기계학습의이해 2025-2

Final Report (assignment 6)

독일어과 202400420 장지수

1. 프로젝트 개요

1). 기존 문제

독일어과 부속 오스트리아도서관의 도서 데이터베이스를 구축하는 과정에서 도서의 제목, 저자, 보관 위치, 분야 4가지의 정보 수집이 하였습니다. 하지만 다른 항목과는 달리 '분야'는 도서 표지에 명시적으로 드러나지 않는 경우가 많아 외국어 서적을 직접 읽거나 인터넷에 검색하는 등의 과정이 필요했고, 이는 업무 효율의 저하를 초래했습니다.

2). 무엇을 만들었나?

오스트리아도서관의 서적을 어학, 문학, 사회과학, 역사 4가지 분야로 분류하는 multiclass classification model을 제작하였다. 독일어뿐만 아니라 영어, 한국어, 라틴어 등의 다국어를 처리할 수 있도록 하였고, 입력 feature로는 도서의 제목을 이용하며, 출력에서 위 네가지 분야를 confidence에 따라 나열합니다.

2. 진행 과정

1). 주제 선정 및 문제 정의

필자는 2025학년도 2학기 현재 한국외국어대학교 독일어과 부속 도서관인 '오스트리아도서관 (본관 301호)'에서 책임자로 근무하고 있습니다. 오스트리아도서관은 1982년 개관하여 독일어권 어학, 문학, 지역학 도서를 소장하고 있는 국내의 몇 안되는 독일어 서적 특화 도서관으로서, 독일어권 인문학 연구에 높은 기여를 하는 가치 높은 공간입니다.

하지만 2025년도 1학기까지도 대출·반납의 전 과정이 1982년도의 아날로그 방식인 수기 장부 사용을 그대로 고수하고 있었으며, 도서 목록은 도서관 데스크톱 내 로컬 엑셀 파일 내에 입력하는 방식이었습니다. 필자는 1학기 때 학부조교를 하며 아날로그 방식에서 기인되는 여러 불편함을 느꼈기에, 2학기 때 도서관 책임자로 승진함에 따라 이와 같은 문제를 해결하고자 도서관 디지털화 프로젝트를 진행하였습니다.

디지털화를 위한 과정 중 도서 목록 데이터베이스 구축의 기존 방식은 다음과 같은 문제점을 포함하고 있었다: 도서 한 권마다 제목, 저자, 언어, 보관 위치, 분야, 청구기호가 기입되는데, 다른 항목들은 비교적 입력이 간편했으나 '분야'의 경우 조교들이 직접 독일어 원서를 읽고 구분해내거나 ChatGPT 등 AI에게 물어보는

방식으로 작업하였습니다. 이와 같은 방식은 업무 난이도 및 소요시간이 상당하였고, 사람에 의한 수작업이기에 오기입도 잦았습니다.

따라서 필자는 입력이 까다로운 '분야' 항목은 아예 수작업 업무 범위에서 제외시켰고, AI 모델을 학습시켜 자동으로 도서 분야를 분류해내는 자동화 과정을 구축하고자 하였습니다.

2). 데이터 수집 및 분석

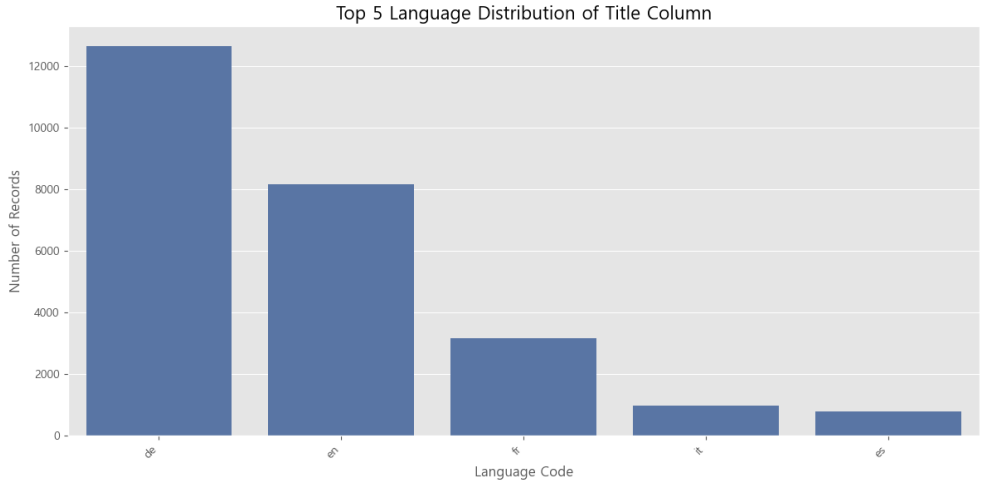
학습 데이터는 베를린 시립 도서관이 open source로 제공하는 도서 목록 dataset을 이용하였습니다. 해당 데이터는 다음 링크에서 확인할 수 있습니다: https://huggingface.co/datasets/SBB/ARK-Metadata

데이터 정제 과정:

1. 기본 제공 형식인 parquet 파일을 csv 파일로 파싱하였습니다.
2. input feature로 사용할 도서 제목만 남기고 기타 정보는 삭제하였습니다.
3. 기존의 DDC 기반 분야 분류 체계(000 총류, 001 철학 등)는 10가지 이상의 세부적인 분류 기준이 포함되어 있어 오스트리아도서관 소장 도서 특성(독일학 중심)에 적절하지 않습니다. 따라서 소장 도서의 4대 핵심 분야인 어학, 문학, 역사, 사회과학으로 새로 분류하였습니다.
4. 새 분류기준에 해당하지 않는 나머지 도서는 데이터 학습의 시간 효율 및 모델 혼란 예방을 위해 삭제하였습니다.
5. 정제 후 데이터 셋은 기존 2,619,397권에서 28,525권으로 축소되었습니다.

언어별 데이터 분포

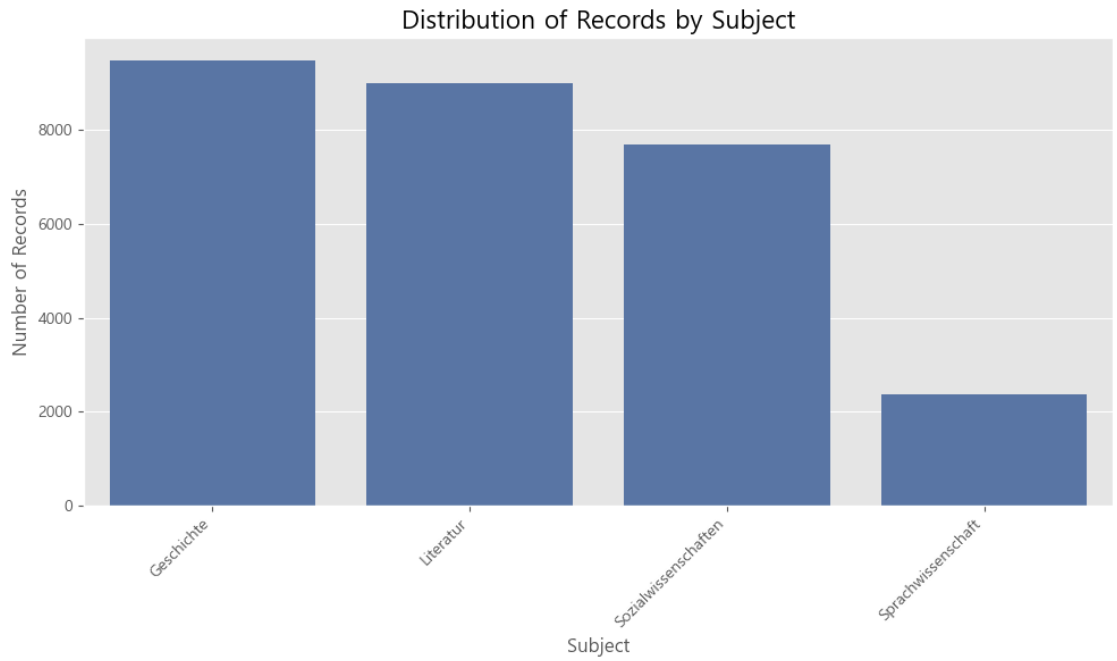
독일어 12,643개 (44.32%), 영어 8,169개 (28.64%), 프랑스어 3,180개 (11.15%), 이탈리아어 985개 (3.45%), 스페인어 805개 (2.82%), 네덜란드어 602개 (2.11%), 라틴어 등을 포함한 기타 언어 2,141개 (7.47%)로서, 독일어 외 언어 비중이 55%를 상회하여 다국어를 다룰 수 있는 모델이 필요함이 확인되었습니다.



좌측부터 독일어, 영어, 프랑스어, 이탈리아어, 스페인어

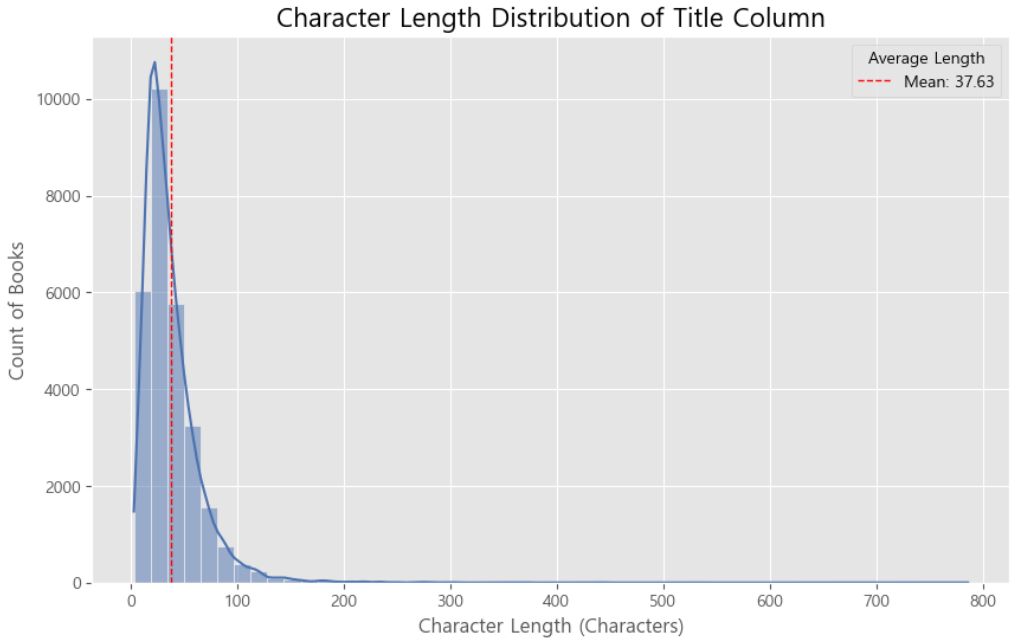
분야별 서적 구성 비율

역사 33.24%, 문학31.51%, 사회과학26.95%, 어학8.29%로 어학에서 클래스 불균형이 존재하여 클래스 가중치 부여 등이 필요함이 관찰되었습니다. DDC 코드를 기반으로 매핑되어 매핑 오류는 희박할 것으로 판단됩니다.



도서 데이터의 제목에서 최대 길이

136문자, 20단어로, pre-trained model의 최대 시퀀스 길이를 고려하여 해당 기준이상으로 토크나이저의 max_seq_length를 설정하는 것이 필요함이 확인되어 실제 fine tuning 과정에서 256으로 설정하였습니다.



3). ML 모델 학습 및 평가

로컬 환경의 하드웨어 한계로 인해 모든 training, evaluation, inference 단계는 google colab 환경에서 진행되었습니다. 저장된 최종 모델 가중치는 다음 링크에서 확인할 수 있습니다:
https://huggingface.co/jsjang0104/book-genre-classifier-bert

모델 구조

104개 언어로 pre-training된 모델인 mBERT를 사용하여 training, evaluation, inference 전 과정에 사용되는 다국어 데이터를 효과적으로 처리하고자 하였습니다. 파인튜닝 방식은 Task-specific Fine-tuning으로, 본 프로젝트가 지향하는 task인 classification에 맞게, mBERT의 마지막 Classification Head만 데이터에 맞추어 fine-tuning하였습니다.

layer 구조는 다음과 같습니다: Pooling Layer (token vector 추출 / 768 features) → Dropout (overfitting 방지 / 768 features) → Linear Layer (최종 4개 class로 매핑)

데이터 셋 및 전처리

데이터 분할 비율은 다음과 같습니다:

		전체 dataset	training set	validation set	test set
데이터 분할 비율		100%	80%	10%	10%
데이터셋 별 클래스 분포	역사	33.24	33.25	33.23	33.23
	문학	31.51	31.51	31.51	31.51
	사회과학	26.95	26.95	26.95	26.95
	어학	8.29	8.29	8.31	8.31

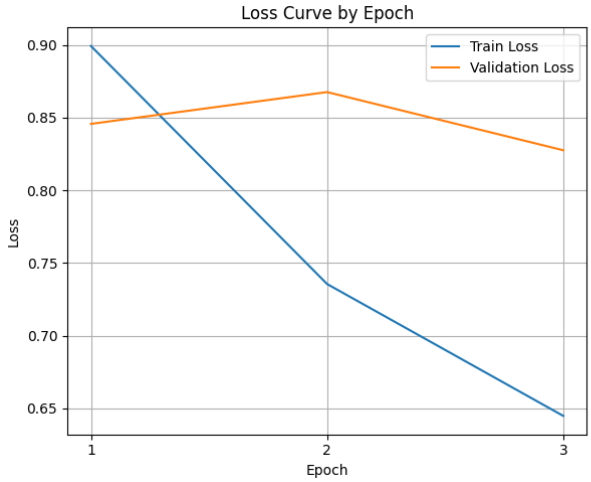
사전 데이터 분석 과정에서 '어학' 클래스에서 불균형을 확인하였기에, 특정 클래스에 대한 bias를 방지하고자 각 클래스의 빈도수에 반비례하는 가중치를 계산하여 CrossEntropyLoss 함수에 적용하였습니다.

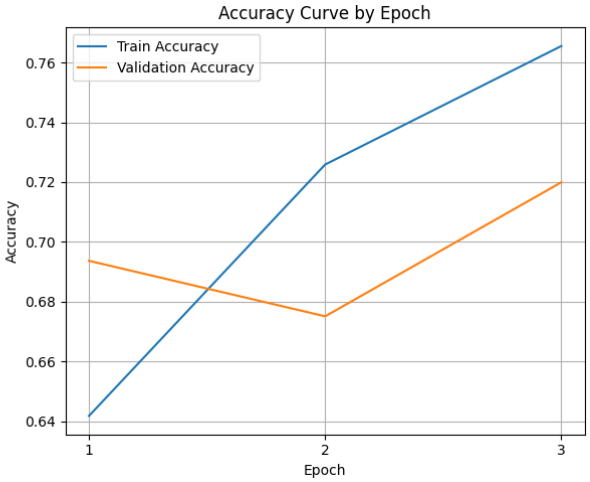
역사	문학	사회과학	어학
0.7520035853632817	0.7933307375681389	0.9277026148042149	3.01405325443787

training

학습 환경: google colab (T4 GPU)

하이퍼 파라미터 설정	최종 training 결과
learning rate: 2e-5 dropout rate: 0.2 random seed: 42 epoch 수: 3 batch size: 12	경과 시간: 1121.53초 (18분 41.53초) Train Loss: 0.6449 Train Acc: 0.7655 Val Loss: 0.8276 Val Acc: 0.7199





Loss		Accuracy	
train loss	validation loss	train accuracy	validation accuracy
0.8993 → 0.7356 → 0.6499	0.8457 → 0.8676 → 0.8276	0.6417 → 0.7259 → 0.7655	0.6937 → 0.6751 → 0.7199

training에서는 전 과정에서 안정적인 학습이 이루어졌지만, validation은 2 epoch 때 일시적으로 불안정성을 띈 모습이 관찰되었습니다. 그러나 3 epoch 때 다시 안정적으로 일반화 성능을 확보하였습니다.

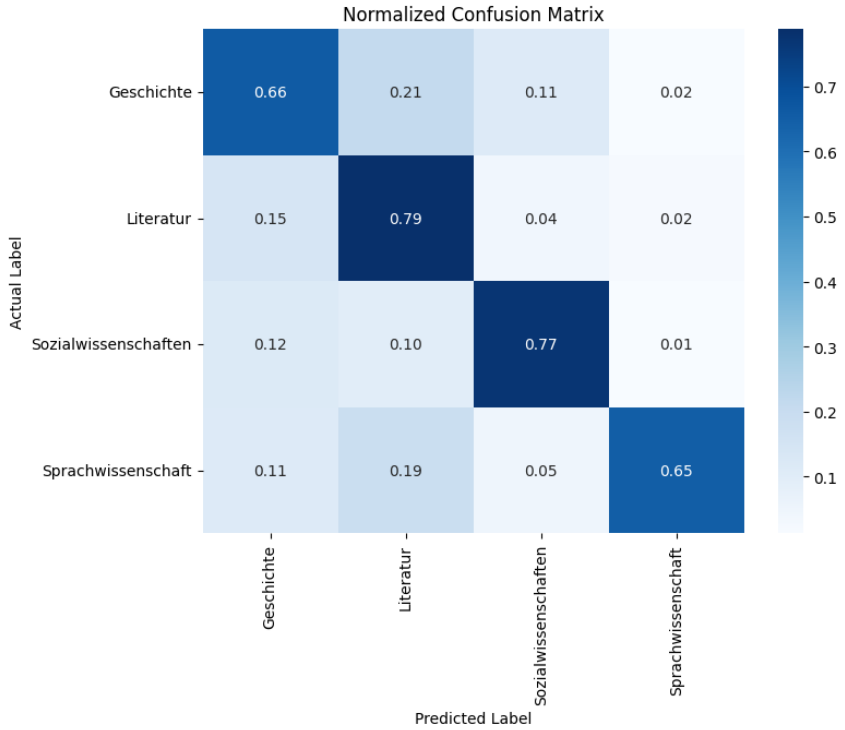
Epoch 선정 근거는 BERT의 원논문(Devlin et al., 2018) 및 일반적인 NLP fine tuning 가이드라인에 따라 권장 epoch 수인 3~4회를 근거로 초기 3회로 설정하여 학습을 시작하였습니다. 이후 3 epoch 시행에서 validation loss 및 accuracy가 적정 수중에 도달한 점, 코랩 환경의 제한된 컴퓨팅 리소스 등을 고려하여 추가 학습 없이 3 epoch에서 학습을 종료하였습니다.

evaluation

최종 성능표		class별 F1-score	
Accuracy (Overall)	0.7291	역사 (Geschichte)	0.6868
F1-Score (Weighted)	0.7284	문학 (Literatur)	0.7348
F1-Score (Macro)	0.7262	사회과학 (Sozialwissenschaften)	0.7800
Precision (Weighted)	0.7314	어학 (Sprachwissenschaft)	0.7032
Recall (Weighted)	0.7291		

최종 성능표 해석은 다음과 같습니다:

- Precision과 Recall이 조화를 이루고 있어 weighted F1-score와 overall accuracy가 비슷한 성능을 기록하였습니다.
- Macro F1-score가 weighted F1-score와 비슷한 값을 기록했으며, 이는 어학 클래스에 부여한 클래스 가중치가 효과적으로 bias를 방지했음을 시사합니다.



Actual Label에서 위측부터, Predicted Label에서 좌측부터: 역사, 문학, 사회과학, 어학

대각선 값 = Recall, 대각선이 아닌 값 = 오분류 비율 (predicted label / actual labe)

class별 해석은 다음과 같습니다:

- 역사 (Geschichte)의 약세: 역사 클래스는 가장 낮은 F1-score(0.69) 및 어학과 비슷한 수준의 Recall(0.66)을 기록하며 전반적으로 가장 약세를 보였습니다. 특히 전체 역사 class의 21%가 문학으로 오분류 된 것은 인문학 계열 내에서 고전, 신화, 특정 시대 연구 등 주제가 겹치면서 발생하는 혼동으로 추정되며, 역사 서적 제목에서 발생할 수 있는 오분류 case를 다음 inference 파트에서 언어학적 관점으로 해석할 것입니다.

- 문학 (Literatur) 강세: 문학 클래스는 F1 Score 0.73, Recall 0.79로 높은 성능을 보였습니다. 약세 클래스(역사, 어학)가 문학으로 넘어오는 노이즈에도 불구하고, 성능이 높게 유지되었다는 것은 문학의 텍스트 특징이 모델에게 잘 학습되었다고 해석됩니다.

- 사회과학 (Sozialwissenschaften) 강세: 사회과학 클래스는 주제의 경계가 명확하여 높은 성능(F1 score 0.78, Recall 0.77)을 보였습니다. 이는 문학, 어학, 역사의 인문학 계열에서는 제목에서 모호함이 다소 존재하나, 사회과학 계열에서는 제목에 사용되는 키워드가 타 클래스에 비해 비교적 명확하여(경제, 정책, 법 등) 성능이 좋게 나타난 것으로 추정됩니다.

- 어학 (Sprachwissenschaft) 약세: 어학 클래스는 F1 Score 0.7로 두번째로 낮고, Recall이 0.65로 가장 낮게 나타났습니다. 문학이나 역사로 오분류 되는 경향이 큰 것으로 보이며, 이는 가중치를 적용했음에도 불구, 데이터 양의 한계로 추정됩니다.

inference

inference 과정에서는 실제 오스트리아도서관의 도서 데이터 일부를 사용하여 모델의 실용성을 확인하였습니다.

input data csv 파일에 실제 도서 데이터인 title, author, subject가 모두 명시되어 있지만, title만 feature로 사용되었습니다. (subject 추후 inference 성능 비교를 위하여 사용됩니다.)

사진의 sample input data는 각 class 별로 5권씩 총 20권이며, 언어별 권수는 독일어 12권, 영어 2권, 한국어 5권, 라틴어 1권입니다.

```
title,author,subject
A study on tones and tonemarks in middle korean,Staffan Rosen,Sprachwissenshaft
Einführung in die Linguistik,Karl-Dieter Bünting,Sprachwissenschaft
Studien zur Texttheorie und zur deutschen Grammatik,Hugo Moser,Sprachwissenschaft
ABC der schwachen Verben,Klaere Meil / Margit Arndt,Sprachwissenschaft
라틴어 문법,프레드릭 M.윌록,Sprachwissenschaft
Austrian foreign policy yearbook,Federal Ministry for Foreign Affairs,Sozialwissenschaften
Tatsachen und Zahlen,Bundespressedienst,Sozialwissenschaften
Außenpolitischer Bericht,Bundesministerium für auswärtige Angelegenheiten,Sozialwissenschaften
Wirtschaftspolitik und Arbeitsmarkt,Hermann Kellenbenz,Sozialwissenschaften
독일 민법전,양창선,Sozialwissenschaften
Emblematium Libellus,Andreas Alciatus,Literatur
Brechts Antigone des Sophokles,Werner Hecht,Literatur
Der Mantel des Darius,Georg Scherig,Literatur
1001 Nacht. Arabische Erzählung mit 63 Illustrationen,Gustav Weil,Literatur
셰익스피어 비극론,A.C. 브라드리,Literatur
Lexikon der antiken Mythen und Gestalten,Michael Grant et al.,Geschichte
Der römische Staat I. Die Republik,Ingemar König,Geschichte
BIBLIOTHEK DER GESCHICHTE UND POLITIK DIE FRANZÖSISCHE REVOLUTION,Horst Günther,Geschichte
서양미술사,E.H.곰브리치,Geschichte
한국외국어대학교 60년사 I. 시대별 외대사,한국외국어대학교 60년사 편찬 위원회,Geschichte
```

sample_input_data

sample output data는 다음과 같습니다. output으로는 'title'과 '분류한 클래스 정보'가 담긴 csv 파일을 출력하며, inference 성능 평가 편의를 위해 actual subject를 함께 출력하도록 하였습니다. 이때 이전까지 과정에서는 confidence가 70% 이하일 때만 클래스별 confidence를 출력했지만, 실제 사용에 있어서는 모든 data에서 출력하도록 수정하였습니다. 또한 confidence 크기 별로 나열하는 Top-k 방식을 사용하여 출력된 confidence 목록을 보고 사람이 직접 선택할 수 있게끔 하였습니다.

```
Title,Actual Subject,Predicted Genre,Confidence,Top 4 Probabilities_CSV
A study on tones and tonemarks in middle korean,Sprachwissenshaft,Sprachwissenschaft,0.9685,"{'Sprachwissenschaft': np.float32(0.96846), 'Liter
Einführung in die Linguistik,Sprachwissenschaft,Sprachwissenschaft,0.9870,"{'Sprachwissenschaft': np.float32(0.9869871), 'Literatur': np.float3
Studien zur Texttheorie und zur deutschen Grammatik,Sprachwissenschaft,Sprachwissenschaft,0.9868,"{'Sprachwissenschaft': np.float32(0.9868412),
ABC der schwachen Verben,Sprachwissenschaft,Literatur,0.8378,"{'Literatur': np.float32(0.8378318), 'Geschichte': np.float32(0.076929726), 'Spra
라틴어 문법,Sprachwissenschaft,Sprachwissenschaft,0.9866,"{'Sprachwissenschaft': np.float32(0.98657817), 'Literatur': np.float32(0.009513384), '
Austrian foreign policy yearbook,Sozialwissenschaften,Sozialwissenschaften,0.9865,"{'Sozialwissenschaften': np.float32(0.98653746), 'Geschichte
Tatsachen und Zahlen,Sozialwissenschaften,Sozialwissenschaften,0.8598,"{'Sozialwissenschaften': np.float32(0.85978097), 'Geschichte': np.float3
Außenpolitischer Bericht,Sozialwissenschaften,Sozialwissenschaften,0.6173,"{'Sozialwissenschaften': np.float32(0.61734205), 'Geschichte': np.fl
Wirtschaftspolitik und Arbeitsmarkt,Sozialwissenschaften,Sozialwissenschaften,0.9868,"{'Sozialwissenschaften': np.float32(0.98683316), 'Geschic
독일 민법전,Sozialwissenschaften,Sozialwissenschaften,0.9849,"{'Sozialwissenschaften': np.float32(0.9848679), 'Geschichte': np.float32(0.0083438
Emblematium Libellus,Literatur,Literatur,0.9404,"{'Literatur': np.float32(0.94039255), 'Geschichte': np.float32(0.025928825), 'Sprachwissencha
Brechts Antigone des Sophokles,Literatur,Literatur,0.9186,"{'Literatur': np.float32(0.91863275), 'Sprachwissenschaft': np.float32(0.064763054),
Der Mantel des Darius,Literatur,Literatur,0.6657,"{'Literatur': np.float32(0.6656973), 'Geschichte': np.float32(0.26812896), 'Sozialwissenschaf
1001 Nacht. Arabische Erzählung mit 63 Illustrationen,Literatur,Literatur,0.8373,"{'Literatur': np.float32(0.83729345), 'Sprachwissenschaft': n
셰익스피어 비극론,Literatur,Literatur,0.9285,"{'Literatur': np.float32(0.9284544), 'Sprachwissenschaft': np.float32(0.056951065), 'Geschichte': n
Lexikon der antiken Mythen und Gestalten,Geschichte,Literatur,0.3900,"{'Literatur': np.float32(0.3900312), 'Sozialwissenschaften': np.float32(0
Der römische Staat I. Die Republik,Geschichte,Geschichte,0.8371,"{'Geschichte': np.float32(0.8371221), 'Sozialwissenschaften': np.float32(0.142
BIBLIOTHEK DER GESCHICHTE UND POLITIK DIE FRANZÖSISCHE REVOLUTION,Geschichte,Literatur,0.5551,"{'Literatur': np.float32(0.5550848), 'Geschichte
서양미술사,Geschichte,Literatur,0.7993,"{'Literatur': np.float32(0.7992857), 'Geschichte': np.float32(0.13989817), 'Sprachwissenschaft': np.float
한국외국어대학교 60년사 I. 시대별 외대사,Geschichte,Sprachwissenschaft,0.9820,"{'Sprachwissenschaft': np.float32(0.9820401), 'Literatur': np.float32
```

sample_output_data

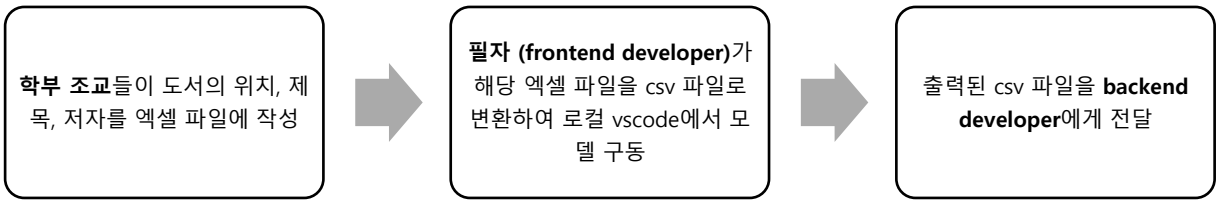
evaluation에서 살펴봤던 것과 유사하게, 문학과 사회과학 클래스에서는 높은 accuracy를 보였지만 어학과 역사에서 오분류가 발생하였습니다. 오분류 case별 해석을 살펴보면 다음과 같습니다:

제목	Actual Label	Predicted Label	confidence	해석
ABC der schwachen Verben	어학	문학	0.8378	문법 교재임에도 'ABC'나 'Verben'이라는 단어가 학습 과정에서 문학 학습 자료와 겹쳐 해석되어 발생한 것으로 추정
Lexikon der antiken Mythen und Gestalten	역사	문학	0.3900	고대 신화 관련 제목이 문학/역사/사회과학 경계에 걸쳐 있어 모델이 가장 낮은 confidence로 예측하며 실패
BIBLIOTHEK DER GESCHICHTE UND POLITIK DIE FRAN...	역사	문학	0.5551	'GESCHICHTE(역사)'가 명시되어 있음에도 문학으로 오분류 'BIBLIOTHEK(뜻: 도서관, 총서)'가 '문학 총서'의 제목 패턴과 유사하게 인식되었거나, '정치'라는 키워드가 사회과학 계열과 인접하게 해석되면서 모델이 모호함에 빠진 것으로 추정
서양미술사	역사	문학	0.7993	'미술사'는 역사 키워드이나, '미술'이라는 주제가 인문학 내에서 문학과 인접한 예술 영역으로 분류되는 경향이 있어 모델이 두 클래스 간의 경계를 혼동한 것으로 추정
한국외국어대학교 60년사 I. 시대별 외대사	역사	어학	0.9820	'70년사'는 역사 키워드이나, 모델이 '외국어'키워드에 더 가중치를 부여하여 어학으로 분류한 것으로 추정

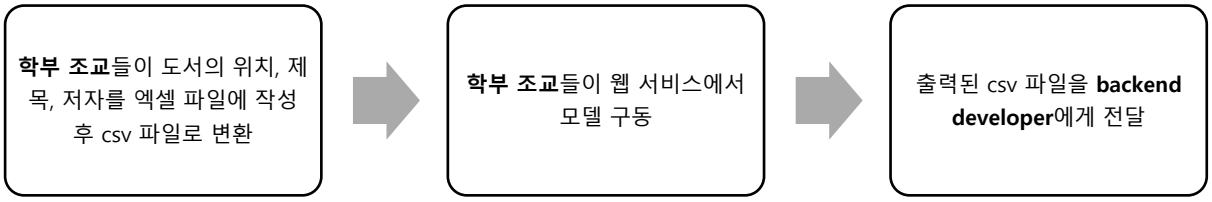
4). 서비스 개발

기존에는 학부조교들이 작성한 엑셀파일을 취합하여 필자 본인이 로컬에서 모델을 동작시켜 분야 라벨링에 이용하려고 했습니다. 하지만 12월 11자 발표를 통한 피드백에서 서비스 배포의 필요성을 느껴 gradio를 통해 구현하였습니다. 사용자 친화적인 UI를 배포하여 추후 독일어과의 다른 IT 비전공자 조교들도 csv 파일과 웹 링크만 있으면 편하게 이용 가능하게 만들고자 하는 목적입니다.

기존 방식: developer가 직접적으로 관여하지 않으면 도서 목록 관리가 힘듦



서비스 개발 이후: frontend developer가 관여하지 않아도 도서 목록 구축 pipeline을 구동할 수 있게 됨



현재는 도서관 홈페이지에 관리자 interface를 따로 구현해놓지 않았지만, 추후 관리자용 UI를 따로 만들어 본 프로젝트의 서비스 웹페이지와 연결되도록 한다면 개발자의 개입 없이 IT 비전공자들끼리도 도서 목록 pipeline을 관리할 수 있습니다.

3. 모델을 서비스로 만든 구조

Hugging Face Spaces를 활용하여 app.py 코드를 배포하였습니다. public URL은 다음과 같습니다: https://huggingface.co/spaces/jsjang0104/book-genre-classifier-service

1). 사용 방법

1 서비스 접속: https://huggingface.co/spaces/jsjang0104/book-genre-classifier-service

2 파일 업로드: 도서의 'title', 'author', 'location'이 담긴 csv 파일을 업로드 후 '분석 시작'을 클릭합니다.

3 검수 및 수정: AI가 분석한 분야 순위를 보고 가장 적절한 클래스를 사용자가 직접 선택합니다: 어학

(Sprachwissenschaft), 문학(Literatur), 역사(Geschichte), 사회과학(Sozialwissenschaften), 기타(Sonstiges)

- 모델은 기타(Sonstiges) 분류를 지원하지 않으며, 기타 선택에 관한 판단은 사용자의 몫입니다.

- 확신도 0.85 이하인 도서는 빨간색으로 표시됩니다.

4 결과 다운로드: 'subject' 열이 자동으로 채워진 csv파일이 다운로드할 수 있도록 제공됩니다.

2). 시스템 구조

1 app.py: gradio 및 모델 코드를 담고 있습니다. 코드가 시행되면 Hugging Face에 저장된 가중치 (https://huggingface.co/jsjang0104/book-genre-classifier-bert)를 불러와서 모델이 로드됩니다. 기존 inference 과정에서는 비교를 위해 actual subject를 같이 출력하도록 했지만, 해당 모델에서는 input으로 actual subject를 받지 않습니다. 또한 기존 training 과정에서는 confidence를 표시할 threshold를 70%로 설정하였지만, 실제 도서관 업무에 적용할 때는 높은 정확성과 사람의 직접적인 검수 필요성 강조를 위하여 85%로 상향 조정하여 그 이하를 기록한 도서는 빨간색으로 표시하였습니다.

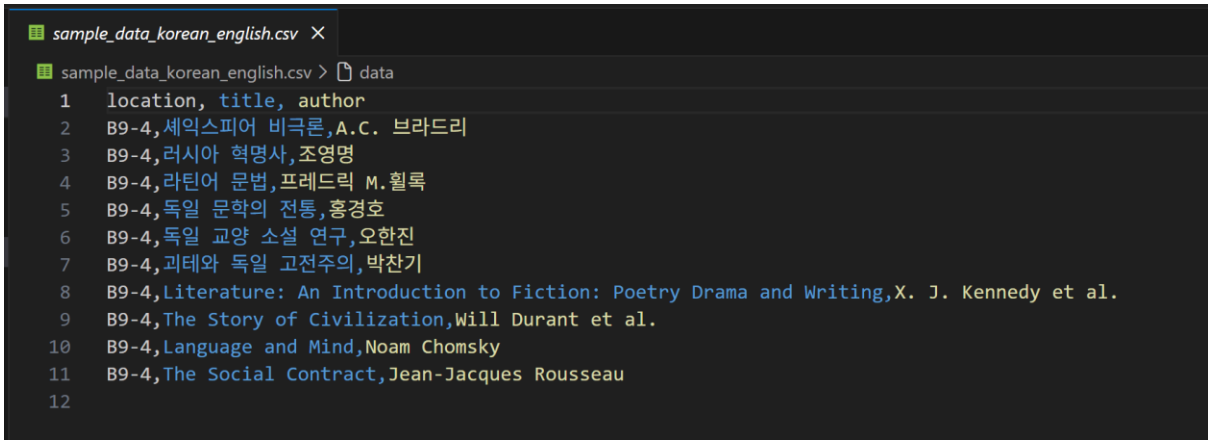
2 Hugging Face 서버: Gradio 코드가 시행되며, 별도의 로컬 서버 구동 없이도 누구나 서비스에 24시간 접근할 수 있습니다.

3 input and output: 사용자는 title, author, location이 담긴 csv 파일을 로드하고, 클래스별 모델의 확신도를 받으며 웹상에서 선택합니다. 선택한 결과물은 기존 도서 정보 옆에 추가되어 csv 파일로 다운로드 할 수 있게 제공됩니다.

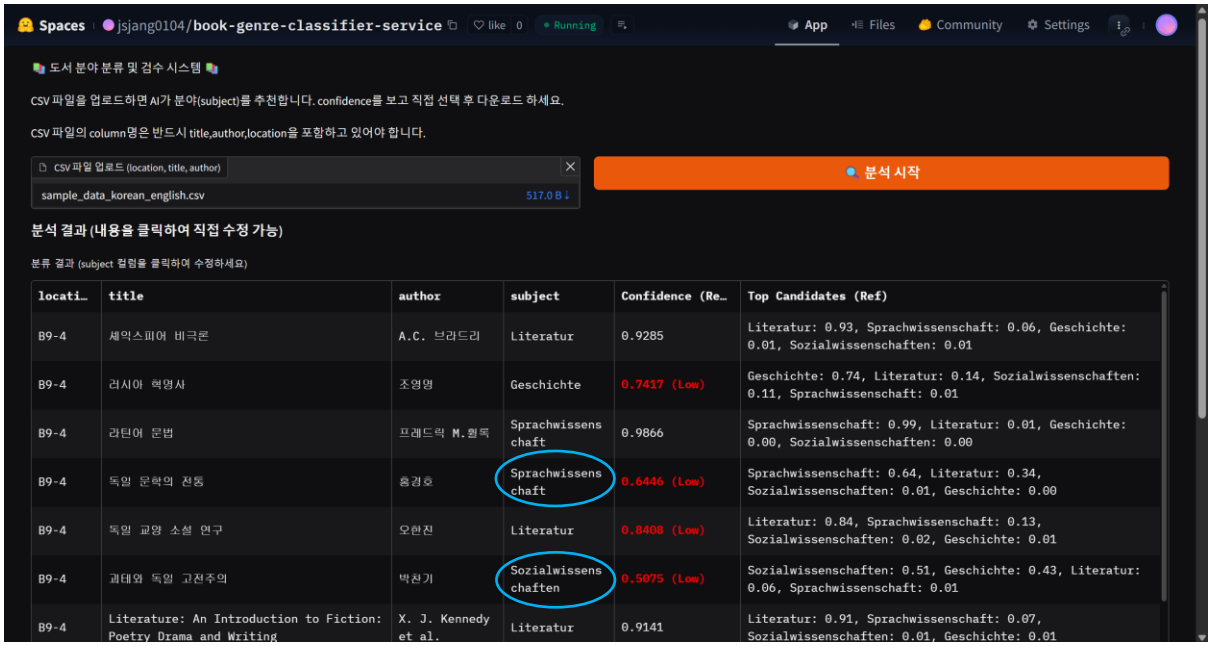
4. 실제 사용 결과

채점 편의를 위해 zip파일 내에 한국어와 영어로만 이루어진 도서 목록 샘플(sample_data_korean_english) 을 포함시켰으며, 해당 파일에 대한 실제 사용 과정입니다:

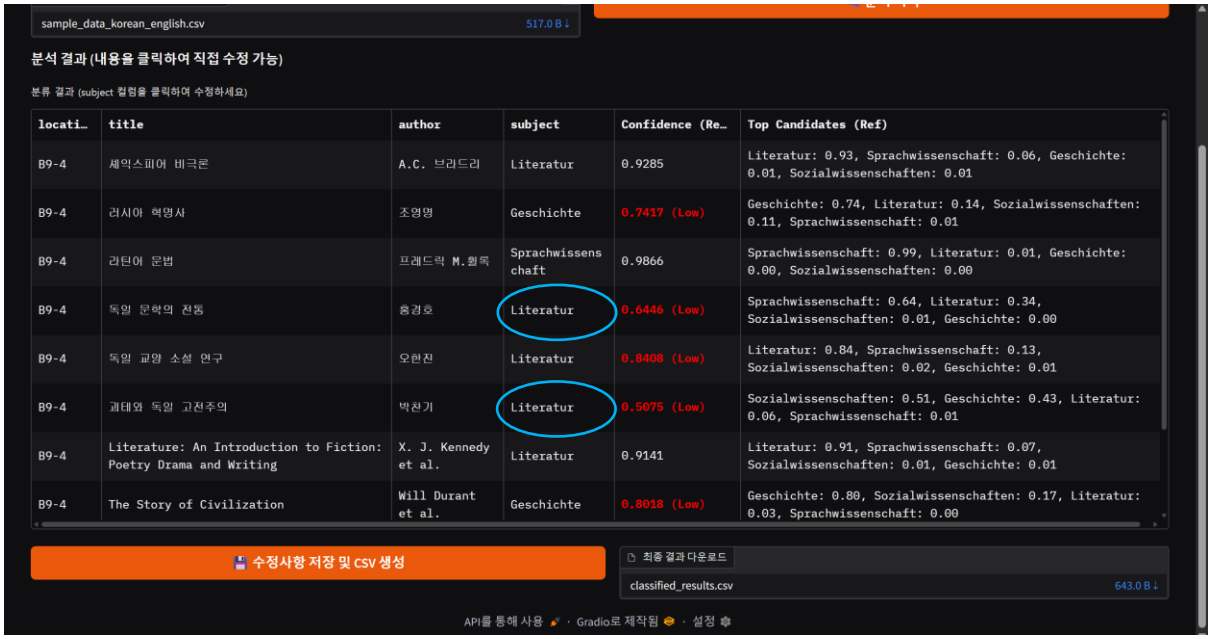
1. 도서 목록 샘플입니다.



2. 웹 상에 csv 파일 업로드 후 분석 시작을 클릭합니다. 일정 시간 후 모델이 subject, confidence, top candidates를 출력합니다. subject는 top candidates 중 첫 번째 클래스입니다.



이때 confidence가 낮게 측정된 두 도서 <독일 문학의 전통>과 <괴테와 독일 고전주의>의 top 1에서 오분류가 발생했습니다. (파란색 동그라미로 표시)



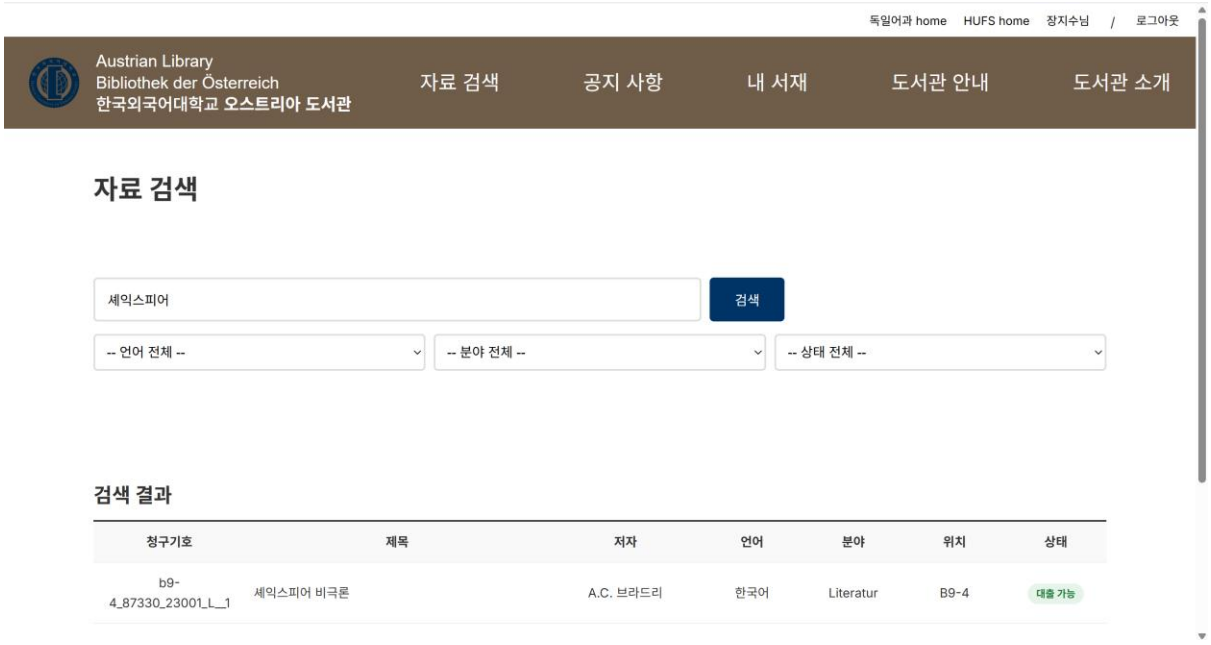
각각 Sprachwissenschaft(어학), Sozialwissenschaft(사회과학)이 아닌 Literatur(문학)으로 바꿔줍니다.

3. '수정사항 저장 및 csv 생성'을 누른 후 최종 결과를 다운로드합니다.

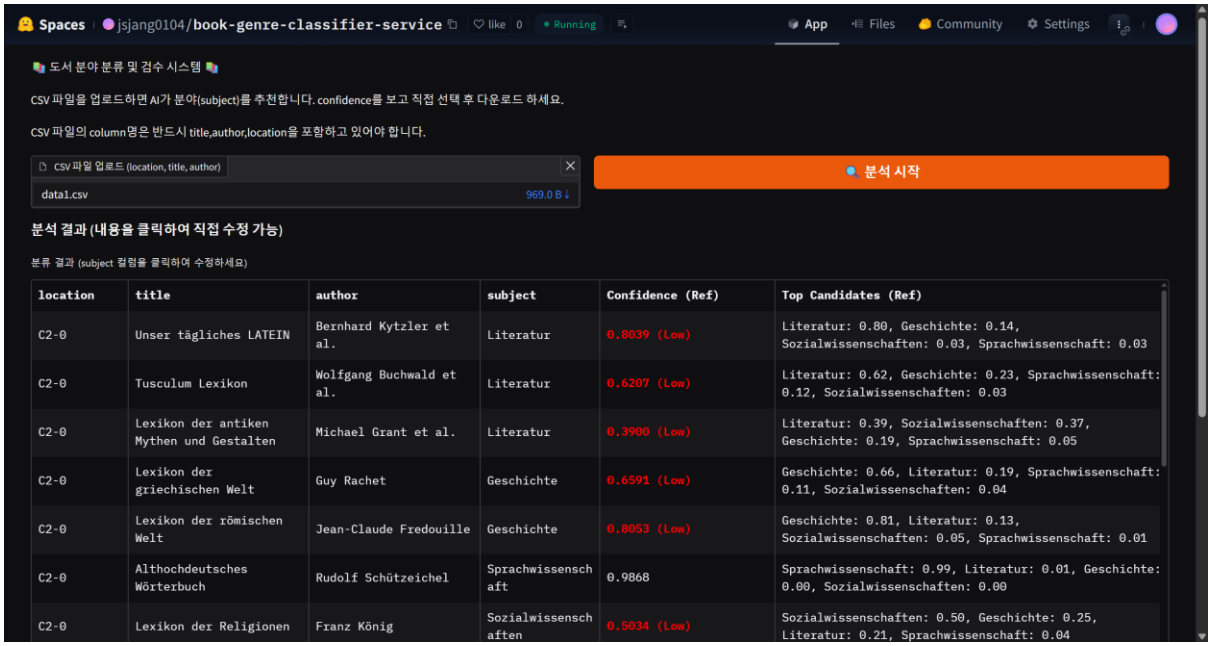
```csv
location,title,author,subject
B9-4,셰익스피어 비극론,A.C. 브라드리,Literatur
B9-4,러시아 혁명사,조영명,Geschichte
B9-4,라틴어 문법,프레드릭 M.휠록,Sprachwissenschaft
B9-4,독일 문학의 전통,홍경호,Literatur
B9-4,독일 교양 소설 연구,오한진,Literatur
B9-4,괴테와 독일 고전주의,박찬기,Literatur
B9-4,Literature: An Introduction to Fiction: Poetry Drama and Writing,X. J. Kennedy et al.,Literatur
B9-4,The Story of Civilization,Will Durant et al.,Geschichte
B9-4,Language and Mind,Noam Chomsky,Sprachwissenschaft
B9-4,The Social Contract,Jean-Jacques Rousseau,Sozialwissenschaften
```

결과 csv 파일입니다.

4. 해당 도서 목록을 데이터 베이스로 옮긴 후 도서관 홈페이지에서 확인합니다.

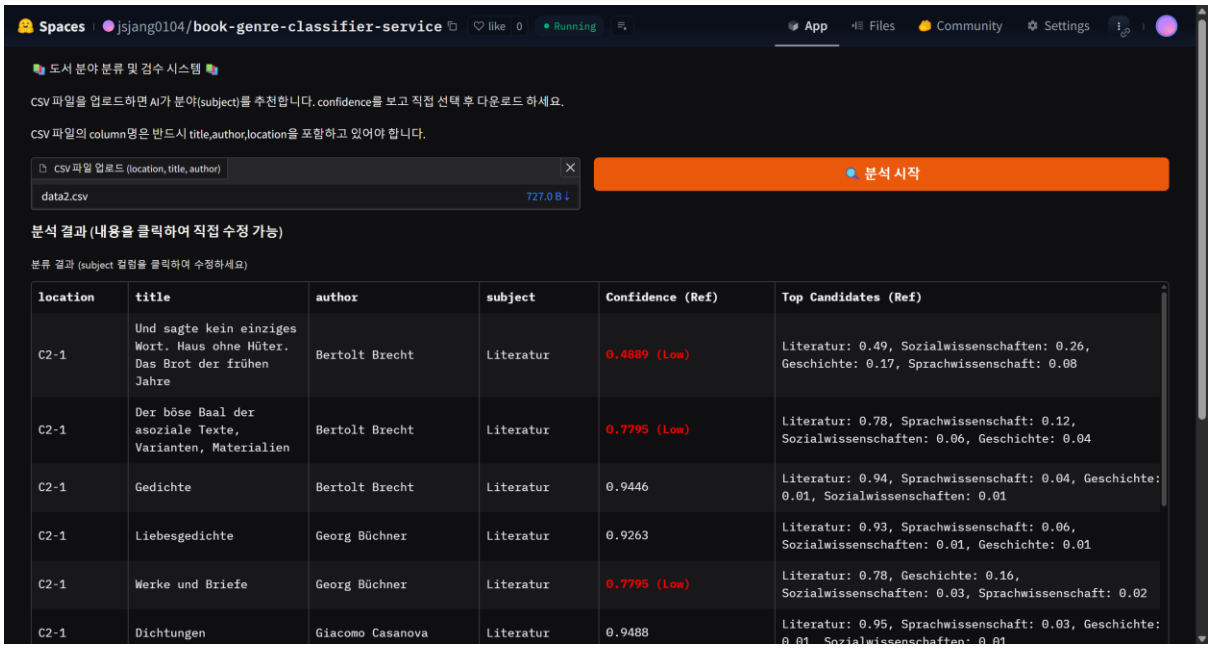


상기 과정에서 사용한 파일 외에 다른 도서 목록 csv 파일 4 건에 대해서도 실제 사용을 진행하였습니다 (총 5 회 사용):



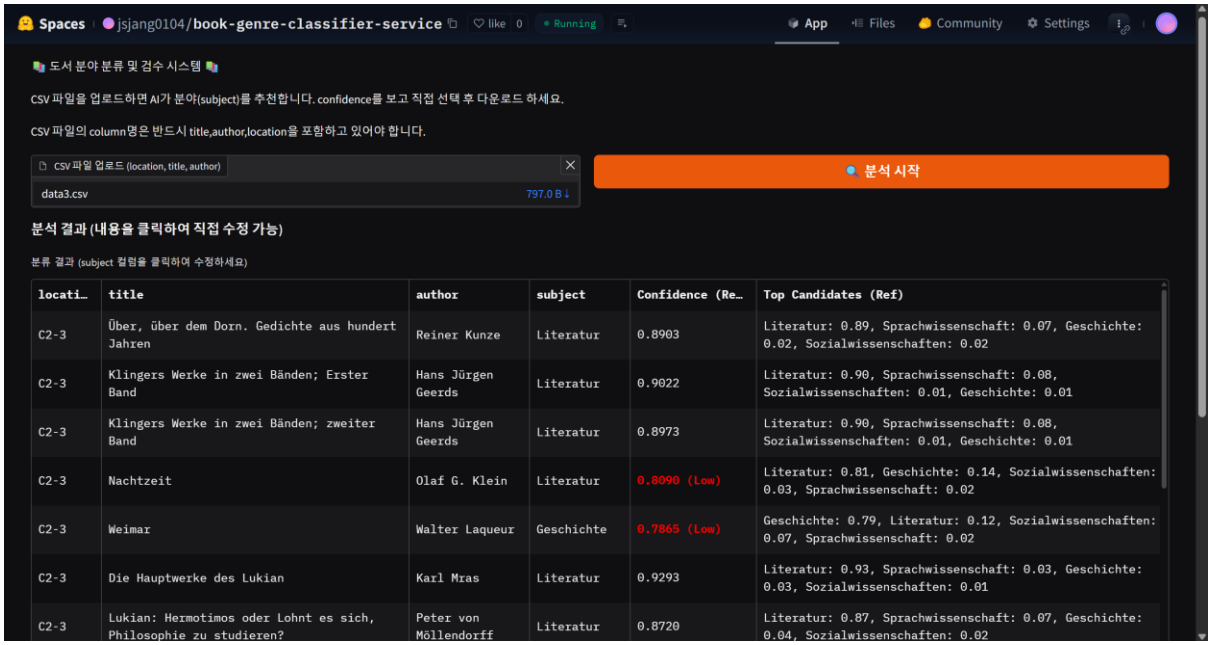
location	title	author	subject	Confidence (Ref)	Top Candidates (Ref)
C2-0	Unser tägliches LATEIN	Bernhard Kytzler et al.	Literatur	0.8039 (Low)	Literatur: 0.80, Geschichte: 0.14, Sozialwissenschaften: 0.03, Sprachwissenschaft: 0.03
C2-0	Tusculum Lexikon	Wolfgang Buchwald et al.	Literatur	0.6207 (Low)	Literatur: 0.62, Geschichte: 0.23, Sprachwissenschaft: 0.12, Sozialwissenschaften: 0.03
C2-0	Lexikon der antiken Mythen und Gestalten	Michael Grant et al.	Literatur	0.3900 (Low)	Literatur: 0.39, Sozialwissenschaften: 0.37, Geschichte: 0.19, Sprachwissenschaft: 0.05
C2-0	Lexikon der griechischen Welt	Guy Rachet	Geschichte	0.6591 (Low)	Geschichte: 0.66, Literatur: 0.19, Sprachwissenschaft: 0.11, Sozialwissenschaften: 0.04
C2-0	Lexikon der römischen Welt	Jean-Claude Fredouille	Geschichte	0.8053 (Low)	Geschichte: 0.81, Literatur: 0.13, Sozialwissenschaften: 0.05, Sprachwissenschaft: 0.01
C2-0	Althochdeutsches Wörterbuch	Rudolf Schützeichel	Sprachwissenschaft	0.9868	Sprachwissenschaft: 0.99, Literatur: 0.01, Geschichte: 0.00, Sozialwissenschaften: 0.00
C2-0	Lexikon der Religionen	Franz König	Sozialwissenschaften	0.5034 (Low)	Sozialwissenschaften: 0.50, Geschichte: 0.25, Literatur: 0.21, Sprachwissenschaft: 0.04

(1)



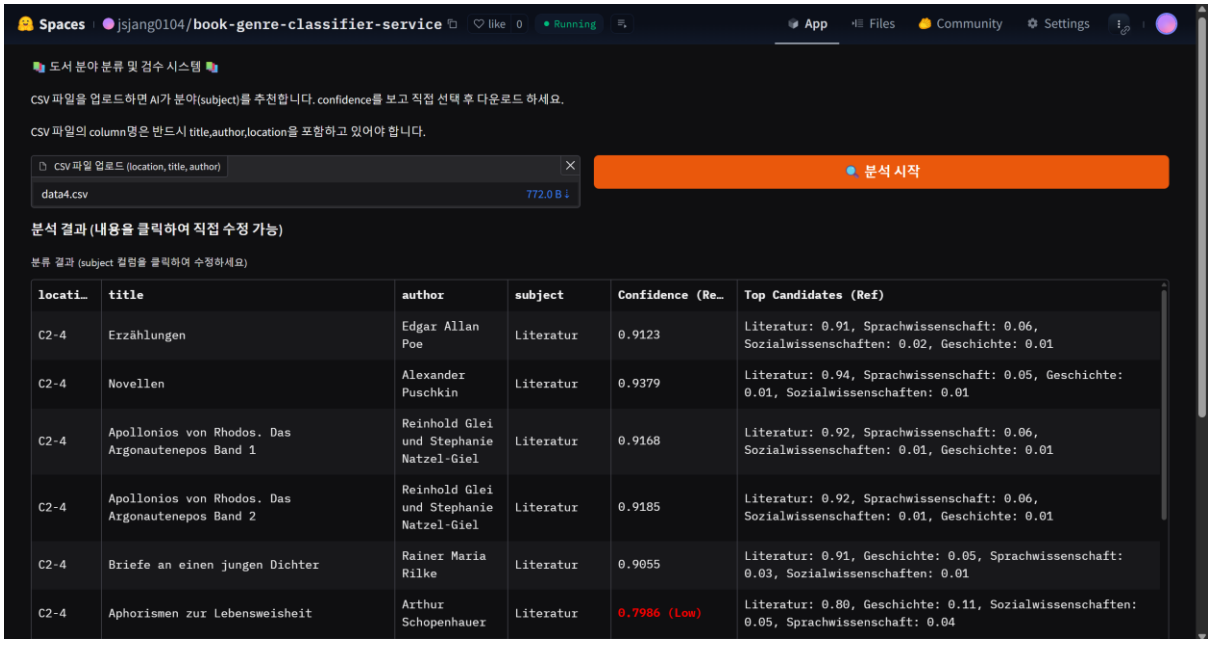
location	title	author	subject	Confidence (Ref)	Top Candidates (Ref)
C2-1	Und sagte kein einziges Wort. Haus ohne Hüter. Das Brot der frühen Jahre	Bertolt Brecht	Literatur	0.4889 (Low)	Literatur: 0.49, Sozialwissenschaften: 0.26, Geschichte: 0.17, Sprachwissenschaft: 0.08
C2-1	Der böse Baal der asoziale Texte, Varianten, Materialien	Bertolt Brecht	Literatur	0.7795 (Low)	Literatur: 0.78, Sprachwissenschaft: 0.12, Sozialwissenschaften: 0.06, Geschichte: 0.04
C2-1	Gedichte	Bertolt Brecht	Literatur	0.9446	Literatur: 0.94, Sprachwissenschaft: 0.04, Geschichte: 0.01, Sozialwissenschaften: 0.01
C2-1	Liebesgedichte	Georg Büchner	Literatur	0.9263	Literatur: 0.93, Sprachwissenschaft: 0.06, Sozialwissenschaften: 0.01, Geschichte: 0.01
C2-1	Werke und Briefe	Georg Büchner	Literatur	0.7795 (Low)	Literatur: 0.78, Geschichte: 0.16, Sozialwissenschaften: 0.03, Sprachwissenschaft: 0.02
C2-1	Dichtungen	Giacomo Casanova	Literatur	0.9488	Literatur: 0.95, Sprachwissenschaft: 0.03, Geschichte: 0.01, Sozialwissenschaften: 0.01

(2)



locati…	title	author	subject	Confidence (Re…	Top Candidates (Ref)
C2-3	Über, über dem Dorn. Gedichte aus hundert Jahren	Reiner Kunze	Literatur	0.8903	Literatur: 0.89, Sprachwissenschaft: 0.07, Geschichte: 0.02, Sozialwissenschaften: 0.02
C2-3	Klingers Werke in zwei Bänden; Erster Band	Hans Jürgen Geerds	Literatur	0.9022	Literatur: 0.90, Sprachwissenschaft: 0.08, Sozialwissenschaften: 0.01, Geschichte: 0.01
C2-3	Klingers Werke in zwei Bänden; zweiter Band	Hans Jürgen Geerds	Literatur	0.8973	Literatur: 0.90, Sprachwissenschaft: 0.08, Sozialwissenschaften: 0.01, Geschichte: 0.01
C2-3	Nachtzeit	Olaf G. Klein	Literatur	0.8090 (Low)	Literatur: 0.81, Geschichte: 0.14, Sozialwissenschaften: 0.03, Sprachwissenschaft: 0.02
C2-3	Weimar	Walter Laqueur	Geschichte	0.7865 (Low)	Geschichte: 0.79, Literatur: 0.12, Sozialwissenschaften: 0.07, Sprachwissenschaft: 0.02
C2-3	Die Hauptwerke des Lukian	Karl Mras	Literatur	0.9293	Literatur: 0.93, Sprachwissenschaft: 0.03, Geschichte: 0.03, Sozialwissenschaften: 0.01
C2-3	Lukian: Hermotimos oder Lohnt es sich, Philosophie zu studieren?	Peter von Möllendorff	Literatur	0.8720	Literatur: 0.87, Sprachwissenschaft: 0.07, Geschichte: 0.04, Sozialwissenschaften: 0.02

(3)



locati…	title	author	subject	Confidence (Re…	Top Candidates (Ref)
C2-4	Erzählungen	Edgar Allan Poe	Literatur	0.9123	Literatur: 0.91, Sprachwissenschaft: 0.06, Sozialwissenschaften: 0.02, Geschichte: 0.01
C2-4	Novellen	Alexander Puschkin	Literatur	0.9379	Literatur: 0.94, Sprachwissenschaft: 0.05, Geschichte: 0.01, Sozialwissenschaften: 0.01
C2-4	Apollonios von Rhodos. Das Argonautenepos Band 1	Reinhold Glei und Stephanie Natzel-Giel	Literatur	0.9168	Literatur: 0.92, Sprachwissenschaft: 0.06, Sozialwissenschaften: 0.01, Geschichte: 0.01
C2-4	Apollonios von Rhodos. Das Argonautenepos Band 2	Reinhold Glei und Stephanie Natzel-Giel	Literatur	0.9185	Literatur: 0.92, Sprachwissenschaft: 0.06, Sozialwissenschaften: 0.01, Geschichte: 0.01
C2-4	Briefe an einen jungen Dichter	Rainer Maria Rilke	Literatur	0.9055	Literatur: 0.91, Geschichte: 0.05, Sprachwissenschaft: 0.03, Sozialwissenschaften: 0.01
C2-4	Aphorismen zur Lebensweisheit	Arthur Schopenhauer	Literatur	0.7986 (Low)	Literatur: 0.80, Geschichte: 0.11, Sozialwissenschaften: 0.05, Sprachwissenschaft: 0.04

(4)

5. 배운 점 및 개선 방향

1). 모델 개선 방향

1. **german BERT 사용**: input sequence 는 독일어, 영어, 한국어, 라틴어 등이 다국어이지만 output sequence 는 독일어로 고정되어 있어 mBERT 사용에도 불구하고 타 언어로의 매핑 과정에서 혼란이 발생한 경황이 자주 포착되었습니다. mBERT 보다 도메인 지식을 더 잘 이해하는 German BERT 등의 모델을 사용하고, 독일어 외 언어는 먼저 번역 후 처리하는 방식을 고려할 수 있습니다.

2. **어학 클래스 데이터 추가 수집**: Weighted Loss 를 사용했음에도 불구하고, minority class 인 어학 관련 도서가 문학 또는 역사로 잘못 예측되는 경우가 많았습니다. 이는 어학 자료가 문학 작품 해설이나 언어 변천사 연구 등 다른 인문학 분야와 제목 키워드를 공유하는 경우가 많음으로 추정되며, 어학 클래스에 대한 추가 데이터를 수집하여 클래스 불균형 문제를 해결할 수 있습니다.

2). 느낀 점

AI 모델의 구상부터 서비스 배포까지 전과정을 경험해보는 실무적 활동이라 매우 유익하게 다가왔습니다. 더군다나 개인적으로 진행하고있는 오스트리아도서관 디지털화 프로젝트와 연계시킬 수 있어서 더욱 실용적으로 느껴졌습니다.

여담으로, 도서관 디지털화 프로젝트를 위해 도서 청구기호 생성 파이썬 모델을 개별적으로 만들어 뒀는데, 이 청구기호 모델 또한 웹서비스에 덧붙이고 자동으로 database 와 연결되기까지 하는 확장성까지 구현한다면 더욱 더 실용적인 all-in-one pipeline 을 구축할 수 있을 것 같습니다.