written Response type questions.

Question 9

Precision at k vs. Average Precision:
Clarify the fundamental difference between Precision at k and Average Precision (AP or AveP) in the context of evaluating search engine results.

1. Precision at k (P@k):
Precision at k measures the fraction of relevant documents retrieved among the top k results only. It evaluates the relevance of results strictly within a limited, fixed scope.

It Does not consider the position of relevant documents within the top-k results (all top-k are treated equally). It Offers a snapshot of precision performance at a predefined cutoff point and Ideal for evaluating user experience when users only look at a small number of top results.

2. Average Precision (AP):
AP calculates precision values at every point a relevant document is retrieved, averaging these values across all relevant documents returned by the query.

It Considers both the position and distribution of relevant documents within the entire retrieved list. It Penalizes systems that place relevant documents lower in the ranked list and Provides a comprehensive single-value measure capturing both precision and ranking effectiveness.

Fundamental Difference:
Precision at k focuses only on the top-k results without considering the ranking order beyond the cutoff.
Average Precision evaluates the entire ranked list, factoring in the ranking order and positions of relevant documents, thus giving a more nuanced measure of retrieval performance.

The correct answer is not displayed for Written Response type questions.

Question 10

Search Engine Evaluation:
You've developed a search engine (A) for the highly popular website xyz.com, aiming to replace the existing system (B). Both engines display top-k results similar to Google. Faced with the question of which engine produces superior rankings, xyz.com suggests conducting an A/B test due to resource constraints. Describe the metric or quantity you would measure to compare the performance of A and B. Offer a detailed explanation of why the chosen measurement aligns with xyz.com's goal in the absence of curated test corpora and human judgments (qrels).

An appropriate metric for comparing engines A and B without curated test corpora or explicit human judgments (qrels) is Click-Through Rate (CTR).

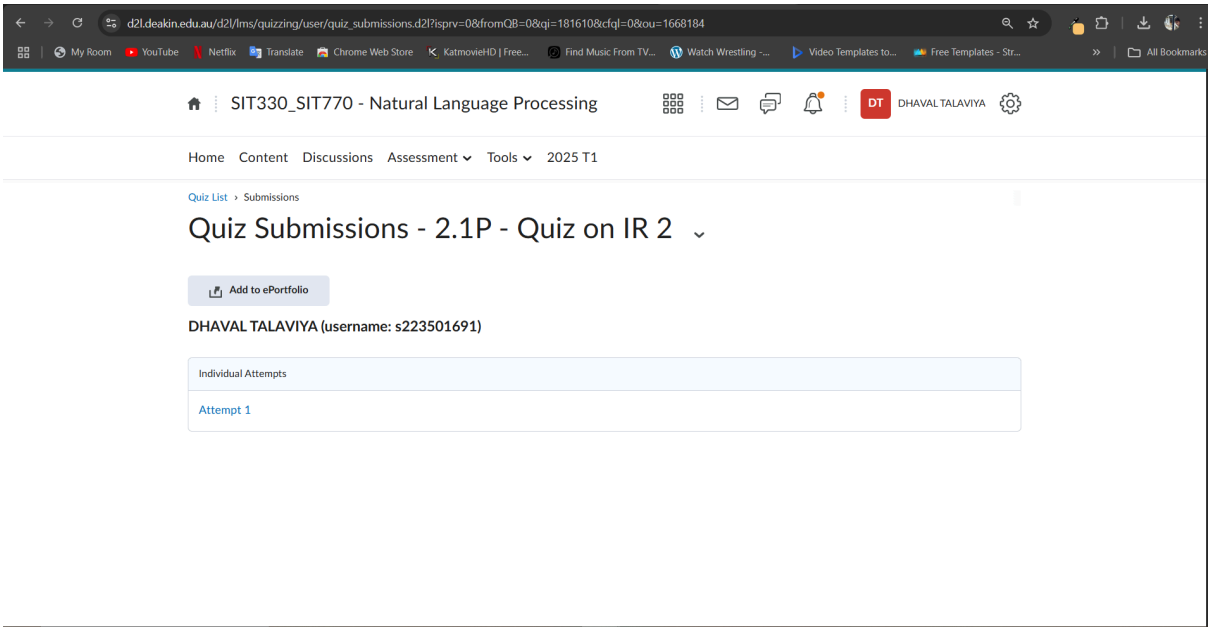
Reasoning:

Explicit relevance judgments (precision, recall) require curated data (qrels), which is unavailable.

CTR leverages real user interactions implicitly indicating relevance.

Higher CTR means users are more satisfied with the results, aligning directly with the website's goal of improving user experience.

It efficiently measures user happiness in practical, real-world scenarios, without extra resources or manual relevance annotations.


d2l.deakin.edu.au/d2l/lms/quizzing/user/quiz_submissions.d2l?isprv=0&fromQB=0&qi=181610&cfql=0&ou=1668184
My Room
YouTube
Netflix
Translate
Chrome Web Store
KatmovieHD | Free...
Find Music From TV...
Watch Wrestling -...
Video Templates to...
Free Templates - Str...
All Bookmarks
SIT330_SIT770 - Natural Language Processing
DT
DHAVAL TALAVIYA
Home
Content
Discussions
Assessment
Tools
2025 T1
Quiz List › Submissions
Quiz Submissions - 2.1P - Quiz on IR 2
Add to ePortfolio
DHAVAL TALAVIYA (username: s223501691)
Individual Attempts
Attempt 1