(개발일지)영어 컨텐츠를 크롤링 하는 과정의 시행착오

일시: 3월 2주차
소프트웨어학부 김동욱

현재 로컬에서 접속되는 youtube는 기본 설정으로 linux dpkg config에 잡힌 asia/seoul 을 사용하는 것 같아 영어 컨텐츠를 읽어오는 방법에 대해서 팀원들과 이야기를 나누었다.

영어 컨텐츠를 가져오는 방법에 대해서 검색어로 검색, 유튜브 trend 영상 등 다양한 의견이 나왔지만, 매일 업데이트 되는 인기 유튜브 영상들을 크롤링 하기로 했다.

selenium driver를 이용해 크롤링 되는 화면을 봤을 때 한글 컨텐츠로 필터링이 되어 나오는 것을 알 수 있었다. 영어 컨텐츠로 필터링 되는 것이 잘 안되어 방법을 찾아보며 고생하였는데 결국 해결할 수 있었다. 유튜브 우측 상단에 톱니 아이콘을 누르면 국가/언어 설정을 할 수 있는데, 이때 잠깐 나오게 되는 url에 국가를 설정할 수 있는 parameter를 힘들게 얻어 해당 parameter를 붙인 url로 크롤링을 하는 방법이었다. 영어 컨텐츠로 필터링이 되어 결과가 잘 나오는 것을 확인했다.

약 80여개의 컨텐츠가 크롤링 되고 처음에는 매일 모두 크롤링 할 예정이었지만, 회의를 통해 서비스 특성 상 하루에 30~40개의 영상이 새로 등록되는 정도면 충분하다는 의견이 많아 이대로 진행하기로 했다.