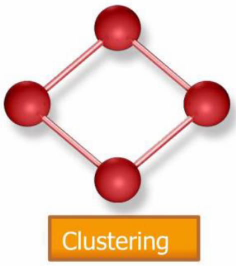
What is Clustering?



Organizing data into *clusters* such that there is:

- ✓ High intra-cluster similarity
- ✓ Low inter-cluster similarity
- ✓ Informally, finding natural groupings among objects.

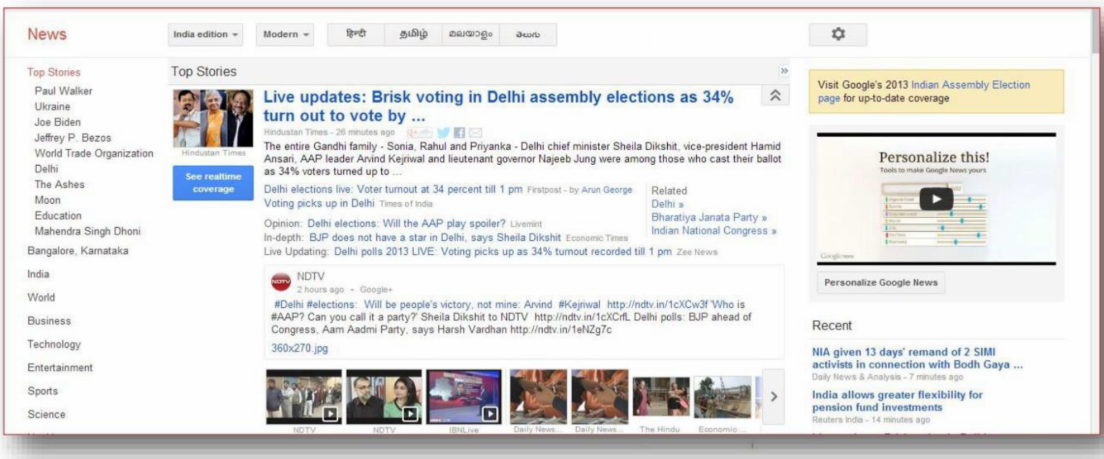
Why do we want to do it??

Why Clustering?

- ✓ Organizing data into clusters shows internal structure of the data
  Ex. Clusty and clustering genes
- ✓ Sometimes the partitioning is the goal
  Ex. Market segmentation
- ✓ Prepare for other AI techniques
  Ex. Summarize news (cluster and then find centroid)
- ✓ Techniques for clustering is useful in knowledge
- ✓ Discovery in data
  Ex. Underlying rules, reoccurring patterns, topics, etc.

Clustering - Example

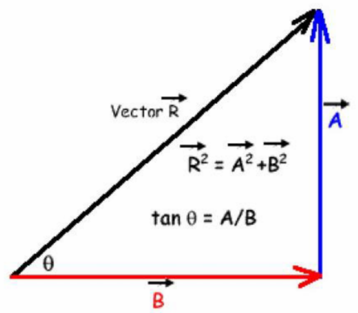
A sample news grouping from Google News:

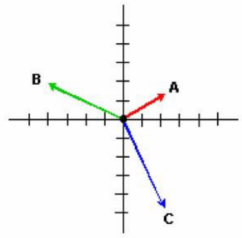


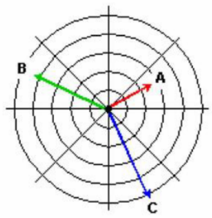
Vector

A **vector** is a quantity or phenomenon that has two independent properties: magnitude and direction.

The term also denotes the mathematical or geometrical representation of such a quantity.







Similarity Measurement

Similarity measurement definition

- Similarity by Correlation
- Similarity by Distance

Distance measures

Similarity by distance

- Euclidean distance measure
- Manhatten distance measure
- Cosine distance measure
- Tanimoto distance measure
- Squared Euclidean distance measure

Euclidean distance measure

Mathematically, Euclidean distance between two n-dimensional vectors

(a1, a2, ... , an) and (b1,b2,...,bn) is:

$$d = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \dots + (a_n - b_n)^2}$$

Manhatten distance measure

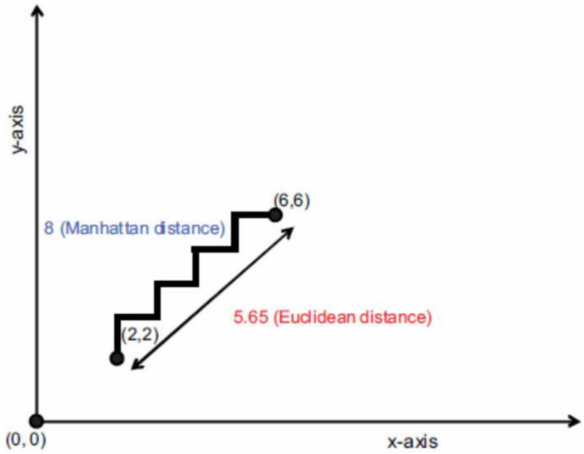
Mathematically, the Manhattan distance between two n-dimensional vectors

(a1, a2, ... , an) and (b1, b2, ... , bn) is

$$d = |a1 - b1| + |a2 - b2| + \dots + |an - bn|$$

Difference between Euclidean and Manhattan

From this image we can say that, The Euclidean distance measure gives 5.65 as the distance between (2, 2) and (6, 6) whereas the Manhattan distance is 8.0



Cosine distance measure

The formula for the cosine distance between n-dimensional vectors (a1, a2, ... , an) and (b1, b2, ...,bn) is

$$d = 1 - \frac{(a_1 b_1 + a_2 b_2 + \ldots + a_n b_n)}{(\sqrt{(a_1^2 + a_2^2 + \ldots + a_n^2)}\sqrt{(b_1^2 + b_2^2 + \ldots + b_n^2)})}$$

Tanimoto distance measure

The formula for the Tanimoto distance between two *n*-dimensional vectors (a1, a2, ... , an) and (b1, b2, ... , bn) is

$$d = 1 - \frac{(a_1 b_1 + a_2 b_2 + \ldots + a_n b_n)}{\sqrt{(a_1^{\,2} + a_2^{\,2} + \ldots + a_n^{\,2})} + \sqrt{(b_1^{\,2} + b_2^{\,2} + \ldots + b_n^{\,2})} - (a_1 b_1 + a_2 b_2 + \ldots + a_n b_n)}$$