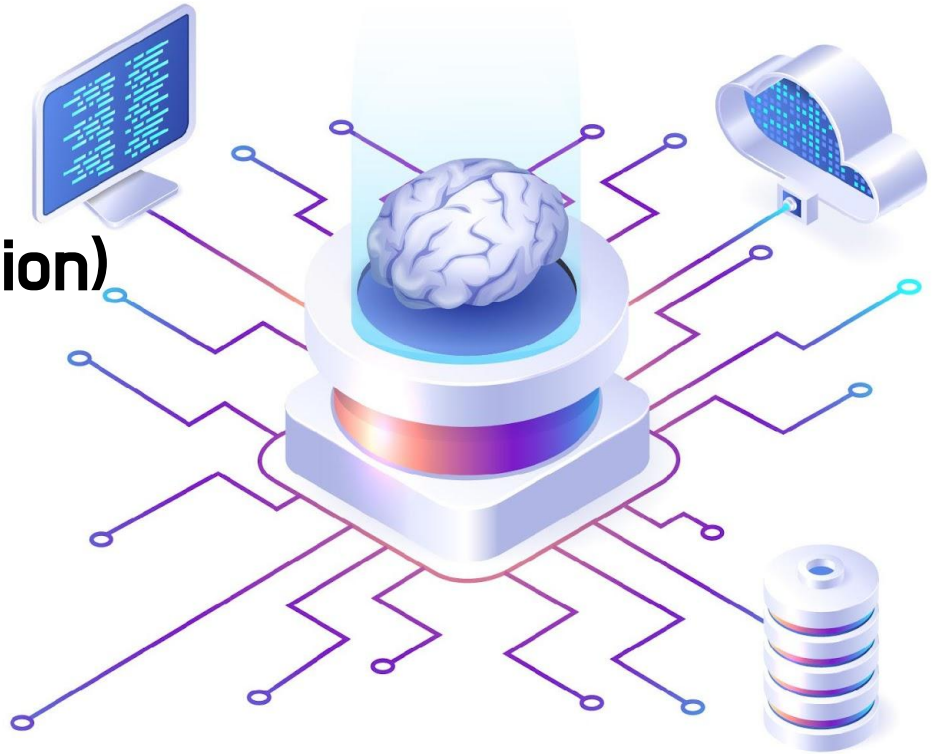
Insight campus SeSAC

문서 요약 (Document Summarization)

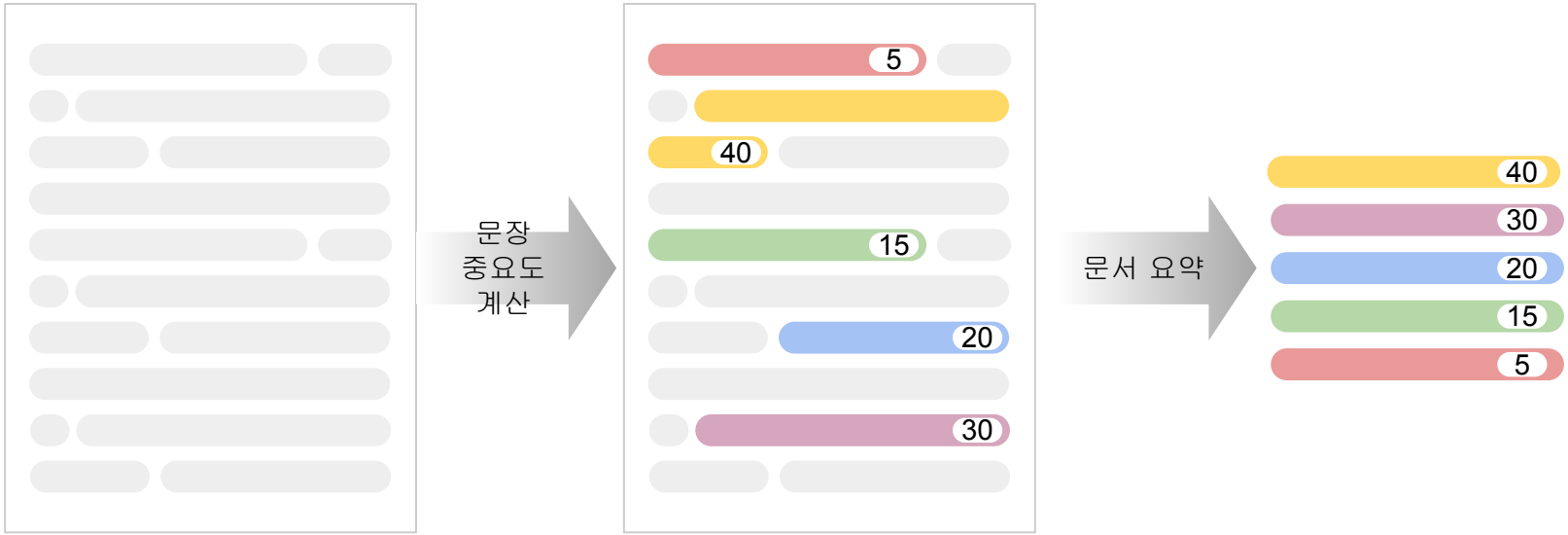
실무형 인공지능 자연어처리



FIN INSIGHT
Copyright FIN INSIGHT. All Right Reserved

문서 요약

- 문서 요약은 문서에서 중요한 문장을 자동으로 추출하는 과정
- 중요한 문장을 추출한다 **=>** 문장의 중요성을 어떻게 판단한 것인가



Insight campus SeSAC

요약 방법

추상적 요약
(Abstractive Summarization)

- 문서를 의미적으로 이해한것을 바탕으로 요약
- 요약 문장에 문서에서 언급되지 않는 단어가 등장하기도 함
- 추상적 요약은 “문서 → 문맥의 이해 → 의미 추출 → 요약 생성”의 과정으로 진행
- 사람이 문서를 읽고 해석하여 자신의 단어로 표현하여 요약하는 것과 같은 맥락

추출 요약
(Extractive Summarization)

- 문장별로 중요도를 계산하여 요약
- 추출 요약은 “문서 → 문장 중요도 계산 → 순위 높은 문장 선택”의 과정으로 진행
- 추상적 요약이 더 좋은 결과를 제공해줄 것이라는 예상을 할수 있지만 추출 요약이 더 나은 결과를 제공하기도 함
- 추상적 요약은 의미 이해, 추론, 자연어 생성과정의 어려움

1 Luhn Summarize를 활용한 문서 요약

Hans Peter Luhn

Hans Peter Luhn은 공학과 정보과학에서의 개척 작업으로 *"정보 검색의 아버지"*로 알려져 있다. 그는 표제어가 문맥에 포함된 채 배열된 색인(KWIC : keyword-in-context) 개발, 정보 선택 제공(SDI), 완전 텍스트 프로세싱, *자동 발췌(요약)*, 단어 시소로스의 최초 현대식 사용으로 신뢰를 얻었다. 오늘날 파생된 지식 대부분에는 KWIC 색인이 있으며 과학의 모든 분야에 SDI시스템이 있다.

1933년 Luhn은 자사인 공학회사 H.P. Luhn & Association을 설립하였고 8년간 자문기술자로 일했다. 1941년 Luhn은 IBM에서 수석 연구기술자로 참여하였고 이후에 정보검색연구 관리자로 일했다. Luhn이 IBM에서 새로운 아이디어를 지속적으로 내놓고 문제를 다르게 접근하여 주목을 받는 동안, 다른 기술자들에게 고차원적인 창조를 하도록 자극하면서 그들의 촉매제 역할을 하여 신뢰를 얻었다.

Insight campus SeSAC

Luhn Summarize 개요

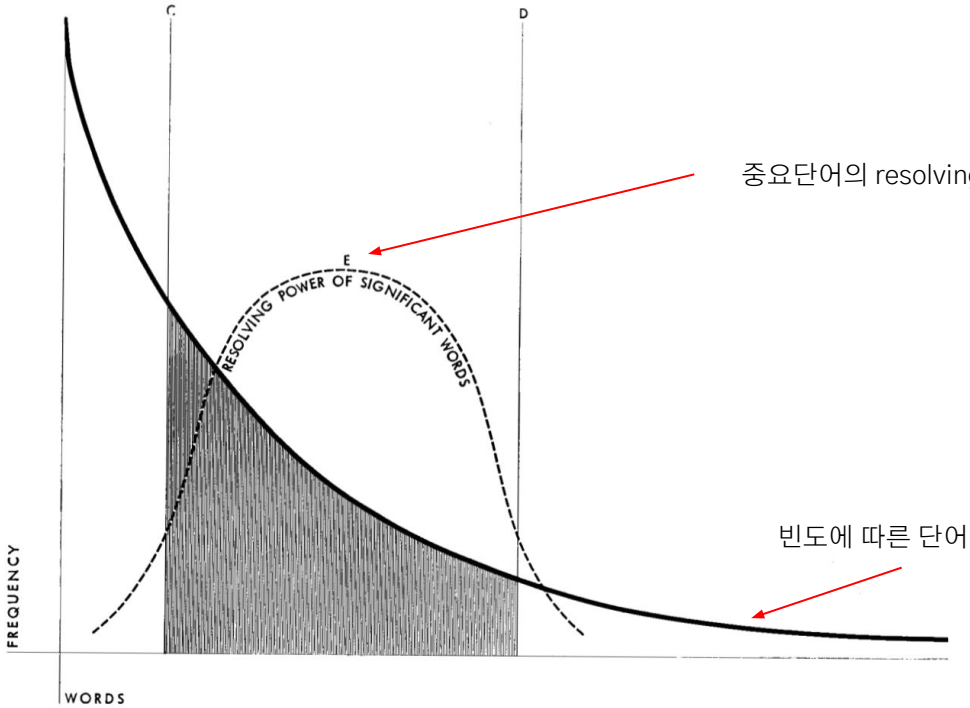
The justification of ***measuring word significance by use frequency*** is based on the fact that a ***writer normally repeats certain words*** as he advances or varies his arguments and as he elaborates on an aspect of a subject. This means of emphasis is taken as an indicator of significance. The more often certain words are found in each other's company within a sentence, the more significance may be attributed to each of these words.

– H.P. Luhn

단어의 중요도는 사용 빈도로 측정. 작가는 중요한 단어를 반복해서 사용한다는 사실에 기반

중요 단어 (Significant Words)

Figure 1 **Word-frequency diagram.**
Abscissa represents individual words arranged in order of frequency.



중요단어의 resolving power

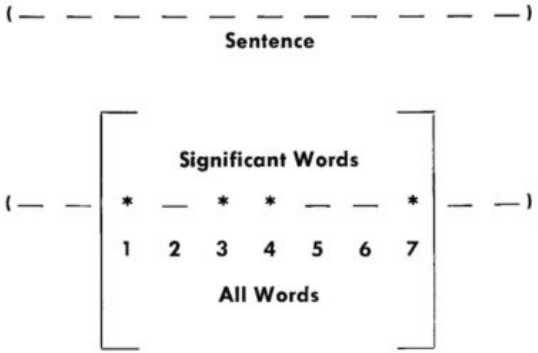
빈도에 따른 단어 목록

문장 중요도 (Significance factor)

문장 중요도

- 중요 단어를 포함하는 경우
- 중요 단어가 등장하는 처음과 끝사이 단어들 중 중요단어의 상대 비율
- 예시 : 중요단어 4개, 윈도내 단어 7개 = $4^2 / 7 = 2.3$

$$\text{문장 중요도} = \frac{\text{윈도내 포함된 중요단어 갯수}^2}{\text{윈도내 포함된 단어갯수}}$$



Portion of sentence bracketed by and including significant words not more than four non-significant words apart. If eligible, the whole sentence is cited.

Figure 2 **Computation of significance factor.** *The square of the number of bracketed significant words (4) divided by the total number of bracketed words (7) = 2.3.*

Luhn Summarize 절차

감사합니다.

Insight campus SeSAC

daniel@fins.ai