☼ TF-IDF의 수식

- $TF_{ij} * IDF_i = TFIDF_{ij}$

☼ TF와 IDF

- TF
  - 단어의 빈도수이고 해당 문서에서 해당 단어가 나타나는 비율
- IDF
  - 역문서 빈도로서 전체 문서에서 해당 단어가 나타나는 문서의 비율의 역수에 log를 취한 값

☼ 한글 형태소 분석기

- Twitter, Komoran, 꼬꼬마 등

☼ 텍스트 추출 라이브러리

- html이나 xml을 파싱하여 순수한 텍스트를 추출함 – BeautifulSoup 등