Overview of topics

- Clustering
  - Agglomerative
  - k-means
- Classification
  - Rule based
  - Support Vector Machines
  - Naive Bayes
- Finding communities (aka Trawling)
- Summarization
- Recommendation systems

Supervised vs. unsupervised learning

- Unsupervised learning:
  - Given corpus, infer structure implicit in the docs, without prior training.
- Supervised learning:
  - Train system to recognize docs of a certain type (e.g., docs in Italian, or docs about religion)
  - Decide whether or not new docs belong to the class(es) trained on

Why cluster documents

- Given a corpus, partition it into groups of related docs
  - Recursively, can induce a tree of topics
- Given the set of docs from the results of a search (say jaguar), partition into groups of related docs
  - semantic disambiguation

Agglomerative clustering

- Given target number of clusters $k$.
- Initially, each doc viewed as a cluster
  - start with $n$ clusters;
- Repeat:
  - **while** there are $> k$ clusters, find the “closest pair” of clusters and merge them.

k-means

- At the start of the iteration, we have $k$ centroids.
  - Need not be docs, just some $k$ points.
  - Axes could be terms, links, etc...
- Loop
  - Each doc assigned to the nearest centroid.
  - All docs assigned to the same centroid are averaged to compute a new centroid;
    - thus have $k$ new centroids.
Classification

- Given one or more topics, decide which one(s) a given document belongs to.
- Applications
  - Classification into a topic taxonomy
  - Intelligence analysts
  - Routing email to help desks/customer service

Accuracy measurement

- Confusion matrix

Image:

```
This (i, j) entry means 53 of the docs actually in topic i were put in topic j by the classifier.
```

Explicit queries

- Topic queries can be built up from other topic queries.

Classification by exemplary docs

- Feed system exemplary docs on topic (training)
- Positive as well as negative examples
- System builds its model of topic
- Subsequent test docs evaluated against model
  - decides whether test is a member of the topic

Vector Spaces

- Each training doc a point (vector) labeled by its topic
- Hypothesis: docs of the same topic form a contiguous region of space
- Define surfaces to delineate topics in space

Support Vector Machine (SVM)

- Quadratic programming problem
- The decision function is fully specified by training samples which lie on two parallel hyper-planes
Naive Bayes

**Training**
- Use class frequencies in training data for \( Pr[c] \).
- Estimate word frequencies for each word and each class to estimate \( Pr[w | c] \).

**Test doc \( d \)**
- Use the \( Pr[w | c] \) values to estimate \( Pr[d | c] \) for each class \( c \).
- Determine class \( c_j \) for which \( Pr[c_j | d] \) is maximized.

Content + neighbors’ classes

- Naïve Bayes gives \( Pr[c_j | d] \) based on the words in \( d \).
- Now consider \( Pr[c_j | N] \) where \( N \) is the set of labels of \( d \)'s neighbors. (Can separate \( N \) into in- and out-neighbors.)
- Can combine conditional probs for \( c_j \) from text- and link-based evidence.

Finding communities on the web

- not easy, since web is huge
- what is a “dense subgraph”?
- define \((i, j)\)-core: complete bipartite subgraph with \( i \) nodes all of which point to each of \( j \) others

**Document Summarization**

- Lexical chains: look for terms appearing in consecutive sentences.
- For each sentence \( S \) in the doc.
  \[ f(S) = a \cdot h(S) - b \cdot t(S) \]
  where \( h(S) = \) total score of all chains starting at \( S \)
  and \( t(S) = \) total score of all chains covering \( S \), but not starting at \( S \)

Recommendation Systems

Recommend docs to user based on user’s context (besides the docs’ content).

Other applications:
- Re-rank search results.
- Locate experts.
- Targeted ads.