Son GiRak

(직무)NLP Developer

Birthday	1996. 11. 18
Email	rlfkr1234@naver.com
Mobile	010-6669-2361
Address	대구광역시 서구 내당동 970-2

소개 / About Me

- NLP 알고리즘을 사용하여 딥러닝 개발 구현
- 데이터 EDA 및 전처리를 활용하여 업무 의사 결정 가능
- 다양한 데이터 분석 Tool 사용 경험 보유 (Python, Pytorch, Excel, SPSS, Qgis)
- 데이터 분석 업무를 통하여 프로세스에 대한 이해

기술 스택 / Skill Set

기능 구현 등의 사용 경험이 있는 Skill Set

구분	Skill
Programing Languages	Python, Excel, SPSS, Qgis
Framework/ Library	Pytorch, Tensorflow, Keras, Pandas, Numpy, Re
Server	-
Tooling/ DevOps	-
Environment	Github, GCP
ETC	Git, HuggingFace

사용경험은 없으나, 이론적 지식이 있는 Skill Set

구분	Skill
Programing Languages	-
Framework/ Library	
Server	-
Tooling/ DevOps	-
Environment	- Unbunto
ETC	- Docker

과제 프로젝트 경험

문자 텍스트 데이터를 사용한 요약문 생성 AI 개발

작업 기간	2022. 11 – 2022. 12
인력 구성	3 명
프로젝트 목적	대화형 텍스트 데이터를 사용하여 상황 요약
주요업무 및 상세역할	1. 데이터 EDA - 대화 카테고리별 특성을 파악하기 위하여 수행 - 텍스트 데이터에서 다음과 특성을 파악하였다. (1) 구어체이기 때문에 문법이 맞지 않은 부분이 많다. (2) 결측치는 존재하지 않으나 중복값은 존재 (3) 문장 길이가 비정상적으로 긴 데이터가 존재 (4) 개인정보 데이터는 라벨링 처리가 되어있음 (5) 연속적인 자음 및 모음이 텍스트를 차지하고 있음 2. 전처리 - 학습 효율을 높이기 위하여 수행 (1) 정규표현식을 사용한 정제 (2) 대화 데이터 유형별 메타데이터 삽입 (3) 개인정보 라벨링 문자 스페셜 토큰에 삽입 3. 모델링 - KoGPT-2 를 사용 - 성능 개선을 위해 아래와 같은 실험을 수행 (1) Beam Search = 5 적용 (2) Sampling 적용 - 수행 결과 Beam 을 적용하는 것이 더 높은 성능을 보여 Beam 을 채택 4. 일정 조율 (1) 매일 팀원의 업무 진행도 확인 (2) 매주 주말 해소해야 할 문제가 무엇인지 토의 진행 (3) 분석 중 발생하는 애로 사항 개선을 위한 자문 요청
사용언어 및 개발환경	사용언어 : Python 프레임워크 : Pytorch 개발환경 : 리눅스, GCP, JupyterNotebook Etc : HuggingFace
느낀 점	1. 다양한 텍스트 처리 전략 필요 - 전처리 단계에서 성능을 높일 수 있는 전략을 더 확인하는 것이 필요 2. 데이터 시각화 능력 향상 필요 - 수 많은 텍스트 데이터 간 특성을 파악하기 위하여 시각화 능력이 필요할 것으로 예상됨. 3. 딥러닝 프레임워크 기본기 강화 - 부트캠프에서 TensorFlow/Keras 만을 배운 상태에서 프로젝트는 Pytorch 로 진행하였기 때문에 코드의 대한 이해도가 부족 - ipynb 파일이 아닌 py 파일로 만들어 실무 환경에서의 적응이 필요할 것으로 보임.
참고자료	1. GitHub : https://github.com/AIFFEL-NLP-PROJECT/Aiffelthon 2. 생성 전략 : https://littlefoxdiary.tistory.com/46 3. 논문 학습 : https://www.notion.so/StudyWithUs-97d1570c7863469eb37a9c405798376b

영화관 임장권 데이터 시각화

작업 기간	2022. 07 – 2022. 07
인력 구성	3 명
프로젝트 목적	시기별 영화 배급의 트렌드 확인
주요업무 및 상세역할	1. 데이터 EDA - 입장권 데이터의 특성을 파악하기 위하여 진행하였고 다음과 특성을 파악하였다. (1) 결측치는 존재 (2) 관객수가 1 이하인 데이터 존재 (3) 특정 장르(성인물) 관객 수 분포 파악 2. 전처리 - 원활한 분석을 위하여 수행 (1) 결측치 제거 (2) 특정 장르(성인물) 데이터 삭제 (3) 관객수 400 이하 데이터 삭제 3. 데이터 시각화 - Python Matplotlib 사용 (1) 각 년도 장르별 관객수 및 상영관 수 차트 (2) 각 년도 장르 키워드 워드 클라우드 (3) 각 년도 분기별 장르 키워드 코사인 유사도 4. 분석 결과 (1) 관객 수가 가장 많을 때 상영 영화 수가 가장 적음 (2) 특정 장르(공포, 액션)은 많이 개봉되는 시기와 관객 수가 동일하게 분포 (3) 각 년도 가장 많이 개봉되는 장르는 분석 기간 내 대부분 유사 (4) 각 년도 분기별 장르 키워드 코사인 유사도는 분석 기간 내 대부분 일정
사용언어 및 개발환경	사용언어 : Python 라이브러리 : Matplotlib, WordCloud 개발환경 : JupyterNotebook
느낀 점	1. 영화 트렌드를 알기 위해서 추가적인 요소를 고려 - 배급의 영향을 주는 다른 요소를 찾아 분석에 활용해야 할 필요가 있다고 생각
참고자료	-

예천군 불법주정차 완화 데이터 분석

작업 기간	2020. 10 – 2020. 12
인력 구성	2 명
프로젝트 목적	예천군 불법주정차 완화를 위한 주차장 최적입지 선정
주요업무 및 상세역할	1. 데이터 정제 및 현황 파악 - 불법주정차 데이터에서 사용이 불가능한 데이터를 정제하고 현황을 파악 (1) 총 4,130 건 중 3,870 건을 사용하고 260 건은 결측치로 지정 (2) 상위 적발 구역, 월, 일 , 시간대별 적발 건수 확인, 전체 주차면수 파악 2. 가설 설정 - 불법주정차가 많이 발생하는 장소의 특성을 고려하여 가설 설정 (1) 건축물 사용 용도에 따른 가점과 불법주정차 단속 건수가 관계가 있다고 판단 (2) 주차관리 핵심 지역 = 건축물 가점 + 불법주정차 적발건수 3. 데이터 분석 및 결과 - 실제로 이런 관계가 발생할 지 여부 파악을 위한 Qgis 를 사용하여 시각화 작업을 거치고 분석 결과 도출 (1) 적발 장소 기준 반경 100m 를 기준으로 주차관리 핵심지수를 계산 (2) 예천군이 소유한 공유지 위치를 주차관리 핵심지역과 함께 나타내어 주차장 후보지 확인 4. 보고서 작성 - 주차장 신설 근거 확보를 위한 보고서를 작성 (1) 분석 개요, 방법, 통계 분석, 최종 입지 선정 근거 작성 (2) 총 5 곳을 최적 입지로 선정하고 137 면의 주차면수 확보가 가능하다는 결론 제시 (3) 현재 불법주정차 단속의 한계점과 데이터 표준화가 필요하다는 시사점 제시
사용언어 및 개발환경	개발툴 : Excel, SPSS, Qgis
느낀 점	1. 정제되지 않은 Raw 데이터 활용의 어려움 - 대부분 지자체는 데이터 분석 관련 전문가가 없기 때문에 데이터가 효율적으로 축적되지 않기에 분석에 활용하려면 까다로운 정제 과정이 필요하다는 것을 느낌
참고자료	1. 문헌 : 공공빅데이터 표준분석 모델 매뉴얼 22p (2018)

학력

- 2015.03 ~2022.02 안동 대학교 경제학 졸업
- 2012.03 ~2015.02 세명 고등학교 졸업

자격증

- 사회조사분석사 2 급 (2020. 11 취득)
- 컴퓨터활용능력 1 급 (2020. 06 취득)
- ADsP 빅데이터 준분석가 (2019. 10 취득)
- 한국사능력검정시험 고급 (2019. 02 취득)
- TOEIC 635 (2021. 07 취득)

교육 내용

- 2022. 06 ~2022. 12 AIFFEL AI 부트캠프 모두의연구소/AIFFEL
- 2020. 08 ~2020. 09 2020 년 공공 빅데이터 청년인턴십 확대운영 행정안전부/한국지능정보사회진흥원

자기소개서

1. (지원 동기 및 입사 포부) AI Data 분석 역량을 가진 지원자 손기락입니다.

- AI 알고리즘 개발과 서비스 기획 역량을 가지고 있어서 지원하게 되었습니다. AI 테크닉 중 자연어 처리 영역에 많은 관심이 생겼고 개발 능력을 기르기 위해 제가 학습한 내용을 GitHub 에 정리하여 꾸준히 기록하고 있습니다. 제가 다룰 수 있는 언어는 기본적으로 Python 이 가능하며, 프레임워크는 Pytorch, TensorFlow/keras 가 있습니다. 특히 Pytorch 는 실무 환경에서의 적응을 위하여 CLI 환경에서 코드를 작동할 수 있도록 능력을 길렀습니다. 부트 캠프에서 최종 프로젝트를 수행할 때 기존 서비스가 가지는 문제점을 분석하고 이를 해소하는 방법이 무엇이 있을지 고민해보고 실제로 프로젝트를 완료하였습니다. 이런 역량을 고려하였을 때 AI M 팀에 지원하여 AI 를 설계하고 서비스를 기획하는 사람이 되고 싶습니다. 만약 입사를 하게 된다면 더 나은 서비스를 제공하는 감성 분석 전문가로 성장하고 싶습니다.

2. (직무를 위한 준비) 과감한 커리어 전환을 도전하였습니다

- 리서치 연구원에서 딥러닝 개발자로 커리어 전환을 시도하기 위해 도전하였습니다. 회사를 퇴사한 바로 다음 날 부트 캠프에 참여하고 기본적인 파이썬 코딩부터 머신러닝, 딥러닝 이론을 공부하였습니다. 짧은 6 개월 동안 후회를 남기고 싶지 않아 매 순간 열심히 학습하였습니다. 항상 제가 배운 코드에서 더 나은 목표를 만들기 위한 방법을 고민하였습니다. 적절한 파라미터를 찾기 위해 HyperOpt 를 적용하여 머신러닝의 성능을 높였고 모델에 Dropout 을 적용하여 성능 개선을 이루었습니다. 또한 학습할 때 배우지 못한 양방향 LSTM 모델을 직접 구현하여 개선된 결과를 도출하였습니다. 그리고 제가 출력한 결과의 성능을 객관적으로 알아보기 위해 Rouge-Score 를 적용하여 다른 교육생들이 시도하지 않은 방법들을 시도하였습니다. 그 결과 NLP

프로젝트 기준별 우수 프로젝트 선정에서 총 3 회 선정되었습니다. 이처럼 항상 보다 더 나은 방법을 제시하는 개발자가 되겠습니다.

3. (직무 성공 경험) 직면한 문제를 해소하여 성능을 향상 시켰습니다.

- 요약문 생성 AI 개발 프로젝트를 수행하면서 성능 개선을 성공시켰습니다. 모델의 출력 결과 성능이 기대에 미치지 못하여 팀원과 회의를 거쳐 다음과 같은 방법을 적용하였습니다. 모델이 텍스트별 대화 유형을 구분하기 쉽게 대화 유형을 본문 데이터 앞에 삽입하는 방법의 논문을 참고하여 수행하였습니다. 그리고 라벨링이 된 개인정보를 삭제하기보다는 스페셜토큰으로 지정하여 대화 흐름의 자연스러움을 유지했습니다. 마지막으로 추가적인 데이터 전처리를 위하여 구어체의 특성을 다시 생각하여 반복되는 특수 기호, 자음, 모음 등을 제거하여 본문 데이터의 길이를 줄였습니다. 이런 방법론을 수행한 결과보다 더 나은 요약문 생성이 출력되었고 Rouge-Score 는 각 분야에서 상승하였습니다. 특히 정밀도가 0.15 에서 0.25 로 가장 크게 상승하여 모델의 출력 결과 품질 개선이 이루어졌습니다. 이처럼 직면한 문제에 대하여 보다 더 나은 방법을 제시하는 개발자가 되겠습니다.

ETS TOEIC

OFFICIAL SCORE CERTIFICATE

한국 TOEIC 위원회

손기락

SON GIRAK

Name

103778	1996/11/18
Registration number	Date of birth (yyyy/mm/dd)
2021/07/11	2023/07/11
Test date (yyyy/mm/dd)	Valid until (yyyy/mm/dd)

LISTENING

Your Score 345

5 (0%) — 495 (99%)

40 Percentile Rank

READING

Your Score 290

5 (0%) — 495 (99%)

47 Percentile Rank

TOTAL SCORE

635

원본확인 QR코드

발급번호 : 078237-0410003801

한국TOEIC위원회는 성적표 위변조 및 부정사용 방지를 위해 스마트폰 앱과 TOEIC 홈페이지를 통해 성적진위확인 서비스를 제공하고 있습니다. "YBM 어학시험" 앱을 스마트폰에 설치한 후, 성적표 좌측의 QR코드로 성적의 진위를 확인할 수 있습니다. 이 서비스는 인터넷 연결을 필요로 하며 자세한 이용방법은 TOEIC 홈페이지를 참조하시기 바랍니다.

< 앱 설치 및 서비스 이용 방법 >

1. "YBM 어학시험" 앱을 앱스토어에서 검색하거나 우측에 보이는 QR코드를 이용해 설치합니다.
2. "YBM 어학시험" 앱을 실행 한 후 앱 화면 우측 상단의 "성적 원본대조" 버튼을 클릭합니다.
3. 스캔 화면에서 성적표에 인쇄된 "원본확인 QR코드"를 인식하면 성적의 진위확인이 가능합니다.

YBM 앱 다운로드 QR코드

안드로이드용 아이폰용

Detach here

LISTENING

Your scaled score is between 300 and 400. Test takers who score around 300 *typically* have the following strengths:

- They can sometimes infer the central idea, purpose, and basic context of *short* spoken exchanges, especially when the vocabulary is not difficult.
- They can understand the central idea, purpose, and basic context of *extended* spoken texts when this information is supported by repetition or paraphrase.
- They can understand details in *short* spoken exchanges when easy or medium-level vocabulary is used.
- They can understand details in *extended* spoken texts when the information is supported by repetition and when the requested information comes at the beginning or end of the spoken text. They can understand details when the information is slightly paraphrased.

To see weaknesses typical of test takers who score around 300, see the Proficiency Description Table. www.toeic.co.kr/table

If your performance is closer to 400, you should also review the descriptors for test takers who score around 400.

ABILITIES MEASURED	PERCENT CORRECT OF ABILITIES MEASURED (Your percentage / Average)
Can infer gist, purpose, and basic context based on information that is explicitly stated in short spoken texts	56 / ▲72
Can infer gist, purpose, and basic context based on information that is explicitly stated in extended spoken texts	60 / ▲70
Can understand details in short spoken texts	93 / ▲84
Can understand details in extended spoken texts	76 / ▲79
Can understand a speaker's purpose or implied meaning in a phrase or sentence	40 / ▲66

READING

Your scaled score is between 250 and 350. Test takers who score around 250 *typically* have the following strengths:

- They can make simple inferences based on a limited amount of text.
- They can locate the correct answer to a factual question when the language of the text matches the information that is required. They can sometimes answer a factual question when the answer is a simple paraphrase of the information in the text.
- They can sometimes connect information within one or two sentences.
- They can understand easy vocabulary, and they can sometimes understand medium-level vocabulary.
- They can understand common, rule-based grammatical structures. They can make correct grammatical choices, even when other features of language, such as difficult vocabulary or the need to connect information, are present.

To see weaknesses typical of test takers who score around 250, see the Proficiency Description Table. www.toeic.co.kr/table

If your performance is closer to 350, you should also review the descriptors for test takers who score around 350.

ABILITIES MEASURED	PERCENT CORRECT OF ABILITIES MEASURED (Your percentage / Average)
Can make inferences based on information in written texts	71 / ▲69
Can locate and understand specific information in written texts	76 / ▲72
Can connect information across multiple sentences in a single written text and across texts	67 / ▲73
Can understand vocabulary in written texts	76 / ▲77
Can understand grammar in written texts	74 / ▲80

HOW TO READ YOUR SCORE REPORT:

Percentile Rank: Percentage of the global TOEIC Secure Program test-takers in **2018** through **2020** scoring below your scaled section score.

Percent Correct of Abilities Measured: Percentage of items you answered correctly on this test form for each one of the Abilities Measured. Your performance onquestions testing these abilities cannot be compared to the performance of test-takers who take other forms or to your own performance on other test forms. The average for each ability is the averaged percentage of items answered correctly by the test-takers of the TOEIC Secure program on this form.

Note : TOEIC scores more than two years old cannot be reported or validated.