Near-Neighbor Search

Finding Variants of a News Article
Finding Similar Restaurants
Shingling Documents → Minhashing → Locality-Sensitive Hashing → Textually Similar Documents

- Set of character sequences of length $k$
- Summaries of sets, preserving Jaccard distance
- Focuses on high-similarity items

Vectors Describing Products → Random Hyperplanes → Locality-Sensitive Hashing → Similar Products

- Summaries of vectors, preserving cosine distance
- Similar Products
Organizing News Articles

- Applications such as Google News see thousands of on-line news articles.
- Many come from the same story, with modifications by the publisher.
- Need to cluster by underlying story.
- Perfect for shingling – minhashing – locality-sensitive hashing.
Restaurant Advisor

◆ Celixis is a Stanford startup trying to create “advisors.”
  ♦ **First application:** restaurant advisor.

◆ Uses data from restaurant reviews plus one or more restaurants you select as examples of what you like.
Chris Manning’s NLP software

Restaurant reviews

Parse

Identify key words and phrases

TF.IDF-like strategy

Sentiment information

Create vectors of positive and negative features

Easier than general case; you don’t care if “big portions” are good or bad, as long as “small portions” is the opposite.

Match user-selected restaurants with local restaurants

Cosine distance used for matches

Celixis Architecture
Further Reading

◆ A new book by Anand Rajaraman and Jeff Ullman, titled *Mining of Massive Datasets*, covers these topics.

◆ See it on-line at infolab.stanford.edu/~ullman/pub/book.pdf