성향기반 추천 시스템

프로젝트 목표 및 가치

1. 협업필터링에서 나타나는 콜드 스타터, 희박성 문제 극복
   - Item 위주의 추천에서 사용자 정보를 기반한 추천 방식

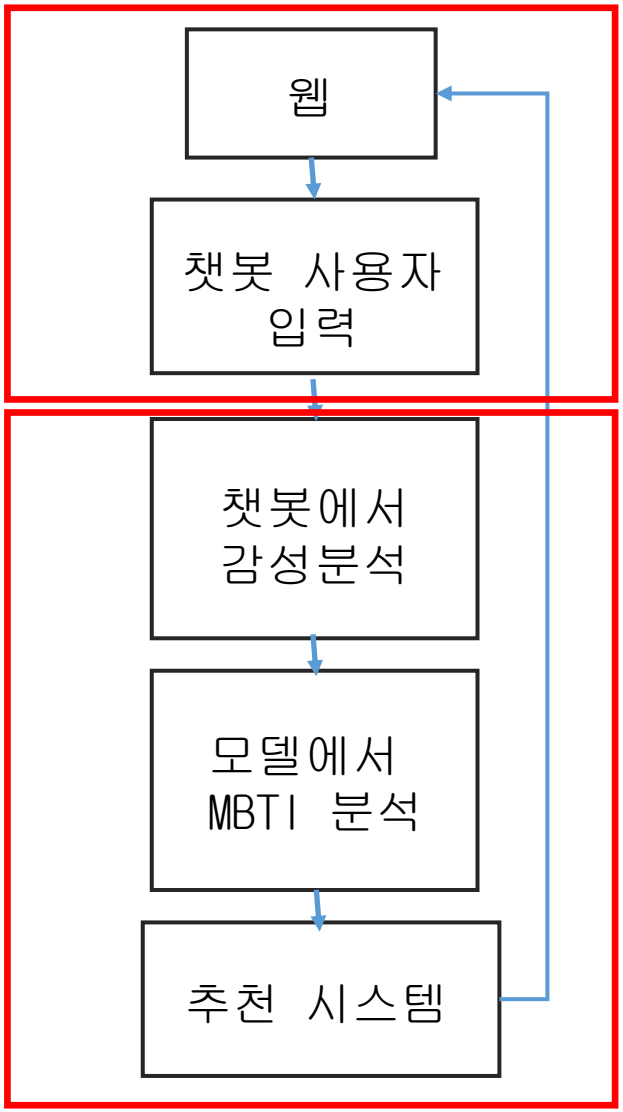
2. 실시간 추천 시스템
   - 실시간으로 유저가 원하는 다른 item 추천이 가능

3. 도메인 지식이 사용 가능한지 시도
   - 질적데이터의 편입은 압도적으로 적은 양의 데이터로 효과적인 가치를 창출

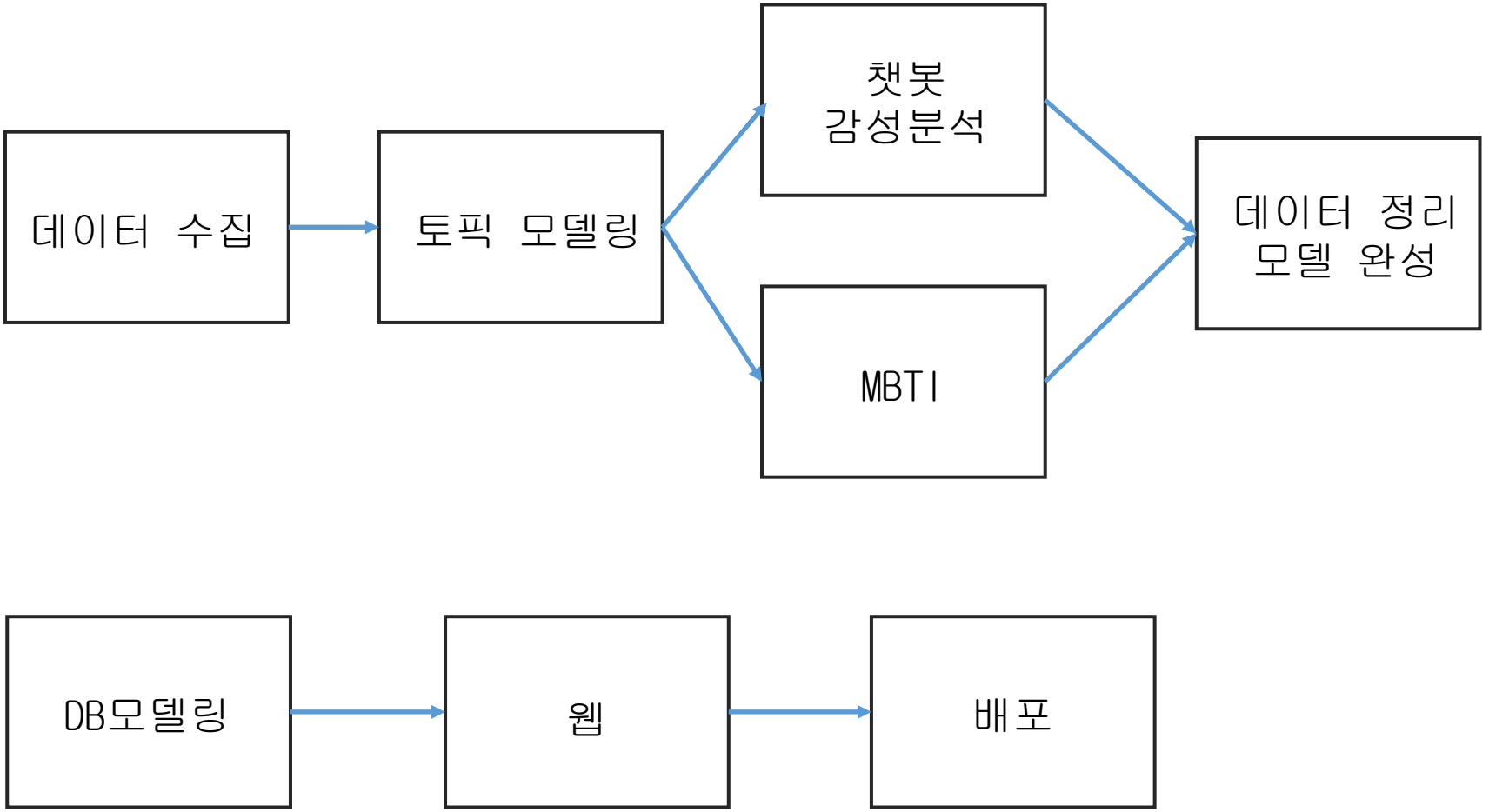
프로젝트 목표

MBTI 및 도메인 지식을 기반으로 각 성향들의
개인화 추천 시스템

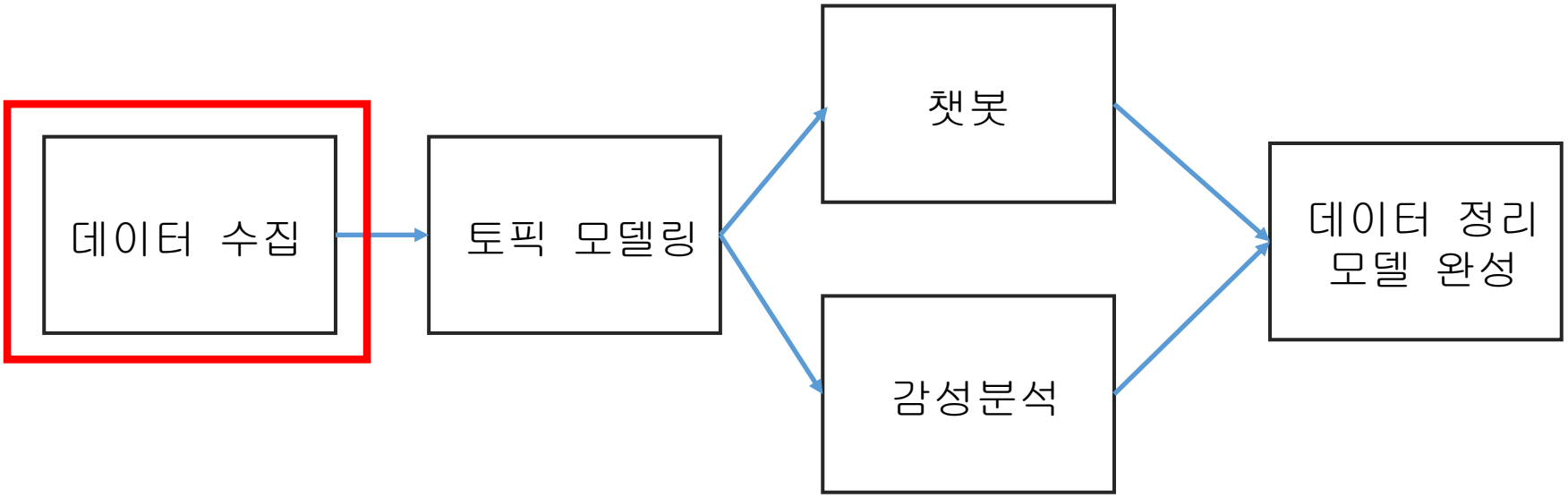
최종 파이프라인 예상



작업 프로세스



데이터 수집



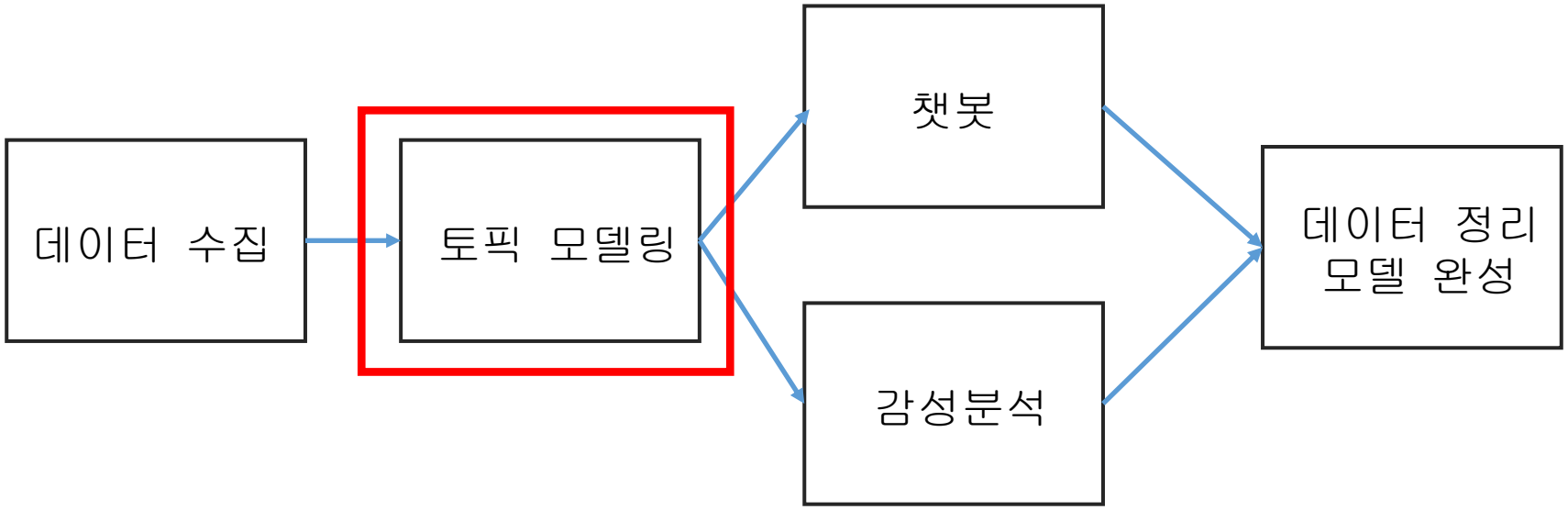
크롤링 및 API 데이터 수집

1. MBTI별 특징 크롤링
    1. 네이버 블로그
    2. 티스토리 블로그
2. RapidAPI를 통한 한국 넷플릭스 정보 가져오기
    1. 한국 넷플릭스 정보
    2. IMDB 리뷰 정보(영문)
    3. 네이버/티스토리 영화 리뷰 정보(한국어)

모델을 위한 데이터

1. 주제별 텍스트 일상 대화 데이터
2. 웹데이터 기반 한국어 말뭉치 데이터
3. 온라인 구어체 말뭉치 데이터
4. 감성 대화 말뭉치
5. MBTI 데이터 셋

토픽 모델링



크롤링한 데이터 키워드 추출

크롤링한 데이터로 얻게 되는 최종적인 메타데이터

1. 각 MBTI별 특징에 관한 한/영 정보

2. 넷플릭스에 존재하는 모든 작품의 정보 및 한/영 리뷰

1. MBTI 데이터 -> 각 MBTI들의 주요 키워드 추출 및 도메인 지식 첨가

- 같은 I라도 어떻게 다른 내향성을 가지는가?

- 동일한 상황에서 문제가 발생했을 때 '나'를 향하는가 '너'를 향하는가

- 만약 INTJ의 키워드가 '내성적', '계획', '논리', '추상' 이라면 이런 사람은 현실도피형일 가능성보다는 현실을 개척하는 개척가형이자 모험가형일 가능성이 높고 위험은 또한 큰 기회라고 인식할 가능성이 있음

2. 넷플릭스 데이터 -> 각 작품들의 주요 키워드 추출 (우울한 분위기, 미래지향적)

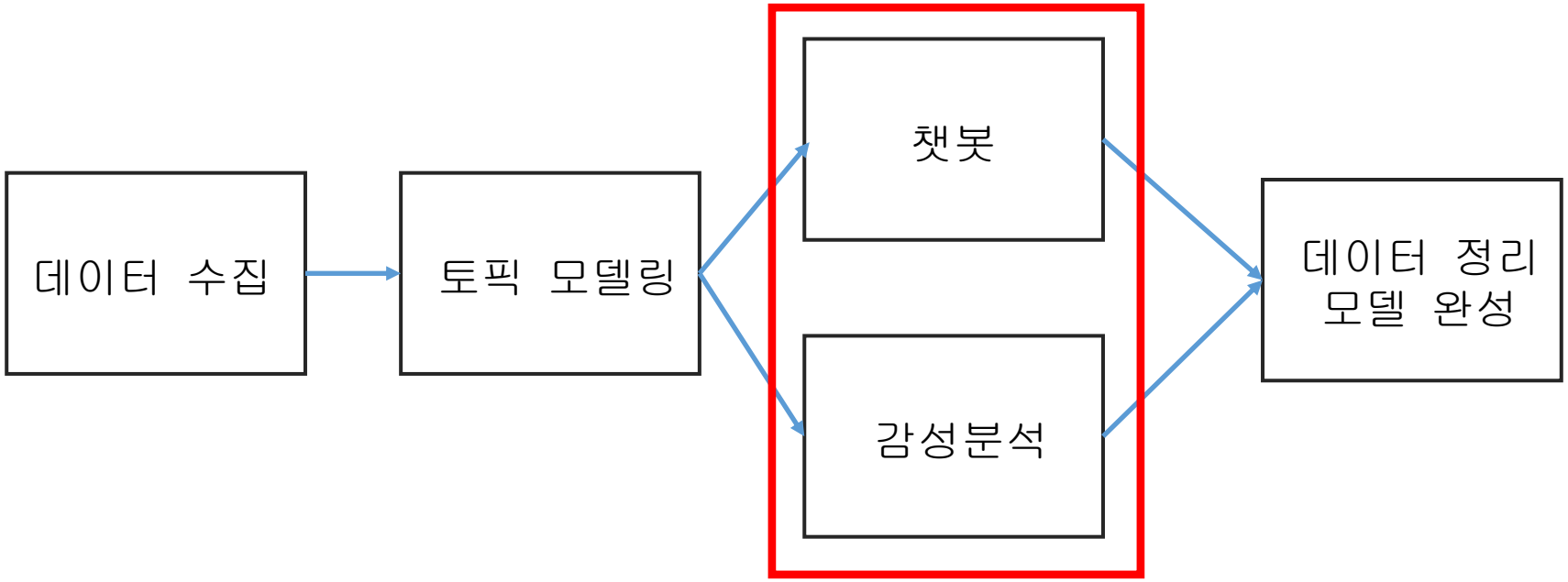
- 작품들의 키워드들이 어떻게 성향과 연관지어질 수 있는가?

- 개척가형이자 모험가형의 INTJ는 영화 메멘토를 선호할 확률은? 반대로 ESFP는 메멘토를 얼마나 선호할까?

토픽 모델링

- TF-IDF
- LSA
- LDA / LDiA
- PCA(SVD)
- Topic Modeling or BERTopic

모델링



모델링

1.챗봇

- BERT 기반 모델을 통해서 기본 학습
- 미리 작성된 각 성향별 설문을 사용자의 입력에 따라서 다르게 뿌려준다.

2. 감성분석 / BERTopic

- BERT 기반 모델을 통해서 기본 학습
- 사용자가 입력한 텍스트 , 작품의 전반적인 분위기를 모델링을 통해 도출
- 이는 추천 때 사용 할 feature 변수로 이용

추천

- 아직 많은 가능성을 열어 두고 있음
- 새로운 이론 + 기존 추천을 기반으로 시도
- 우리가 해결해야 할 문제는 '희박성' 즉 'Cold Start'
- 수집한 데이터를 기반으로 only 데이터 기반 추천
- 지식 도메인 + 데이터 기반 추천

MovieLens Data

- Grouplens에서 제공하는 데이터셋
- Full version의 경우 28만명의 유저, 2700만 평점, 110만 태그, 5.8만의 영화 데이터 제공
- 많은 연구 분석에서 해당 데이터를 추천 평가지표로 사용
- EDA를 통해서 고전적인 추천론 방식과 도메인 지식 기반의 링크 확인

추천

- 기존 추천 모델은 SVD / Model Based
  - SVD
  - NFC
  - DeepFM
- 지식 도메인 기반은 네트워크 분석 및 이웃 기반 클러스터링
  - KNN
  - Link Prediction

추천

- 추천 평가 지표

  1. 사용자 연구 평가
     - 실제 사용자들에게 추천시스템 평가 피드백

  2. 오프라인 평가
     - movielens 데이터를 기반으로 오프라인 평가

End.