DTM 의 한계

중요한 단어와 중요하지 못한 단어를 구분 할수 없다

그것을 해결한게 = TFIDF 를 개발 하였다

# **Term Frequency-Inverse Document Frequency**

왜 사용하는것인가

빈도수 기반은 중요한 단어와 불필요한 단어

많이 사용하는 단어는 불용어 지만 의미가 크게 없다 근데 이런 단어가 문장에서 제일 많다

근데 이게 중요한 단어는 절대 아니다

불필요한 단어는 중요도를 줄이고 문장을 구분할수 있는 가장 중요한 단어를 중요도를 높인다

1.Definition

TFIDF(t,d) = TF(t,d) \* IDF(t)

TF(t,d) = d에서 t (문서에서 용어 가 증장하는 빈도 를 나타낸다)

DF(t) = 특정 단어 t 가 등장한 문서의 수

흔한단어는 전체 문서에서 빈도수가 높고

흔하지 않은 단어는 전체 문서에서 빈도수가 낮다 = 이게 문장을 구분할수 있는 중요도가 높다고 할수 있다

IDF(t) = inverse 역수를 취해준다

log(n/1+DF(t)) = 선택적이긴 한데 문서의 개수에 따라 밑도 커진다

로그를 쓰는 이유 흔하지 않은 단어에 부여되는 가중치가 너무 커짐

빈도가 작으면 역수가 기하 급수적 으로 idf 가 증가하는것을 방지 하기 위해 = 정수 overflow 방지

분모에 +1 은 분모가 0 이 되는것을 방지

흔하지 않는 단어가 문장을 구분하는 요소인가? = 아닐수도 있지만 그렇다고 가정해야 여러가지 고려를 할수 있다

n 은 전체 문서의 개수 한 책에 문장이 10개 면 n 은 10 n이 너무크고 dt 가 너무 작으면 값이 너무 커지는 현상 이다

T가 빈도수가 낮으면서 d 문서에서는 많이 출연 하면 그 문서(문장) 에서 중요한 요소 이다.