# Similarity-based Clustering: General Idea

- Explicitly define a similarity function to measure similarity between two text objects (i.e., providing “clustering bias”)
- Find an optimal partitioning of data to
  - maximize intra-group similarity and
  - minimize inter-group similarity
- Two strategies for obtaining optimal clustering
  - Progressively construct a hierarchy of clusters (hierarchical clustering)
    - Bottom-up (agglomerative): gradually group similar objects into larger clusters
    - Top-down (divisive): gradually partition the data into smaller clusters
  - Start with an initial tentative clustering and iteratively improve it (“flat” clustering, e.g., k-Means)