TF-IDF

- description : TF-IDF
- author : 오션
- email : shlim@repia.com
- lastupdate : 2020-11-09

의미

여러 문서로 이루어진 무서군에서 어떤 단어가 특정 문서에서 얼마나 중요한지를 나타내는 통계적 수치

용어

TF (Term Frequency, 단어 빈도)

- 특정 단어가 문서 내에 얼마나 자주 등장하는지를 나타냄
- 각 문서마다 모든 단어의 빈도수를 계산
- TF가 높다는 것은 많이 사용되는 단어라는 의미

DF (Docuemnt Frequency, 문서 빈도)

- 특정 단어가 문서군(文書群)에서 얼마나 자주 등장하는지를 나타냄
- 문서군에서 각 단어가 등장한 문서의 수를 계산
- 단어가 등장하는(포함된) 문서 수

IDF (Inverse Document Frequency, 역문서 빈도)

- 각 단어의 IDF는 log (전체 문서수 / 단어의 DF)로 계산

==== TD - IDF (Term Frequency - Inverse Document Frequency 텀 프리퀀시 - 인버스 다큐먼트 프리퀀시, 단어 빈도와 역문서 빈도)

- 특정 단어의 TF와 IDF의 역수갑(IDF)를 곱한 값
- 결과가 가장 높은 각 문서의 단어들이 유의미하다고 봄

의미

1. 정보 검색이나 텍스트 마이닝에서 주로 사용하는 것으로, 해당 엔티티(entity - 실체, 여기서는 '단어'의 의미)가 가지는 중요도를 파악할 수 있다.
2. TF - 단어빈도(수) : 특정 단어가 문서내에 얼마만큼의 빈도로 등장하는지를 나타낸다.
3. IDF - 역문서빈도(수) : 전체 문서 갯수를 해당 엔티티가 포함된 문서의 갯수로 나누는 것이며, 해당

문서군의 특징을 보여준다.

Ref

TF-IDF(Term Frequency - Inverse Document Frequency)/단어 빈도와 역문서 빈도

참조

빈도 : 같은 현상이나 일이 반복되는 도수, 빈도수
inverse : 역의, 정반대의
entity : 엔티티, 실체, 독립체

오션, TF-IDF, Term, Frequency, -, Inverse, Document, Frequency

From:
http://125.132.25.164/dokuwiki/ - **문제를 잘 정의하는 것은 문제를 절반 해결한 것이다.** - 2023.12

Permanent link:
http://125.132.25.164/dokuwiki/doku.php?id=wiki:miscellaneous:tf-idf&rev=1604889347

Last update: **2022/03/10 19:52**