NLP 파트 - 새기능 #6527

새기능 # 6552 (신규): [정보검색] 정보검색 Metric 조사

[정보검색] TF-IDF

2022-07-26 10:30 - 신재영

상태:	신규	**시작시간:**	2022-07-26
우선순위:	보통	**완료기한:**	
담당자:	신재영	**진척도:**	0%
범주:		**추정시간:**	0.00 시간
목표버전:		**소요 시간:**	0.00 시간
설명			

이력

#1 - 2022-07-26 15:27 - 신재영

- 파일에 clipboard-202207261521-xhtfb.png이(가) 추가되었습니다.

- 파일에 clipboard-202207261522-jsmd1.png이(가) 추가되었습니다.

- 파일에 clipboard-202207261526-fszqw.png이(가) 추가되었습니다.

TF-IDF (Term Frequency-Inverse Document Frequency)

(1) 개념

- 하나의 문서에서 단어가 여러번 반복되면 관련성은 높아지지만 전체 문서에서 단어가 자주 반복된다면 오히려 관련성은 낮아지게 되는 알고리즘
- 즉, 문서에서는 등장하지 않지만 특정 문서에서만 자주 등장하는 단어를 찾아내 문서 내 "중요한" 단어의 가중치를 계산하는 방법
- 주로 문서의 유사도를 구하는 작업, 검색 시스템에서 검색 결과의 중요도를 정하는 작업, 문서 내에서 특정 단어의 중요도를 구하는 작업 등에 쓰임

(2) TF (Term Frequency)

$$tf(t,d) = 0.5 + 0.5 \cdot \frac{freq(t,d)}{(\text{문서 내 단어들의 } freq(t,d)\text{값 중 최대 값})}$$

- TF는 단어 빈도를 의미하며 전체 문서에서 특정 단어가 얼마나 자주 등장하는 지를 의미
- Term이 문서에 등장하는 횟수로, 빈도수가 높을수록 문서의 검색 점수는 높아짐
- 문서 내에 같은 단어가 여러번 등장하면 그 단어에 가중치를 주는 알고리즘

(3) IDF (Inverse Document Frequency)

$$idf(t,D) = \log\left(\frac{|D|}{1 + |\{d \in D : t \in d\}|}\right) = \log\left(\frac{\text{전체문서의 수}}{1 + \text{ 단어 } t\text{가 포함된 문서의 수}}\right)$$

- IDF는 역문서 빈도를 의미하며 너무 자주 나오는 단어는 덜 중요하다고 간주하여 패널티를 주는 것
- DF의 역수, 즉 DF에 반 비례하는 수를 뜻함
- 전체 문서에서 등장하는 Term의 빈도수가 작을수록 문서의 검색 점수는 높아지며, 여러 문서에 등장할수록 낮은 가중치를 주는 알고리즘
- 총 문서의 수(N)가 커지면 커질 수록 값이 너무 커지기 때문에 log를 취함
- 특정 단어가 전체의 문서 안에 존재하지 않을 경우, 분모가 0이 되는 것을 방지하기 위해 분모에 1을 더해줌

(4) Score 계산

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D)$$

- TF-IDF Score 계산은 위의 그림과 같이 TF 값과 IDF 값을 곱함

1) TF 계산

	양념	맛	최고	배달
문서1	0	0	0	**1**
문서2	0	0	0	0
문서3	**1**	**1**	**2**	0
문서4	**1**	0	0	0

- 각 단어의 빈도 수를 구하여 문서 단어 행렬로 표현(DTM : Document-Term Matrix)

2) IDF 계산

단어	IDF
양념	ln(4/(2+1)=0.288
맛	ln(4/(1+1)=0.693
최고	ln(4/(2+1)=0.288
배달	ln(4/(1+1)=0.693

- 전체 문서의 수(N)가 4이기 때문에 ln 안에서 분자는 4로 동일
- 분모는 '양념'이라는 단어가 전체 문서에서 2번 등장했기 때문에 2를 더해줌

3) TF*IDF

	양념	맛	최고	배달
문서1	0	0	0	**0.693**
문서2	0	0	0	0
문서3	**0.288**	**0.693**	**0.576**	0
문서4	**0.288**	0	0	0

- 2)에서 구한 IDF 값을 가지고 각 단어의 빈도수만큼 곱함
- '최고'라는 단어의 IDF 값은 0.288 이지만 문서3에서 총 2번 등장했기 때문에 0.288*2를 해줌

#2 - 2022-07-27 19:04 - 최한준

- 상위 일감을(를) #6520에서 #6552(으)로 변경되었습니다.

파일

clipboard-202207261521-xhtfb.png	15.9 KB	2022-07-26	신재영
clipboard-202207261522-jsmd1.png	23.7 KB	2022-07-26	신재영
clipboard-202207261526-fszqw.png	8.44 KB	2022-07-26	신재영