**TF-IDF 알고리즘**

* TF-IDF
* Term Frequency – Inverse Document Frequency
* 정보 검색과 텍스트 마이닝에서 이용하는 가중치
* 여러 문서로 이루어진 문서군이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치
* 문서의 핵심어를 추출하거나, 검색 엔진에서 검색 결과의 순위를 결정하거나, 문서들 사이의 비슷한 정도를 구하는 등의 용도로 사용
* TF (Term Frequency)
* 어떤 단어가 특정 문서에 얼마나 많이 쓰였는지 빈도를 나타내 줌
* 많이 쓰인 단어가 중요하다는 가정을 전제로 한 수치
* DF (Document Frequency)
* 특정 단어가 나타내는 문서의 수를 의미

ex. A 단어가 말뭉치 전체에서 문서1, 문서3에만 등장했다고 가정하면 DF = 2 (DF가 클수록 다수의 문장에서 쓰이는 단어라고 할 수 있음)

* TF = 단어의 빈도, DF = 문서의 빈도
* IDF (Inverse Document Frequency)
* 전체 문서 수를 해당 단어의 DF로 나눈 뒤 로그를 취해준 값
* 값이 클수록 특이한 단어임을 의미
* IDF는 DF에 Inverse를 해주는 의미로 DF의 역수를 의미
* 역수를 취하게 되면 원래의 큰 값들을 작은 값으로 줄이는 효과를 낼 수 있음
* TF-IDF는 각각의 문서에서 자주 등장하는 단어에 높은 가중치를 주고, 모든 문서에서 자주 등장하는 단어에 대해서는 페널티를 주는 방식으로 값을 부여하게 되는 것

텍스트, 폰트, 화이트, 타이포그래피이(가) 표시된 사진

자동 생성된 설명

* 단어 출현빈도(TF) \* 역 문서 빈도(IDF) = TF-IDF 가중치 부여

폰트, 타이포그래피, 서예, 친필이(가) 표시된 사진

자동 생성된 설명

DF 수식

* DF = 전체 문서 D에서 특정 단어 t가 등장한 문서의 개수

폰트, 텍스트, 화이트, 서예이(가) 표시된 사진

자동 생성된 설명

IDF 수식

* 파이썬을 활용한 TF-IDF 계산

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

텍스트, 폰트, 그래픽, 스크린샷이(가) 표시된 사진

자동 생성된 설명

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

DTM, TF 계산

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

TF 계산결과

텍스트, 폰트, 스크린샷, 그래픽이(가) 표시된 사진

자동 생성된 설명

DF 계산

텍스트, 폰트, 영수증, 화이트이(가) 표시된 사진

자동 생성된 설명

DF 계산 결과

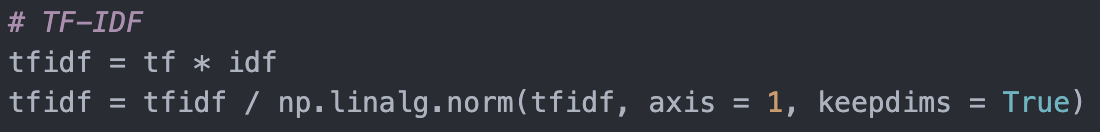
텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

텍스트, 폰트, 화이트, 타이포그래피이(가) 표시된 사진

자동 생성된 설명

IDF 계산 결과



텍스트, 폰트, 스크린샷, 화이트이(가) 표시된 사진

자동 생성된 설명

TF-IDF 계산 결과