• Chapter 10. Social Search
  – Search within a social environment, Web 2.0
  – User tags: social media tagging, types of user tags (content, context, attribute, subjective, organizational), tag expansion, inferring missing tags (TF/IDF, Classification, MMR)
  – Searching with communities: online community, finding communities, clustering, community-based question-answering (QA) systems, HITS
  – Document filtering: static vs. adaptive filtering, user profiles

• Chapter 9. Classification and Clustering
  – Classification (also called categorization): a supervised learning approach
    * Naive Bayes Classifier (Naive Bayes’ rule): estimating $P(c)$, $P(d)$, $P(d|c)$
    * Multiple Bernoulli model: binary vector document representation, Laplacian smoothed estimate
    * Multinomial model: term-frequency document vector representation, Laplacian smoothing
    * Support Vector Machines (SVM): a vector-space-based ML classifier, hyperplanes, support vectors, maximum margin, (non-)linear separable case, kernel functions
  – Evaluation: accuracy, precision, recall, F-measure
  – Feature selection (entropy/information gain), sentiment classifier
  – Clustering: an unsupervised learning approach
    * Hierarchical clustering: Divisive (top-down) vs. Agglomerative (bottom-up), clustering costs (Single/Complete/Average/Average group linkage)
    * K-Means clustering: $K$, centroids
    * K-Nearest Neighbor clustering: $K$, overlapping clusters

• Chapter 8. Evaluating Search Engines
  – Evaluating the effectiveness, efficiency, and cost of performing searches using search engines
  – Effectiveness measures: precision, recall, F-measure/Harmonic mean, false positive/negative
  – Ranking effectiveness: Mean Average Precision (MAP), (interpolated) Recall-precision graph, Precision at R (P@R), Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (nDCG)

• Chapter 7. Retrieval Models
  – Topic relevance vs. user relevance, binary vs. multi-valued relevance
  – Boolean Model: exact-matching retrieval, no ranking, pros and cons
  – VSM: document/query vectors with term weights, Cosine similarity measure, TF/IDF
  – Probability Model: probability of relevance, Bayes Classifier, binary (term) independence model ($p_i$ vs. $s_i$), BM25 ranking algorithm
  – Fuzzy Set Model: degree of membership, term-term correlation matrix, normalized correlation factor, the $\mu$ function, $q_{dnf}$
• Chapter 6. Queries and Interfaces
  – Keyword queries: simple natural language, stemming, word variants
  – Spell checking: similarity measure of words in dictionary, edit distance, Soundex Code
  – Noisy Channel Model: estimate probability of correction
  – Snippet Generation, Sentence Significant Factor

• Chapter 4. Processing Text
  – Converting documents to index terms
    * tokening: hyphenated vs. non-hyphenated, capitalized vs. lower-cased words, special characters, numbers, punctuation
    * stopword removal: function words (articles, prepositions, etc.)
    * stemming: reduce morphological variations of words to their common stem - inflectional (plurals, tenses), derivational (verbs, nouns, etc.), Porter stemmer
    * link analysis: link/anchor text, PageRank algorithm
  – Distribution of word frequencies: Zipf’s law and Heaps’ law
  – Result set size estimation: word co-occurrence

• Chapter 2. Architecture of a Search Engine
  – Requirements: effectiveness (quality) and efficiency (throughput)
  – Indexing process
    * text acquisition: crawler, data stream feeds, document (format) conversion
    * text transformation: parser (tokenizer), information extraction, classifier
    * index creation: word counts/positions, weighting (TF/IDF), inversion
  – Query process
    * user interaction: Boolean/phrase/keyword queries, query expansion/suggestion/spell checking, relevance feedback
    * ranking: document scores
    * query evaluation: ranking/performance analysis

• Chapter 1. Search Engines and Information Retrieval
  – Goals and challenges of information retrieval (IR)
    * Representing user’s information needs
    * Query refinement
    * User interface
    * Result representation
  – Unstructured documents versus database records
  – IR tasks: ad-hoc search, filtering, classification, QA
  – IR systems versus search engines
  – Evaluations of IR systems: relevance

Breakdown on Chapters
1. Chapters not covered in midterm exams: Chapters 9 and 10 (∼70%)
2. Midterm Exam 2: Chapters 7 and 8 (∼15%)
3. Midterm Exam 1: Chapters 1, 2, 4, and 6 (∼15%)