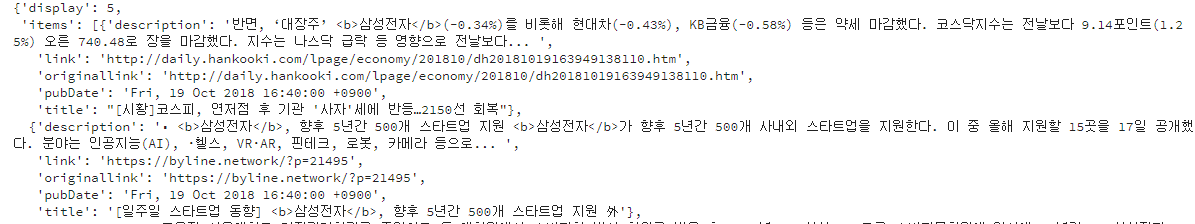
**국내 기업에 대한 텍스트 분석**

**고동형**

**1. 자료 수집**

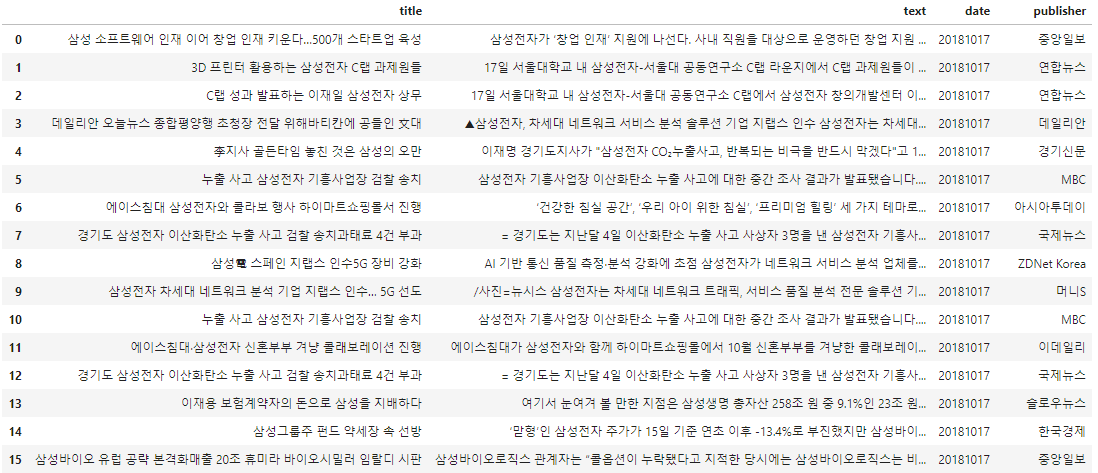
* 1. **네이버 API**

네이버에서 제공하는 API를 사용한 결과, 제목, Publish된 날짜, 언론사, 본문요약 정도의 정보를 json과 xml 형태로 얻을 수 있습니다. 그러나, 하나의 검색어에 대해 수집할 수 있는 자료의 수가 1,000개로 한정되어 있다는 단점이 있습니다.



* 1. **네이버 뉴스 크롤링**

Selenium을 통해 뉴스검색 페이지를 움직이며 자료를 수집합니다. 네이버는 한 검색어에 대해 열람할 수 있는 뉴스 자료를 4천개로 제한하고 있습니다. 제목, 언론사, 날짜, 본문 요약 정도를 추출할 수 있습니다. 원문을 모두 가져오고 싶은 경우, 해당 원문과 연결되는 url에 접속하면, html 구조가 달라서 현재로선 크롤링이 불가능합니다. 모든 언론사에 대한 예외처리를 하는 것은 어려울 것으로 보입니다.

**Issue**

1. 수집할 수 있는 자료의 수가 한정적 (API 1,000개, 크롤링 4,000개)

2. 원문을 모두 가져올 수 없음. 본문 요약이 한계 (기사 당 약 2~30단어)

**Altenatives**

1. 크롤링 소스 변경: 네이버 뉴스가 아닌, 특정 언론사를 선정하여, 해당 언론사에서 각각의 키워드로 검색한 결과를 수집한 뒤, 종합한다.

2. 논의 필요