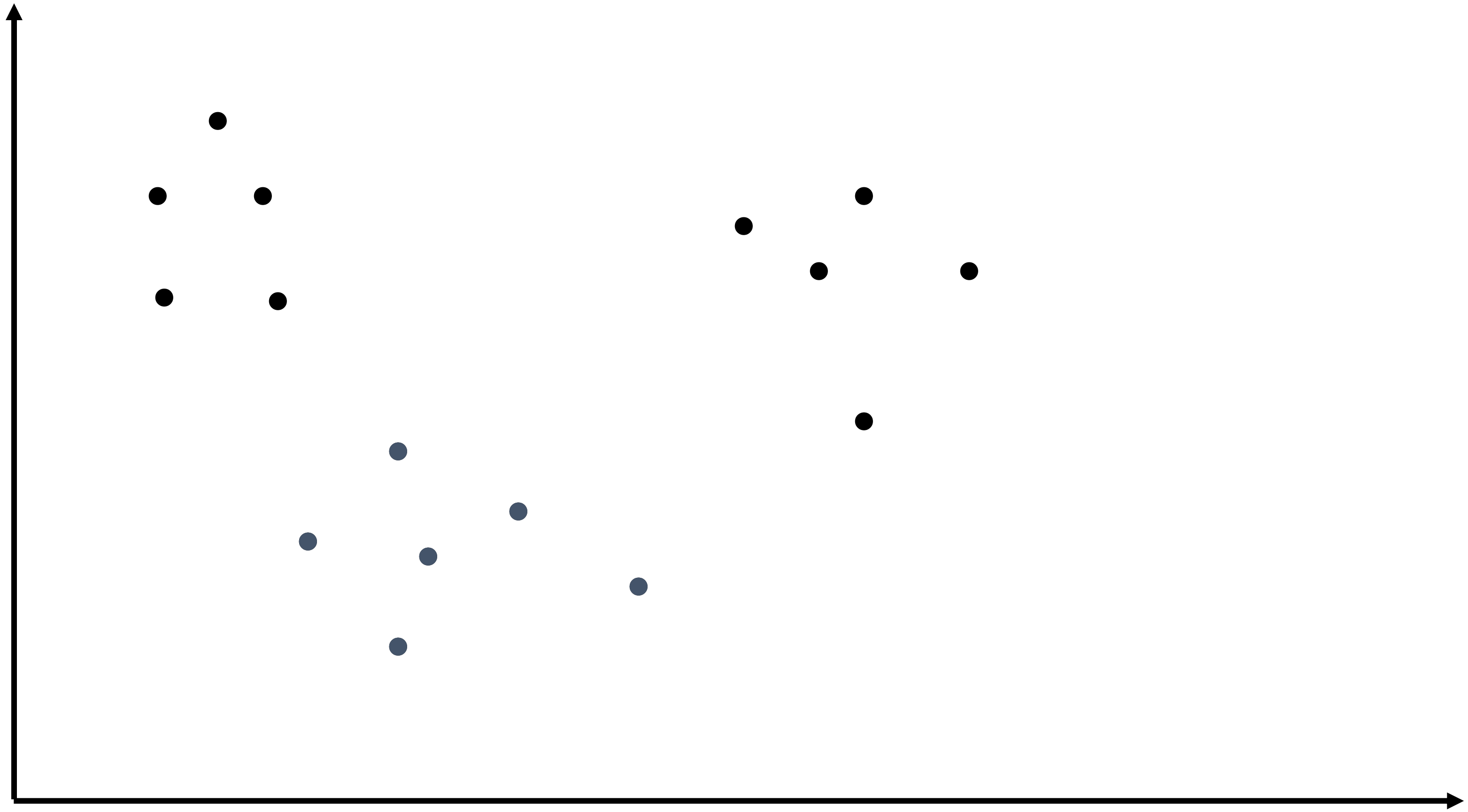
ISE 540
Text Analytics

Mayank Kejriwal
Research Assistant Professor/Research Lead
Department of Industrial and Systems Engineering
Information Sciences Institute
USC Viterbi School of Engineering
kejriwal@isi.edu

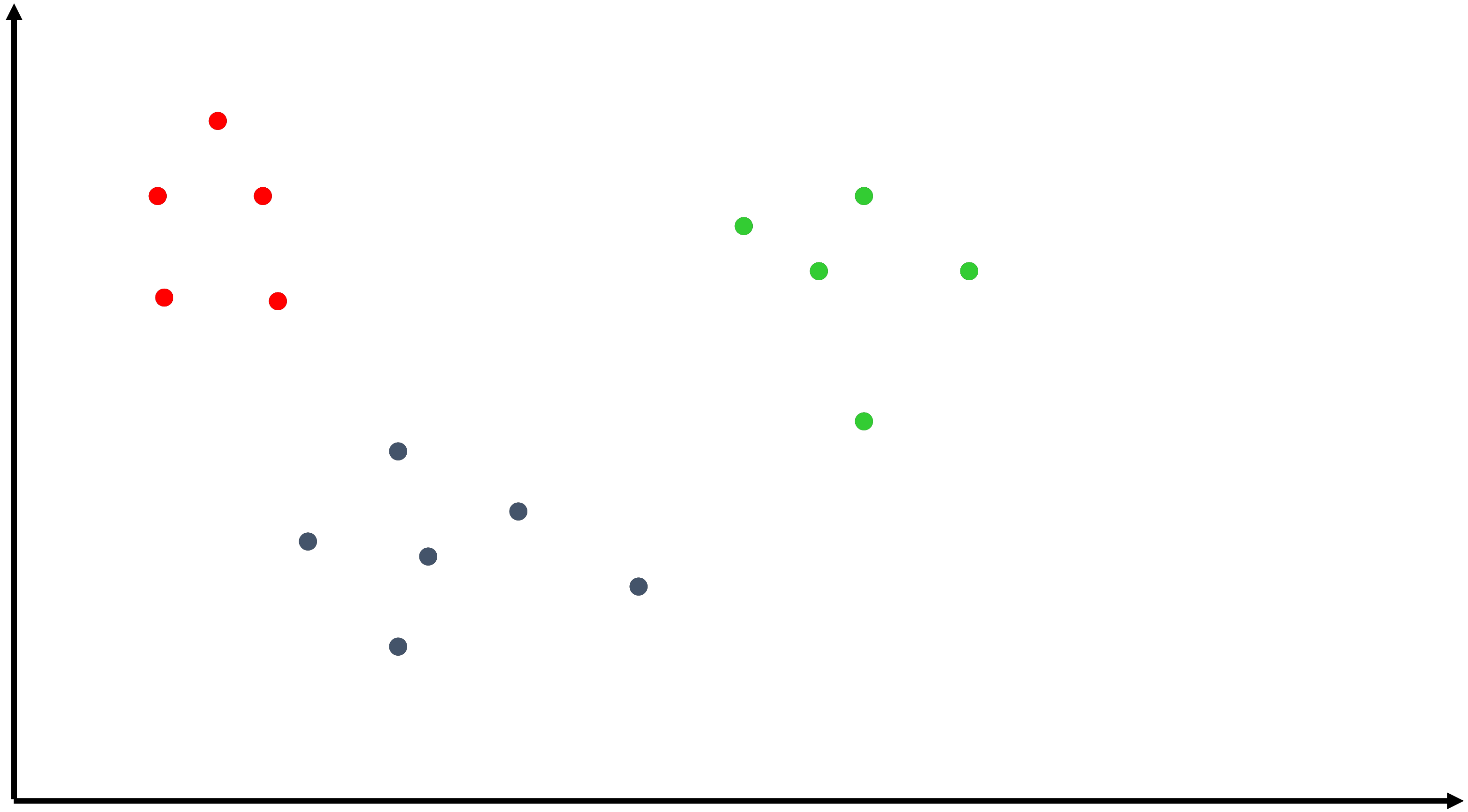
Clustering

- Partition unlabeled examples into disjoint subsets of *clusters*, such that:
  - Examples within a cluster are very similar
  - Examples in different clusters are very different
- Discover new categories in an *unsupervised* manner (no sample category labels provided).

Clustering Example



Clustering Example



Direct Clustering Method

- *Direct clustering* methods require a specification of the number of clusters, *k*, desired.
  - Hierarchical and agglomerative clustering (not covered in this class) don't need k
- A *clustering evaluation function* assigns a real-value quality measure to a clustering.
- The number of clusters can be determined automatically by explicitly generating clusterings for multiple values of *k* and choosing the best result according to a clustering evaluation function.

K-Means

- Assumes instances are real-valued vectors.
- Clusters based on *centroids*, *center of gravity*, or mean of points in a cluster, *c*:

$$\vec{\mu}(\mathrm{c}) = \frac{1}{|c|} \sum_{\vec{x} \in c} \vec{x}$$

- Reassignment of instances to clusters is based on distance to the current cluster centroids.

Distance Metrics

- Euclidian distance (L_2 norm): $$L_2(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$$
- L_1 norm: $$L_1(\vec{x},\vec{y}) = \sum_{i=1}^{m}\left|x_i - y_i\right|$$
- Cosine Similarity (transform to a distance by subtracting from 1):

Predominantly used for tf-idf, word embeddings etc.

$$1 - \frac{\vec{x} \bullet \vec{y}}{\left|\vec{x}\right| \cdot \left|\vec{y}\right|}$$

K-Means Algorithm

Let d be the distance measure between instances.
Select k random instances $\{s_1, s_2, \ldots s_k\}$ as seeds.
Until clustering converges or other stopping criterion:

For each instance x_i:

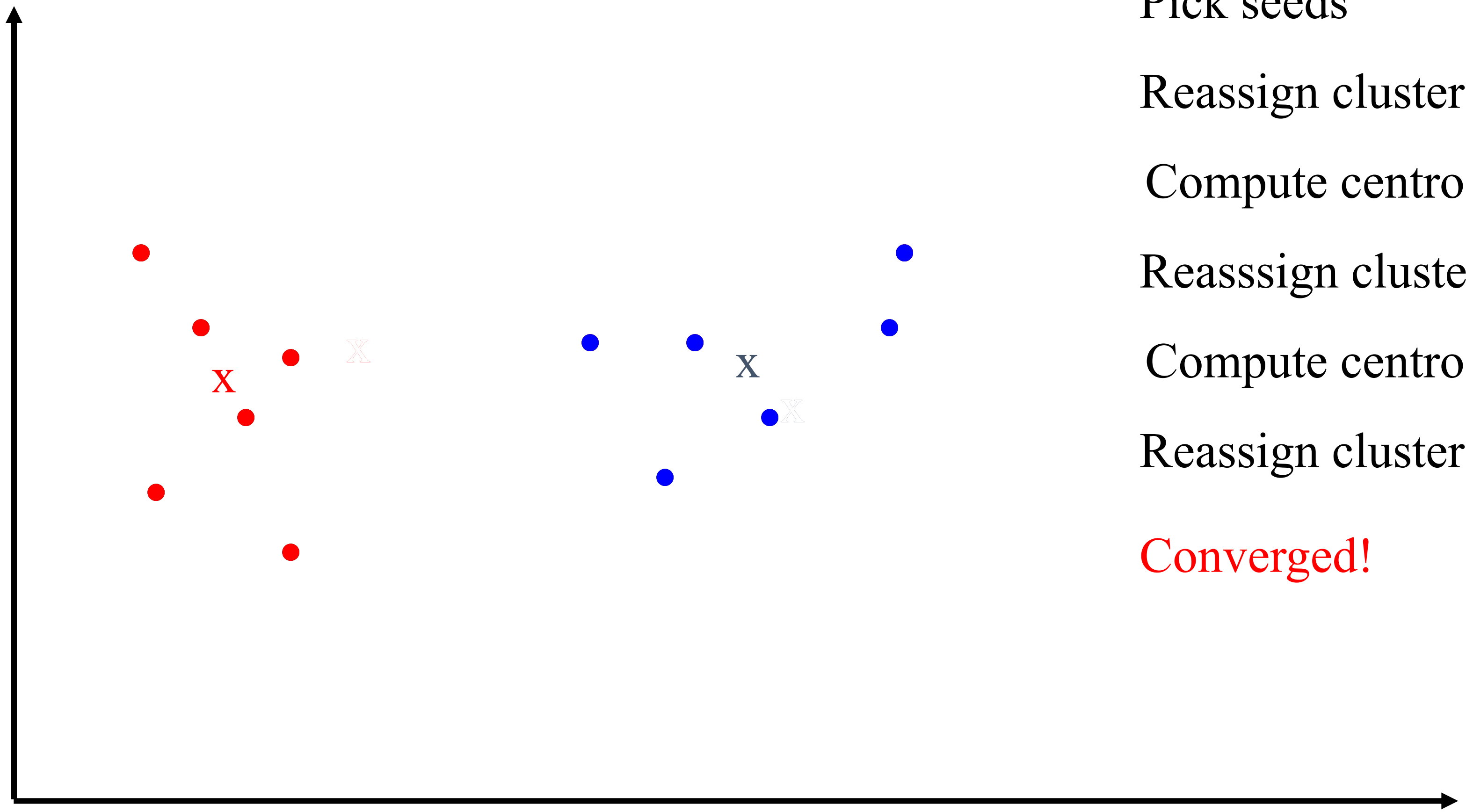
Assign x_i to the cluster c_j such that $d(x_i, s_j)$ is minimal.

(Update the seeds to the centroid of each cluster)

For each cluster c_j

$s_j = \mu(c_j)$

K Means Example
(K=2)



Pick seeds

Reassign clusters

Compute centroids

Reasssign clusters

Compute centroids

Reassign clusters

Converged!

Time Complexity

- Assume n poins and that computing distance between two instances is O(m) where m is the dimensionality of the vectors.
- Reassigning clusters: O(kn) distance computations, or O(knm).
- Computing centroids: Each instance vector gets added once to some centroid: O(nm).
- Assume these two steps are each done once for I iterations: O($Iknm$).

Seed Choice

- Results can vary based on random seed selection.
- Some seeds can result in poor convergence rate, or convergence to sub-optimal clusterings.
- Select good seeds using a heuristic or the results of another method.

More on Text Clustering

- Typically use ***normalized***, TF/IDF-weighted vectors and cosine similarity.
- Optimize computations for sparse vectors.
- Many applications: can you name some?
- Clustering is a valuable tool in class projects

Soft Clustering

- Clustering typically assumes that each instance is given a “hard” assignment to exactly one cluster.
- Does not allow uncertainty in class membership or for an instance to belong to more than one cluster.
- *Soft clustering* gives probabilities that an instance belongs to each of a set of clusters.
- Each instance is assigned a probability distribution across a set of discovered categories (probabilities of all categories must sum to 1).

Issues in Clustering

- How to evaluate clustering?
  - Internal:
    - Tightness and separation of clusters
    - Fit of probabilistic model to data
  - External
    - Compare to known class labels on benchmark data
- Improving search to converge faster and avoid local minima.
- Overlapping clustering.