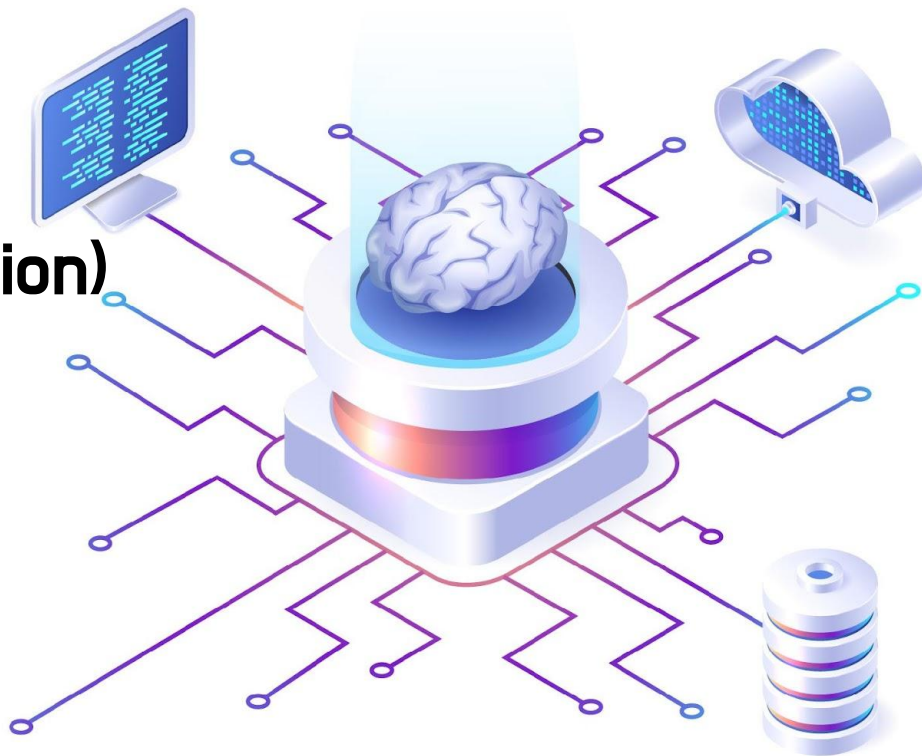


문서 요약 (Document Summarization)

실무형 인공지능 자연어처리



문서 요약 (Document Summarization)

통계기반 자연어 처리

2

TextRank를 활용한 문서 요약



4 Sentence Extraction

The other TextRank application that we investigate consists of sentence extraction for automatic summarization. In a way, the problem of sentence extraction can be regarded as similar to keyword extraction, since both applications aim at identifying sequences that are more “representative” for the given text. In keyword extraction, the candidate text units consist of words or phrases, whereas in sentence extraction, we deal with entire sentences. TextRank turns out to be well suited for this type of applications, since it allows for a ranking over text units that is recursively computed based on information drawn from the entire text.

- 문장을 추출하는 것은 키워드 추출과 유사
- 두 방법 (키워드 추출, 문장 추출) 모두 텍스트를 대표하는(representative) 시퀀스를 식별하는 것을 목표로 함

4.1 TextRank for Sentence Extraction

The co-occurrence relation used for keyword extraction cannot be applied here, since the text units in consideration are significantly larger than one or few words, and “co-occurrence” is not a meaningful relation for such large contexts. Instead, we are defining a different relation, which determines a connection between two sentences if there is a “similarity” relation between them, where “similarity” is measured as a function of their content overlap. Such a relation between two sentences can be seen as a process of “recommendation”

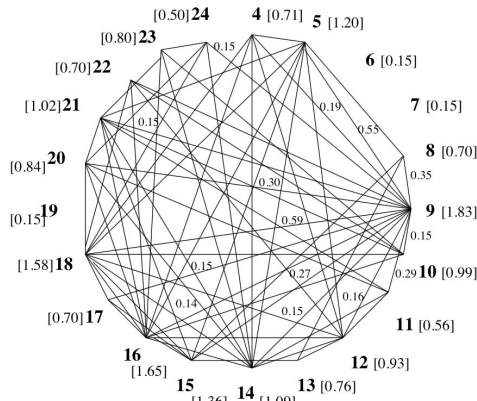
$$\text{Similarity}(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{\log(|S_i|) + \log(|S_j|)}$$

The resulting graph is highly connected, with a weight associated with each edge, indicating the strength of the connections established between various sentence pairs in the text. The text is therefore represented as a weighted graph...After the ranking algorithm is run on the graph, sentences are sorted in reversed order of their score, and the top ranked sentences are selected for inclusion in the summary.

- 문장내 co-occurrence에 기반한 관계 정의는 적용할 수 없음 (문장이기 때문에)
- 문장 간 “유사성”이 있는 경우 connection 있다고 정의. 유사성은 content overlap 함수로 측정
- 텍스트 내 다양한 문장사이의 관계 강도가 결정되고, 이를 역순으로 정렬하여 텍스트를 요약

4.1 TextRank for Sentence Extraction

- 3: BC--Hurricane Gilbert, 09-11 339
- 4: BC--Hurricane Gilbert, 0348
- 5: Hurricane Gilbert heads toward Dominican Coast
- 6: By Ruddy Gonzalez
- 7: Associated Press Writer
- 8: Santo Domingo, Dominican Republic (AP)
- 9: 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.
- 10: The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph.
- 11: "There is no need for alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly after midnight Saturday.
- 12: Cabral said residents of the province of Barahona should closely follow Gilbert's movement.
- 13: An estimated 100,000 people live in the province, including 70,000 in the city of Barahona, about 125 miles west of Santo Domingo.
- 14: Tropical storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.
- 15: The National Hurricane Center in Miami reported its position at 2 a.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.
- 16: 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.
- 17: The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday.
- 18: Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds, and up to 12 feet to Puerto Rico's south coast.
- 19: There were no reports on casualties.
- 20: San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsided during the night.
- 21: On Saturday, Hurricane Florence was downgraded to a tropical storm, and its remnants pushed inland from the U.S. Gulf Coast.
- 22: Residents returned home, happy to find little damage from 90 mph winds and sheets of rain.
- 23: Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane.
- 24: The first, Debby, reached minimal hurricane strength briefly before hitting the Mexican coast last month.



TextRank extractive summary

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 National Hurricane Center in Miami reported its position at 2 a.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. Strong winds associated with Gilbert brought coastal flooding, strong southeast winds and up to 12 feet to Puerto Rico's south coast.

Manual abstract I

Hurricane Gilbert is moving toward the Dominican Republic, where the residents of the south coast, especially the Barahona Province, have been alerted to prepare for heavy rains, and high wind and seas. Tropical storm Gilbert formed in the eastern Caribbean and became a hurricane on Saturday night. By 2 a.m. Sunday it was about 200 miles southeast of Santo Domingo and moving westward at 15 mph with winds of 75 mph. Flooding is expected in Puerto Rico and in the Virgin Islands. The second hurricane of the season, Florence, is now over the southern United States and downgraded to a tropical storm.

Manual abstract II

Tropical storm Gilbert in the eastern Caribbean strengthened into a hurricane Saturday night. The National Hurricane Center in Miami reported its position at 2 a.m. Sunday to be about 140 miles south of Puerto Rico and 200 miles southeast of Santo Domingo. It is moving westward at 15 mph with a broad area of cloudiness and heavy weather with sustained winds of 75 mph gusting to 92 mph. The Dominican Republic's Civil Defense alerted that country's heavily populated south coast and the National Weather Service in San Juan, Puerto Rico issued a flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday.

5 Why TextRank Works

Intuitively, TextRank works well because it does not only rely on the local context of a text unit (vertex), but rather it takes into account information recursively drawn from the entire text (graph) ... The sentences that are highly recommended by other sentences in the text are likely to be more informative for the given text, and will be therefore given a higher score... Through its iterative mechanism, TextRank goes beyond simple graph connectivity, and it is able to score text units based also on the “importance” of other text units they link to. The text units selected by TextRank for a given application are the ones most recommended by related text units in the text, with preference given to the recommendations made by most influential ones, i.e. the ones that are in turn highly recommended by other related units. The underlying hypothesis is that in a cohesive text fragment, related text units tend to form a “Web” of connections that approximates the model humans build about a given context in the process of discourse understanding.

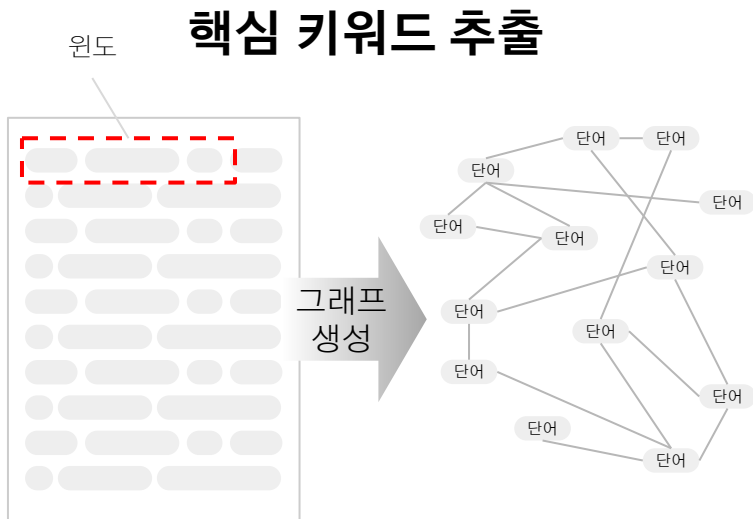
- 텍스트 단위(버텍스)의 로컬 컨텍스트를 고려할뿐만 아니라, 전체 텍스트(그래프)에서 재귀적으로 정보를 고려하기 때문에 TextRank가 잘 작동
- 링크하는 다른 텍스트 단위의 "중요성"을 바탕으로 텍스트 단위를 평가

6 Conclusions

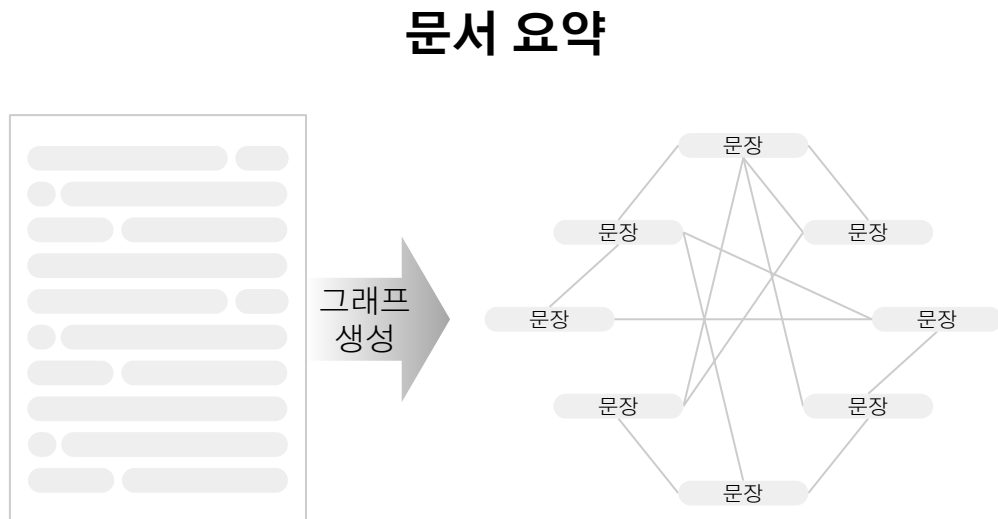
In this paper, we introduced TextRank – a graph- based ranking model for text processing, and show how it can be successfully used for natural language applications. In particular, we proposed and evaluated two innovative unsupervised approaches for keyword and sentence extraction, and showed that the accuracy achieved by TextRank in these applications is competitive with that of previously proposed state-of-the-art algorithms. An important aspect of TextRank is that it does not require deep linguistic knowledge, nor domain or language specific annotated corpora, which makes it highly portable to other domains, genres, or languages.

- TextRank는 깊은 언어지식이나 도메인 별 corpora를 필요로 하지 않고, 다른 도메인, 장르, 언어에 적용할 수 있음

키워드 추출 vs 문서요약



원도가 이동하며
그래프 생성



모든 문장간 유사도를 기준으로
그래프 생성

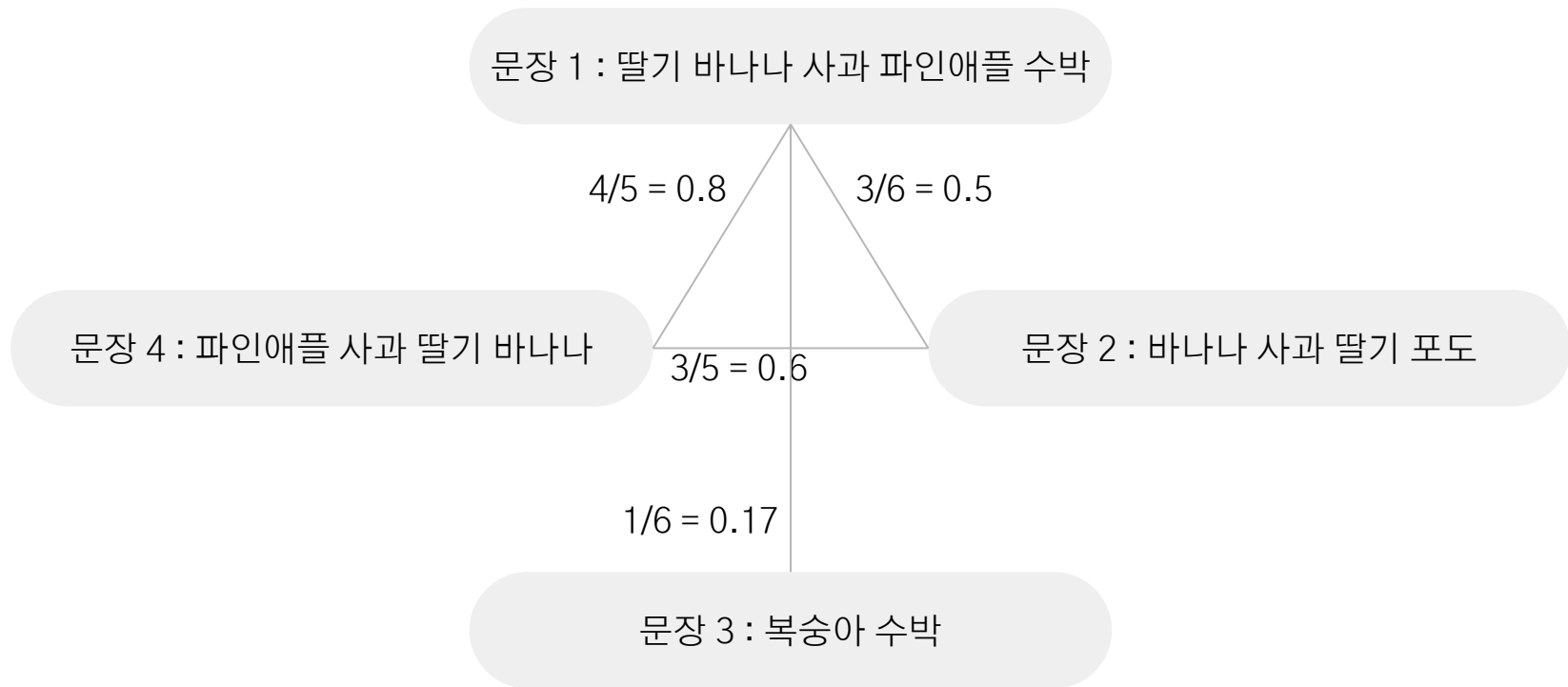
TextRank 과정 : 그래프 생성

딸기 바나나 사과 파인애플 수박. 바나나 사과 딸기 포도. 복숭아 수박.
파인애플 사과 딸기 바나나.

문장 토큰화

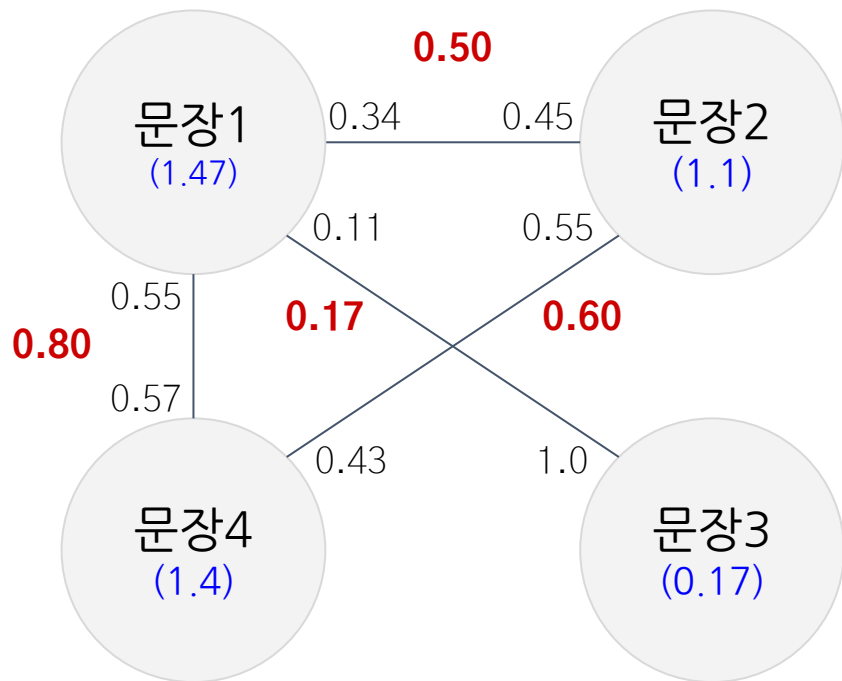
문장 1 : 딸기 바나나 사과 파인애플 수박
문장 2 : 바나나 사과 딸기 포도
문장 3 : 복숭아 수박
문장 4 : 파인애플 사과 딸기 바나나

Textrank 과정 : 그래프 생성



TextRank 과정 : 그래프 생성

$$0.5 / (0.5 + 0.6) = 0.45$$



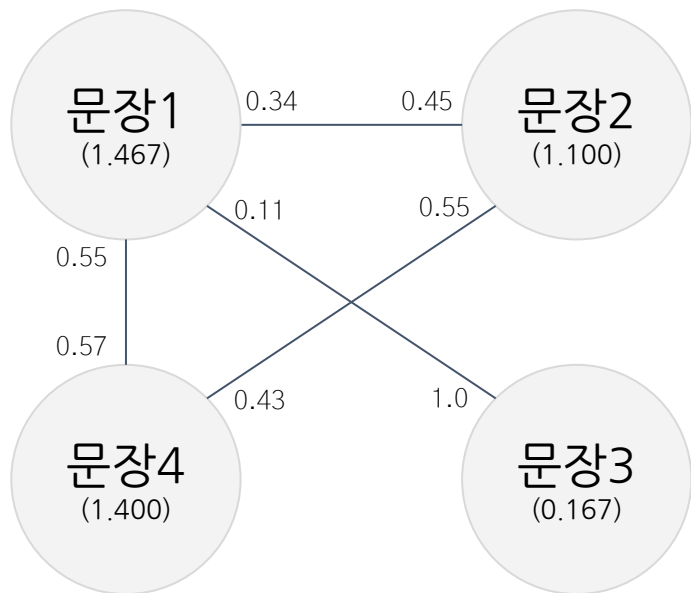
문장간 유사도

| | |
|-----------|------|
| 문장1 - 문장2 | 0.50 |
| 문장1 - 문장3 | 0.17 |
| 문장1 - 문장4 | 0.80 |
| 문장2 - 문장4 | 0.60 |

엣지 가중치

| | 문장1 | 문장2 | 문장3 | 문장4 |
|-----|------|------|------|------|
| 문장1 | 0.00 | 0.34 | 0.11 | 0.55 |
| 문장2 | 0.45 | 0.00 | 0.00 | 0.55 |
| 문장3 | 1.00 | 0.00 | 0.00 | 0.00 |
| 문장4 | 0.57 | 0.43 | 0.00 | 0.00 |

Textrank 과정 : 행렬로 계산하기



| 노드 | 최초 스코어 |
|-----|--------|
| 문장1 | 1.467 |
| 문장2 | 1.100 |
| 문장3 | 0.167 |
| 문장4 | 1.400 |

$$S(\text{문장1}) = (1-0.85) + 0.85 \times (0.5 + 0.167 + 0.8) = \mathbf{1.397}$$

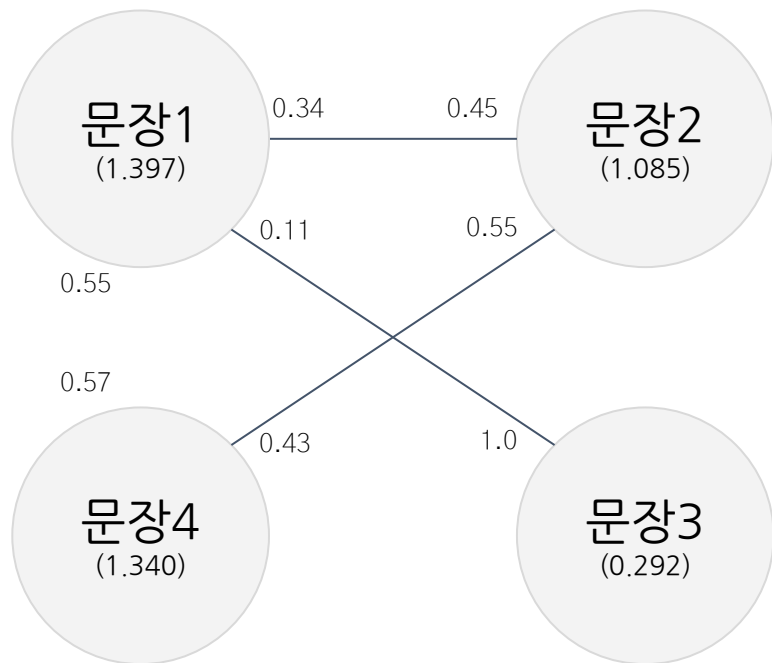
$$S(\text{문장2}) = (1-0.85) + 0.85 \times (0.5 + 0.6) = \mathbf{1.085}$$

$$S(\text{문장3}) = (1-0.85) + 0.85 \times (0.167) = \mathbf{0.292}$$

$$S(\text{문장4}) = (1-0.85) + 0.85 \times (0.8 + 0.6) = \mathbf{1.340}$$

| | 문장1 | 문장2 | 문장3 | 문장4 |
|-----|-------|-------|-------|-------|
| 문장1 | 0.000 | 0.500 | 0.167 | 0.800 |
| 문장2 | 0.500 | 0.000 | 0.000 | 0.600 |
| 문장3 | 0.167 | 0.000 | 0.000 | 0.000 |
| 문장4 | 0.800 | 0.600 | 0.000 | 0.000 |

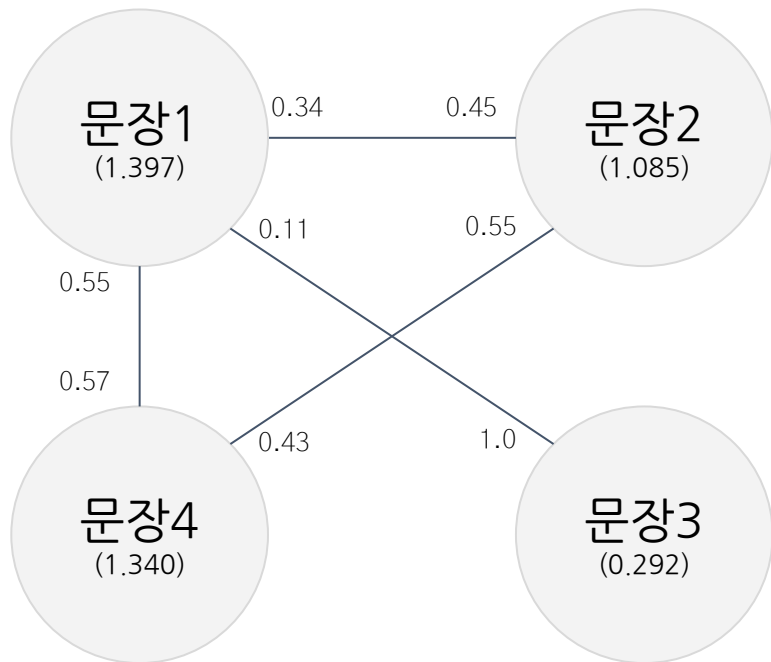
Textrank 과정 : 행렬로 계산하기



| 노드 | 이전 스코어 | 현 스코어 |
|-----|--------|-------|
| 문장1 | 1.397 | 1.468 |
| 문장2 | 1.085 | 1.043 |
| 문장3 | 0.292 | 0.285 |
| 문장4 | 1.340 | 1.301 |

| | 문장1 | 문장2 | 문장3 | 문장4 |
|-----|-------|-------|-------|-------|
| 문장1 | 0.000 | 0.476 | 0.159 | 0.762 |
| 문장2 | 0.493 | 0.000 | 0.000 | 0.592 |
| 문장3 | 0.292 | 0.000 | 0.000 | 0.000 |
| 문장4 | 0.766 | 0.574 | 0.000 | 0.000 |

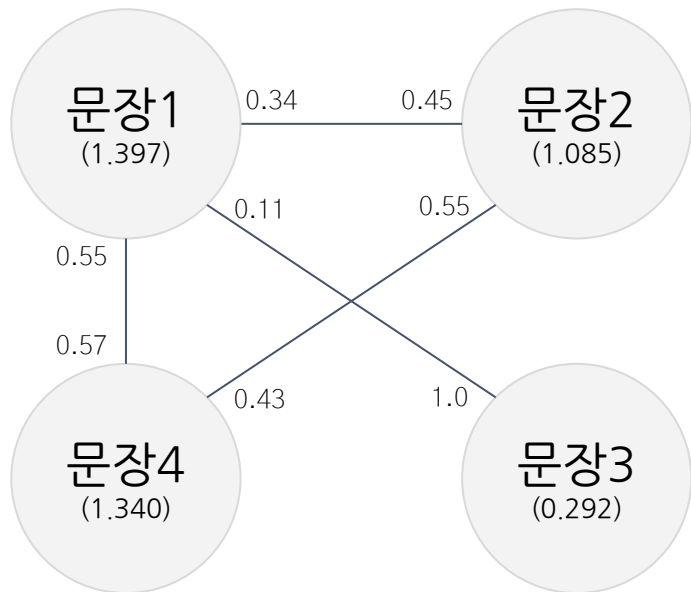
Textrank 과정 : 행렬로 계산하기



| 노드 | 이전 스코어 | 현 스코어 |
|-----|--------|-------|
| 문장1 | 1.468 | 1.427 |
| 문장2 | 1.043 | 1.049 |
| 문장3 | 0.285 | 0.292 |
| 문장4 | 1.301 | 1.314 |

| | 문장1 | 문장2 | 문장3 | 문장4 |
|-----|-------|-------|-------|-------|
| 문장1 | 0.000 | 0.500 | 0.167 | 0.801 |
| 문장2 | 0.474 | 0.000 | 0.000 | 0.569 |
| 문장3 | 0.285 | 0.000 | 0.000 | 0.000 |
| 문장4 | 0.743 | 0.557 | 0.000 | 0.000 |

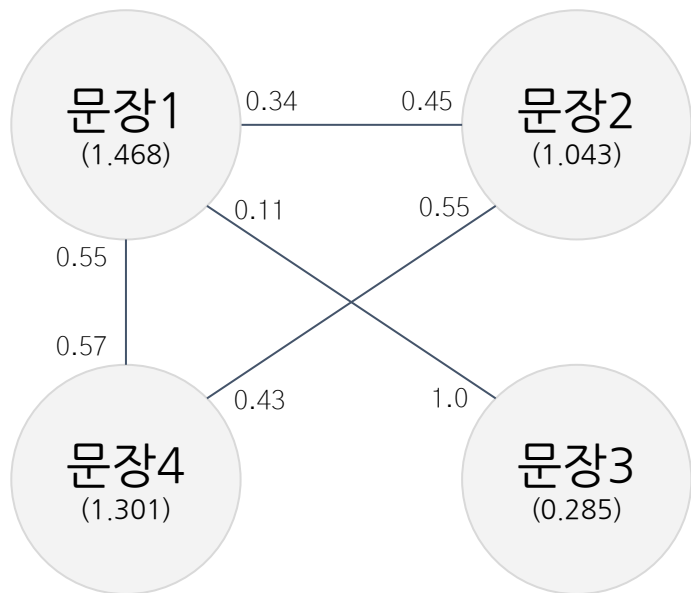
Textrank 과정 : 행렬로 계산하기



| Vertex | TR score |
|--------|----------|
| 문장1 | 1.397 |
| 문장2 | 1.085 |
| 문장3 | 0.292 |
| 문장4 | 1.340 |

| | 문장1 | 문장2 | 문장3 | 문장4 | |
|-----|-------|-------|-------|-------|-------|
| 문장1 | 0.000 | 0.476 | 0.159 | 0.762 | 1.397 |
| 문장2 | 0.493 | 0.000 | 0.000 | 0.592 | 1.085 |
| 문장3 | 0.292 | 0.000 | 0.000 | 0.000 | 0.292 |
| 문장4 | 0.766 | 0.574 | 0.000 | 0.000 | 1.340 |
| | 1.468 | 1.043 | 0.285 | 1.301 | 4.113 |

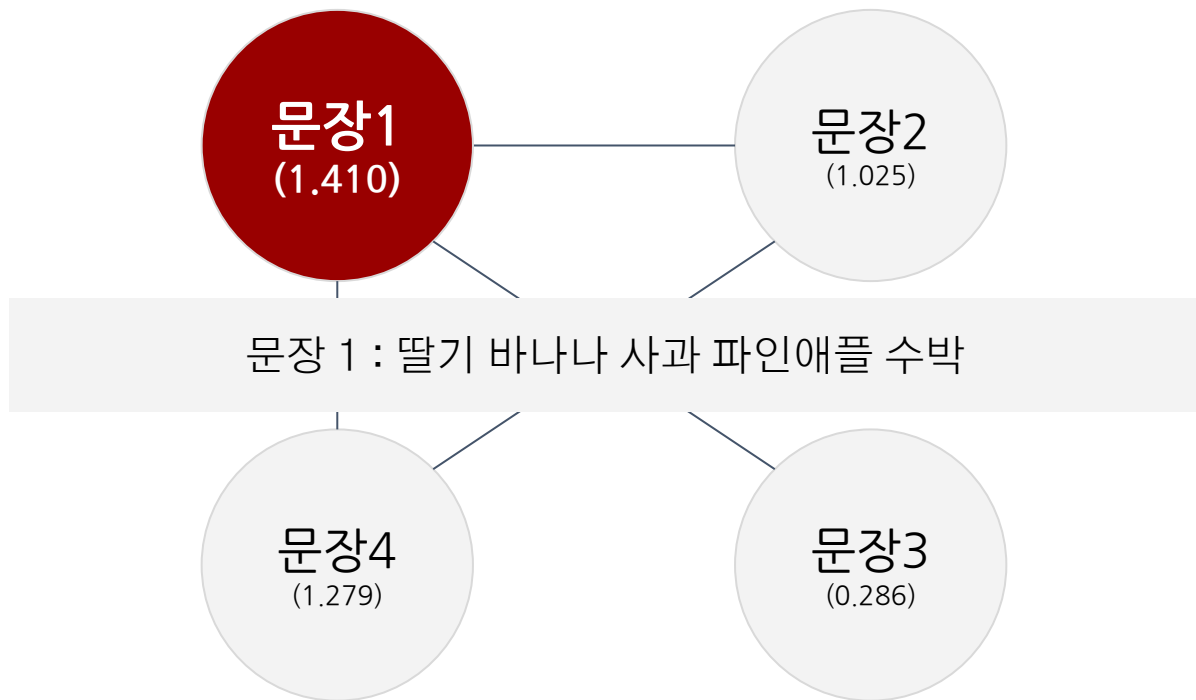
Textrank 과정 : 행렬로 계산하기



| Vertex | TR score |
|--------|----------|
| 문장1 | 1.468 |
| 문장2 | 1.043 |
| 문장3 | 0.285 |
| 문장4 | 1.301 |

| | 문장1 | 문장2 | 문장3 | 문장4 | |
|-----|-------|-------|-------|-------|-------|
| 문장1 | 0.000 | 0.500 | 0.167 | 0.801 | 1.468 |
| 문장2 | 0.474 | 0.000 | 0.000 | 0.569 | 1.043 |
| 문장3 | 0.285 | 0.000 | 0.000 | 0.000 | 0.285 |
| 문장4 | 0.743 | 0.557 | 0.000 | 0.000 | 1.301 |
| | 1.427 | 1.049 | 0.292 | 1.314 | 4.096 |

Textrank 과정 : 행렬로 계산하기



감사합니다.

Insight⁺campus

