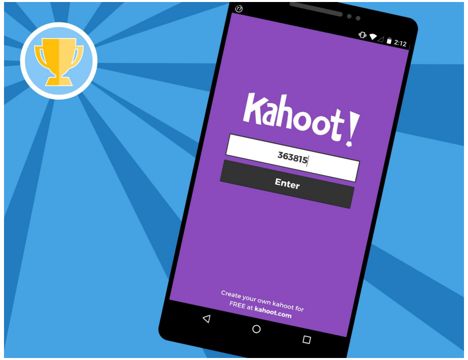
Clustering

Install Kahoot! App on your smartphone



Agenda

- Clustering
- K-means Algorithms
- Text Summarization

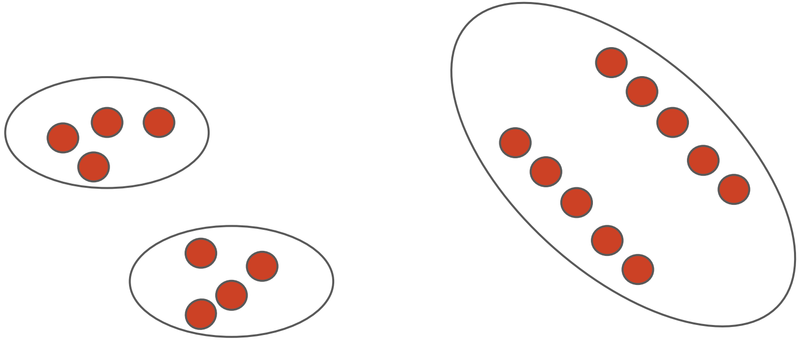
Clustering

Clustering

- Unsupervised learning
- Requires data, but not labels
- Detect patterns
  - Group emails or search results
  - Customer shopping patterns
  - Regions of images
- Userful when do not know what you are looking for
- But: can get gibberish

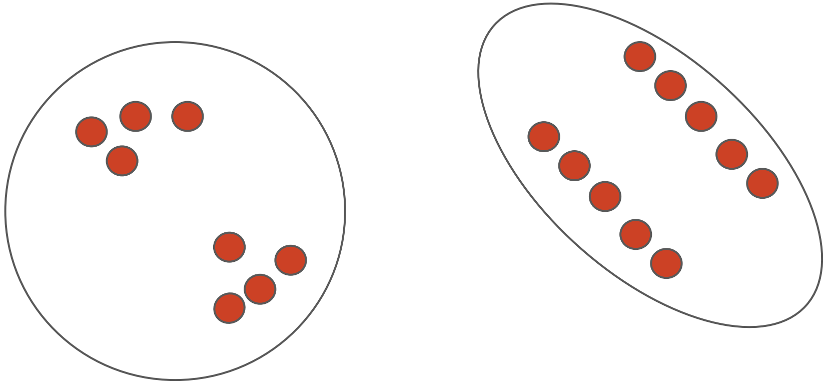
Clustering

- Basic idea: group similar instances together
- Examples: 2D points patterns



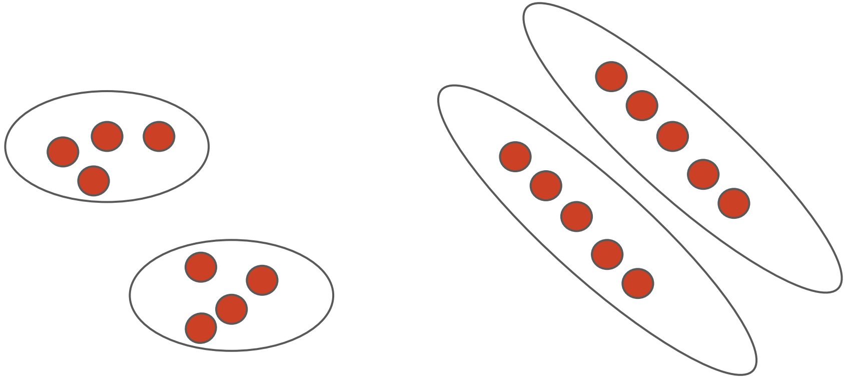
Clustering

- Basic idea: group similar instances together
- Examples: 2D points patterns



Clustering

- Basic idea: group similar instances together
- Examples: 2D points patterns



Similarity

- How to define similar
  - The measures of similarity (or distance) betweens data samples are key components for clustering results
  - One option: small Euclidean distance (squared)

$$dist(\overrightarrow{x}, \overrightarrow{y}) = ||\overrightarrow{x} - \overrightarrow{y}||^2$$

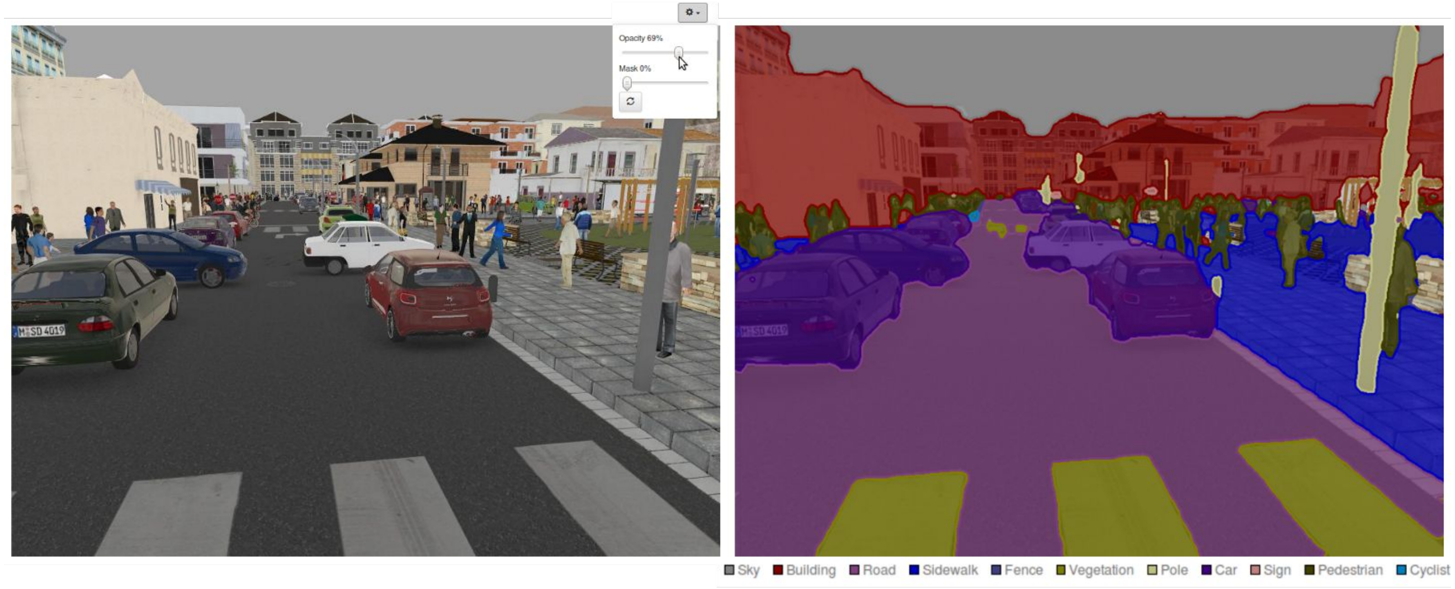
  - Similarity measures should match problem definition

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

For example: two dimensional data

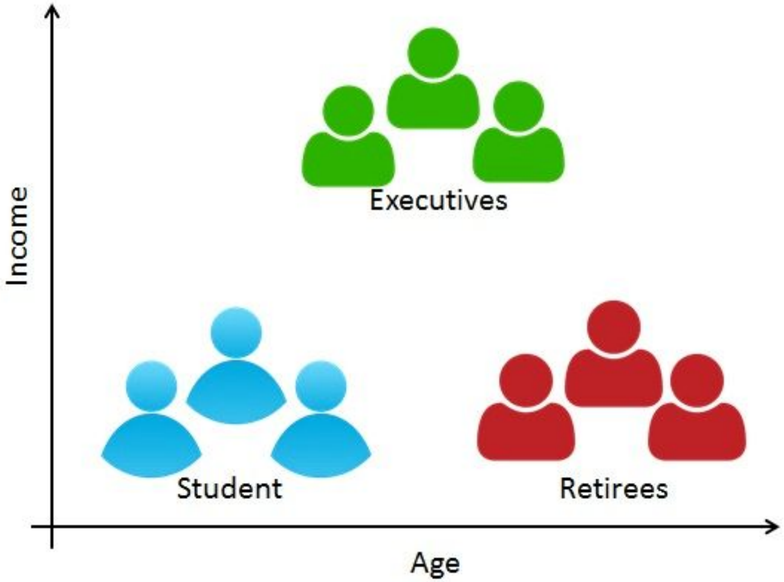
Clustering Applications

- Image Segmentation



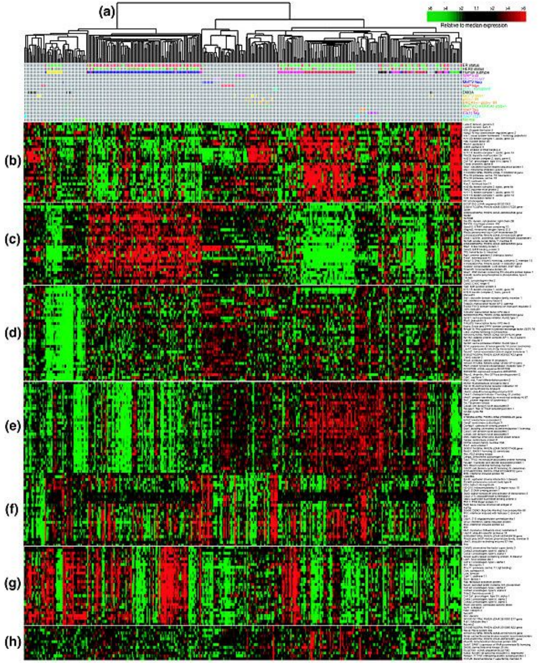
Clustering Applications

- Customer Segmentation



Clustering Applications

- Gene expression data clustering



source: Genome Biology 2007

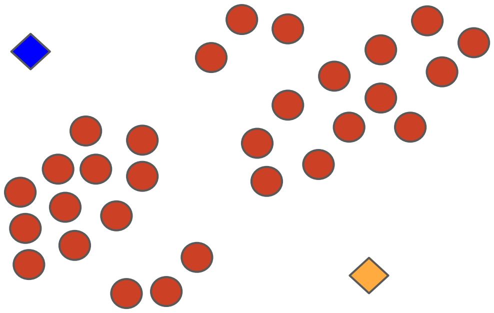
K-means

K-means

- An iterative clustering
    - Initialize: select K random points as cluster centers
    - Iteration process:
        - Assign data points to closet cluster center
        - Change the cluster center to the average of its assigned points
    - Stop when no points assignments change

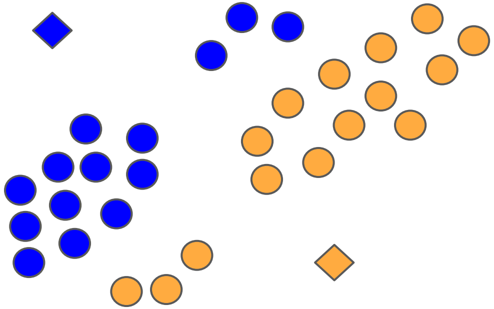
K-means clustering: examples

- Initialize 2 random points as cluster centers



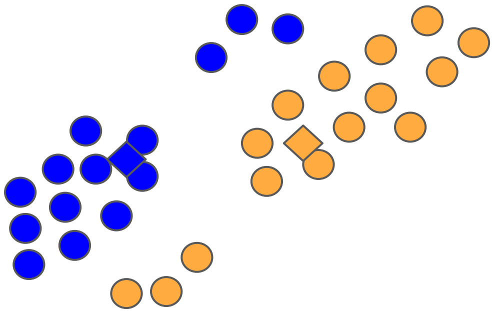
K-means clustering: examples

- Iteration one: Assign data points to closest cluster center



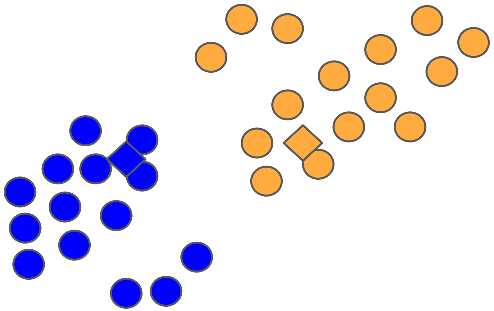
K-means clustering: examples

- Iteration one: Update the cluster center



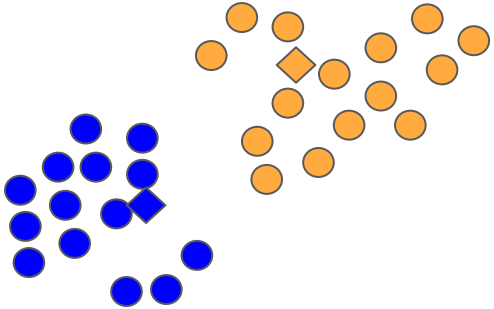
K-means clustering: examples

- Iteration two: Assign data points to closest cluster center



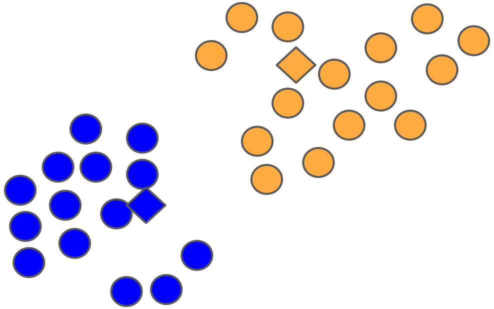
K-means clustering: examples

- Iteration two: Update the cluster center



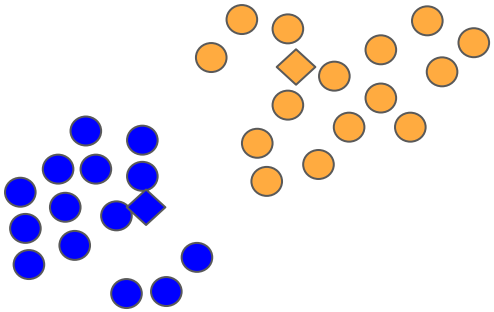
K-means clustering: examples

- Repeat until convergence



K-means clustering: examples

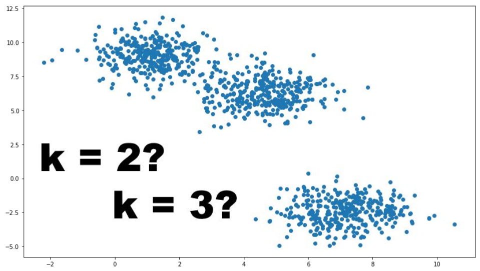
- Repeat until convergence



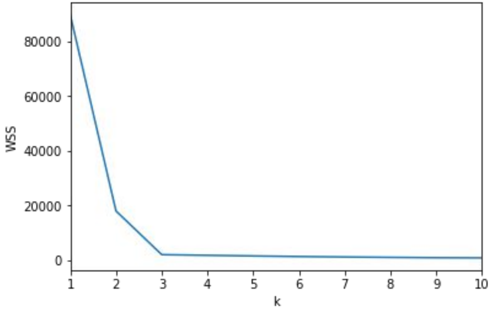
Stopping criteria

- How to define convergence?
  - No data points change clusters
  - Sum of the distances is minimized
  - Some maximum number of iterations is reached
- This algorithm is guaranteed to converge to a result (but maybe a local optimum)

How to find K?



- The number of cluster should be pre-defined
- One of the metrics can be the mean distance between data points and their cluster centroids
  - Draw the figure with the mean distance and the number of centroids
- Elbow point: ***Within-Cluster-Sum of Squared*** *Errors (WSS)*

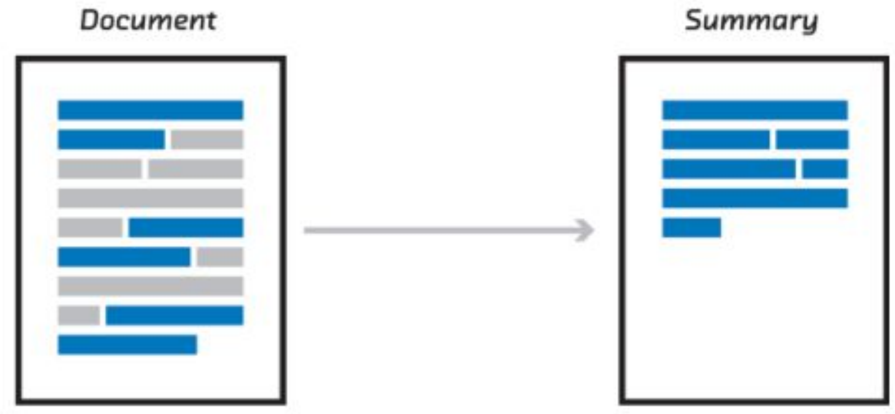


Source: Analytics Vidhya

Text Summarization

Text Summarization

- The process of shortening a text document, in order to create a summary of the major points of the original document.



source: Medium

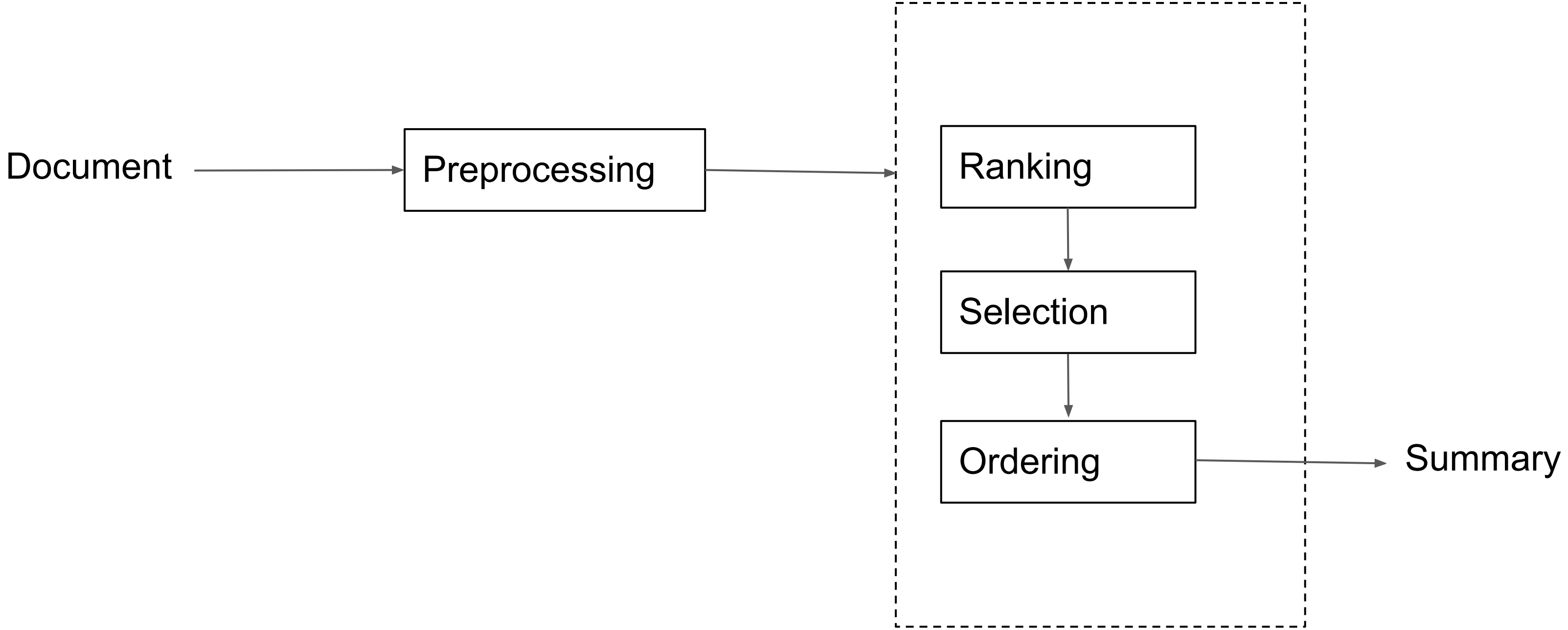
Why Automatic Summarization

- Algorithm for reading in many domains is:
    - Read summary
    - Decide whether relevant or not
    - If relevant: read whole document
- Summary is gate-keeper for large number of documents
- Information overload
- Human-generated summaries are expensive

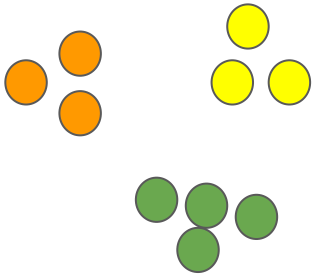
Summarization Algorithms

- Keyword summaries
  - Display most significant keywords
  - Easy to do
  - Hard to read
- Extractive summaries
  - Extract key sentences
  - Medium hard
  - Summaries often do not read well
  - Good representation of content
- Abstractive summaries
  - Build knowledge representation of text
  - Generate sentences summarizing content
  - Hard to do well

Extractive summarization



K-means clustering



- Schemes:
  - Sentences as data points
  - Divide into clusters
  - Select sentences from each cluster
  - Diverse summaries

Feature extraction for sentences

- TF-IDF Model for sentences
- Word embeddings