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 (e.g., Lucene)

General preprocessing

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

Adding annotations

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

'Under the hood'

- distributed word representations
- hidden Markov models (HMMs)
- n-gram models
- string 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), e.g., with vector semantics

- topic modeling (latent Dirichlet allocation)

Supervised techniques

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

Extra step

- turn audio into text!

Practice yourself

- you'll use the whole pipeline for hw4