010-7371-3774
서울특별시 도봉구 창동
ktu9682@gmail.com

김태욱 | Taeuk Kim

1996.08.22
병역: 필(육군, 2016.09 - 2018.06)
GitHub: taeukkkim

데이터 분석을 전공하고 NLP 도메인에 관심이 많으며 추천시스템에도 관심있는 호기심 많은 주니어 개발자 입니다. 부스트캠프 동료들의 평가와 함께 저를 소개 드리자면, **'영특하고 소통 능력이 뛰어난 편안한 사람'** 입니다.
개발자로서 커뮤니케이션을 가장 중요하게 생각하고 있으며 함께 일하는 사람에게 편안함을 주고 전공자가 아닌 사람도 쉽게 알아들을 수 있도록 설명하는 능력을 가진 사람이 되고자 합니다. 다양한 분야에 대한 지식과 커뮤니케이션 능력을 모두 발전시켜 제너럴리스트가 되는 것이 목표입니다.

학력 사항

국민대학교, 빅데이터경영통계전공 **2015.03 — 2021.08**
Seoul, Korea

- 전체 평점: 3.67/4.5
- 전공 평점: 3.75/4.5

Experience

LG유플러스 (아르바이트) 2020.03.27 — 2020.08.31

업무 내용: 데이터분석 지원 및 과제 수행

주요 프로젝트

- **홈 위면해지 사유 텍스트 분석** - 위면해지 현장 개통 불가 사유의 코드화를 위해 진행된 과제. 개통 불가 사유에서 전문 용어를 뽑아 형태소 분석기 단어사전에 추가 했으며, 토크나이징 해 word2vec으로 임베딩 시켜 정확도 90% 이상의 분류모델을 만들어 업무 자동화에 기여함
- **이상탐지 알고리즘 개발** - 룰베이스 알고리즘 중 하나를 맡아 지표에 대한 변화율 임계치를 정하고 임계치 이상의 변화가 생길 시 탐지 및 시각화 하는 알고리즘 구현
- **Acc PKG 적용 전/후 성능저하국소 조기인지 및 불만콜 예방** - 소프트웨어 패키지 업데이트 후 성능이 저하된 국소를 조기 인지 하기 위해 R언어를 사용해 two sample t test를 적용함
- **SQM 데이터저장 자동화** - 크롤링 코드를 통한 서버 내 데이터 저장 자동화

보유 기술

Languages	Python
Tools	pytorch, huggingface

프로젝트

읽거나 보고, 아는 것만 답변하는 지혜로운 기영이봇 2021.11.29 — 2021.12.24

(부스트캠프 AI Tech 2기 최종 프로젝트)

목적: 학습 측면으로는 *Boostcamp* 학습 내용 최종 정리와 사업 측면으로는 다양한 형태의 입력 정보에 대한 질의응답 서비스 제공
내용:사용자 입력 문서, 또는 이미지에 대한 질의응답과 일반 상식(위키피디아)에 대한 질의응답, 간단한 일상대화를 결합한 챗봇

- ODQA의 요소 중 MRC 단계에서 Noanswer(정답이 없는 경우)에 대한 처리를 위해 Noanswer가 포함되어 있는 데이터로 학습하고 모델이 추출한 start, end logit 값과 null score 값의 차이를 구해 threshold를 기준으로 noanswer도 출력하도록 코드 수정하였음. koelectra-small model을 finetuning 시켜 82.3, 85의 Exact Match, F1 score를 기록, huggingface hub에 공유
- ODQA, VQA 전후로 이어지는 간단한 일상 대화와 Noanswer가 정답으로 나올 시 적용하는 Fallback speech를 뱉어주는 simple chatbot을 sklearn model을 통해 설계, huggingface hub에 공유
- Frontend로부터 유저의 query를 받았을 때 해당 query가 QA model에 들어갈 지 chatbot에 들어갈지 결정해주는 Intent classifier를 약 5만건의 데이터를 이용해 sklearn model로 학습, huggingface hub에 공유. validation f1 score가 98% 이상 나왔지만, QA query의 경향성 등 데이터의 한계로 inference 시 intent classifier와 simple chatbot은 성능에 한계가 있었음

Open-Domain Question Answering **2021.10.12 — 2021.11.04**

*(*부스트캠프 *AI Tech 2*기 *P Stage level-2)*

주제*:* 주어진 지문을 이해하고*,* 주어진 질의의 답변을 추론하는 태스크

결과*: 19*팀 중 *1*위 *(EM* 기준 *0.7556)*

내용*:* 발표자료

- Huggingface hub를 활용한 데이터셋과 모델의 버전 관리
- 최적의 Elastic search 세팅을 위해 내부 파라미터 검색 및 retrieval 성능 실험
- roberta-large model의 output head를 linear layer에서 N-gram Convolution layer로 교체해 Exact Match 기준 약 1, 2% 의 성능 향상을 보임
- 사람처럼 학습하는 AI를 만들기 위해 BART 논문에서 등장하는 denoising 방식 중 하나인 sentence permutation을 활용해 문장 간의 순서를 바꿔주는 정도를 조절하여 난이도를 상, 중, 하로 나누고 이를 토대로 augmentation 및 난이도별 학습 방법인 Curriculum learning을 적용
- Extractive MRC model의 output에 대해 형태소 분석기 5개를 활용해 조사를 제거하고, 결과를 종합하여 앙상블 하는 post processing 과정에 기여

문장 내 개체간 관계 추출 Relation Extraction **2021.09.27 — 2021.10.07**

*(*부스트캠프 *AI Tech 2*기 *P Stage level-2)*

주제*:* 문장의 단어*(Entity)*에 대한 속성과 관계를 예측하는 인공지능 만들기

결과*: 19*팀 중 *1*위 *(micro f1* 기준 *0.7573)*

내용*:* 발표자료

- Data Manager 역할을 맡아 전처리 및 Relation Extraction 관련 SoTA 방법론들을 사용하기 위해 필요한 entity type에 대해 EDA 과정에서 찾은 부적절한 entity pair에 대한 서치 및 수정 작업 진행, rule base 수정을 통한 새로운 데이터셋 버전 생성
- 모델의 다양성을 위해 Electra, Roberta, XLM Roberta의 classifier head를 Linear에서 LSTM 구조로 변경하여 Roberta 기준 micro f1이 68.32 -> 68.55가 되었고 앙상블 시 성능 상승에 기여
- OPTUNA를 활용한 random hyperparameter search 코드 구현

화장품 추천 시스템 **2020.05 — 2020.07**

Make Up For U

주제*:* 유저의 피부타입*,* 제품 성분 등을 고려한 화장품 추천 시스템

결과*:* 투빅스 제 *10*회 컨퍼런스 발표 발표영상

내용*:* 발표자료

- 팀장의 역할을 맡아 프로젝트 전반을 관리, 주도하였음
- 웹 크롤링을 통한 화장품 성분, 리뷰, 평점 등의 데이터 수집 및 리뷰데이터에 대한 전처리(cleansing) 작업 진행
- 새롭게 가입한 유저는 데이터가 없고 활동 기록이 남아 있는 유저는 데이터가 존재하기 때문에 각 유형의 유저에 따라 신규유저, 기존유저로 나누어 추천모델을 제작하였으며 여기서 기존유저에 대한 추천을 맡아 기존유저가 사용했던 화장품 성분, 리뷰 등을 토대로 추천시스템 구현

PGN 기반 텍스트 생성 요약 **2019.09 — 2019.11**

Make Up For U

주제*:* 텍스트 생성 요약 프로젝트

결과*:* 국민대학교 학회 *D&A* 컨퍼런스 발표

내용*:* 발표자료

- 웹 크롤링을 통한 블로그 포스트, 뉴스기사 등 제목, 본문으로 구성된 데이터 약 70만건 수집
- ETRI의 KorBERT의 토크나이징 방식과 유사한 khaiii 형태소 분석기를 통해 토크나이징한 후 학습을 진행하여 약 0.45의 F1 score 기록

수상 및 공모전 참여

[NH투자증권] 'AI야, 진짜 뉴스를 찾아줘!' 2020.11.23 — 2021.02.10(2020.12.31)

주제: 뉴스 데이터 내 가짜 뉴스 분류
결과: *192*팀 중 *4*위*(*입선상*)*
내용:발표자료

- 팀장의 역할을 맡아 프로젝트 일정 관리 및 프로젝트 전반을 관리
- 진짜 뉴스에는 조금 등장하고 가짜 뉴스에 자주 등장하는 종목, 테마주, 무료 등을 BAD token으로 선정하고 해당 토큰들이 등장하는 빈도를 머신러닝 모델을 위한 파생변수로 생성
- 진짜 뉴스인지 가짜 뉴스인지 파악하기 위해 문맥을 고려한 분석이 필요하므로, 양방향 탐색을 통해 단어 간의 관계 파악이 용이한 BERT를 선정하고 Mecab 형태소 분석기를 사용해 형태소 기반의 임베딩을 한 ETRI KorBERT 모델을 finetuning하여 LB기준 0.98206의 성능을 냄
- 가짜뉴스 분류 모델을 통해 만들어진 신뢰도 있는 데이터를 기반으로 한 실시간 투자 관련 뉴스 생성요약, 기사에서 중요 키워드를 추출해 관련 투자 종목 추천 등의 서비스 제언

[LG] '시스템 품질 변화로 인한 사용자 불편 예지 AI 경진대회' 2021.01.06 — 2021.02.24

주제: 비식별화 된 시스템 기록*(*로그 및 수치 데이터*)*을 분석하여 시스템 품질 변화로 사용자에게 불편을 야기하는 요인을 진단, 에러가 발생한 로그에 대한 데이터*(*에러 데이터*)*와 펌웨어 및 퀄리티에 대한 데이터*(*품질 데이터*)*로 구성되어 있음
결과: *418*팀 중 *7*위
내용:발표자료

- 품질 데이터 지표 간의 연관성을 발견해 분석해야 할 지표의 수를 추린 후 해당 지표들과 에러 데이터 간의 관계에 대한 분석 진행. 그 결과 특정 품질 지표와 에러 타입간에 선후관계가 존재함을 파악하고 각 유저마다 해당 관계쌍이 등장한 횟수를 피쳐로 사용해 모델의 성능 개선
- 성능이 좋은 5개의 모델 선정 후 해당 모델들에 대한 Voting 앙상블 적용으로 AUC 기준 0.8435 -> 0.8463으로 성능 개선, 일반화

EXTRACURRICULAR

D&A (전공학회) **2019.03 — 2019.12**
- 딥러닝 세션을 통한 교육 및 컨퍼런스 진행

투빅스(Tobigs) **2019.07 — 2020.07**
- 딥러닝 교육, 심화 세미나(NLP, 추천시스템) 및 컨퍼런스 진행

[청년취업아카데미] 빅데이터를 활용한 파이썬 프로그래밍 (멀티캠퍼스) **2019.01 — 2019.02**

네이버 커넥트재단 부스트캠프 AI Tech 2기 **2021.08 — 2021.12**