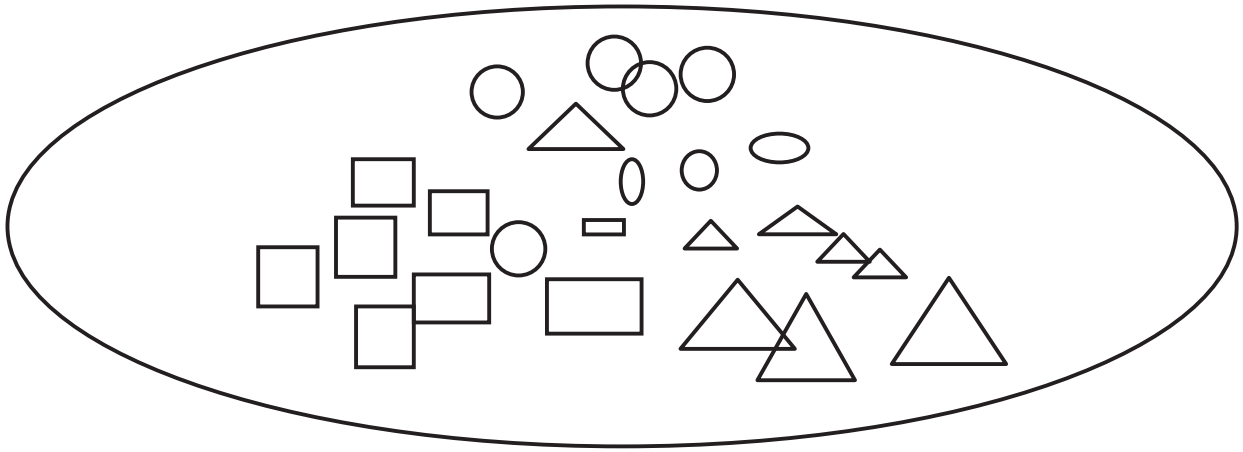
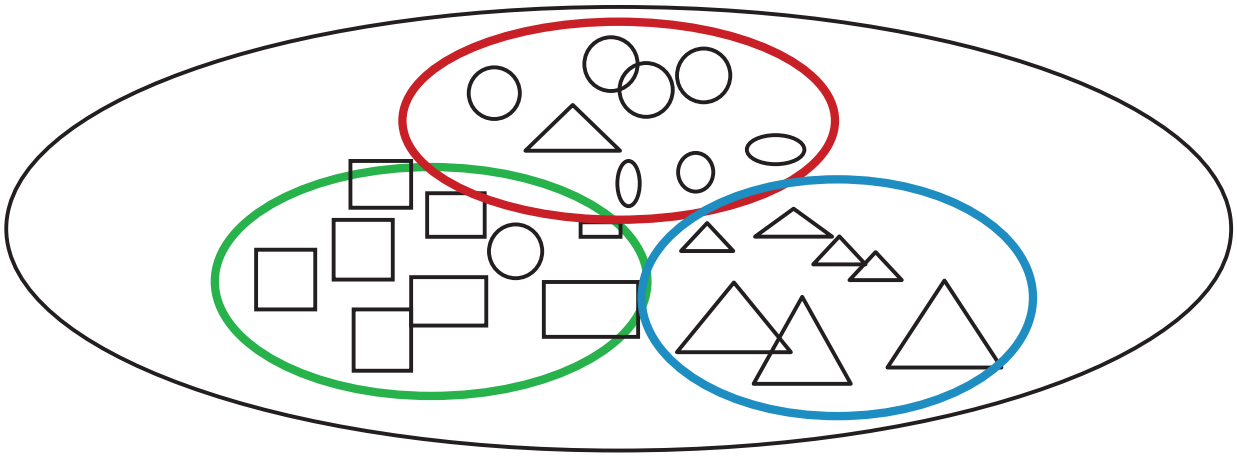
Week 6
Text Clustering

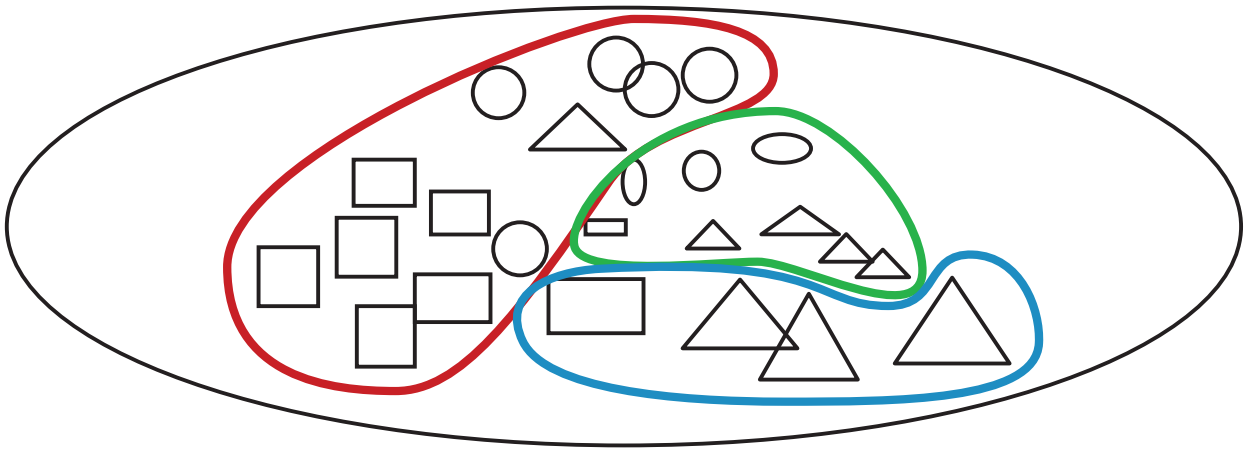
Reference and Slide Source: ChengXiang Zhai and Sean Massung. 2016. Text Data Management and Analysis: a Practical Introduction to Information Retrieval and Text Mining. Association for Computing Machinery and Morgan & Claypool, New York, NY, USA.

Find three clusters in the objects given below!





Clustering based on Shape!



Clustering based on Size!

Text Clustering

- Clustering the patterns in the data, e.g. clustering objects based on color
- Objects in one cluster have a similar property
- The core of clustering is the similarity measure between documents/terms

Similarity Function Properties

- It must be symmetric, i.e., $S(d_1, d_2)$ should be the same as $S(d_2, d_1)$.
- It should be normalized on some range, usually [0, 1].

Similarity Measure 1: Dot product

Similarity Measure 2: Cosine Similarity

$$\text{sim}_{\text{cosine}}(x, y) = \frac{x \cdot y}{||x|| \cdot ||y||}$$

$$= \frac{\sum_i x_i y_i}{\sqrt{\sum_i (x_i)^2}\sqrt{\sum_i (y_i)^2}}$$

Similarity Measure 3: Jaccard similarity

$$\text{sim}_{\text{Jaccard}}(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

X and Y are sets rather than vectors in this case!
Will work with bit vector representation only

Term Clustering

- Groups similar words together
- Can be used to refine query by adding similar words
- Can also be used to reduce the feature vector size for the document

Challenge: you need to define semantic similarity which is very difficult!

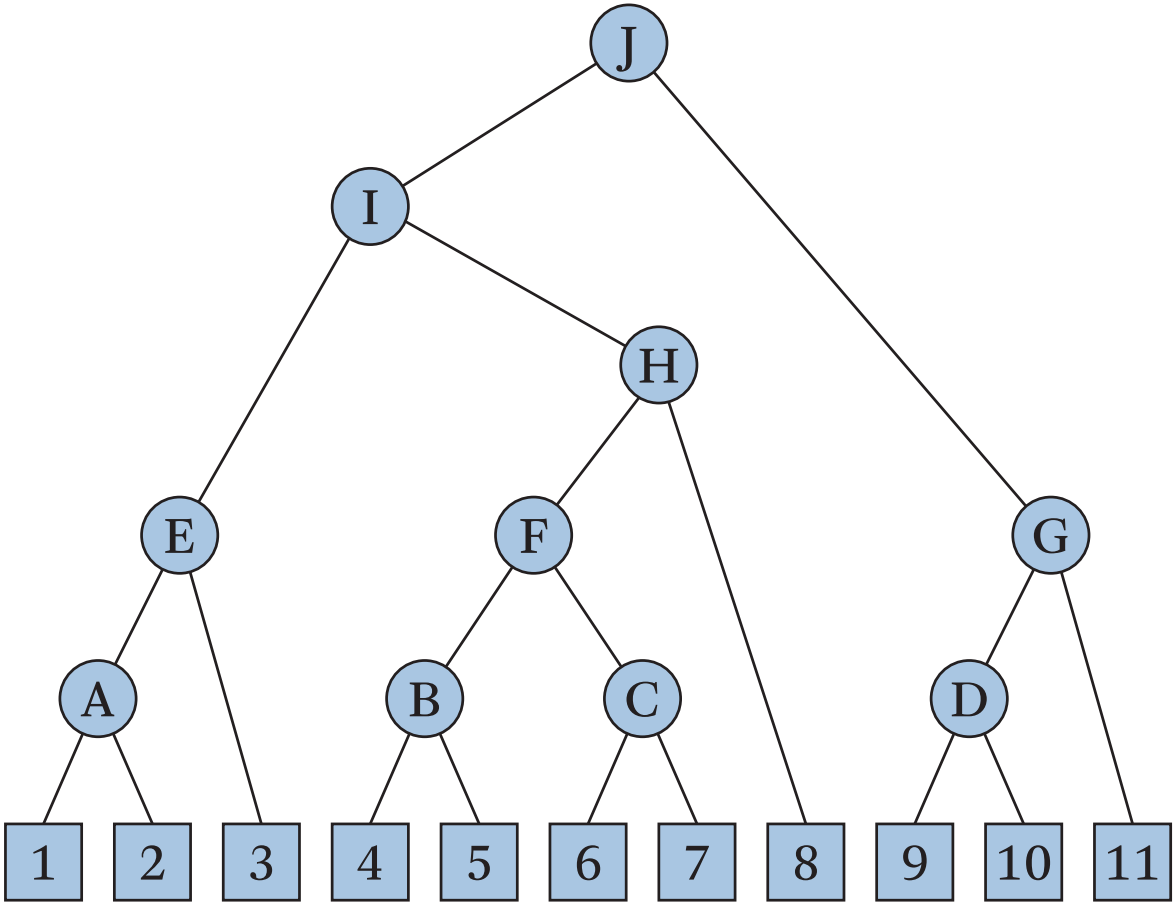
Document Clustering

- Represent documents as vectors
- If number of clusters is known, use k-means (LBG) to obtain clusters
- If number of clusters not known, use bottom-up or top-down clustering

Bottom-up Clustering

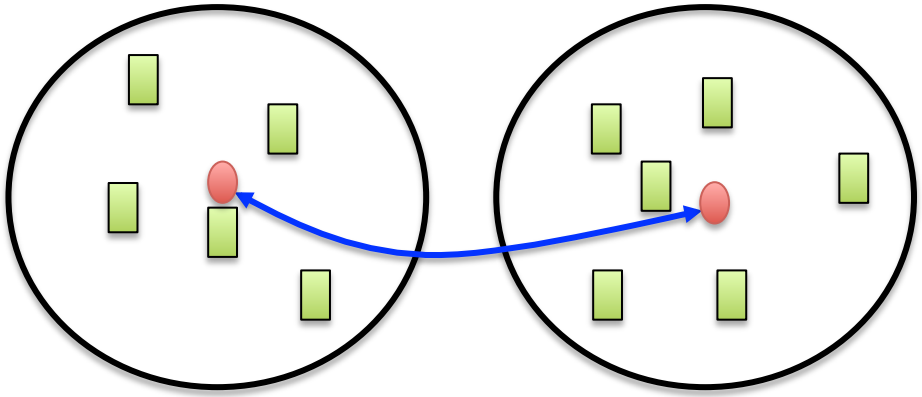
- Also called Agglomerative Hierarchical clustering
- Start by assuming each element a cluster
- Merge two closest clusters, and keep repeating this process
- The process will continue until we have 1 cluster or when we have desired number of clusters

The final tree is Called a Dendogram



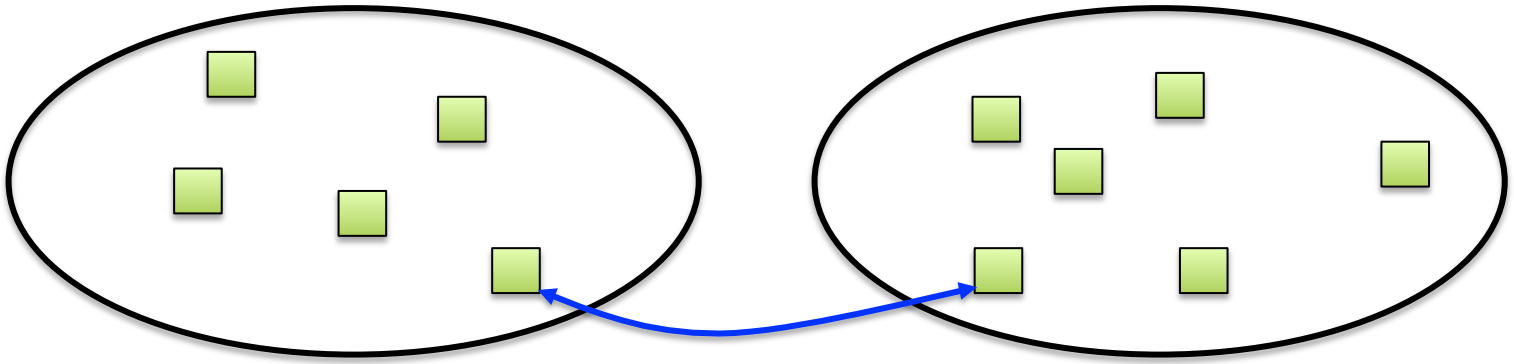
How do you measure similarity of two clusters?

Distance Between Centroids



Problem – Ignores the distribution of elements within cluster.

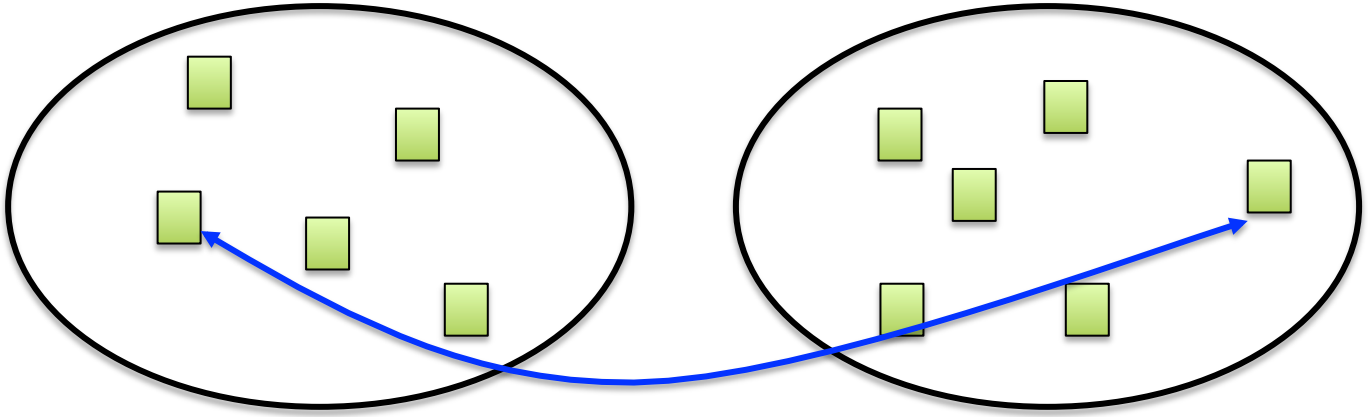
Single Link Measure



Minimum distance between pairs

Pros: Can handle non-elliptical shapes
Cons: Sensitive to noise and outliers

Complete Link Measure



Maximum distance between pairs

Complete Link Measure

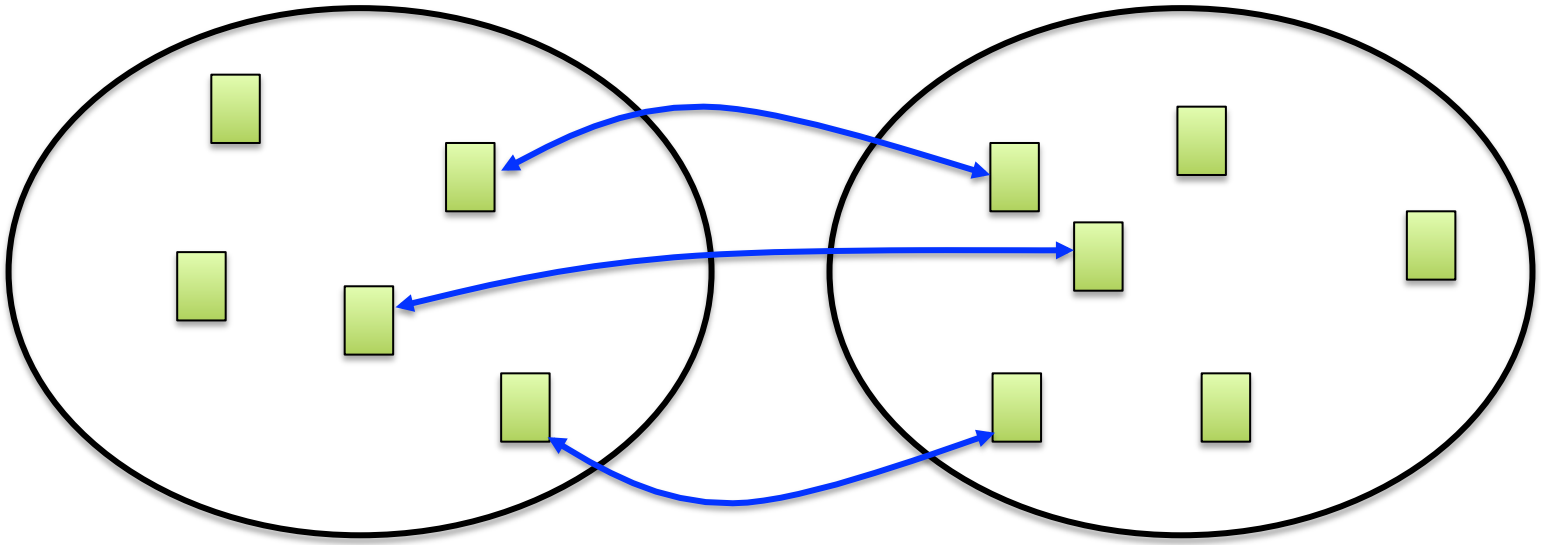
Pros:

1. Less susceptible to noise
2. More Balanced clusters
3. Results in similar size clusters

Cons:

1. Breaks large clusters
2. Small clusters are merged with large clusters

Average Link Measure



Average distance between pairs

When to Stop Merging?

- We can also stop when we have desired number of clusters
- If number of clusters is not know, use Elbow method:
  - Calculate total distortion of all clusters
  - It will increase as the number of clusters decreases
  - Stop when there is an elbow of sharp increase

Top-Down (Divisive) Clustering

- Start with a single cluster for all documents
- Find distance between all pairs of documents
- Find the largest distance pair and partition the cluster into 2.
- Use the pair as seeds for the new clusters
- Stopping criteria:
  - Desired number of clusters
  - Distance threshold