Natural Language Processing Applications

Week 6: Text Similarity

fit@hcmus
KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

- ❑ Introduction to Similarity
- ❑ Text Similarity
- ❑ Similarity Evaluation

fit@hcmus

Applied Natural Language Processing- Text Similarity

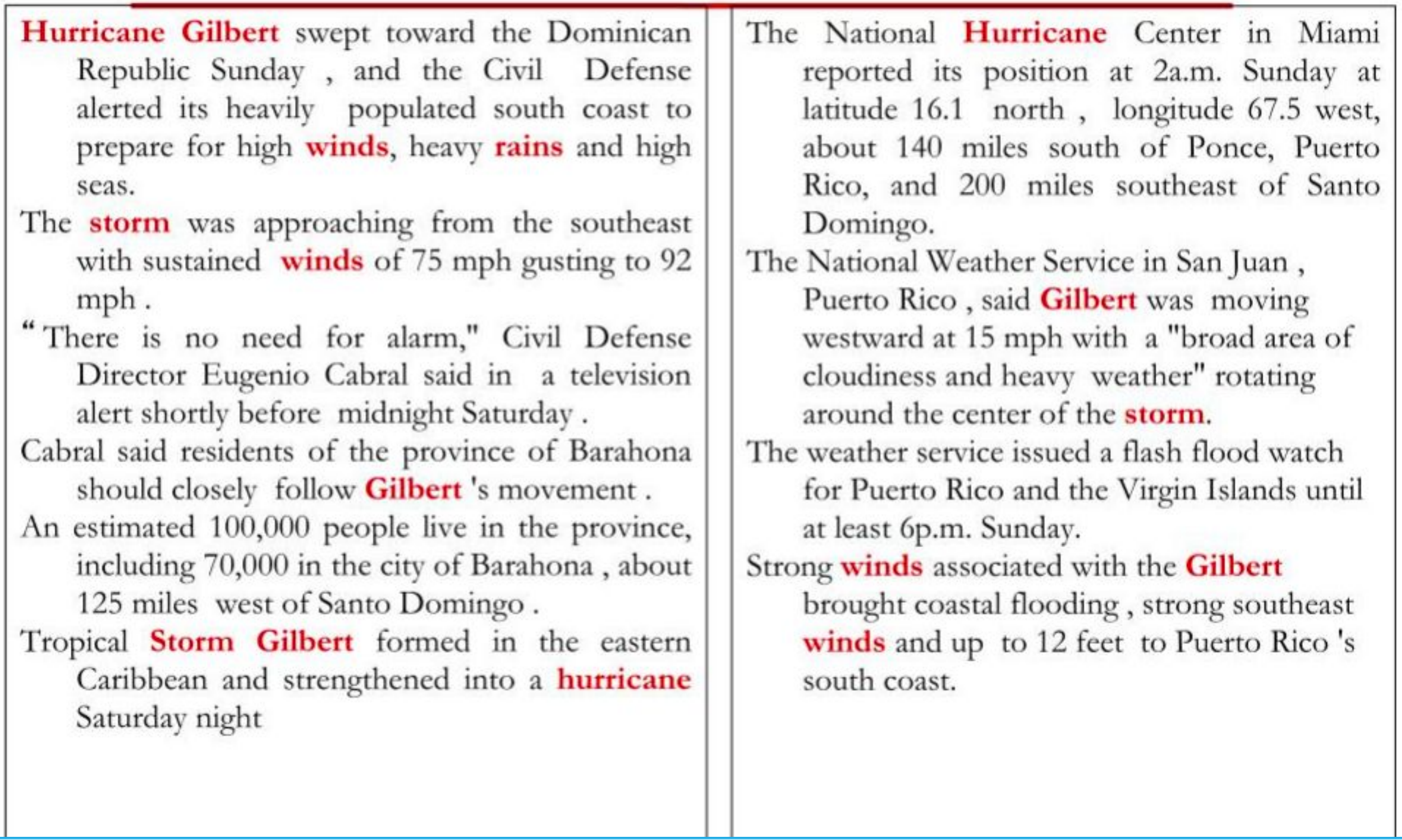
INTRODUTION TO SIMILARITY

Introdution to Similarity

- Word similarity: finding similarities between words is a fundamental part of text similarity
- Words are considered similar if they :
  - Have the same meaning(Synonyms)
  - Have opposite meanings(Antonyms)
  - Are used in the same way(For example: red, green…)
  - Are used in the same context(For example: doctor, hospital…)
  - Are type of another word (For example: fluffy dog, dog, animal…)

Introdution to Similarity(continued)

- Text Similarity:



Hurricane Gilbert swept toward the Dominican Republic Sunday , and the Civil Defense alerted its heavily populated south coast to prepare for high **winds**, heavy **rains** and high seas.

The **storm** was approaching from the southeast with sustained **winds** of 75 mph gusting to 92 mph .

"There is no need for alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly before midnight Saturday .

Cabral said residents of the province of Barahona should closely follow **Gilbert** 's movement .

An estimated 100,000 people live in the province, including 70,000 in the city of Barahona , about 125 miles west of Santo Domingo .

Tropical **Storm Gilbert** formed in the eastern Caribbean and strengthened into a **hurricane** Saturday night

The National **Hurricane** Center in Miami reported its position at 2a.m. Sunday at latitude 16.1 north , longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo.

The National Weather Service in San Juan , Puerto Rico , said **Gilbert** was moving westward at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the **storm**.

The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6p.m. Sunday.

Strong **winds** associated with the **Gilbert** brought coastal flooding , strong southeast **winds** and up to 12 feet to Puerto Rico 's south coast.

Introduction to Similary(continued)

- Text Similary:
  - Document1
    - Gilbert: 3
    - Hurricane: 2
    - Rains: 1
    - Storm: 2
    - Winds: 2
  - Document2
    - Gilbert: 2
    - Hurricane: 1
    - Rains: 0
    - Storm: 1
    - Winds: 2

Cosine similarity: 0.9439

Introduction to Similary(continued)

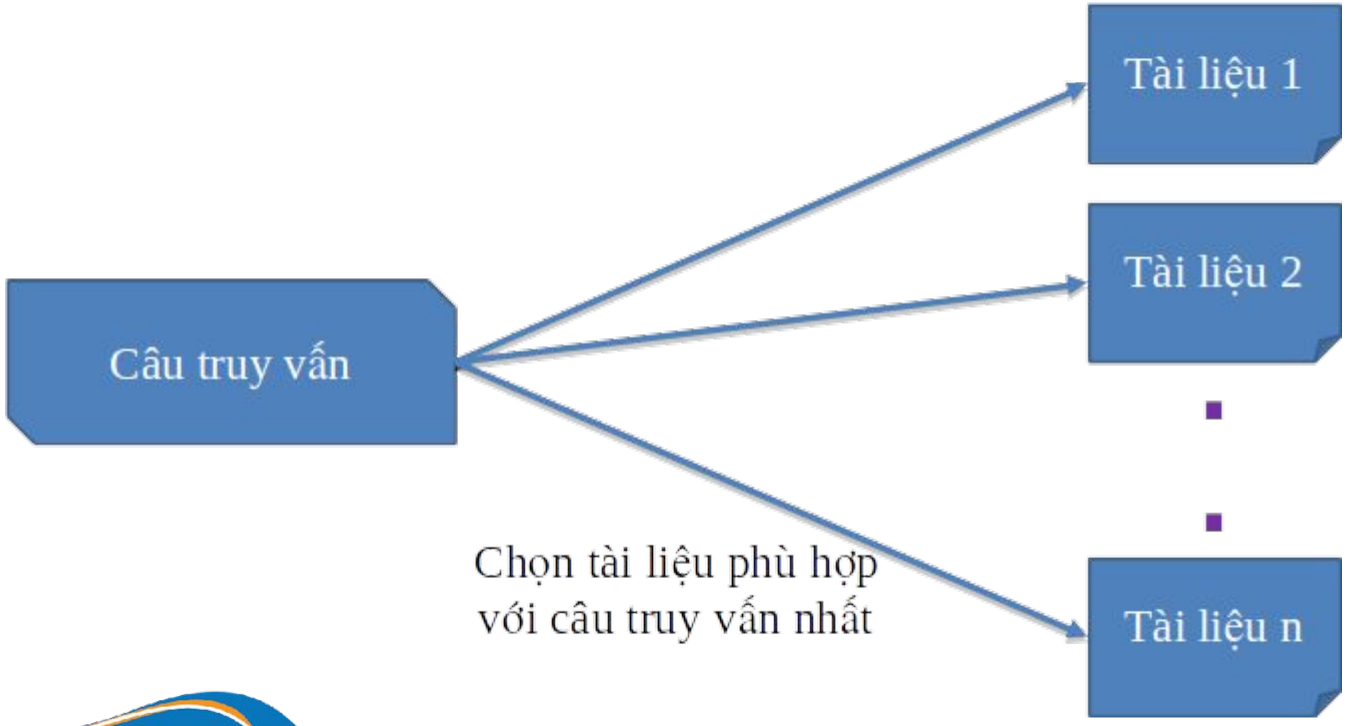
- According to John Philip McCrae:
  - “Semantic textual similarity is the task of deciding if two sentences express a similar or identical meaning and requires a deep understanding of a sentence and its meaning in order to achieve high performance.”

Introduction to Similarity(continued)

- Applications of Text Similarity:
    - Information Retrieval
    - Text Summarization
    - Machine Translation
    - Plagiarism Detection
    - ...

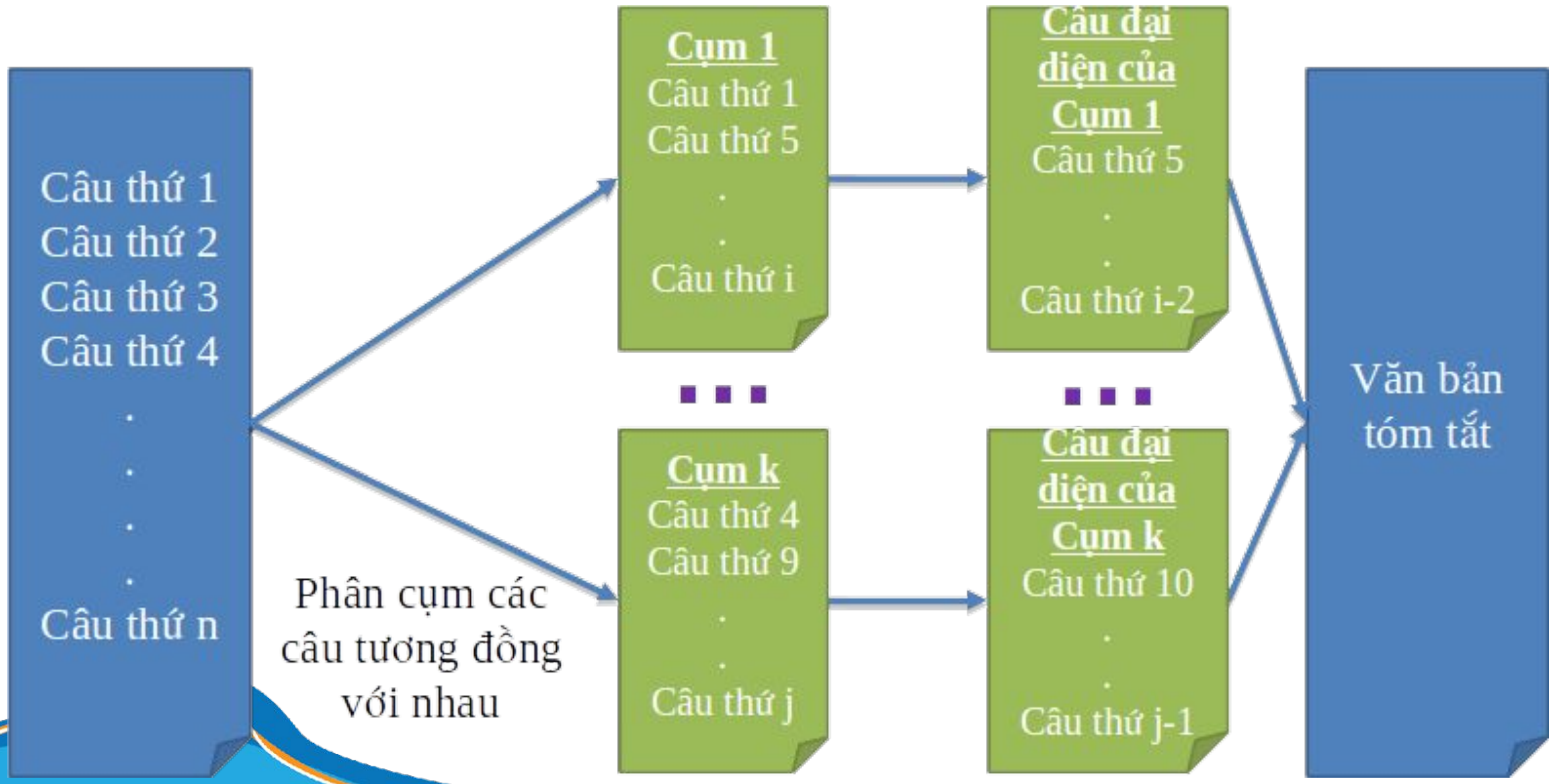
Introduction to Similarity(continued)

- Information Retrieval



Introduction to Similary(continued)

Text Summarization



Câu thứ 1
Câu thứ 2
Câu thứ 3
Câu thứ 4
.
.
.
.
Câu thứ n

Phân cụm các câu tương đồng với nhau

Cụm 1
Câu thứ 1
Câu thứ 5
.
.
Câu thứ i

...

Cụm k
Câu thứ 4
Câu thứ 9
.
.
Câu thứ j

Câu đại diện của Cụm 1
Câu thứ 5
.
.
Câu thứ i-2

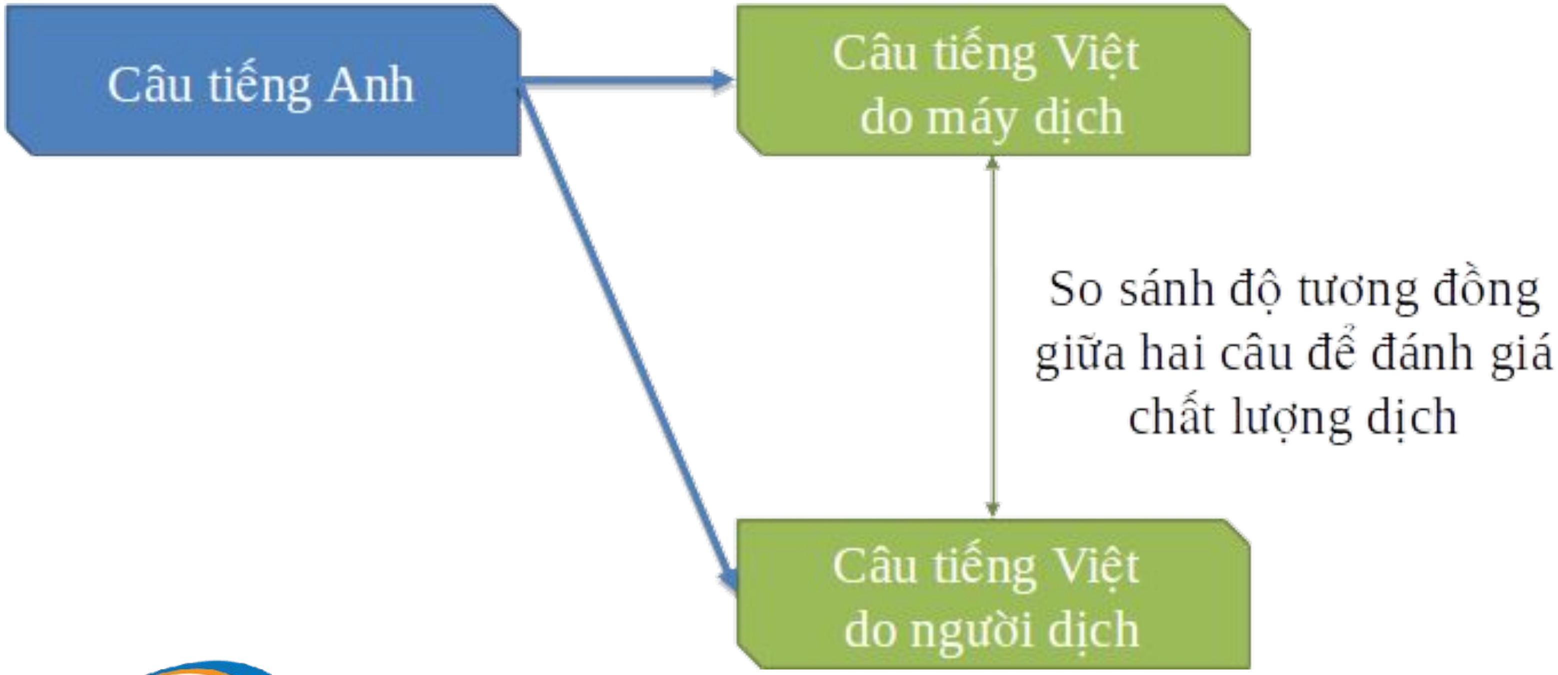
...

Câu đại diện của Cụm k
Câu thứ 10
.
.
Câu thứ j-1

Văn bản tóm tắt

Introduction to Similarity(continued)

❑ Machine Translation (Quality evaluation)



Câu tiếng Anh

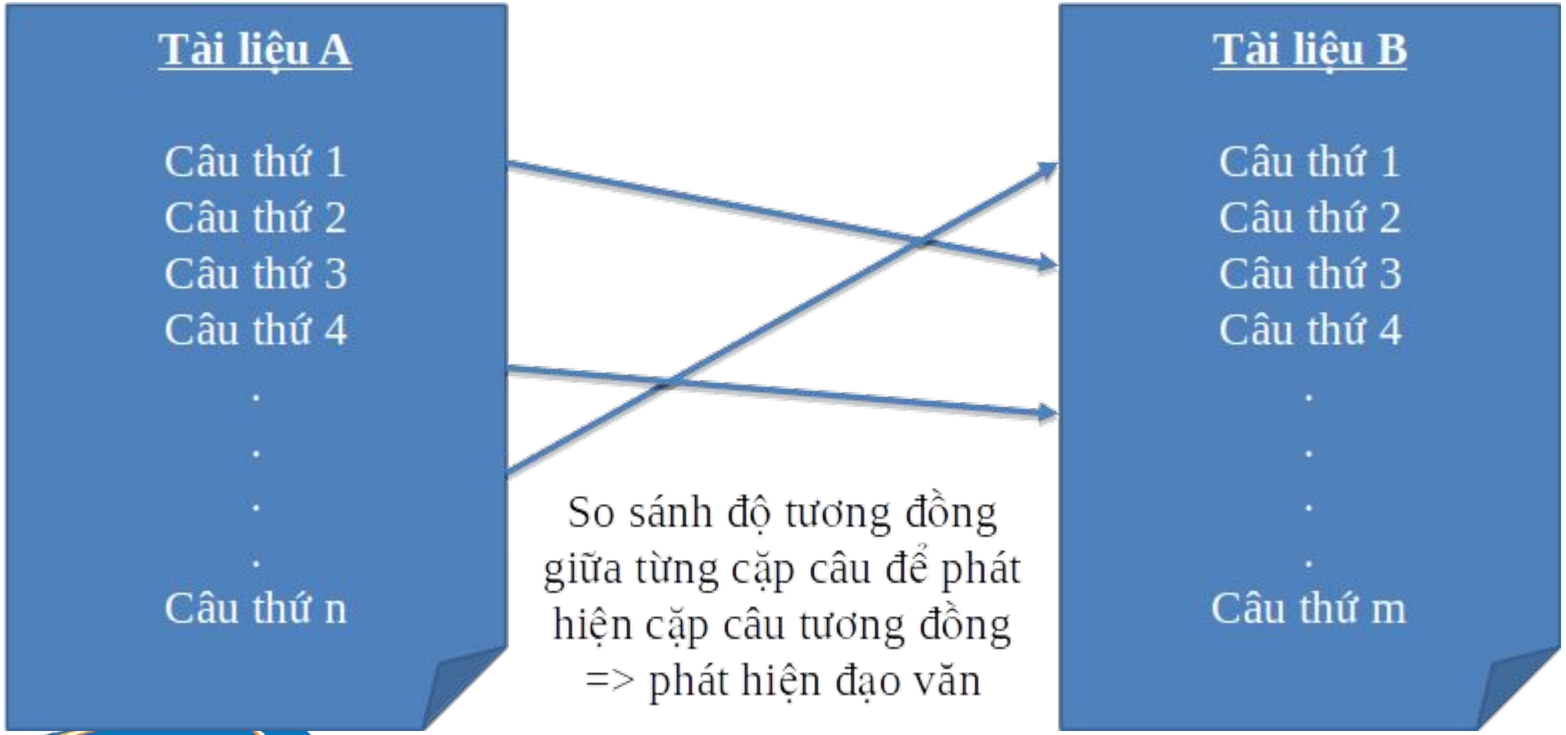
Câu tiếng Việt do máy dịch

So sánh độ tương đồng giữa hai câu để đánh giá chất lượng dịch

Câu tiếng Việt do người dịch

Introduction to Similarity(continued)

- Plagiarism Detection



Tài liệu A

Câu thứ 1
Câu thứ 2
Câu thứ 3
Câu thứ 4
.
.
.
.
Câu thứ n

Tài liệu B

Câu thứ 1
Câu thứ 2
Câu thứ 3
Câu thứ 4
.
.
.
.
Câu thứ m

So sánh độ tương đồng giữa từng cặp câu để phát hiện cặp câu tương đồng => phát hiện đạo văn

fit@hcmus

Introduction to Similary(continued)

- There are 2 popular scales :
  - Scale 0 to 1 is used in detecting rewrites text while retaining full meaning(paraphrase identification)
  - Scale 0 to 5:

Introduction to Similary(continued)

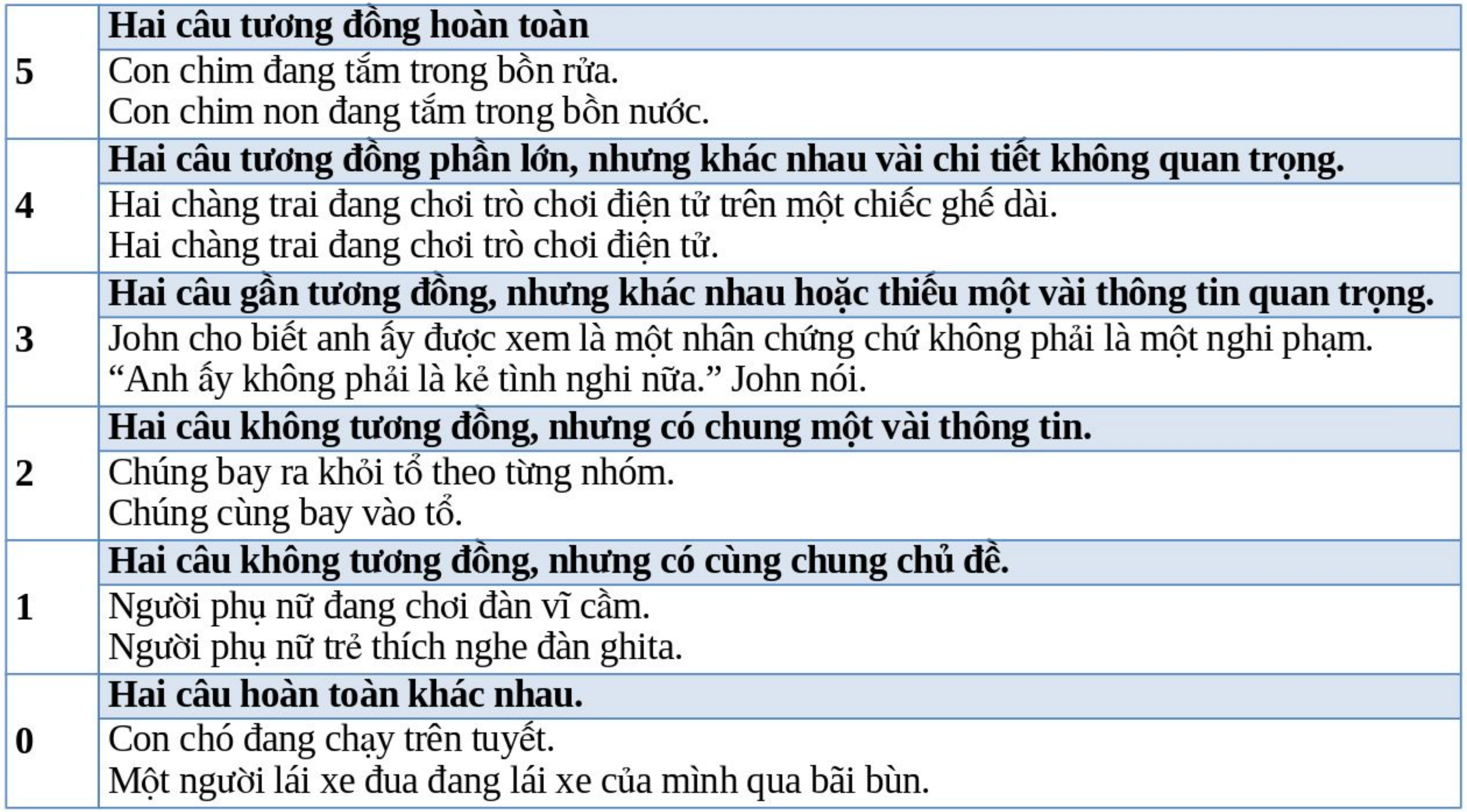
Scale 0 to 1:

- The following pair of sentences is labeled as 1
  - Sentance 1: Customers will have to use a decoder card from the cable TV provider to plug in the set.
  - Sentance 2: To watch pay TV, customers will plug in the television set a decoder card provided by the cable TV provider.
- And the following pair of sentences is labeled as 0
  - Sentance 1: With an interpreter like you everything will be fine.
  - Sentance 2: Today's interpreter is Mr. Nam.

fit@hcmus

Introduction to Similary(continued)

- Scale 0 to 5:



5	**Hai câu tương đồng hoàn toàn**
	Con chim đang tắm trong bồn rửa. Con chim non đang tắm trong bồn nước.
4	**Hai câu tương đồng phần lớn, nhưng khác nhau vài chi tiết không quan trọng.**
	Hai chàng trai đang chơi trò chơi điện tử trên một chiếc ghế dài. Hai chàng trai đang chơi trò chơi điện tử.
3	**Hai câu gần tương đồng, nhưng khác nhau hoặc thiếu một vài thông tin quan trọng.**
	John cho biết anh ấy được xem là một nhân chứng chứ không phải là một nghi phạm. “Anh ấy không phải là kẻ tình nghi nữa.” John nói.
2	**Hai câu không tương đồng, nhưng có chung một vài thông tin.**
	Chúng bay ra khỏi tổ theo từng nhóm. Chúng cùng bay vào tổ.
1	**Hai câu không tương đồng, nhưng có cùng chung chủ đề.**
	Người phụ nữ đang chơi đàn vĩ cầm. Người phụ nữ trẻ thích nghe đàn ghita.
0	**Hai câu hoàn toàn khác nhau.**
	Con chó đang chạy trên tuyết. Một người lái xe đua đang lái xe của mình qua bãi bùn.

Applied Natural Language Processing- Text Similarity

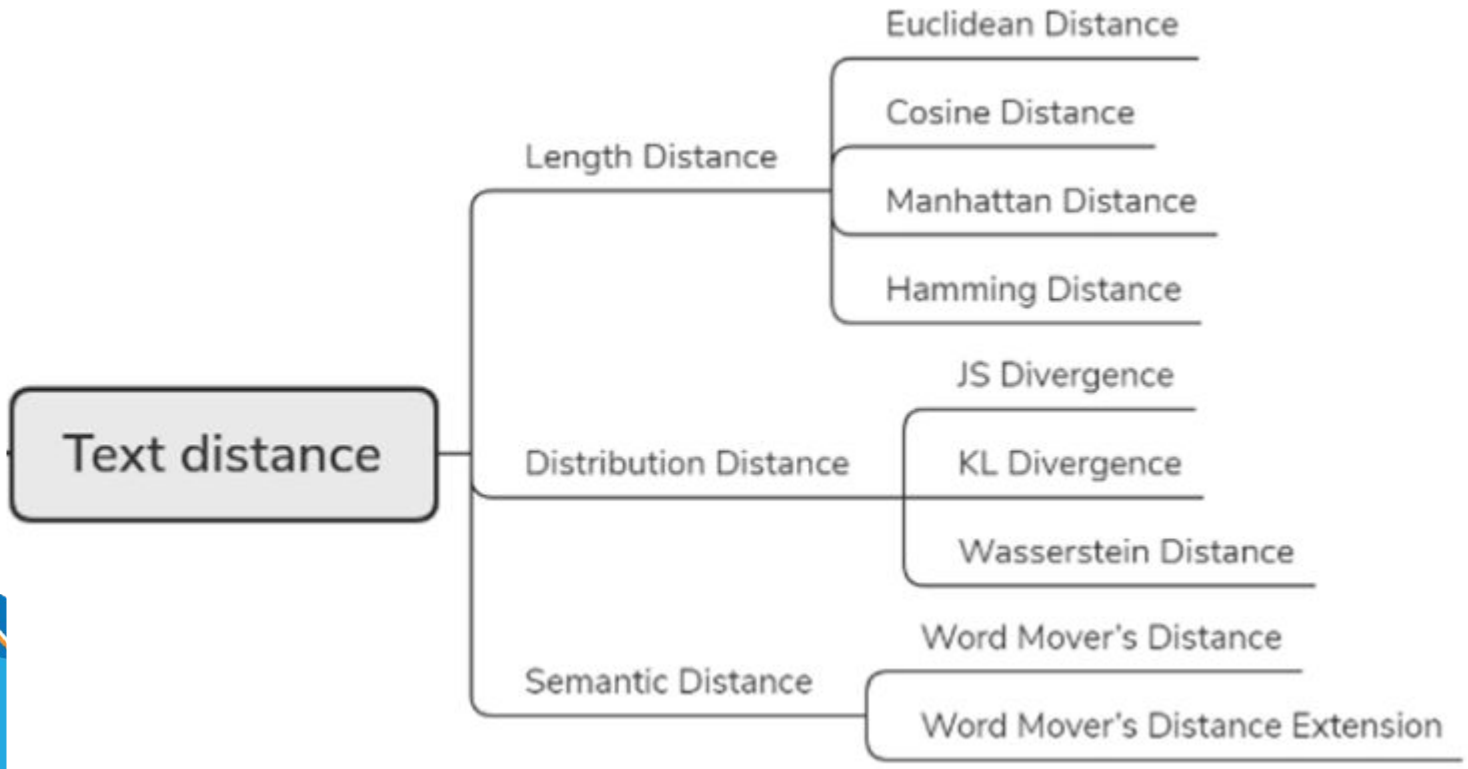
TEXT SIMILARITY METHODS

Text Similarity methods

- Text Distance
- Text Representation

Text Similarity methods

- Text distance: describes the semantic proximity of two text words from the perspective of distance

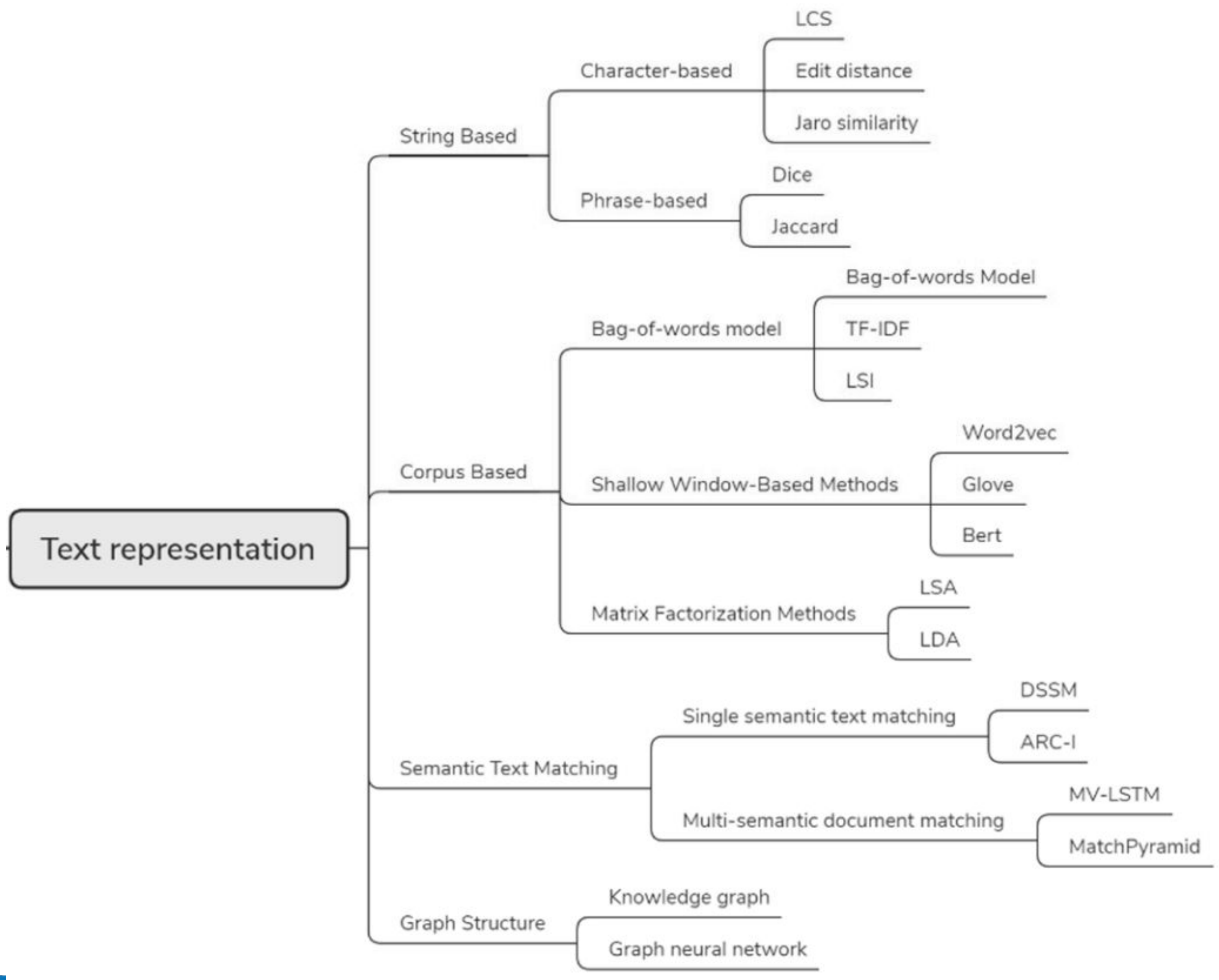


Text distance

- Length Distance
  - Euclidean Distance
  - Cosine Distance
  - Manhattan Distance
  - Hamming Distance
- Distribution Distance
  - JS Divergence
  - KL Divergence
  - Wasserstein Distance
- Semantic Distance
  - Word Mover's Distance
  - Word Mover's Distance Extension

Text Similarity methods

- Text representation: numerically represent the unstructured text documents to make them mathematically computable.



- Text representation
  - String Based
    - Character-based
      - LCS
      - Edit distance
      - Jaro similarity
    - Phrase-based
      - Dice
      - Jaccard
  - Corpus Based
    - Bag-of-words model
      - Bag-of-words Model
      - TF-IDF
      - LSI
    - Shallow Window-Based Methods
      - Word2vec
      - Glove
      - Bert
    - Matrix Factorization Methods
      - LSA
      - LDA
  - Semantic Text Matching
    - Single semantic text matching
      - DSSM
      - ARC-I
    - Multi-semantic document matching
      - MV-LSTM
      - MatchPyramid
  - Graph Structure
    - Knowledge graph
    - Graph neural network

Text Distance

- Main methods:
  - Length Distance
  - Distribution Distance
  - Semantic Distance

Text Distance

- Length Distance:
  - Euclidean Distance

$$d(S_a, S_b) = \sqrt{\sum_{i=1}^{n} \left(S_a^{(i)} - S_b^{(i)}\right)^2}$$

  - Cosine Distance

$$\mathrm{Sim}(S_a, S_b) = \cos\Theta = \frac{\vec{S_a}\cdot\vec{S_b}}{||S_a||\cdot||S_b||}$$

  - Manhattan Distance

$$\mathrm{Sim}(\mathrm{x},\ \mathrm{y}) = |x_1 - x_2| + |y_1 - y_2|$$

Text Distance

- Length Distance:
  - Suitable for symmetrical problems
    - Sim(A, B) = Sim(B, A)
    - => But for question Q to retrieve answer A, the corresponding similarity is not symmetrical.
  - Lack of statistical characteristics of the data

Text Distance

- Distribution Distance:
  - Kullback–Leibler Divergence

$$\mathrm{d}(\mathrm{p}||\mathrm{q}) = \sum_{i=1}^{n} p(x) log \frac{p(x)}{q(x)}$$

  - Jensen–Shannon Divergence

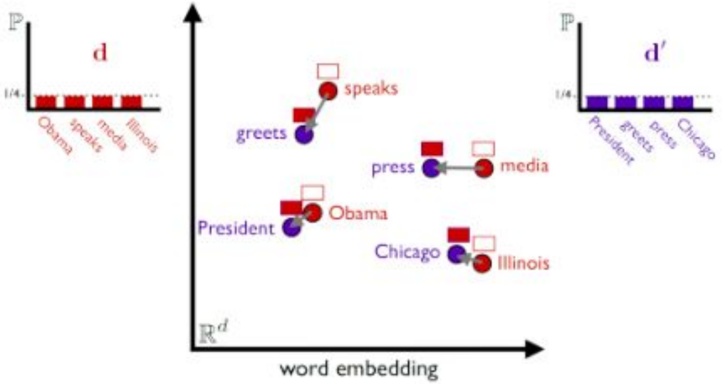
$$\mathrm{JS}(P_1||P_2) = \frac{1}{2}KL(P_1||\frac{P_1+P_2}{2}) + \frac{1}{2}KL(P_2||\frac{P_1+P_2}{2})$$

  - Wasserstein Distance

$$W(p_r, p_g) = \inf_{\gamma\sim\prod(p_r,p_g)} E_{(x,y)\sim\gamma}[||x-y||]$$

Text Distance

- Semantic Distance :
  - Word Mover’s Distance



  - Word Mover’s Distance Extension
    - Use the Mahalanobis distance instead of the Euclidean distance

Text representation

- String-based:
  - Operate on string sequences and character composition
  - Includes 2 methods:
    - Character-based
    - Phrase-based

Text representation

- String based:
  - Longest common substring:

$$\mathrm{LCS}(S_a, S_b) = \begin{cases} 0,\ if\ S_a = 0\ or\ S_a = 0 \\ 1 + LCS(S_a - 1, S_b - 1), if\ x[S_a] == y[S_b] \\ max\begin{cases} LCS(S_a, S_b - 1) \\ LCS(S_a - 1, S_b) \end{cases} if\ x[S_a] \neq y[S_b] \end{cases}$$

  - Jaro Similarity:

$$\mathrm{Sim} = \begin{cases} 0, if\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|S_a|} + \frac{m}{|S_b|} + \frac{m-t}{m}\right) \end{cases}$$

Text representation

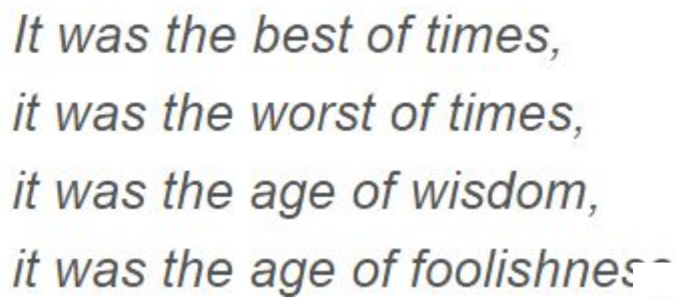
- String based:
    - Edit Distance:
        - Levenshtein distance (L distance)
            - The minimum number of single-character edits (insertions, deletions or substitutions) required to change S1 into S2.
        - Damerau–Levenshtein Distance (D distance)
            - Like L distance, but with the addition of the transposition operation
        - Optimal String Alignment
            - Like D distance
            - No substring/subsequence is edited more than once.

Text representation

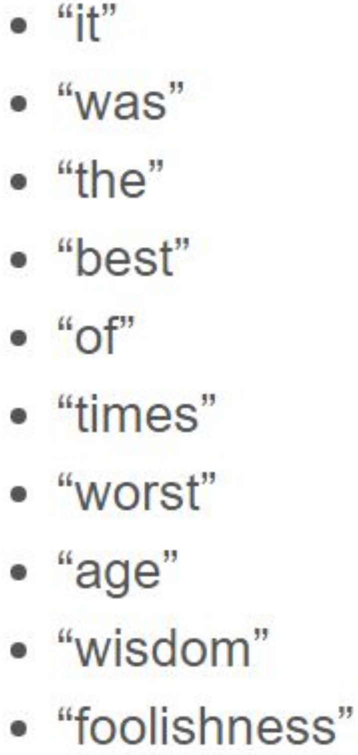
- Corpus based:
  - Use data from corpus
    - Textual feature
    - Co-occurrence probability
  - Includes 3 methods
    - Bag-of-words
    - Distributed representation
    - Matrix factorization

Text representation

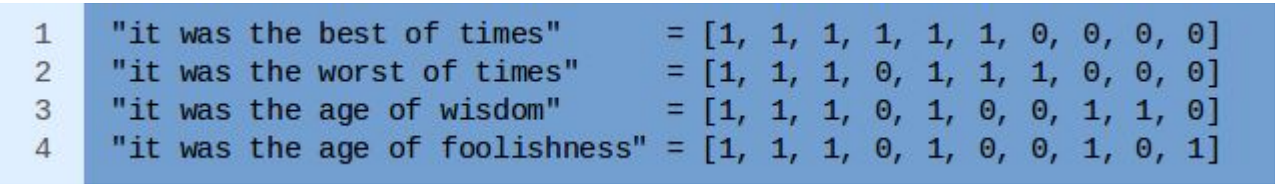
- BOW:
  - Count the number of times each word appears



It was the best of times,
it was the worst of times,
it was the age of wisdom,
it was the age of foolishnes



- “it”
- “was”
- “the”
- “best”
- “of”
- “times”
- “worst”
- “age”
- “wisdom”
- “foolishness”



```
"it was the best of times"     = [1, 1, 1, 1, 1, 1, 0, 0, 0, 0]
"it was the worst of times"    = [1, 1, 1, 0, 1, 1, 1, 0, 0, 0]
"it was the age of wisdom"     = [1, 1, 1, 0, 1, 0, 0, 1, 1, 0]
"it was the age of foolishness" = [1, 1, 1, 0, 1, 0, 0, 1, 0, 1]
```

Text representation

- Term frequency–inverse document frequency (TF - IDF):
    - To measure how important a word is to a document in a collection (or corpus) of documents
        - TF: the ratio of a word's occurrence in a document

$$tf\ (w, d) = Freq\ (w, d)$$

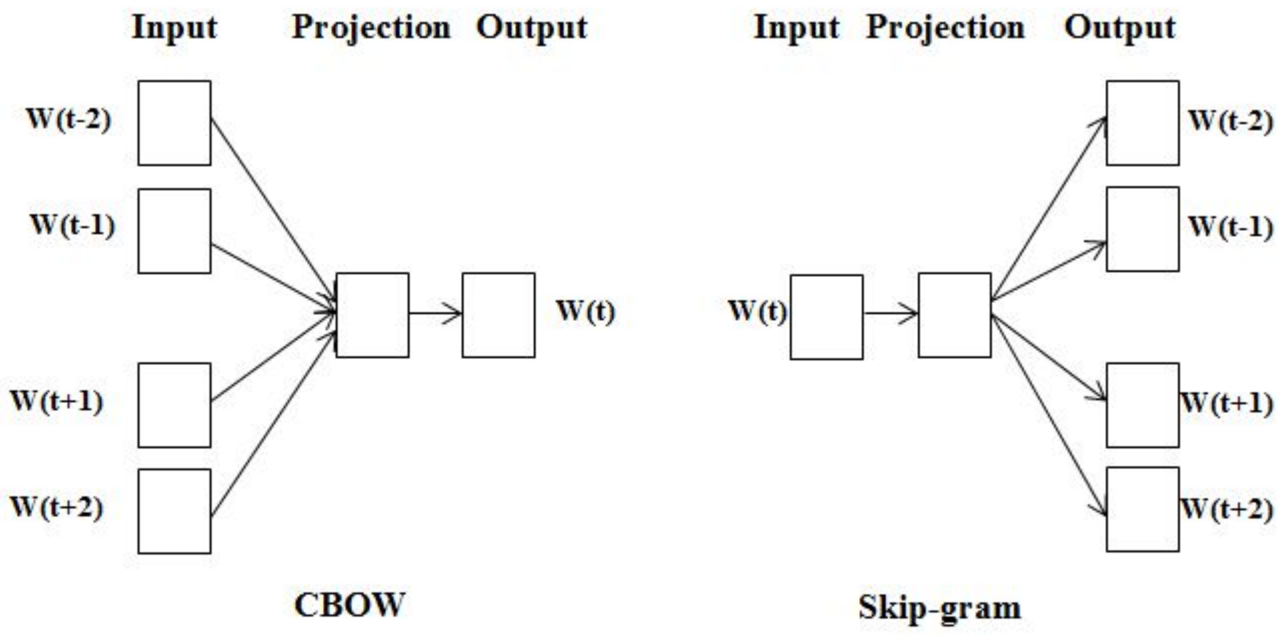
        - IDF: indicates the amount of information provided

$$idf\ (w, D) = log \frac{|D|}{N(w)}$$

$$tf\text{-}idf\ (w, d, D) = tf\ (w, d) \times idf\ (w, D)$$

Text representation

- word2vec:
  - Continuous Bag-of-words (BOW)
  - Word-skip grams (skip-gram)



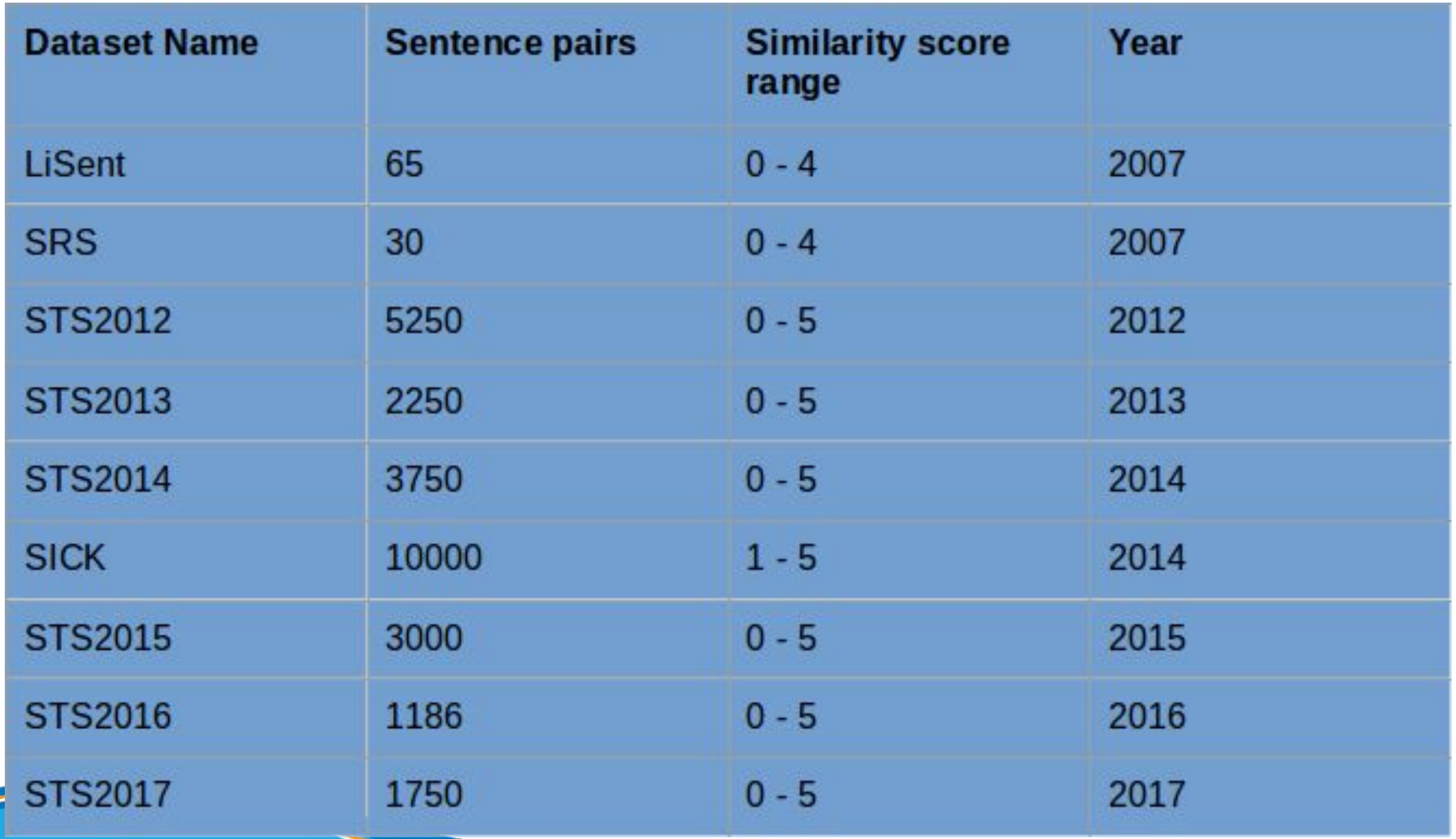
Text representation

- Glove:
  - Words with similar meanings tend to appear in similar contexts
  - Encode the ratios of co-occurrence probabilities with vector differences
- Bert: Bidirectional Encoder Representations from Transformers
  - Already pre-trained on massive datasets

Corpus

- STS Benchmark (STSb)
  - Includes 8,628 sentence pairs :
    - Train: 5,749
    - Develop: 1,500
    - Test: 1,379
  - Three categories : captions, news, and forums

Corpus



Dataset Name	Sentence pairs	Similarity score range	Year
LiSent	65	0 - 4	2007
SRS	30	0 - 4	2007
STS2012	5250	0 - 5	2012
STS2013	2250	0 - 5	2013
STS2014	3750	0 - 5	2014
SICK	10000	1 - 5	2014
STS2015	3000	0 - 5	2015
STS2016	1186	0 - 5	2016
STS2017	1750	0 - 5	2017

Applied Natural Language Processing- Text Similarity

EVALUATING

Similarity Evaluation

- Pearson correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$