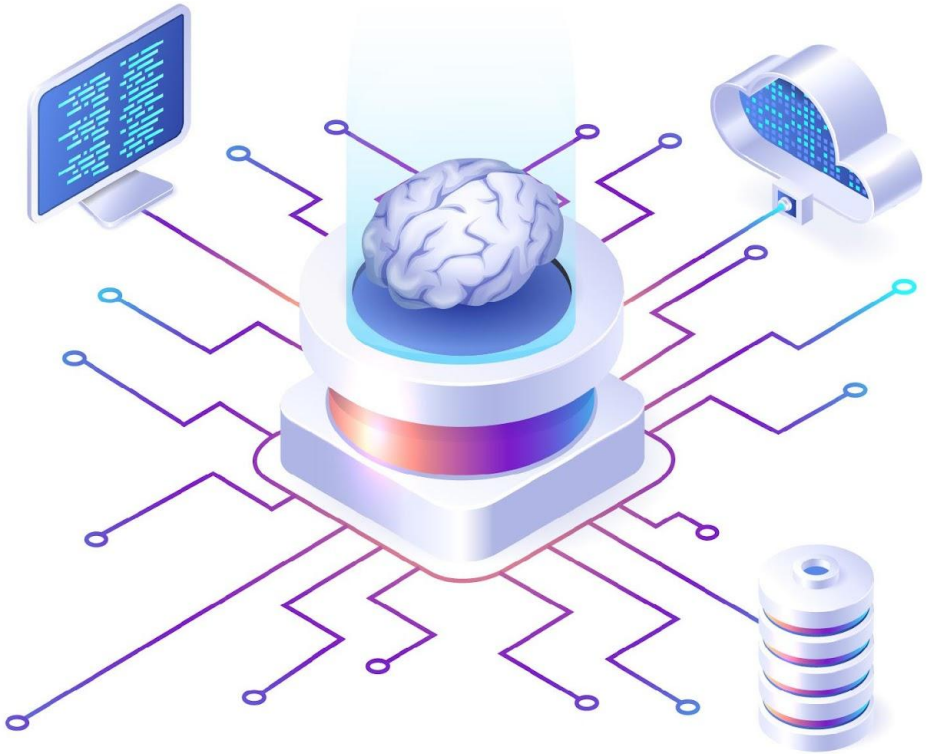
Insight campus SSAC

핵심 키워드 추출 (Keyword Extraction)

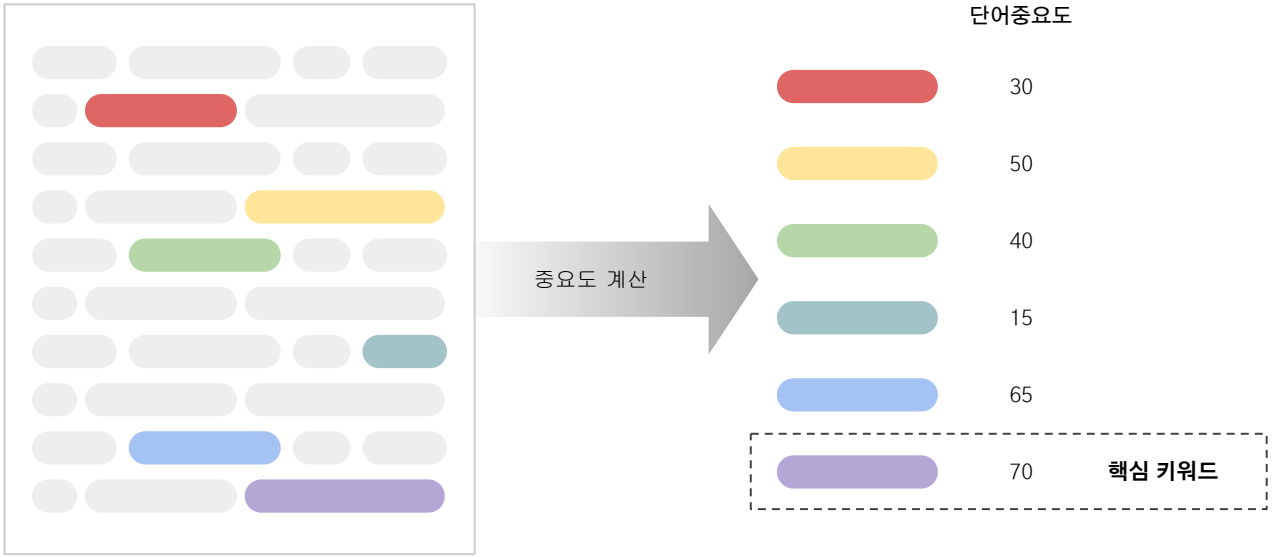
실무형 인공지능 자연어처리



Insight campus SSAC

핵심키워드 추출

- 키워드 추출은 문서에서 가장 중요한 단어를 자동으로 추출하는 과정
- 중요한 단어를 추출한다 **=>** 단어의 중요성을 어떻게 판단한 것인가



핵심키워드 추출의 필요성

대량 데이터 처리 가능
키워드 추출을 활용해 대량데이터를 분석할 수 있다. 직접 문서를 읽고 주요 용어를 수동으로 식별 할 수 있지만 시간이 많은 시간이 소요된다. 이 작업을 자동화하면 다른 더 중요한 작업에 집중할 수 있다.

추출의 일관성
키워드 추출은 규칙과 사용자가 정의한 매개변수를 기반으로 작동한다. 따라서 텍스트 분석을 수동으로 수행 할 때 나타나는 불일치를 고려할 필요가 없다.

실시간 분석 가능
소셜 미디어 게시물, 고객 리뷰, 설문 조사 또는 고객 지원에 대한 키워드 추출을 실시간으로 수행하고 제품에 대한 의견을 얻을 수 있다.

통계적 접근

단어빈도

- 단어의 등장 빈도를 활용하여 중요 단어를 추출
- 단어 빈도 접근 방식은 문서를 단순한 단어 모음으로 간주
- 단어의 의미, 구조, 문법 및 단어 순서를 고려하지 않음

연어 / 동시발생

- 단어의 의미구조를 이해하기 위해서 N-gram과 같은 통계기법을 활용하여 연어나 동시발생 단어를 하나의 단어로 처리 할 수 있음
- 연어는 연이어 함께 자주 등장하는 단어 묶음. 예, “고객 서비스”
- 동시 발생(co-occurrence)은 동일 코퍼스 내에 함께 등장하는 단어 묶음. 연어와 다르게 반드시 단어가 인접할 필요 없음.

핵심 키워드 추출 (Keyword Extraction)

통계기반 자연어 처리

1 TF-IDF 활용 핵심 키워드 추출

단어의 중요도 판단

TF-IDF

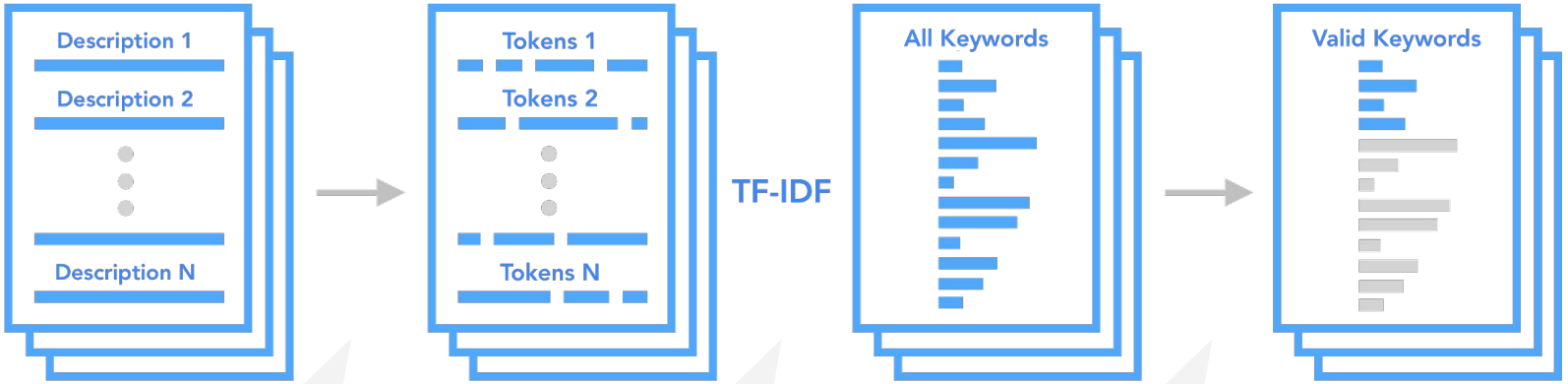
TF-IDF

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

tf(d,t)	특정 문서 d에서의 특정 단어 t의 등장 횟수
df(t)	특정 단어 t가 등장한 문서의 수
idf(d, t)	df(t)에 반비례하는 수

TF	IDF	설명
높	높	특정 문서에 많이 등장하고 타 문서에 많이 등장하지 않는 단어 (중요 키워드)
높	낮	특정 문서에도 많이 등장하고 타 문서에도 많이 등장하는 단어
낮	높	특정 문서에는 많이 등장하지 않고 타 문서에만 많이 등장하는 단어
낮	낮	특정 문서에 많이 등장하지 않고 타 문서에만 많이 등장하는 단어

TF-IDF 활용 핵심 키워드 추출 절차



토큰화

문서 집합을
문서별로 토큰화

TF-IDF 계산

문서의 단어별
TF-IDF score 계산

TF-IDF score가 높은순으로 추출

TF-IDF score가 높은
순으로 상위부터 추출

예제 1 : 토큰 Index 생성

문서1 : d1 = "The cat sat on my face I hate a cat"
문서2 : d2 = "The dog sat on my bed I love a dog"

	Index
The	0
cat	1
sat	2
on	3
my	4
face	5
I	6
hate	7
a	8
dog	9
bed	10
lov	11

예제 : TF 계산

문서1 : d1 = "The cat sat on my face I hate a cat"
문서2 : d2 = "The dog sat on my bed I love a dog"

$$f_{t,d} \Big/ \sum_{t' \in d} f_{t',d}$$

$f_{t,d}$ = 문서내 토큰 빈도
SUM($f_{t,d}$)= 문서내 전체 토큰빈도

문서1

	문서내 토큰 빈도	문서내 전체 토큰빈도	TF
The	1	10	0.1
cat	2	10	0.2
sat	1	10	0.1
on	1	10	0.1
my	1	10	0.1
face	1	10	0.1
I	1	10	0.1
hate	1	10	0.1
a	1	10	0.1
dog	0	10	0
bed	0	10	0
lov	0	10	0

문서2

	문서내 토큰 빈도	문서내 전체 토큰빈도	TF
The	1	10	0.1
cat	0	10	0
sat	1	10	0.1
on	1	10	0.1
my	1	10	0.1
face	0	10	0
I	1	10	0.1
hate	0	10	0
a	1	10	0.1
dog	2	10	0.2
bed	1	10	0.1
love	1	10	0.1

예제 : IDF 계산

문서1 : d1 = "The cat sat on my face I hate a cat"
문서2 : d2 = "The dog sat on my bed I love a dog"

$$\log \frac{N}{n_t} = -\log \frac{n_t}{N}$$

N = 문서수

n_t = 토큰이 등장한 문서수

	문서수	토큰이 등장한 문서수	IDF
The	2	2	0
cat	2	1	0.301
sat	2	2	0
on	2	2	0
my	2	2	0
face	2	1	0.301
I	2	2	0
hate	2	1	0.301
a	2	2	0
dog	2	1	0.301
bed	2	1	0.301
love	2	1	0.301

Insight campus SSAC

예제 : TF-IDF 계산

문서1 : d1 = "The cat sat on my face I hate a cat"
문서2 : d2 = "The dog sat on my bed I love a dog"

문서1

	TF	IDF	TF-IDF
The	0.1	0	0
cat	0.2	0.301	0.060
sat	0.1	0	0
on	0.1	0	0
my	0.1	0	0
face	0.1	0.301	0.301
I	0.1	0	0
hate	0.1	0.301	0.301
a	0.1	0	0
dog	0	0.301	0.301
bed	0	0.301	0.301
lov	0	0.301	0.301

문서2

	TF	IDF	TF-IDF
The	0.1	0	0
cat	0	0.301	0
sat	0.1	0	0
on	0.1	0	0
my	0.1	0	0
face	0	0.301	0.301
I	0.1	0	0
hate	0	0.301	0.301
a	0.1	0	0
dog	0.2	0.301	0.601
bed	0.1	0.301	0.301
love	0.1	0.301	0.301

TF-IDF

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

tf(d,t)	특정 문서 d에서의 특정 단어 t의 등장 횟수
df(t)	특정 단어 t가 등장한 문서의 수
idf(d, t)	df(t)의 역수

TF	IDF	TF-IDF	설명
높	높	높	특정 문서에 많이 등장하고 타 문서에 많이 등장하지 않는 단어 (중요 키워드)
높	낮	-	특정 문서에도 많이 등장하고 타 문서에도 많이 등장하는 단어
낮	높	-	특정 문서에는 많이 등장하지 않고 타 문서에만 많이 등장하는 단어
낮	낮	낮	특정 문서에 많이 등장하지 않고 타 문서에만 많이 등장하는 단어

감사합니다.

Insight campus

daniel@fins.ai