Clustering and Topic Models

Grouping Documents Together

- The **responses to a search query** give you a group documents
- If we represent documents as points in a space, we can try to **identify regions**
  - **Clustering:** Nearby regions of points
  - **Support Vector Machine:** Chop space up into parts
  - (Gaussian) **Mixture Models**: A type of fuzzy clustering
  - **K-Nearest Neighbors** (if have examples)
- Alternatively we can determine "hidden meaning" with a **topic model** (defined by dynamic small bags of words)
  - Latent Semantic Indexing
  - Latent Dirichlet Allocation
  - With lots of variants of these methods to find "**latent factors**"

Topic Models

- Illustrated by Google News
- These try to group documents by Topics such as "Presidential Election" and not by inclusion of particular phrases
- You imagine each document is a set of topics (the latent factors) and each topic is a bag of words.
- Find the best set of topics and best set of words in topics

Yet Another Example

- Reuters-21578 collection
  - 21578 short newswire messages from 1987
- Top-3 results when querying for **taxes reagan** using LSI:

FITZWATER SAYS REAGAN STRONGLY AGAINST TAX HIKE
WASHINGTON, March 9 - White House spokesman Marlin Fitzwater said President Reagan's record in opposing tax hikes is "long and strong" and not about to change.

ROSTENKOWSKI SAYS WILL BACK U.S. TAX HIKE, BUT DOUBTS PASSAGE WITHOUT REAGAN SUPPORT

WHITE HOUSE SAYS IT OPPOSED TO TAX INCREASE AS UNNECESSARY

A Latent Factor Finding Method

- **The last document doesn't mention the term "reagan"!**

http://www.ifis.cs.tu-bs.de/teaching/ss-11/irws

An example of DA-PLSI/PLSA

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
percent	stock	soviet	bush	percent
million	market	gorbachev	dukakis	computer
year	index	party	percent	aids
sales	million	i	i	year
billion	percent	president	jackson	new
new	stocks	union	campaign	drug
company	trading	gorbachevs	poll	virus
last	shares	government	president	futures
corp	new	new	new	people
share	exchange	news	israel	two

Top 10 popular words for each of 5 topics found from the AP news dataset divided into 30 topics. Processed by DA-PLSI and showing only 5 of 30 topics