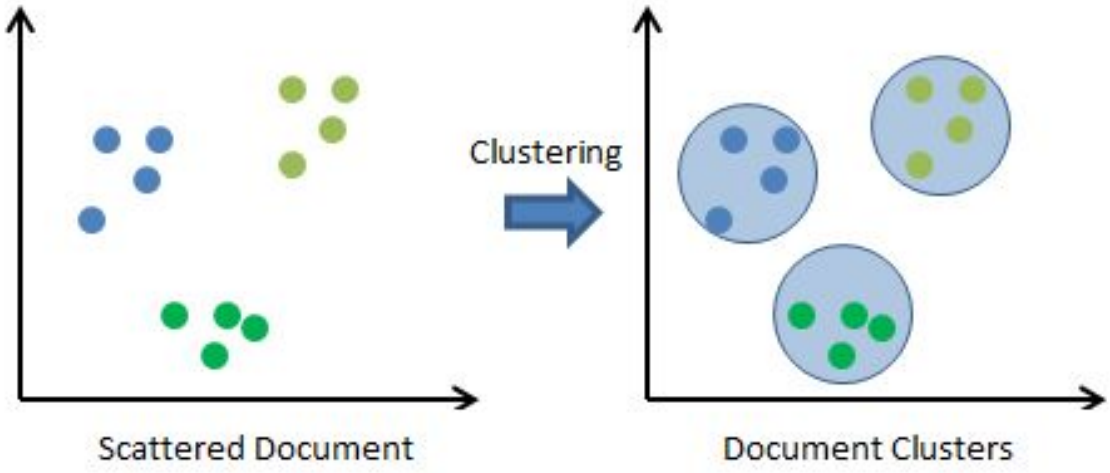
실습: Clustering

Text Clustering

- Feature Vectors
- Similarity, Distance
- K-means algorithm

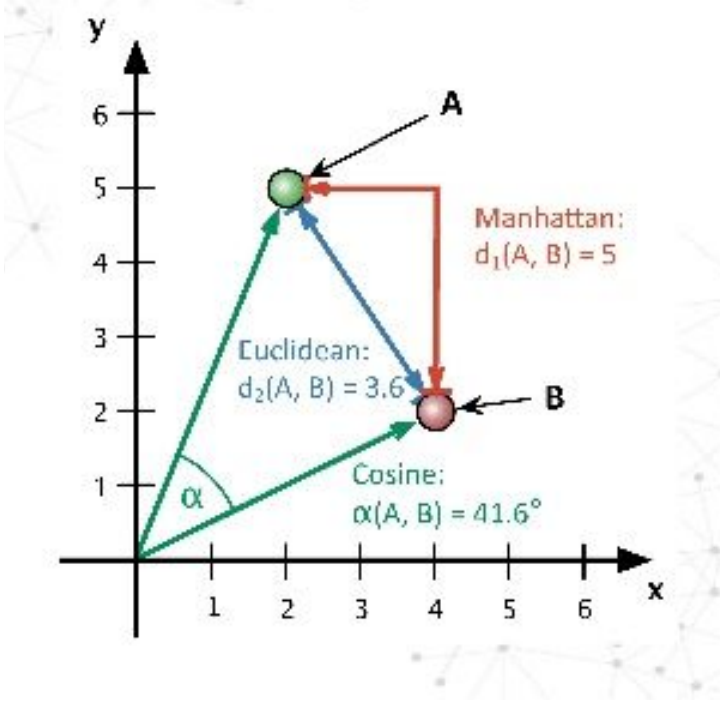


Similarity Measures

- Euclidean distance
- Manhattan distance
- Hamming distance

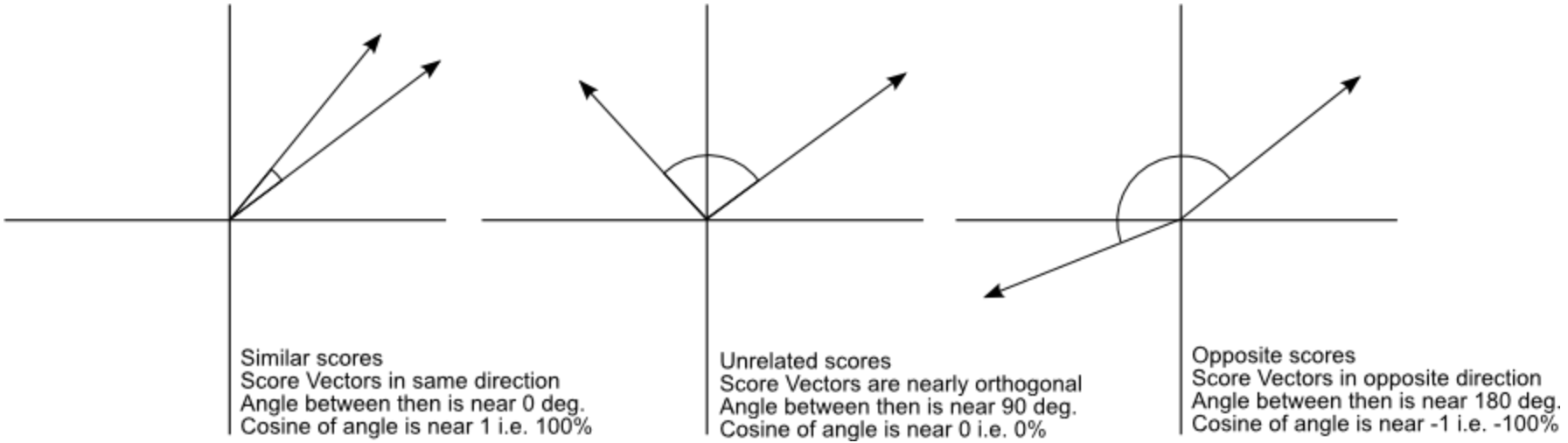
Hamming distance between two strings of equal length is the number of positions at which the corresponding symbols are different

- "karolin" and "kathrin" is 3.
- "karolin" and "kerstin" is 3.
- **1011101** and **1001001** is 2.
- **2173896** and **2233796** is 3.



Cosine distance and similarity

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\|\vec{b}\|}$$



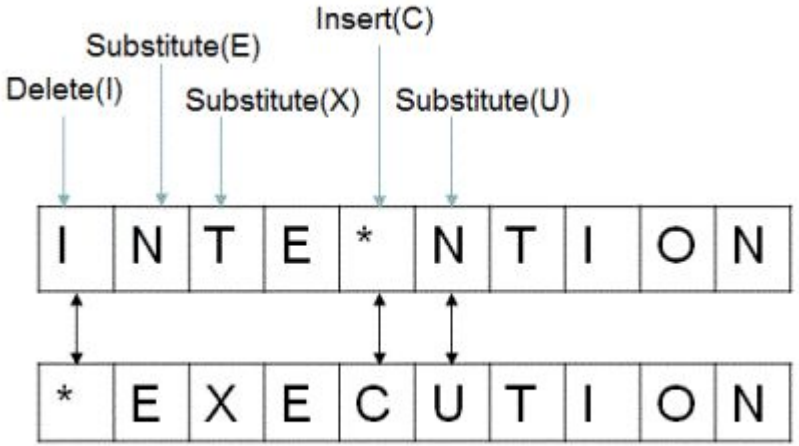
Similar scores
Score Vectors in same direction
Angle between then is near 0 deg.
Cosine of angle is near 1 i.e. 100%

Unrelated scores
Score Vectors are nearly orthogonal
Angle between then is near 90 deg.
Cosine of angle is near 0 i.e. 0%

Opposite scores
Score Vectors in opposite direction
Angle between then is near 180 deg.
Cosine of angle is near -1 i.e. -100%

Levenshtein edit distance

- edit distance: minimum number of operations required to transform one string into the other
- an operation is something like adding/removing or substituting letters from the strings



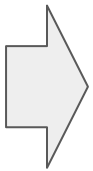
Levenshtein edit distance

- Dynamic programming algorithm

Step	Description
1	Set n to be the length of s. Set m to be the length of t. If n = 0, return m and exit. If m = 0, return n and exit. Construct a matrix containing 0..m rows and 0..n columns.
2	Initialize the first row to 0..n. Initialize the first column to 0..m.
3	Examine each character of s (i from 1 to n).
4	Examine each character of t (j from 1 to m).
5	If s[i] equals t[j], the cost is 0. If s[i] doesn't equal t[j], the cost is 1.
6	Set cell d[i,j] of the matrix equal to the minimum of: a. The cell immediately above plus 1: d[i-1,j] + 1. b. The cell immediately to the left plus 1: d[i,j-1] + 1. c. The cell diagonally above and to the left plus the cost: d[i-1,j-1] + cost.
7	After the iteration steps (3, 4, 5, 6) are complete, the distance is found in cell d[n,m].

$$\operatorname{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \operatorname{lev}_{a,b}(i-1,j) + 1 \\ \operatorname{lev}_{a,b}(i,j-1) + 1 \\ \operatorname{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

		G	U	M	B	O
	0	1	2	3	4	5
G	1					
A	2					
M	3					
B	4					
O	5					
L	6					



		G	U	M	B	O
	0	1	2	3	4	5
G	1	0	1	2	3	4
A	2	1	1	2	3	4
M	3	2	2	1	2	3
B	4	3	3	2	1	2
O	5	4	4	3	2	1
L	6	5	5	4	3	2

https://people.cs.pitt.edu/~kirk/cs1501/Pruhs/Spring2006/assignments/editdistance/Levenshtein%20Distance.htm

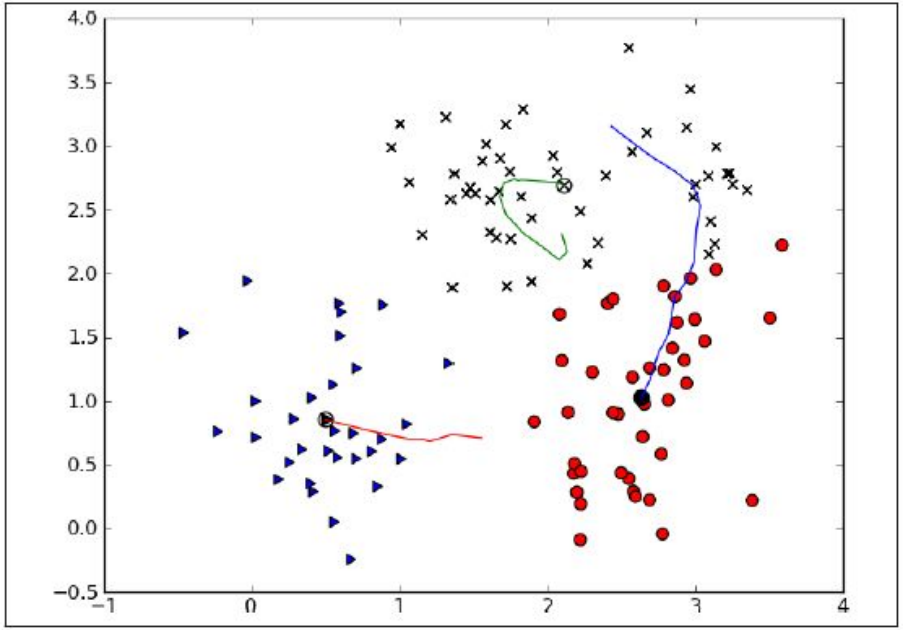
Jaccard Index

- looks for strings whose set of letters match
  - “Dynamo” and “yDnamo” as being identical.
- Jaccard also rates “Dyno” as being a better match than “Dinomo”, because “Dyno” is only four letters long and shares more letters in common.

Jaccard Index	Dynamo	dynamo	yDnamo	Dyno	Dymamo	Dinomo
Dynamo	1.00	0.71	1.00	0.67	0.83	0.57
dynamo	0.71	1.00	0.71	0.43	0.57	0.38
yDnamo	1.00	0.71	1.00	0.67	0.83	0.57
Dyno	0.67	0.43	0.67	1.00	0.50	0.50
Dymamo	0.83	0.57	0.83	0.50	1.00	0.43
Dinomo	0.57	0.38	0.57	0.50	0.43	1.00

k-means algorithm

1. randomly initialize k cluster centroids.
2. **Cluster assignment step** goes through each of the data points and assigns the data points to their closest cluster centroids
3. **Centroid update step** moves the centroids to the average of the points in a cluster
4. Repeat 2-3 until there is no change in the clusters



실습: 문서 clustering

- Data 준비: reuters corpus에서 카테고리 3개를 골라 해당 문서 5개씩을 선택하여 리스트에 저장
- 전처리: 단어로 Tokenization, Stemming Stopword removal, 소문자화
- 각 문서의 tf-idf 벡터 생성
- k-means 알고리즘 구현
- 3개로 클러스터링
- 결과 확인: 각 클러스터에 맞는 문서가 포함되었는지 확인

REUTERS 코퍼스 활용 함수

```
>>> from nltk.corpus import reuters
>>> print(reuters.fileids()[-10:])
 ['training/9982', 'training/9984', 'training/9985', 'training/9988',
'training/9989', 'training/999', 'training/9992', 'training/9993',
'training/9994', 'training/9995']
>>> reuters.categories()
['acq', 'alum', 'barley', 'bop', 'carcass', 'castor-oil', 'cocoa',
'coconut', 'coconut-oil', 'coffee', 'copper', 'copra-cake',
'corn','cotton', 'cotton-oil', 'cpi', 'cpu', 'crude', 'dfl', 'dlr', ...]
>>> reuters.categories(fileids=['test/14829'])
['crude', 'nat-gas']
>>> reuters.sents( categories='acq')
[['SUMITOMO', 'BANK', 'AIMS', 'AT', 'QUICK', 'RECOVERY',
'FROM', 'MERGER', 'Sumitomo', 'Bank', 'Ltd', '&', 'lt', ';', 'SUMI',
'.', 'T', '>', 'is', 'certain', 'to', 'lose', 'its', 'status', 'as', 'Japan', "'", 's',
'most', 'profitable', 'bank', 'as', 'a', 'result', 'of', 'its', 'merger', 'with',
'the', 'Heiwa', 'Sogo', 'Bank', ',', 'financial', 'analysts', 'said', '.'],
['Osaka', '-', 'based', 'Sumitomo', ',', 'with', 'desposits', 'of',
'around', '23', '.', '9', 'trillion', 'yen', ',', 'merged', 'with', 'Heiwa',
'Sogo', ',', 'a', 'small', ',', 'struggling', 'bank', 'with', 'an',
'estimated', '1', '.', '29', 'billion', 'dlrs', 'in', 'unrecoverable', 'loans',
',', 'in', 'October', '.'], ...]
```

Coffee, cotton, crude 카테고리의 문서 3개씩을 실험한 결과

cluster 0 :
sentence 6 : JAPAN TO REVISE LONG-TERM **ENERGY DEMAND DOWNWARDS** The Ministry of International Trade and...
sentence 7 : ENERGY/U.S. PETROCHEMICAL INDUSTRY Cheap oil feedstocks, the weakened U.S. dollar and a plant utilization rate approaching 90 pct...
sentence 8 : TURKEY CALLS FOR DIALOGUE TO SOLVE DISPUTE Turkey said today its disputes with...
cluster 1 :
sentence 0 : COLOMBIA BUSINESS ASKED TO DIVERSIFY FROM COFFEE A Colombia government trade official has.. urged the business community to aggressively diversify its activities and stop relying so heavily on coffee.
sentence 1 : COFFEE COULD DROP TO 70/80 CTS, CARDENAS SAYS International coffee prices could drop to
sentence 2 : COLOMBIA COFFEE REGISTRATIONS REMAIN OPEN Colombia's coffee export registrations
cluster 2 :
sentence 3 : PAKISTAN COTTON CROP SEEN RECORD 7.6 MLN BALES Pakistan is likely to produce a record
sentence 4 : CERTIFICATED COTTON STOCKS Certificated cotton stocks deliverable on the New York Cotton Exchange No 2 cotton futures contract as
sentence 5 : BRAZIL COTTON CROP LOWER -- USDA REPORT Brazil's 1986/87 cotton crop estimate has been reduced to 710,000 from 735,000 tonnes (lint basis),

예제 입력, 출력

```
sentences = ["Nature is beautiful","I like green apples",
        "We should protect the trees","Fruit trees provide fruits",
            "Green apples are tasty", 'Life is beautiful',
            'Pineapples are my favorite fruits']
  nclusters= 3
  clusters = cluster_sentences(sentences, nclusters)
  for cluster in range(nclusters):
          print ("cluster ",cluster,":")
          for i, s in enumerate(clusters[cluster]):
                  print ("\tsentence ",s,": ",sentences[s])
```

```
[result]
cluster  0 :
        sentence  2 :  We should protect the trees
        sentence  3 :  Fruit trees provide fruits
        sentence  6 :  Pineapples are my favorite fruits
cluster  1 :
        sentence  0 :  Nature is beautiful
        sentence  5 :  Life is beautiful
cluster  2 :
        sentence  1 :  I like green apples
        sentence  4 :  Green apples are tasty
```

제출

구현 후, aimecca@skku.edu 로 제출

기한은 내일 실습시간 중 답을 확인하기 전까지. 할 수 있는 데까지 해서 내면 됩니다.