CS/INFS780 Midterm Topics:

1. Data mining concepts
    a. Classification
    b. Clustering
    c. Association Rules
2. Multidimensional Indexing
    a. Common indexing methods for RDBMS
        i. B+ Tree
    b. Indexing for multidimensional data
        i. Grid File
        ii. Kd-tree
        iii. Space-filling curves (z-ordering, Hilbert)
        iv. R-tree
    c. For all indexing techniques, know how
        i. data are stored in the index
        ii. different types of queries are processed
            1. point queries
            2. range queries
            3. k-nn queries
            4. spatial joins
    d. For R-tree, know how different data operation works
        i. Insertion
        ii. Deletion
3. GEMINI Framework
    a. How multimedia data can be indexed using multidimensional indexes
    b. Curse of dimensionality
    c. Dimensionality reduction
    d. Lower-bounding property
4. Text Mining
    a. Indexing
        i. Inverted index
        ii. Vector space model
        iii. LSI/SVD
            1. What does it do?
            2. What does each decomposed matrix mean?
            3. How is dimensionality reduction achieved?
            4. Advantages/disadvantages
            5. How to compute similarity between:
                a. Document and document (including query and document)
                b. Document and term
                c. Term and term
    b. Random projection
        i. How does it work?
        ii. Compare to LSI (advantages/disadvantages
    c. Other representations

        i. Bigram Proximity Matrix (BPM)
5. Web Mining
    a. PageRank algorithm
    b. HITS algorithm
    c. Web spam