### tfidf

문자열에서 중요한 단어를 추출합니다. 빈도가 적어 상대적으로 특이성이 높은 단어에 더 높은 스코어를 부여하여 의미 있는 정보를 찾아냅니다. 이 명령어는 각 단어의 빈도(TF)와 역문서 빈도(IDF)를 결합하여 TF-IDF 점수를 부여합니다.

#### 문법

필드 목록에 의해 입력받은 문자열에서 단어 빈도를 계산

tfidf [delimiter=STR\_DELIMITER] [numeric=t] [op=build] [threshold=INT] FIELD\_LIST

필드 목록에 의해 입력받은 문자열에서 단어별 TF-IDF 점수를 계산

tfidf [delimiter=STR\_DELIMITER] [numeric=t] [op=query] [threshold=INT] FIELD\_LIST

서브쿼리문으로 단어 빈도 데이터를 가져온 후, 필드 목록에 의해 입력받은 문자열에서 단어별 TF-IDF 점수를 계산

tfidf [delimiter=STR\_DELIMITER] [numeric=t] [op=query] [threshold=INT] FIELD\_LIST [ SUBQUERY ]

**delimiter=STR\_DELIMITER**

문자열 구분자 목록(기본값: 공백문자). 구분자를 세미콜론(;)이나 슬래시(/) 등으로 변경하면 각 단어를 더 세분화하거나 특정 패턴으로 나누어 처리할 수 있습니다. 구분자로 인식할 문자들을 띄어쓰기 없이 붙여서 입력하세요.

예를 들어 다음과 같이 입력할 수 있습니다:

delimiter=" ;&/:-=\_.,[](){}\\n\\t"

1. 위 문자열에서 가장 앞에 있는 특수문자는 공백문자입니다.
2. 탭(\t)이나 줄바꿈(\n) 같은 특수 문자는 "\n", "\t"와 같이 이스케이프 처리해야 합니다.

**numeric=t**

동일한 길이의 숫자 문자열로 구성된 토큰을 동일하게 처리하도록 지정합니다. 예를 들어, "123"과 "456"은 길이가 같기 때문에 동일하게 처리됩니다(기본값: f).

일정 길이의 숫자가 반복될 때 해당 숫자들을 동일하게 처리하고자 할 때 사용하세요. 부수적으로 성능 개선 효과가 있습니다.

**op={build|query}**

명령어의 동작을 지정합니다(기본값: query).

1. query: 단어의 빈도와 역문서 빈도를 계산합니다.
2. build: 단어의 빈도만 계산합니다.

op=build로 계산된 단어 빈도는 따로 저장되지 않습니다. 후속 쿼리로 import 명령을 이용해 계산된 단어 빈도값을 테이블에 저장하세요. 주기적인 업데이트가 필요한 경우, 행위 프로파일이나 예약된 쿼리의 사용을 고려해보세요.

**threshold=INT**

단어의 빈도가 지정된 값 이하일 때 해당 단어를 무시합니다(기본값: 0. "0"일 때, 무시하지 않음). 예를 들어, threshold=5는 단어의 빈도가 5 이하인 단어를 무시합니다.

**FIELD\_LIST**

여러 필드를 하나의 문자열로 취급하여 분석합니다. 예를 들어, 필드 a에 "abc def", 필드 b에 "ghi jkl"이 있을 경우, 총 4개의 토큰("abc", "def", "ghi", "jkl")을 갖는 문자열("abc def ghi jkl")로 간주합니다.

**[ SUBQUERY ]**

단어의 빈도를 저장한 테이블이나 행위 프로파일에서 단어의 빈도 데이터를 가져오는 쿼리문을 대괄호 쌍([ ]) 안에 입력하세요. 이 서브쿼리문은 op=query일 때, 사용할 수 있습니다.

#### 설명

단어 빈도(TF, Term Frequency)는 문서 내에서 특정 단어의 반복 횟수를 측정하는 지표입니다. 문자열 데이터에서 단어의 중요도를 계산할 때, 해당 단어가 문서 내에서 얼마나 자주 나타나는지 평가하는 데 사용됩니다. 단어가 문서에서 자주 등장하면 TF 값이 커집니다. 문서가 길어질수록 단순 빈도를 사용할 경우 비교가 어려울 수 있으므로, 이를 문서의 전체 단어 수로 나눠서 정규화합니다.

역문서 빈도(IDF, Inverse Documant Frequency)는 문자열 데이터에서 단어의 중요도를 측정하는 지표입니다. 어떤 단어가 많은 문서에서 자주 등장한다면, 그 단어는 정보량이 적고 중요도가 낮은 단어로 간주됩니다. 예를 들어, "the", "is"와 같은 영어의 일반적인 단어는 거의 모든 문서에 등장하기 때문에 IDF 값이 작습니다. 반면, 특정 문서에서만 주로 등장하는 단어는 정보량이 많다고 볼 수 있으므로 높은 IDF 값을 갖습니다.

TF-IDF는 단어 빈도와 역문서 빈도를 결합해 문자열 데이터에서 단어의 중요도를 계산하는 기법으로, 단어의 빈도(TF)와 역문서 빈도(IDF)를 곱해서 계산합니다. TF가 높은 단어는 문서 내에서 자주 사용되는 단어로, IDF가 높은 단어는 문서 집합에서 특이성이 높은 단어로 강조됩니다. 이를 통해 흔히 등장하는 단어(불용어)는 중요도를 낮게, 문서마다 고유한 단어는 중요도를 높게 평가할 수 있습니다.

한국어와 같은 교착어는 조사(예: "~은", "~는")와 어미(예: "~합니다", "~하는")를 포함한 다양한 접사가 포함됩니다.접사는 단어의 의미를 결정짓는 데 중요한 요소가 아니지만 불필요한 단어(불용어)가 많으면 TF-IDF 계산 결과에 부정적인 영향을 줄 수 있으므로 형태소 분석을 통해 "어간"만 추출하거나, 불용어를 제거하는 전처리가 필요합니다.

성능 지표

다음은 tfidf 명령의 성능 지표입니다.

1. **빌드**: 초당 10 ~ 20만 건 처리
2. **서브쿼리를 사용하지 않을 때**: 초당 5 ~ 10만 건 처리
3. **서브쿼리를 사용할 때**: 초당 10 ~ 20만 건 처리

스트림 쿼리로 사용할 때는 사전 빌드된 데이터를 서브쿼리로 활용하세요. 미리 빌드된 데이터를 이용하면 쿼리 시점에 계산 부하를 줄임으로써 실시간 처리 속도를 높일 수 있습니다.

출력 필드

출력 필드는 다음과 같습니다.

|  |  |  |
| --- | --- | --- |
| 필드 | 타입 | 설명 |
| term | 문자열 | 단어 |
| \_idf | 64비트 정수 | 단어의 역문서 빈도 |
| \_tfidf | 64비트 정수 | 단어의 TF-IDF 점수. op=build일 때, 출력하지 않습니다. |

#### 사용 예시

기본 사용 예시

\_tfidf 필드에 각 문자열 라인의 TF-IDF 점수를 부여합니다.

table iis\_result | tfidf numeric=t delimiter=" ;&/:-=\_.,\\n[](){}\\t" op=query line

빈도 테이블 빌드

\_idf 필드에 각 문자열의 IDF 점수를 계산하여 iis\_idf 테이블에 저장합니다. 자주 등장할수록 점수가 낮습니다.

table iis\_result | tfidf numeric=t delimiter=" ;&/:-=\_.,\\n[](){}\\t" op=build line | import iis\_idf

서브쿼리 사용

사전에 빌드된 IDF 테이블을 서브쿼리로 불러와 점수를 계산합니다. 성능이 향상됩니다.

table iis\_result | tfidf numeric=t delimiter=" ;&/:-=\_.,\\n[](){}\\t" op=query line [ table iis\_idf ]