AI 사서

소프트웨어학과 정찬호 201521001
소프트웨어학과 서상원 201720750
소프트웨어학과 성진호 201820808
소프트웨어학과 이동현 201920811
소프트웨어학과 김관주 202022307

1. What kind of problem you are solving

도서관에서는 도서하고자 하는 책들을 KDC(한국십진분류법)에 기반해 책의 장르를 나누고 그에 맞는 청구기호를 매겨 분류하고 있다. 책의 장르나 종류에 따라 도서관의 알맞은 위치에 책을 배치하고 있다.

본 프로젝트는 책의 제목에 사용된 단어들은 책의 내용 및 장르를 함의하고 있음을 전제로 하고 있다. 우리는 기계학습을 통해 책의 제목과 장르간의 연관성을 찾아내고, 더 나아가 책의 제목과 아주대 도서관 책의 위치간의 상관관계가 있는지를 기존 도서들의 데이터를 통해 알아보고자 한다.

이번 프로젝트를 통해 책을 쓰는 사람은 책 제목에 어떤 단어를 써야 사람들에게 내용을 잘 전달할 수 있을지 알 수 있게 되고, 책을 분류하는 사람 및 읽을 책을 찾는 사람은 책의 제목만으로도 어떤 내용일지 예측할 수 있게 되고, 도서관 내에서의 책의 위치도 예측할 수 있게 된다.

또한 우리는 각 단어가 가지고 있는 사전적 의미뿐만 아니라 사전에는 등록되지 않았지만 실질적으로 사용되고 있는 단어의 새로운 의미도 기계학습을 통해 발견할 수 있을지 기대할 수 있다.

2. What kind of data you will use

아주대학교 중앙도서관 측에 문의해 도서관 내 현존하는 책들의 제목, 소장 위치, 청구기호 등이 담긴 목록을 받아 학습에 사용한다. 만약 협조가 어려울 경우, 중앙도서관 홈페이지에서
얻을 수 있는 데이터셋 약 만 개를 이용하여 기계학습에 사용할 것이다. 학습된 알고리즘의 성능은 신규 출간도서목록 데이터를 이용하여 테스트 해볼 것이다.

다음은 청구기호를 부여하는데 쓰이는 KDC(한국십진분류법)의 십진분류표의 일부이다. 청구기호에 책의 장르 정보가 포함되어 있음을 알 수 있다.

000 총류	100 철학	200 종교	300 사회과학	400 자연과학
010 도서학, 서지학	110 형이상학	210 비교종교	310 통계학	410 수 학
020 문헌정보학	120 인식론,인과론,인간학	220 불 교	320 경제학	420 물리학
030 백과사전	130 철학의 체계	230 기독교	330 사회학,사회문제	430 화 학
040 강연집,수필집,연설문집	140 경 학	240 도 교	340 정치학	440 천문학
050 일반연속간행물	150 동양철학,사상	250 천도교	350 행정학	450 지 학

다음은 중앙도서관 홈페이지에서 예시로 다운받은 파일의 일부이다.

도서명	저자	출판사	출판년도	소장위치	청구기호
Stochastic processes in engineering systems	Wong, Eugene	Springer-Verlag	1985	중앙도서관 4층.자연	519.202462 W872ss

3. How you would solve your problem (main approach, algorithms)

2번의 데이터 중 서명, 소장위치, 청구기호가 학습에 사용될 예정이다. 제목은 단어와 조사로 분류하여 각 단어마다 해당하는 장르에 맞게 가중치를 부여하여 Random Forest로 자연어를 처리한다.
소장위치의 경우 층수와 세부 위치를 나누어 사용한다.

청구기호의 경우에는 소숫점 앞 세 자리만 사용한다. 청구기호 앞 세 자리는 책의 장르를 나타내는 부분으로 라벨의 역할을 하게된다. 세 자리는 아래 자리 일수록 세부 장르로 나뉘어지며, classification의 결과가 너무 다양할 경우 정확한 판단이 어려울 것으로 예상되어 진행 상황에 따라 앞 두자리 또는 한 자리만 사용하게 될 수도 있다.
분류기호, 도서기호, 부차적기호 중 분류기호 정수부분만 사용함.

이외에도 프로젝트 진행 중 더 많은 데이터가 필요하다고 생각될 경우 책의 목차나 저자, 출판사, 출판년도 등의 정보도 추가로 사용할 예정이다.

4. How you will evaluate your result

장르라는 것은 추상적인 개념으로 어떤 feature가 결과에 크게 영향을 주게 될지 정확히 알지 어렵기 때문에 Random forest를 사용하여 최적의 feature 들을 선정한다.

training 과 validation과정을 거쳐 얻어낸 test set이 제목과 장르간의 뚜렷한 상관관계를 보이지 않아 결과가 유의미하지 않을경우 다른 feature들을 추가하여 훈련 및 검증 과정을 거쳐본다. 유의미한 결과값이 나온경우 어떠한 방식으로 알고리즘을 활용하였고 test set 및 validation set을 설정하였는지 기록한다.

프로젝트 진행 도중 새로운 도서가 입고 될 경우, 이 또한 활용하여 test 의한 결과와 실제 도서관에서 정한 결과를 비교해서 알고리즘의 성능을 평가할 수 있다.