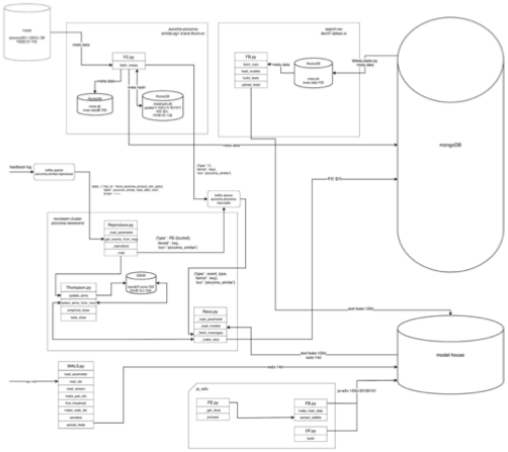
Piccoma 연관추천 개선 프로젝트(3주)



대외비일지도 몰라서 해상도를 많이 낮췄습니다.

프로젝트 개요

- Piccoma의 추천 구좌 중 연관추천 구좌의 모델을 개선하는 프로젝트
- 일차적으로 위의 데이터 플로우를 이해하였습니다.

프로젝트 아이디어

- read after click이 1회만 발생하는 작품은 표지만 매력적인 소위 표지 사기 작품일 확률이 높을 것이라는 아이디어에서 시작했습니다.(read after click : 추천 구좌를 통해서 발생하는 모든 클릭 log, 연속적인 작품의 열람도 포함)

eda 결과

- 전체 유료(1화부터 유료) 작품의 경우, 1 개의 작품만 보고 끝내는 비율이 2화 이상을 보는 경우의 2배가 넘었습니다.
- 1개 이상 무료로 공개된 작품들에서는 1개의 작품만 보는 경우와 2개 이상을 보는 작품의 비율이 비슷했습니다.
- 이는 전체 유료 작품에서 표지사기로 유입되는 케이스가 상당히 많다는 것을 유추할 수 있습니다.

결과 예측

- 이러한 결과를 바탕으로 생각해보았을 때, 표지만 매력적인 작품은 해당 로직을 통해 없어질 것이며, 2화 이상 지속할만한 작품들을 노출 시킬 것으로 예상했습니다.
- 이는 한 번의 클릭 이 후에 연속적인 작품의 열람 가능성이 높아질 것으로 예상합니다.
- 하지만 표지라도 매력적인 작품들이 없어질 것이므로 이로 인한 클릭이 줄어들 수 있습니다.

결과

- CVR과 CTR둘 다 떨어졌지만, CTR이 더 많이 떨어졌습니다.
- 이는 작품을 들어가서 연속적으로 보는 비율이 더 높아졌음을 시사합니다.
- 하지만 표지사기 작품이 없어지고, 없어진 자리를 더 매력적인 작품이 채워주지는 못했습니다.

해결한 이슈

- 과정
  - 카프카 관련 이슈를 겪었습니다.
  - 카프카의 리스너는 정해진 시간(max poll time) 안에 다시 메세지를 fetch 해와야 합니다.

    - 이를 제대로 인지하지 못했고, 여러 개의 실험이 돌아가면서 max poll time을 넘기며, process가 주기적으로 꺼지는 이슈였습니다.
- 해결
    - fetch messages를 5000 → 3000으로 조정하였습니다.
    - 줄어든 처리량 때문에 렉이 계속 쌓였습니다.
    - 프로세스를 1개 더 띄우면서 이를 해결할 수 있었습니다.

배운 점

- 데이터 파이프라인 간의 의존성을 낮추는 것이 효율적이라는 것을 배웠습니다.
- 이슈를 해결하며, 협업의 중요성을 깨달았습니다.
    - 혼자서 원인 분석을 하는 것보다 훨씬 빠르게 분석할 수 있었고, 성공적으로 이슈를 해결할 수 있었습니다.
- 현재 데이터엔지니어를 꿈꾸는 계기가 되었습니다.

기술 keyword :

MAB(multi arm bandit), Tomson Sampling, CF(Collaborative Filtering), CB(Contents based filtering), reciprocal rank fusion, tf-idf, ALS(Alternating Least Square)

RocksDB, SSDB, mysql, kafka