멘티 활동 일지

작성일자: 2023년 9월 4일

도메인	빅데이터 (분산)
진행 일시	2023-09-04 13:40 ~ 14:10

내용	비즈니스 관점 피드백
	
	• 법적 문제
	  ◦ 아카이브 서비스 시 아카이빙 대상의 동의가 필요할 수 있다.
	  ◦ 비즈니스로 확장할 것이라면 법적인 문제를 확인해야 한다.
	• 서비스 관점
	  ◦ 추가적으로 제시 가능한 서비스가 있는지 고민해보면 좋겠다.
	
	기술 피드백
	
	• 맵
	  ◦ 맵에서 장르만을 가지고 유사도를 계산하면 0 혹은 1 이 나올 수 있으므로, 어떻게 유사도를 측정할 것인지 더 고민해봐야 한다.
	  ◦ 장르로 유사도를 계산하는 대신 배우 중심으로 유사도를 측정해보는 것도 좋은 방법이다. 그 다음 순서로 장르나 다른 피처를 넣어보는 방법도 있다.
	• 데이터
	  ◦ 데이터 품질이 중요하므로 아카이브 서비스에 결측값이나 이상치가 있으면 안된다.
	  ◦ 수집한 양의 크기보다는 데이터 컬럼을 다 채웠는지에 중점을 두는 것이 더 좋다.
	  ◦ 네이버에 영화/드라마 관련 자료가 많으므로 네이버에서 1 차적으로 자료를 수집하고 나머지 수집되지 않은 정보를 크롤링을 이용해 채우면 프로젝트 진행을 빨리 할 수 있을 것 같다.
	• 설계
	  ◦ 데이터 수집부터 정규화, 서비스에 서빙하는 과정까지 전체 파이프라인을 설계할 때 어떤 점에 집중해서 하면 좋을까요?
	  ▪ 비즈니스 로직을 먼저 만들고, 그 방향성에 필요한 데이터 목록을 리스트업 하는 순서로 가는 것이 좋겠다. 그리고 나서 리스트업할 데이터를 찾으면 된다.
	  ◦ 영화/드라마에서 정해진 컬럼에 데이터를 채울 때, 크롤링 한 자료에서 전처리를 거친다면 Hadoop 보다 전처리 단계에 집중해 프로젝트를 수행하는 방법도 있다.
	  ◦ 아카이브 서비스는 정형데이터를 이용하므로 RDB 를 이용할 수밖에 없다.
	
	빅데이터 관련 피드백
	
	• 어느 정도의 데이터 양부터 빅데이터로서의 의미가 있을까요?
	  ◦ 빅데이터 필드에서 정해진 양적 기준은 없다. 중요한 것은 데이터 자체의 품질이다.