**CS F469 Information Retrieval**

Assignment - 1: Boolean Retrieval System

Design Document

**Introduction**

Boolean retrieval models are one of the most simplistic information retrieval models which evaluate a bracketed boolean query consisting of operators like AND, OR and NOT to return a set of documents.

Boolean retrieval models can be made tolerant by allowing for wildcard (\*) queries and spelling correction.

**Preprocessing**

The preprocessing step includes stop word removal, followed by tokenization and stemming operations. This is followed by creation of an inverted index of the stemmed terms which has the count, rotations and postings for every word that stems from a stemmed term. An additional two-gram index is also created with all the stems of words which contain a two-gram.

Data structures: Inverted index and two-gram index, both of which are Python dictionaries

Average running time: *O(N2)* where N is the number index terms

**Query Parsing**

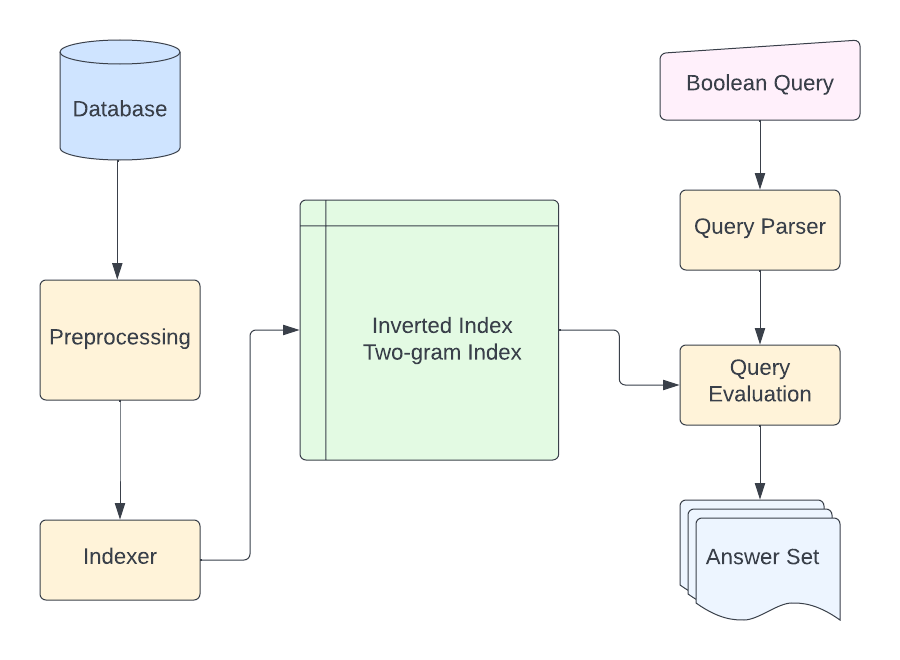
The boolean query is parsed using a stack that evaluates bracketed expressions.

Data structures: Stack

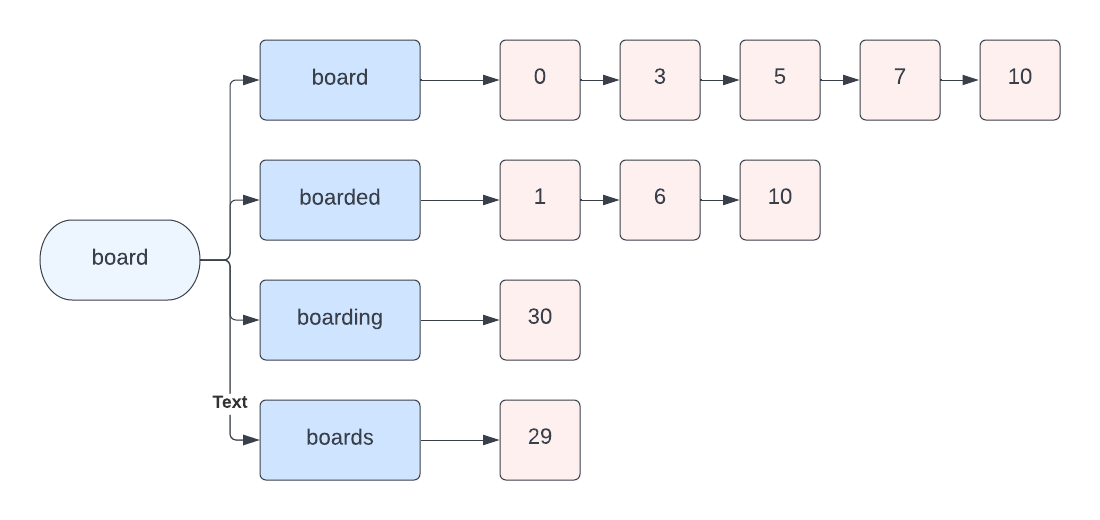
**Query Evaluation**

For wildcard queries, the query term is rotated to form a trailing wildcard and the rotated term is searched for in the rotations lists and the corresponding posting lists are added to the result. Query words are identified as misspelled if they do not occur in the dictionary. The two-gram index is used to find the closest word(s) to the misspelled word(s), followed by calculation of Levenshtein distance to find accurate matches. All words in the dataset are assumed to be correctly spelled.

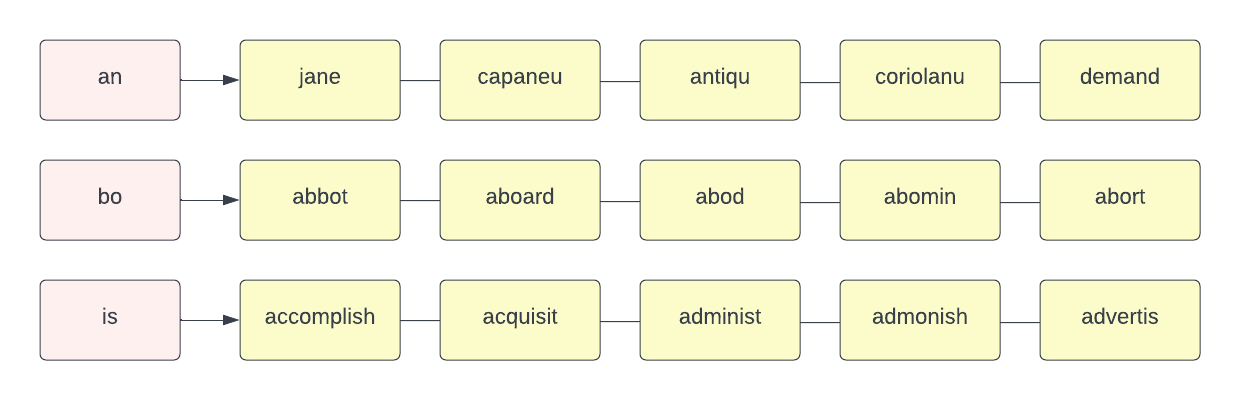
Average running time: *O(L2)* where L is the number of characters in the query term.



*Fig. System architecture*

**

*Fig. Inverted index*

**

*Fig. Two-gram index*