Putting it together

Text analytics pipeline

- search
- general preprocessing
- add annotation (~ 'features')
- search (again)
- unsupervised techniques
- supervised techniques

Search

- regular expressions (FSAs)

- indexing and retrieval (Lucene)

General preprocessing

- tokenization
- sentence tokenization
- text normalization: stemming & lemmatization (FSTs)

Adding annotations

- part-of-speech tagging
- named entity recognition
- relation extraction
- coreference
- syntactic parsing
- etc.

'Under the hood'

- distributed word representations
- hidden Markov models
- n-gram models
- edit distance
- Viterbi algorithm
- maximum entropy models
- formal grammars
- noisy channel models

Search (again)

- searching annotations

- either more regular expressions, or special purpose code

Unsupervised techniques

- frequency / cooccurrence analysis
  (words / bigrams / annotated features)
- similarity and clustering (your favorite techniques)
- topic modeling (latent Dirichlet allocation)

Supervised techniques

- relevant or not?
- sentiment analysis (is also a CoreNLP annotator)
- arbitrary questions
- naive bayes (easy to implement)
- or your favorite supervised technique