CS 322: Natural Language Processing
Exam Overview: Monday, May 13th *May 9, 2019*

This list is to give you a starting point for studying for the first exam. **Note that this is not a contract, nor do I promise that this is an exhaustive list of everything that could be on the exam.** For the exam, you may bring a crib sheet on standard 8.5x11 paper with handwriting on one side. If this poses a problem due to accommodations registered with Disability Services, please let me know and I'm happy to make an exception in that case.

While I have not written the exam yet, I am anticipating 6-8 questions with a mix of short answer, derivations, always-sometimes-nevers, probability computations from word counts, etc. There will be no explicit code-writing on the exam (e.g., I won't ask you to "write a python function") nor will there be any proofs. The arithmetic will be simple enough so that a calculator will not be required.

Major topics:

- Tokenization: whitespace, "smarter" whitespace, statistical.
- Discrete Probability: conditional/joint probabilities, independence, bayes rule.
- Linear Algebra Operations: vector-vector multiply, matrix-vector multiply, element-wise operations, reduction operations.
- N-gram Language Models: maximum likelihood estimation, sparsity, parameter counting and memory requirements, dealing with unknown words, determining vocabularies from corpora, smoothing.
- Machine learning experiment concepts: train/val/test sets, overfitting, cross-validation, hyperparameters vs. "normal" parameters.
- Document Classification with Naive Bayes: ability to derive the classifier from bayes rule, understanding where "naive" comes from, understanding the difference between multinomial and binary NB, connection with language models.
- Evaluation of classifiers: precision/recall/F-measure, understanding the context-specific nature of precision and recall
- Document Classification with Logistic Regression: understanding how the model makes probability estimates, understanding why the sigmoid function is used, gradient derivation for each of the parameters, gradient descent, L_1 and L_2 regularization.
- Vector semantics: representing words and documents as vectors, computation on word/document embeddings (e.g., cosine similarities, addition), truncated SVD.