CSC 483/583: FINAL REVIEW OUTLINE

MIHAI SURDEANU

Before we begin:

- The final exam will be Wednesday 5/10/2017 10:30am – 12:30pm, in this room (Gould Simpson 906).
- Similar to the midterm, you can bring a 2-page sheet with your notes.
- You are allowed a simple, self-contained hand-calculator. Internet-connected devices are **not allowed** under any circumstances.

Topics to know for the final:

1. Lecture 8: Evaluation
   A. Result summaries, static or dynamic
   B. Criteria for good dynamic summaries
2. Lecture 9: Relevance feedback & query expansion
   A. Centroid
   B. Rocchio algorithm – theoretical version and the SMART implementation
   C. Query expansion using global resources
   D. Query expansion at search engines
3. Lecture 11: Probabilistic information retrieval
   A. Basic probability theory:
      - What is probability?
      - What is conditional probability?
      - Chain rule, partition rule, Bayes' rule, law of total probability (partition rule), odds
      - Ability to construct and interpret contingency tables, and probability trees

Date: Spring 2017.

- Independent events; detect if two events are truly independent from data

B. Probability ranking principle

C. Binary independence model: how to derive the ranking function for terms; the formula for c_t, with smoothing; BIM after simplifying assumptions

D. Okapi BM25 – formula, what the weights mean

4. Lecture 12: Language models for IR
   A. How to compute $P(q|d)$, smoothing
   B. n-gram language models (see lecture discussion)
5. Lecture 13: Text classification and naive Bayes
   A. Why is text classification useful for IR
   B. How to compute $P(c|d)$, smoothing – in the multinomial naive Bayes
   C. Bernoulli naive Bayes (grad students only)
   D. Positional-dependent NB (see HW4) (grad students only)
   E. n-gram NB (see lecture discussion)
   F. The three ways of messing up the implementation of naive Bayes
   G. The two independence assumptions in naive Bayes
   H. Evaluating classification
   I. Feature selection: frequency thresholding, mutual information, *not* Chi-square (grad students only)
6. Lecture 14: Vector space classification
   A. Rocchio: algorithm, limitations
   B. kNN: implementation, probabilistic kNN
7. Lecture 15: Brief introduction to support vector machines (SVM)
   A. Intuition behind SVM: where to place the separating hyperplane. Why?
   B. How to manually set the separating hyperplane
   C. How to implement a n-class classifier using binary classifiers
   D. Intuition behind active learning
8. Lecture 16: Flat clustering
   A. Classification vs. clustering

   B. Applications of clustering in IR
   C. K-means: algorithm, RSS, convergence proof, time complexity, how to initialize, K-means++
   D. Clustering evaluation: purity, Rand index, F measure
   E. How to choose number of clusters
9. Lecture 20: Link analysis
   A. The simple crawler and its limitations
   B. What a crawler must do
   C. robots.txt
   D. A real crawler: distributed architecture; managing priorities with the two queues (Mercator algorithm)
   E. Duplicate detection: shingles, Jaccard, not sketches
10. Lecture 21: Link analysis
   A. Anchor text: what it is, how to index, how to search
   B. Google bombs
   C. PageRank: random walk problem, how to construct the probability matrix, teleportation probability, how to compute the steady state vector using the power method, issues