Text Clustering

BAF675 금융 빅데이터 분석

이재훈, Week 8

Distance Metrics

- Euclidean distance
- Manhattan distance
- Mahalanobis distance
- Jaccard distance
- Edit distance
- Cosine distance

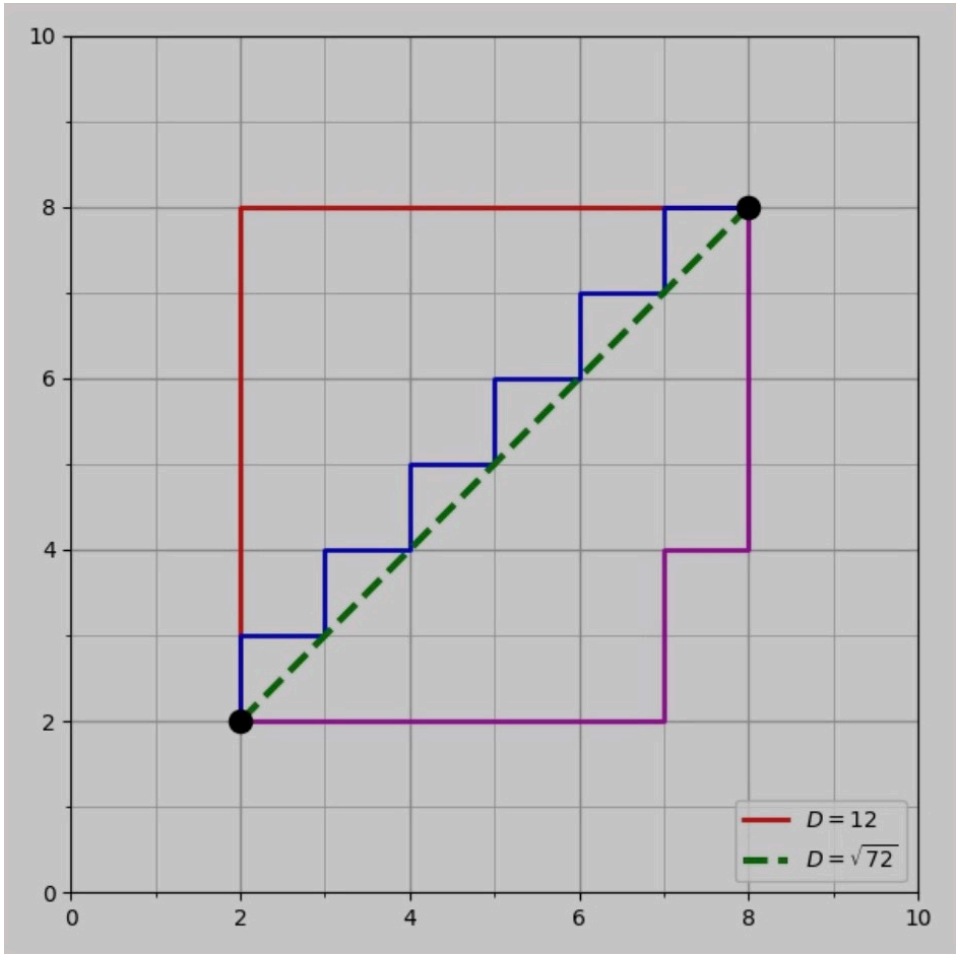
Distance Metrics

String Matching	Distance Metrics	Relational Matching	Other Matching
Edit Distance - Levenstein - Smith-Waterman - Affine **Alignment** - Jaro-Winkler - Soft-TFIDF - Monge-Elkan **Phonetic** - Soundex - Translation	- Euclidean - Manhattan - Minkowski **Text Analytics** - Jaccard - TFIDF - Cosine similarity	**Set Based** - Dice - Tanimoto (Jaccard) - Common Neighbors - Adar Weighted **Aggregates** - Average values - Max/Min values - Medians - Frequency (Mode)	- Numeric distance - Boolean equality - Fuzzy matching - Domain specific **Gazettes** - Lexical matching - Named Entities (NER)

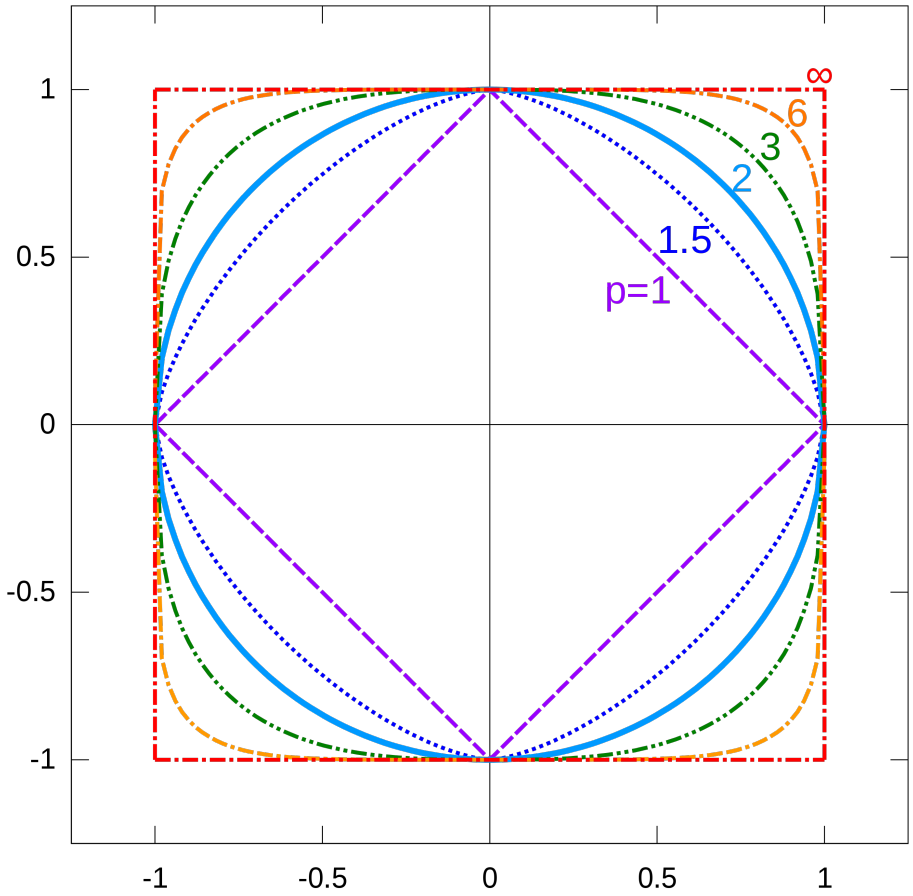
Euclidean & Manhattan distance

- Euclidean distance
  - $\sqrt{\sum \left(a_i - b_i\right)^2}$
- Manhattan distance
  - $\sum \left|a_i - b_i\right|$
- p-norm
  - $\left\| a - b \right\|_p = \left(\sum \left| a_i - b_i \right|^p \right)^{1/p}$

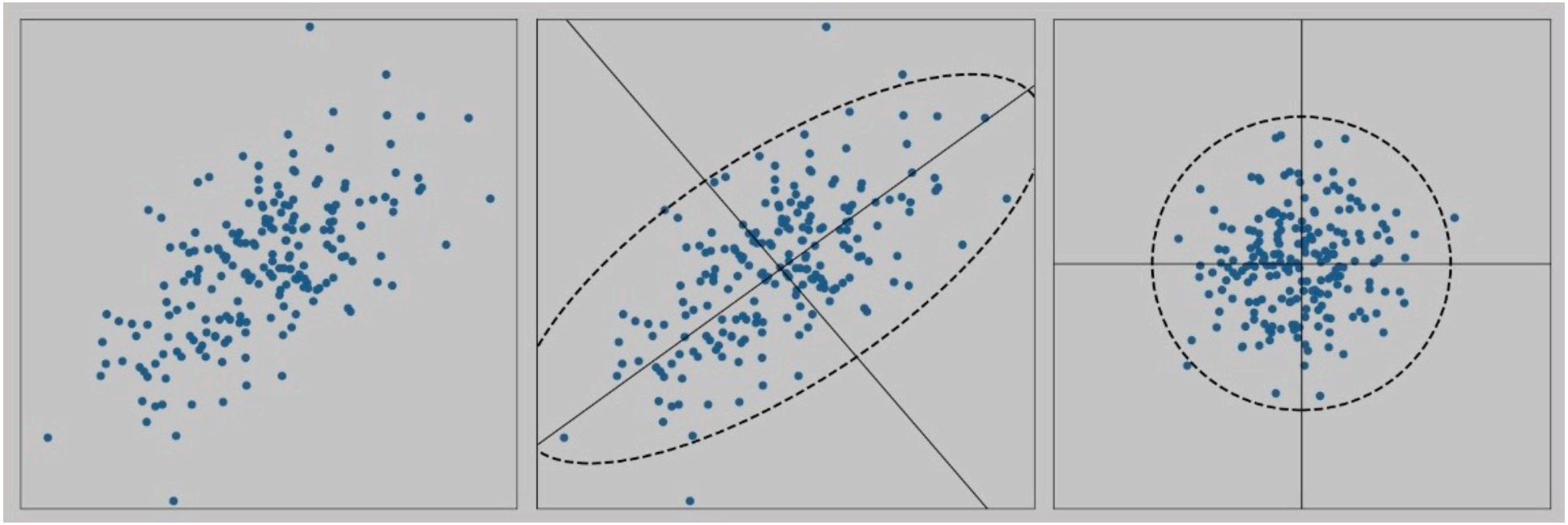
Euclidean & Manhattan distance



p-norm



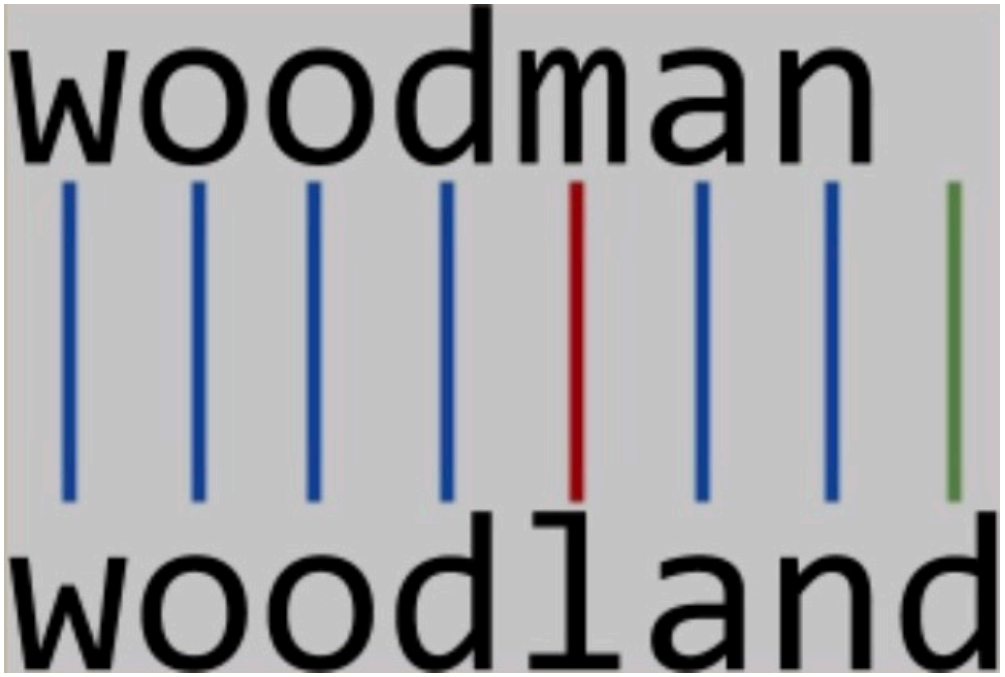
Mahalanobis distance



Jaccard distance

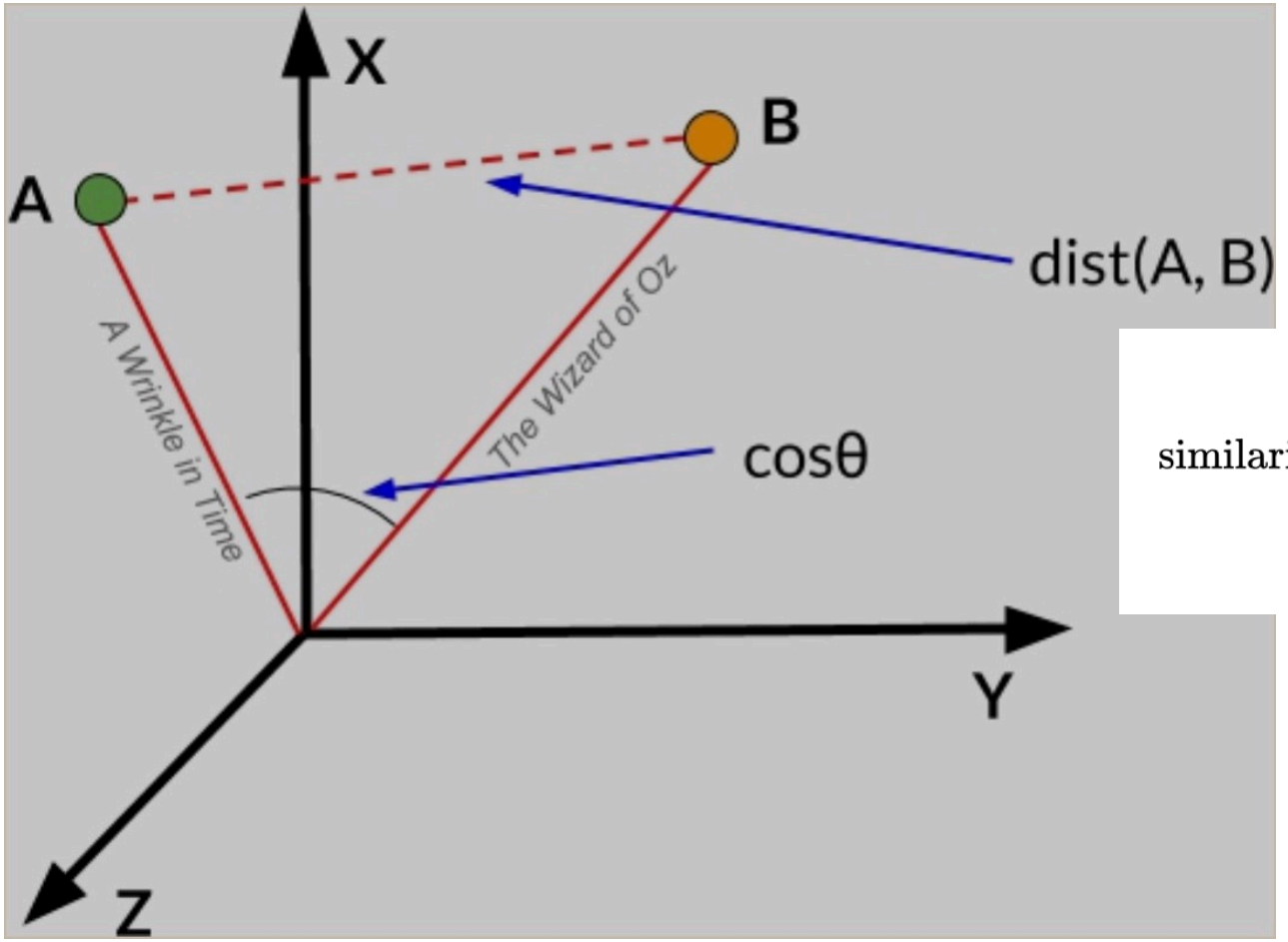
$$J(A,B) = \frac{A \cap B}{A \cup B}$$

Edit distance



- Example: one substitution and one insertion

Cosine distance



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Distance Metrics

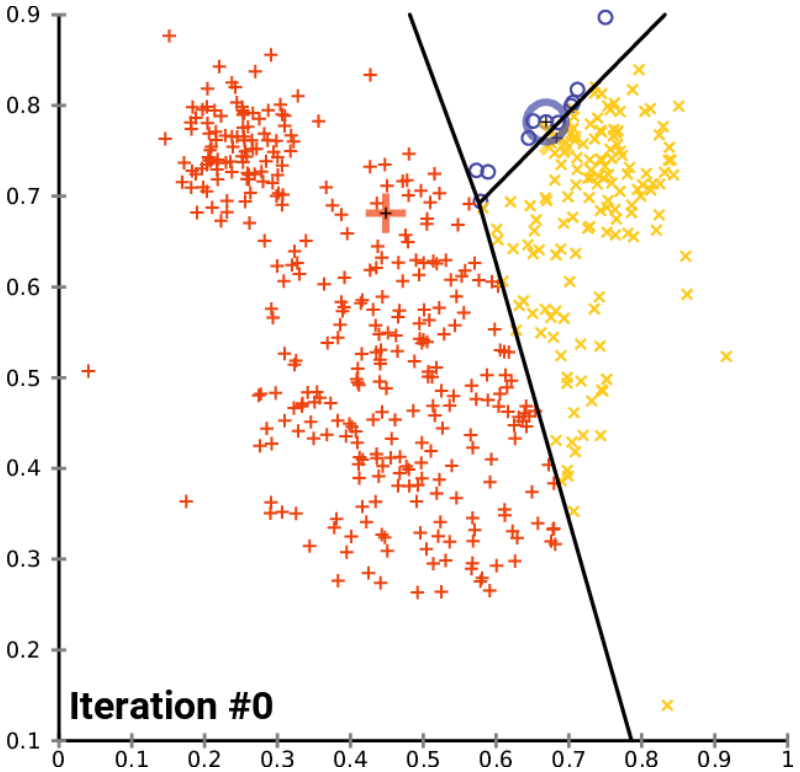
- Euclidean distance is often the default metric in clustering model.
- Cosine distance seems to frequently offer the most success.
- Jaccard distance is useful as an ad-hoc treatment on frontend app.

k-means Clustering

- Given an initial set of k-means,
  - **Assignment step**: assign each observation to the cluster of the nearest mean
  - **Update step**: recalculate means for each cluster
- Repeat the process until the assignments no longer change.

k-means Clustering

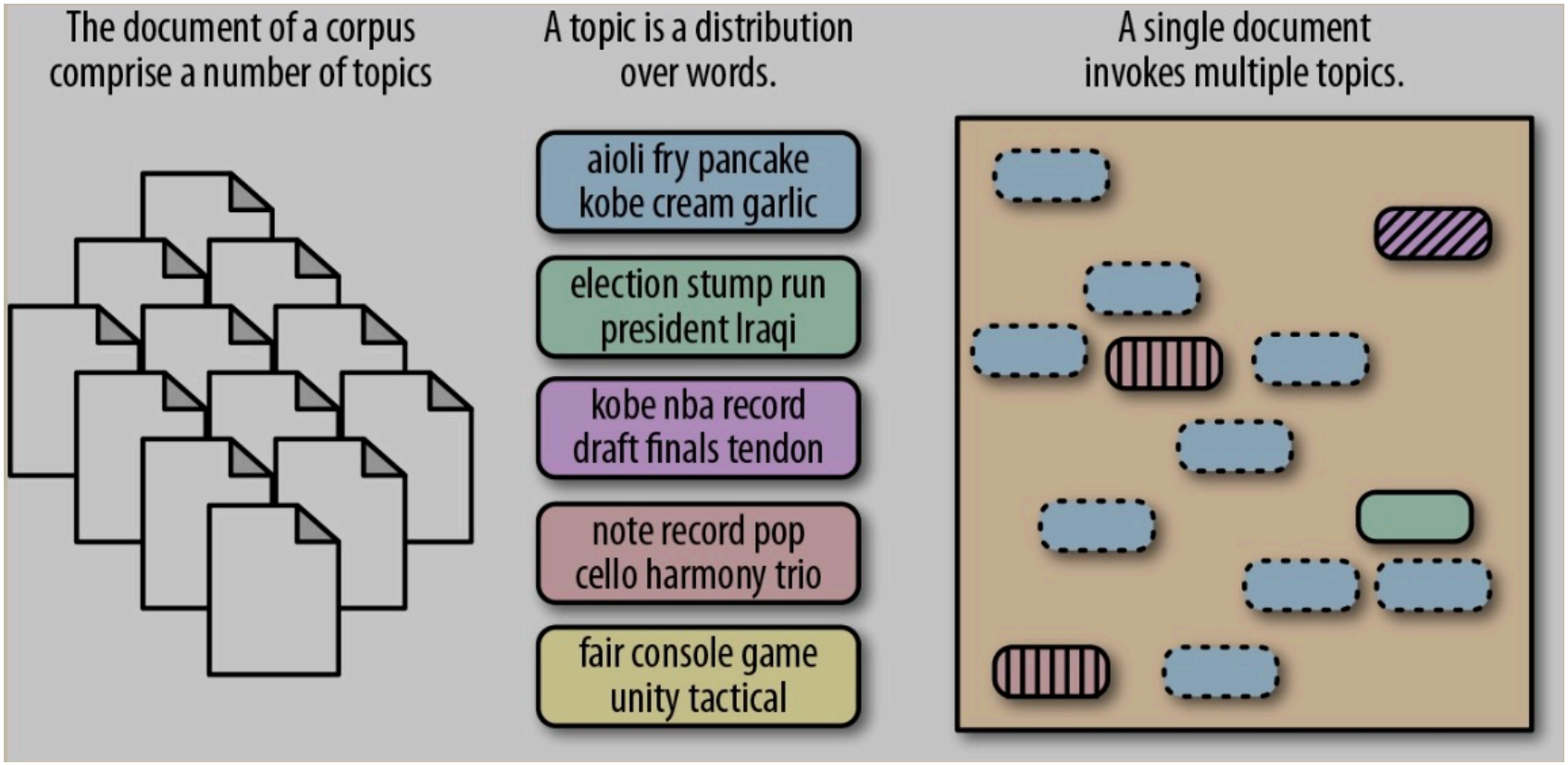
- https://en.wikipedia.org/wiki/K-means_clustering#/media/File:K-means_convergence.gif



Latent Dirichlet Allocation (LDA)

- **k-means**: hard clustering
  - assigns a document to a single disjoint cluster.
- **LDA**: soft clustering
  - assigns a document to multiple topics with probability distribution.

Latent Dirichlet Allocation (LDA)



실습: 셰익스피어 소네트 분류하기

- sonnets.ipynb 설명