

CS780 Final Topics:

Topics pre-midterm:

1. Introduction
 - a. Distance metric - requirements
2. Time Series
 - a. Representations
 - i. Lower-bounding
 - ii. DFT, Wavelet, PAA, SAX, iSAX, APCA, PLA, SVD
 - b. Distance Measures
 - i. Shape-based
 1. Euclidean Distance
 2. Dynamic Time Warping
 3. LCSS
 - ii. Structural similarity
 1. Compression-based
 2. histogram-based (bag-of-patterns)
 3. model-based
 - c. Clustering
 - i. Whole clustering vs. subsequence clustering
 - d. Motif discovery
 - i. Random projection
 - ii. VizTree
 - iii. Grammar-based motif discovery
3. Text Mining
 - a. Indexing
 - i. Inverted index
 - ii. Vector space model
 - iii. LSI/SVD
 - b. Text classification
 - i. Naïve bayes
 - ii. KNN
 - iii. Rocchio
4. Web Mining
 - a. PageRank algorithm
 - b. HITS algorithm

Topics post-midterm

5. Web Mining, con't
 - a. Web spam – up to slide #98
 - b. Reading for web mining (posted at the end of the slides):
<http://nlp.stanford.edu/IR-book/pdf/21link.pdf>
6. Sentiment Analysis

- a. Opinion mining
 - i. Document-level
 - ii. Sentence-level
 - iii. Aspect-based
 - iv. Reading: first 3 chapters of *Sentiment Analysis and Opinion Mining* by Professor Bing Liu
- b. Social Network Graph Mining
 - i. Girvan-Newman algorithm
 - ii. We also used the following slides in class:
 - 1. <http://snap.stanford.edu/class/cs246-2012/slides/11-graphs.pdf>
(corresponds to reading assignment #1 below)
 - 2. <http://cs.stanford.edu/people/jure/talks/networks-icdm-dec12.pdf>
(up to slide 36)
 - iii. Reading list:
 - 1. <http://infolab.stanford.edu/~ullman/mmds/ch10.pdf>
 - 2. http://www.public.asu.edu/~pgundeck/book_chapter/smm.pdf
- c. Image
 - i. Basics for content-based retrieval: features, similarity measures, color quantization
 - ii. Reading:
<http://infolab.stanford.edu/~wangz/project/imsearch/review/JOUR/datta.pdf>

More emphasis will be on the post-midterm topics. In general, anything covered in class as well as reading assignments (including topics in reading that are not covered in class) could be on the exam. The reading assignments were either on the slides or posted on Piazza.