CSC 483/583: FINAL REVIEW OUTLINE

MIHAI SURDEANU

Before we begin:

- The final exam will be Monday 12/14 from 6PM to 8PM, in this room (Gould Simpson 906).
- The exam will be open book. You are welcome to bring any textbooks, notes, etc.
- You are allowed (and recommended!) a simple, self-contained hand-calculator. You can also bring a tablet or laptop with PDFs of the slides, textbook, and your notes, as long as it is **not** connected to the Internet. Internet-connected devices are **not allowed** under any circumstances.

Topics to know for the final:

1. Lecture 6: Vector space model
   A. Feast of famine for Boolean queries
   B. Jaccard coefficient: where else is this useful? Limitations
   C. tf-idf
   D. Vector space model
   E. Cosine similarity
   F. Different ways of encoding: term frequency, document frequency, normalization
2. Lecture 7: Complete search system
   1. Exact top K retrieval using min heap
   2. Inexact top K retrieval: document at a time, term at a time, cluster pruning
3. Lecture 8: Evaluation
   A. Unranked evaluation: Precision, Recall, F score
   B. Accuracy. Why is Accuracy not a good measure?

Date: Fall 2015.

   C. Ranked evaluation: P@1, precision-recall curve, mean average precision (MAP), mean reciprocal rank (MRR)
   D. Inter-annotator agreement: Kappa measure
   E. Real-world evaluations: A/B testing
   F. Result summaries, static or dynamic
   G. Criteria for good dynamic summaries

4. 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

5. 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
      - 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

6. Lecture 12: Language models for IR
   A. How to compute $P(q|d)$, smoothing
   B. n-gram language models (see lecture discussion)

7. Lecture 13: Text classification and naive Bayes
   A. Why is text classification useful for IR

B. How to compute $P(c|d)$, smoothing

C. Multinomial vs. Bernoulli naive Bayes

D. Positional-dependent NB (see HW4)

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

8. Lecture 14: Vector space classification

   A. Rocchio: algorithm, limitations

   B. kNN: implementation, probabilistic kNN

9. 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

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

    D. HITS: *not required*