Mid-term Review Guide

Covering topics Week 1 through Week 5:

- Please review lectures and assignments.
- The focus of the exam will be on theory and concepts,
- But you can expect some related questions on Elastic Search.

Main topics include:

- Overview of Information Retrieval (IR)
    1. IR definitions and focus
    2. What does it do and involve?
- Basic concepts and boolean retrieval
    1. Term-document matrix
    2. Inverted index
    3. Dictionary, and postings
    4. Boolean query processing and optimization
    5. Boolean query on Elastic Search
- Bag of words approach
    1. Text tokenization
    2. Stopword removal
    3. Normalization and related techniques
        - Casefolding
        - Stemming, etc.

    4. Phrase queries and positional index
    5. Elastic Search related techniques:
        - Field data types: text vs. keywords
        - Mappings, analyzers
- Term weighting and scoring (similar measures):
    1. Binary weights 0/1
    2. Term frequencies
    3. Document frequecies
    4. TF*IDF term weighting
    5. Similarity based on sets
- Vector Space Model
    1. Vector: magnitude and direction
    2. Query and documents as vectors
    3. Cosine similarity
- Probabilistic Information Retrieval
    1. Probability basics, and how to combine them
    2. Probability ranking principle
    3. Probabilities and notations in probabilistic IR:
        - Notations and meanings, e.g. P(X|R), P(R|X), P(X|NR), P(NR|X)
        - P(X|R) as a function of P(x_i|R) in the binary independence

model

4. TF*IDF and BM25 term weight/scoring
5. Normalization of TF in BM25:
   - Saturation function with pivot `k`
   - Document length normalization with `b`