2020-06-01 멋쟁이터틀

특허 20만개

1. 특허 단어 추출
각 특허 마다 사용된 모든 단어 추출
제외 단어 = [공백, 숫자, 구두점, 불용어]

2. 특허성과지표 추출 - 각 특허마다 아래의 데이터를 각각 추출해낸다.

특허성과지표 = [Forward Citation / IPC / Claim / Family Patent / PQI / bi-PI / ter-PI]

청구항에 기술 내용이 기재된 특허는 출원된 뒤에 IPC 코드 할당되며, 신규성, 권리성, 기술성 등의 관점에서 평가. 평가 뒤 등록된 특허는 다른 특허에 인용되기도 하고, 다른 국가에서돋 법적 권리 범위를 행사하기 위해 패밀리 특허를 출원

특허는 주로 청구항 수, 피인용 수, 패밀리 특허 수(국가 수)등에 의해 평가
청구항 수는 특허의 권리성과, 피인용 수는 특허의 기술성과,
패밀리 특허는 특허의 시장성과 관련
(특허청의 가이드라인에 제시된 대표적인 질적 우수성 지표는 권리성, 기술성, 시장성, 기술 다양성)

=> 상위의 전처리 과정 후 csv의 포멧으로 전달

ex) 특허번호, [각 특허성과지표], [각 단어]

patent1, 1, 2, 3, 4, 5, 6, 1, 2, man, science, car, tree, water .....