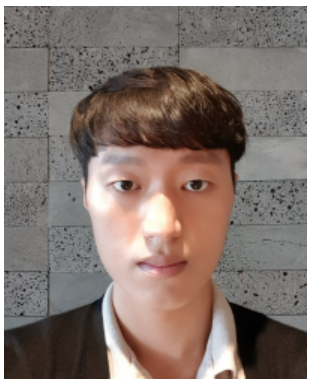
RESUME | 이력서



개인정보

이 름	노경주	성 별	남
생년월일	1996.10.26	이메일	nkj36@naver.com
주 소	경기 성남시 분당구		
전화번호	010-7637-3696		

학력사항

2012 - 2015	정광고등학교 졸업	문과 계열
2015 - 2022	조선대학교 졸업	컴퓨터공학 전공

수상내역

2019	SW 제작 아이디어 공모전	(대상)
2019	SNS 기반 빅데이터 분석 프로젝트	(우수상)
2019	프로그래밍 대회	(동상)

교육 활동

2022	카카오 엔지니어와 데이터 엔지니어링 입문 on cloud 3기	한국통신학회
2020	텍스트 분석을 위한 머신러닝	패스트캠퍼스
2019	기계학습의 기초부터 응용까지	한국통신학회
2019	SNS 기반 빅데이터 분석 교육	스마트인재개발원

경력사항

2022.01 - 재직 중	비트망고	BI/DE, 백엔드
2019.08 - 2021.02	(주)마이데이터랩	ML/AI , 백엔드

위 기재 사항은 사실과 다름이 없습니다.

Index :

경력 사항

경력 기술서

프로젝트 경험 상세 내용

[데이터 엔지니어링 / 백엔드 관련 프로젝트 경험]

[사내 프로젝트]
Data-Driven 환경 구축

- 자체 이벤트 로그 수집 파이프라인 개선 프로젝트 리딩
- 의사결정을 돕기 위한 DW/DM 운영
- 다양한 마케팅 네트워크 데이터의 DW 적재 파이프라인 개발

[개별 프로젝트]
유저 유입을 위한 백엔드 서비스 개발

- 고가치 유저의 유입을 위한 MMP 네트워크의 캠페인 S2S 서비스 개발
- 앱 푸시 메시지를 전송하는 FirebaseCloudMessage 서비스 개발

[ML\AI 및 개인 프로젝트 경험]

[사내 자체 프로젝트]
자기소개서 작성 서비스

- **자기소개서 항목 취지 분류 모델 개발**
- **뉴스 요약 서비스 개발**
- **텍스트 데이터 라벨링 프로세스 설계**

[개별 프로젝트]
교내 및 개인 프로젝트

- 맛집 데이터 수집 파이프라인 개발 중
- 국민청원 2차 카테고리 분류 모델 개발
- K-means 및 유사 문장 탐색 알고리즘 구현

교육 관련 증명 서류

텍스트 분석을 위한 머신러닝

패스트 캠퍼스 주관

SNS기반 빅데이터 분석

스마트인재개발원 주관

Statistic Learning : 머신러닝의 기초부터 응용까지

한국통신학회 주관

카카오 엔지니어와 데이터 엔지니어링_ 입문 on Cloud 3기

러닝스푼즈 주관

수상 관련 증명 서류

SNS기반 빅데이터 분석 우수상/모범상

팀장 - 최종 프로젝트

프로그래밍 대회 동상

팀장

SW 제작 아이디어 공모전 대상

팀원 - 죠스펙

추가 정보

GitHub / 블로그

Notion

Topcit 395

경력 기술서

비트망고 BI Junior

BI 및 DE 업무 / 백엔드 업무

2022-01-03 ~ 재직중

주요 업무 - BI 및 DE 업무

- IDC내 온프레미스 서버 환경의 이벤트 로그 수집 파이프라인 개발 및 구축
- 사내 의사결정을 위한 DW/DM 운영 및 유저 정보를 저장하기 위한 데이터베이스 제안 및 구축
- 활용중인 마케팅 네트워크(Unity, Vungle, Ironsource 등)의 데이터를 API로 불러와 DW에 적재

주요 업무 - 백엔드 업무

- 유저에게 앱 푸시 메세지를 전송하는 FCM(Firebase Cloud Message) 서비스를 설계 및 개발
- 수집한 로그 정보를 활용하여 고가치 유저 마케팅 고도화를 위한 S2S 서비스를 설계 및 개발
- 유니티 광고 소재 업로드 자동화 프로세스를 설계 및 개발

주요 성과

- 일간 약 1억 5천만의 로그 데이터를 읽고 처리 할 수 있는 Fluentd 파이프라인 설계 및 개발
- 로그 수집 파이프라인 개선 프로젝트의 전체 프로세스의 도식화 및 간소화하여 프로젝트 리딩
- 15개의 네트워크에 대한 광고의 성과 데이터 수집 및 4개의 플랫폼(Google, Apple 등) 데이터를 DW 적재
- 기존의 MariaDB의 DM 환경에서 유저의 개별 데이터를 적재하기 위한 MongoDB 환경을 직접 제안 및 개발하여 새로운 데이터 환경 구축
- 앱 푸시 메시지를 전송하는 Firebase Cloud Messaging 시스템을 구축하여 7일 이상 미접속 유저가 새로 접속하는 비율이 15% 증가

경력 기술서

(주)마이데이터랩 R&D 주임연구원

텍스트 데이터 분석 / 백엔드 업무 2019-08-01 ~ 2021-02-28

주요 업무 - 텍스트 데이터 분석 업무

- Sklearn을 이용하여 자기소개서 질문 취지 관련 Multi Lable 분류 모델 개발
- 텍스트 데이터의 Multi Labeling 프로세스 수립 및 구현
- textrank를 이용한 문서 요약 서비스 구현

주요 업무 - 웹 백엔드 개발 업무

- Spring Framework / MariaDB / Git 을 이용한 자기소개서 작성 서비스 백엔드 개발
- 텍스트 데이터 관련 서비스를 제공하기 위한 별도의 Flask API 서버 구현
- aws를 이용한 서버 관리

주요 성과

- 86%의 정확도를 가진 질문 취지 분류기 개발
- ML 모델을 서비스하는 API 개발 및 적용
- 텍스트 데이터의 라벨링 과정 최소화
- 자기소개서 작성 서비스 인증/인가 서비스 개발
- 자기소개서 작성 서비스 중 구직자 성향 분석 서비스 개발

BI - IDC내 온프레미스 서버 환경의 이벤트 로그 수집 파이프라인 개발 및 구축

프로젝트 소개 2022-10 ~ 2023-04

사내 프로젝트

1. 프로젝트 제목 :

IDC내 온프레미스 서버 환경의 이벤트 로그 수집 파이프라인 개발 및 구축

2. 개발 배경 :

GA(구글 애널리틱스)의 업데이트로 인해 Firebase 이벤트 API로 수신받아 GA를 거쳐 빅쿼리로 받은 데이터를 Impala로 일 2회 마이그레이션하며 사용하고 있는 데이터 웨어하우스의 유지에 문제가 발생 할 수 있음을 사전에 발견했습니다.

기존의 Nginx - PHP parser로 동작하던 낙후된 자체 로그 시스템을 개선하는 프로젝트 참가하여 새로운 로그 수집기인 fluentd의 도입을 프로토타입과 도식화 된 설계도를 바탕으로 제안하여 프로젝트를 리드했습니다.

3. 수행 한 상세 업무 :

1. 기존의 로그 수집 파이프라인을 도식화하여 문제점과 개선해야 할 방향성을 정리하여 팀원과 공유했습니다.
- 이 과정에서 Fluentd를 사용한 로그 수집기 개발을 제안하였고, 기존의 PHP 언어로 구현된 방법을 고수하고자 하는 팀원들을 설득하였습니다.
- 효과적으로 공유하기 위해 Fluentd를 이용한 로그 수집기의 프로토타입을 개발하여 함께 공유드렸습니다.

2. 1일 단위로 로그를 수집하던 Nginx 서버의 access.log를 시간 단위로 나누어 access.log rotate 압축 과정에 발생 가능한 유실을 방지했습니다.

3. 원하는 성능의 수집기 서버를 구축하기 위해 멀티 프로세스 환경으로 구축했습니다.
- 약 1억 5천만건의 로그를 정상적으로 수집하는 파이프라인을 개발했으며, 추후 이벤트 확장을 고려하여 4배 정도의 트래픽을 견뎌 낼 수 있도록 확장시켰습니다.
- 네트워크 환경에 따라 최대 3일까지의 과거 데이터가 수집 될 가능성이 있었기 때문에 로그 내 시간 정보로 데이터를 이용해 필터를 추가했습니다.

4. IDC에 구축되어 있는 HDFS 환경이 Fluentd의 WebHDFS Plugin과 적합하지 않은 부분을 발견하여 WebHDFS Plugin 오픈소스를 수정하여 활용했습니다.
- WebHDFS는 1개의 Standby namenode를 기준으로 구축되어 있었지만, 현재 서버의 상황은 2개의 Standby namenode를 사용하고 있었습니다. 이 부분의 코드를 수정하여 활용했습니다.
- 위 수정 내용을 MR했습니다.

5. 데이터 입력의 어느 과정에서라도 재입력이 가능하도록 코드를 별도 작성하였습니다.

4. 사용 한 기술 스택 :

Fluentd, HDFS, Nginx, Impala

BI - 사내 의사결정을 위한 DW/DM 운영 및 유저 정보를 저장하기 위한 데이터베이스 제안 및 구축

프로젝트 소개 2022-01 ~

사내 프로젝트

1. 프로젝트 제목 :

사내 의사결정을 위한 DW/DM 운영 및 유저 정보를 저장하기 위한 데이터베이스 제안 및 구축

2. 개발 배경 :

업무 시간 중 데이터 웨어하우스에 많은 부하가 발생하였으며, 주요 지표의 경우 고정된 방식을 활용하기에 DW 적재 배치 시점에 맞추어 DM에 옮겨 부하를 줄일 필요가 있었습니다.
뿐만아니라, 데이터의 책임과 활용 목적에 따라 구분하는 거버넌스를 강화하기 위해 데이터 마트를 활용했습니다.

앱 내에서 유저가 발생시키는 데이터를 이벤트 단위로 수집하는 데이터 웨어하우스에서 데이터 마트로 운영중인 MariaDB와 유저 데이터를 적재하는 MongoDB에 각 팀에서 원하는 데이터를 별도의 테이블로 종합하여 적재하는 프로젝트를 진행했습니다.

이벤트 소싱 방식으로 쌓이는 데이터 웨어하우스에서 특정 유저에 대한 쿼리가 자주 발생함을 인지하고 효율적인 의사결정을 위해 스스로 유저의 데이터를 저장하는 데이터베이스의 필요성을 어필하여 프로젝트를 리드했습니다.

3. 수행 한 상세 업무 :

1. 의사결정자 및 분석가와의 회의를 통해 필요한 데이터를 종합하여 MariaDB에 적재했습니다.
- 퍼즐게임의 레벨 난이도 관련 데이터의 기초 데이터를 주 단위로 주기적으로 적재하는 Airflow Dag를 개발했습니다.
- 마케팅 전략의 지표로 유저의 Active Retention과 유저 복귀 관련 지표를 수집하는 Dag를 개발했습니다.

2. 유저의 데이터베이스가 MariaDB에 적재하기에 적합하지 않다는 것을 알게되어 MongoDB 구축이 필요함을 설득했습니다.
- 이 과정에서 각 팀의 니즈를 종합하고, COO님 앞에서 프레젠테이션하여 프로젝트화 하였습니다.

3. 유저 단위의 분석에 필요한 데이터를 MariaDB 및 MongoDB에 적재했습니다.
- 1일 단위로 유저의 데이터를 최신화하고 일간 종합 정보를 추가했습니다.
- 특정 조건에 적합한 유저에 대해 세그멘테이션을 수행하여 고가치 유저 그룹을 묶어주었습니다.

4. 사용 한 기술 스택 :

Impala, MongoDB, MariaDB, Airflow, Python

BI - 활용중인 마케팅 네트워크(Unity, Vungle, Ironsource 등)의 데이터를 API로 불러와 DW에 적재

프로젝트 소개 2022-01 ~

사내 프로젝트

1. 프로젝트 제목 :

활용중인 마케팅 네트워크(Unity, Vungle, Ironsource 등)의 데이터를 API로 불러와 DW에 적재

2. 개발 배경 :

앱으로의 유저 유입을 위해 다양한 광고 마케팅 네트워크를 활용하고 있는 상황에서, 광고 및 캠페인에 대한 성과를 정확히 측정하기 위해서는 해당 네트워크에서 제공하는 지표가 필요했습니다.

Unity, remerge, snapchat, google, apple 등의 광고 네트워크에서 집계하고 있는 데이터를 API로 수집하여 분석 및 시각화 할 수 있는 동일한 형태로 변환하여 컬럼기반 데이터베이스에 적재하는 ETL 프로젝트를 진행했습니다.

3. 수행 한 상세 업무 :

1. Unity, snapchat, remerge 등의 네트워크에서 제공하는 API 데이터를 Shell Script와 python을 활용하여 파이프라인을 개발하였습니다.
- 각 네트워크에서 제공하는 성과 지표에 관련 내용을 통합, 연산하여 동일한 형태로 변환했습니다.
- 성과 지표에 활용하는 DW인 columnstore에 bulk insert하는 로직을 개발했습니다.

2. 파일로 제공하는apple, amazon과 같은 플랫폼에서 제공하는 매출 데이터를 사내 API, Airflow를 활용하여 DW에 적재했습니다.
- apple에서 제공하는 salesReports API를 활용하여 데이터를 불러 온 후 csv 형태로 제공하는 사내 서버로 구성된 API 서버에 인증 절차를 포함하여 개발했습니다.
- amazon 에서 제공하는 Reports 파일을 별도 임시 공간에 저장하여 MariaDB로 입력하는 파이프라인을 개발하였습니다. 이후 MariaDB에서 데이터 웨어하우스에 적재하도록 했습니다.

3. 입력 파이프라인을 crontab으로 스케줄링 되도록 초기 개발했습니다. 이후에는 Airflow를 이용하여 스케줄링 되도록 Dag를 개발 프로젝트를 진행했습니다.

4. 사용 한 기술 스택 :

MariaDB, ColumnStore, Airflow, Bash, Python, REST API

BE - 유저에게 앱 푸시 메세지를 전송하는 FCM(Firebase Cloud Message) 서비스를 설계 및 개발

프로젝트 소개 2022-08 ~ 2023-05

사내 프로젝트

1. 프로젝트 제목 :

유저에게 앱 푸시 메세지를 전송하는 FCM(Firebase Cloud Message) 서비스를 설계 및 개발

2. 개발 배경 :

광고 수익의 비율이 높은 도메인의 특성 상 유저의 DAU가 매출에 큰 비중을 가지고 있습니다. 이러한 수익 구조 상 유저의 유입이 중요 지표로 활용되고 있었고, 이탈한 유저들 다시 접속 유저로 불러들이는 방안이 필요했습니다.

기존에는 Firebase를 통해 수집한 이벤트로 플랫폼 내에서 전송하는 서비스를 활용하고 있었지만, 이 경우 다양한 조건에 맞는 유저를 타겟하여 보내기 어려운 한계점이 있었습니다.

이러한 문제를 해결하기 위해 자체적으로 앱 푸시 메시지를 전송하는 프로젝트를 진행하게 되었습니다.

3. 수행 한 상세 업무 :

1. 데이터 웨어하우스에서 유저의 정보를 데이터 마트에 적재합니다.
- 1일 배치로 적재하여 유저의 정보를 최신화 합니다.
- 초기에는 MariaDB로 개발하였으나, 이 후 스키마의 유연한 확장을 위해 MongoDB를 활용했습니다.

2. 메시지의 타겟을 쿼리 형태로 조회 할 수 있도록 데이터 마트에 정보를 저장합니다.
- 초기 개발에서는 메시지 브로커 대신 MariaDB에 적재 한 후 스케줄러에 따라 메시지를 전송하는 방식으로 개발했습니다. 이 경우 1시간마다 해당 시간에 대한 메시지를 검색하여 전송했습니다.
- 이후 메시지 브로커를 추가하여, 기존 MariaDB의 메시지 전송은 CDC 프로세스를 활용하여 유지하고 원하는 메시지를 즉시 전송 할 수 있는 환경을 구축했습니다.

3. 전송 결과를 별도 저장하여 지표로써 활용 할 수 있도록 했습니다.
- 메시지를 전송 후 기록을 유저 데이터에 남겨 메시지 수신 후 이벤트를 추적 할 수 있도록 했습니다.

4. 사용 한 기술 스택 :

MariaDB, MongoDB, RabbitMQ, Kafka, Debezium, Airflow, Bash, Python, REST API

BE - 수집한 로그 정보를 활용하여 고가치 유저 마케팅 고도화를 위한 S2S 서비스를 설계 및 개발

프로젝트 소개 2022-06 ~ 2022-07

사내 프로젝트

1. 프로젝트 제목 :

수집한 로그 정보를 활용하여 고가치 유저 마케팅 고도화를 위한
S2S 서비스를 설계 및 개발

2. 개발 배경 :

사내 전략이 많은 유저의 유입에서 가치있는 유저의 유입으로 전략을 변경하면서, 고가치 유저를 유입 할 수 있는 캠페인 생성에 중요도가 높아졌습니다.

고가치 유저를 유입하기 위한 캠페인 생성 및 MMP 네트워크의 대시보드에서 특정 이벤트를 달성 한 유저의 지표를 보기위해 Appsflyer의 S2S API를 활용한 프로젝트를 진행했습니다.

3. 수행 한 상세 업무 :

1. 유저의 정보를 입력하는 테이블에 보내고자하는 이벤트에 대한 컬럼을 추가했습니다.
- 접속 시간, 사용 코인 등 다양한 정보를 유저 단위로 종합하여 저장했습니다.

2. 1일 단위로 입력되는 배치 프로세스의 종료 시점에 S2S API 전송 프로세스를 추가했습니다.
- 수집한 데이터를 기반으로 S2S API를 요청하는 파이프라인을 개발했습니다.

4. 사용 한 기술 스택 :

MariaDB, Bash, Python

BE - 유니티 소재 업로드 자동화 프로세스 개발

프로젝트 소개 2022-05 ~ 2022-06

사내 프로젝트

1. 프로젝트 제목 :

유니티 소재 업로드 자동화 프로세스 개발

2. 개발 배경 :

유니티 네트워크에 광고 소재를 업로드하는 방식이 단순하지만 일일이 작업하기에 상당한 시간이 소요되었습니다. 이러한 문제로 소재 업로드를 자동화하면 업무 효율이 높아질 것 이라는 요청을 받게 되어 프로젝트를 시작했습니다.

팀에서도 처음 시도하는 프로젝트 였기때문에, 앞으로 소재 업로드에 관련한 가이드라인을 포함해서 프로젝트를 설계한 후 개발을 진행하였습니다.

3. 수행 한 상세 업무 :

1. 개발 과정 및 사용 방법을 도식화 하여 협업하는 팀에 공유하였습니다.
- 추후에 유사한 소재 업로드 프로젝트 시 가이드라인으로 활용되고 있습니다.

2. 구글 드라이브 API를 활용하여 드라이브 내 파일의 경로와 파일 리스트를 읽어 온 후 임시 폴더에 다운로드 받습니다.
- Airflow 내 임시 폴더를 volumn으로 설정하여 파일을 임시로 다운받았습니다.

3. 받아온 파일을 유니티 네트워크의 API를 적절히 활용하여 소재 업로드를 개발했습니다.

4. 프로세스 중간에 업로드 실패 한 소재의 경우 슬랙을 통해 알림을 보내도록 개발했습니다.

4. 사용 한 기술 스택 :

MariaDB, Airflow, Python

AI - 자기소개서 작성 서비스 질문 취지 분류 모델 개발

프로젝트 소개 2020-03 ~ 2020-04

사내 프로젝트

1. 프로젝트 제목 :

자기소개서 취지 분류 모델 개발

2. 프로젝트 개발 배경 :

자기소개서 작성을 도와주는 서비스를 개발하는 스타트업에서 개발을 진행하며 관련 경험이 없는 학생들이 각 질문에 따른 답변에 대해 어려움을 겪는 것을 알게 되었습니다.

자기소개서 작성에 대한 정보 제공 및 서비스 제공을 위해서는 질문의 문항에 대한 취지를 분류하는 작업이 선행되어야 함을 알게 되었으며, 취지를 분류하는 모델 개발 프로젝트를 진행했습니다.

3. 수행 한 상세 업무 :

1. Bag Of Words방식을 적용하여 단어를 임베딩 벡터 차원에 표현했습니다.
- 각 항목 당 등장하는 단어가 극히 한정되어 있으며, 길이가 짧아 Distributed Representation 방식을 사용하지 않았습니다.

2. 다중 라벨 분류 기법인 Binary Relevance를 사용하여 분류 모델을 개발했습니다.
- Scikit-multilearn에서 제공하는 라이브러리로 다중 라벨 분류기입니다.
- 학습 시간과 성능을 고려하여 기법을 선택했습니다.

3. 다중 라벨 분류의 특징인 "부분 정답"을 고려하여 Hamming Loss를 접목하여 테스트 하였습니다.

4. 검증은 CrossValidation(K-fold 5) 방식을 사용하여 적정 파라미터를 선정했습니다.
- 단어의 N-gram 및 최소 등장 빈도 등을 선정했습니다.

4. 사용 한 기술 스택 :

Spring, Python, Sklearn, NLP, Flask, Jupyter

AI - 자기소개서 작성 서비스 뉴스 요약 기능 개발

프로젝트 소개 2020-07 ~ 2020-09

사내 프로젝트

1. 프로젝트 제목 :

뉴스 요약 기능 개발

2. 프로젝트 개발 배경 :

자기소개서 작성 과정 중 많은 사람들이 기업의 정보를 수집하는 과정에서 어려움을 겪는다는 것을 알게 되었습니다. 특히 수집 과정 중 지원하는 모든 기업의 긴 뉴스와 사업 내용등을 읽기 어렵기 때문에 요약하는 서비스의 개발이 필요하다는 생각을 하게 되었습니다.

사용자가 자기소개서를 작성 할 때 지원하는 기업의 정보를 쉽게 얻을 수 있도록 기업의 최근 이슈를 포함한 뉴스 정보를 요약할 수 있는 서비스 직접 제안했습니다.

3. 수행 한 상세 업무 :

1. 기업과 관련 된 뉴스의 정보를 얻기 위해 네이버 뉴스 API를 사용했습니다.
- 사용 시 API 명세에 따라 기간 및 정보를 필터링했습니다.

2. 기업 및 기관에 따른 다르게 사업 보고서에 접근하는 과정을 개발했습니다.
- Alio(공공기관) / Dart(기업) 의 정보에 쉽게 접근 할 수 있도록 API를 이용하여 서비스를 개발했습니다.

3. PageRank를 텍스트에 적용한 TextRank 기법을 이용하여 요약하는 서비스를 개발했습니다.
- Gensim 라이브러리 내 TextRank를 사용했습니다.
- Gensim 라이브러리 내의 Tokenizer가 단어를 추출해내는 과정에서 1글자의 단어를 추출하지 못하는 문제를 발견하였고, Tokenizer를 직접 구현하여 문제를 해결했습니다.

4. 사용 한 기술 스택 :

Spring, Python, Sklearn, NLP, Flask, Gensim, Jupyter

AI - 자기소개서 작성 서비스 텍스트 데이터 라벨링 프로세스 설계

프로젝트 소개 2020-01 ~ 2020-03

사내 프로젝트

1. 프로젝트 제목 :

텍스트 데이터 수집 및 라벨링 프로세스 설계

2. 프로젝트 개발 배경 :

텍스트 데이터를 분류 모델을 개발하기 위해 진행한 데이터 수집 단계에서 수집 된 데이터가 라벨링이 되어있지 않은 Raw데이터라는 것을 알게 되었습니다. Supervised 문제의 특성 상 라벨이 필요하기 때문에 라벨링의 필요성을 인식했습니다.

데이터의 특성 상 2개 이상의 라벨을 가질 수 있는 데이터였으며 데이터의 수가 10만개 이상이였기 때문에 효율적으로 라벨링을 하기 위해 프로세스를 설계하여 직접 제안했습니다.

3. 수행 한 상세 업무 :

1. 데이터를 수집하는 크롤러를 개발했습니다.
- 이 과정에서 robots.txt 및 정책을 확인하여 수집 가능한 데이터를 확인했습니다.

2. 데이터를 전처리하는 과정을 설계했습니다.
- 불용어 탐색 및 제거 과정을 추가했습니다.
- 중복되는 문장을 제거하는 과정을 추가했습니다.
- 형태소를 분석하는 과정을 추가했습니다.

3. 단어의 상대적 출현 비율을 통해 라벨링 방식을 제안하여 직접 구현했습니다.
- 1개월로 예상되던 라벨링 문제를 2일로 단축시킬 수 있었습니다.
- 시각적인 자료로 정리하여 동료분들과 총결정권자 분을 설득하여 채택 될 수 있었습니다.

4. 사용 한 기술 스택 :

Spring, Python, Sklearn, NLP, Jupyter

DE - 맛집 데이터 수집 파이프라인 개발

프로젝트 소개

개인 프로젝트

1. 프로젝트 제목 :

맛집 데이터 수집 파이프라인 개발

2. 프로젝트 개발 배경 :

주변 맛집을 찾기위해 특정 서비스에 검색하면 너무나 많은 음식 종류와 맛집 정보가 분별없이 검색되었습니다. 이러한 정보 환경에서 "내가 원하는 맛집"에 대한 정보를 얻기 위해 프로젝트화 하여 직접 개발하고자 했습니다.
또한 서버를 직접 구축하면서 실력을 쌓는 것이 큰 도움이 된다는 조언을 듣게 되어 관심 주제로 프로젝트를 진행하고 싶었습니다.

3. 요구 사항 :

1. 맛집과 관련 된 홈페이지를 크롤링합니다.
- 크롤링하는 과정에서 웹 페이지 정보를 GCS에 영구 저장해야합니다.
- 검색에 사용된 키워드와 페이지의 중복 검색을 방지해야 합니다.

2. 음식점 리스트와 음식점 상세 페이지의 크롤링과 파싱 과정을 분리해야합니다.
- 각 과정을 분리함으로써 MSA 환경을 구축하도록 해야 합니다.

3. 데이터를 검색에 용이한 데이터베이스에 저장해야 합니다.
- 목표가 원하는 맛집을 찾기 위함이기 때문에 기본 검색 엔진이 포함된 데이터베이스를 사용하고자 합니다.

4. 수행 한 상세 업무 :

1. 음식점 관련 홈페이지의 크롤러를 구현했습니다.
- 망고플레이트 홈페이지의 음식점 데이터를 크롤링했습니다. (3초 간격)
- 검색 한 지역의 해시값 및 페이지를 redis에 저장하여 중복 검색을 방지했습니다.

2. 크롤링 된 페이지에서 음식점 리스트의 주소 정보를 추출하여 메시지 브로커에 발행합니다.
- 각 로직을 수행하는 과정에서의 결과인 웹 페이지 정보를 GCS에 영구 저장했습니다.
- 각 과정에서 메시지 브로커의 메시지를 소비 및 발행하도록 설계하여 각자의 프로세스를 병렬적으로 처리 할 수 있도록 했습니다.

3. 상세 음식점 URL을 크롤링하여 데이터를 추출합니다.
- 정형화 되지 않은 데이터의 경우 Raw한 형태로 수집하여 추후에 변환합니다.
- 상세 음식점 URL을 redis에 저장하여 중복 검색을 방지했습니다.
- elasticsearch에 데이터를 저장합니다.

4. 이후 로직은 추가 개발 예정입니다.

5. 프로젝트 설계

프로젝트 과정

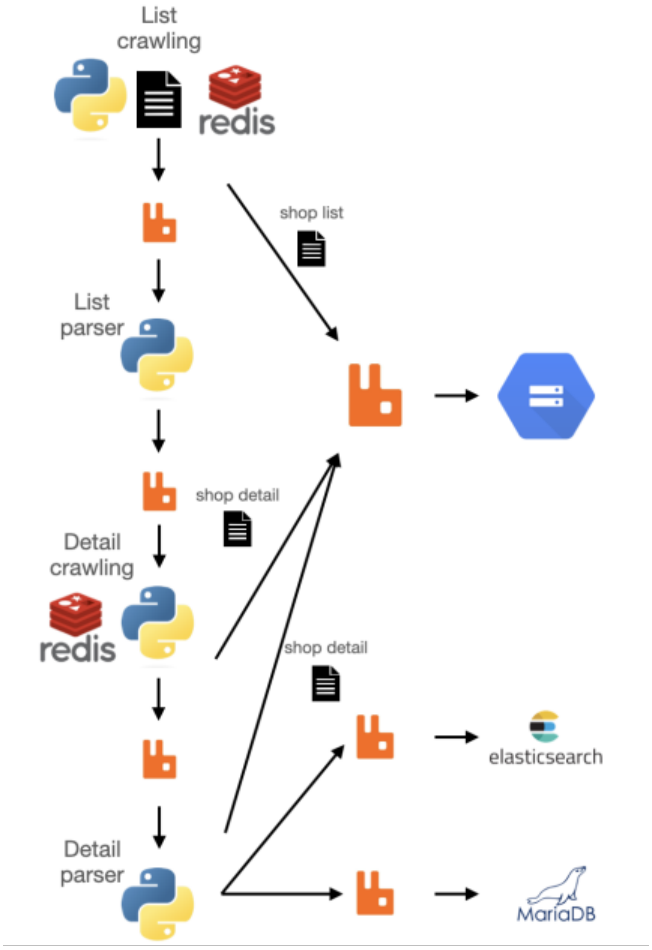
0. 프로젝트의 전체 프로세스를 설계 및 구체화 했습니다.

1. 데이터 수집
- Python을 이용한 크롤링 (redis를 통한 중복 검색 방지)
- GCS 등록 용 queue, 다음 프로세스의 queue에 fanout 전송

2. 상세 음식점 URL 수집
- 다음 프로세스의 queue에 전송

3. 상세 음식점 데이터 수집
- Python을 이용한 크롤링 (redis를 통한 중복 검색 방지)
- GCS 등록 용 queue, 다음 프로세스의 queue에 fanout 전송

5. elasticsearch에 데이터 저장



프로젝트에 기여 한 점

1. 프로젝트의 전체 프로세스를 설계 및 구체화 단계
2. 각 데이터 수집 파이프라인을 개발
3. GCS, elasticsearch, redis 등 다양한 데이터 플랫폼 활용

5. 사용 한 기술 스택 :

Python, Redis, RabbitMQ, GCS, elasticsearch

AI - 개별 프로젝트 국민청원 데이터 2차 카테고리 분류 모델

프로젝트 소개

교내 프로젝트

1. 프로젝트 제목 :

국민청원 데이터 2차 카테고리 분류 모델

2. 프로젝트 개발 배경 :

국민청원 페이지에서 "촉법소년 범죄"에 대한 청원에 동의를 표시하면서 관련 청원이 있는지 찾아보던 중 많은 청원 중 원하는 청원을 찾기 어려운 문제를 알게 되었습니다.

관심이 있는 청원에 빠르게 도달하기 위해 국민청원에서 구분하고 있는 청원 카테고리를 통해 접근을 시도했지만 [제목과 내용이 일치하지 않는 점 / 진행중인 청원의 전체 리스트를 확인해야 한다는 점] 에서 어려움을 느꼈기 때문에 이러한 프로젝트를 진행했습니다.

3. 요구 사항 :

1. 청원 데이터를 수집 할 수 있는 크롤러를 개발해야 합니다.
- 수집 정보는 제목/내용/카테고리/청원시각으로 한정합니다.

2. 데이터 형태에 맞는 전처리 과정이 필요합니다.
- 청원 데이터는 자연어에 가까운 데이터 형태를 가지고 있습니다.

3. 청원의 제목과 내용을 통해 관련성이 높은 청원을 같은 카테고리로 분류하는 모델을 개발해야 합니다.
- 기존에 분류되어있는 1차 카테고리에 따라 같은 청원도 다른 의미를 가질 수 있으므로 1차 카테고리를 기준으로 데이터를 라벨링해야 합니다.

4. 개발 과정 :

1. 크롤러를 구현했습니다.
- 청와대 국민청원 페이지를 크롤링하여 청원 데이터를 수집했습니다.

2. 수집 된 데이터를 전처리하는 과정을 거쳤습니다.
- 분석 할 텍스트를 [제목+내용]으로 결합시켰습니다.
- 단어의 수가 극단적으로 많거나 적은 텍스트를 제거했습니다.
- 형태소 분석을 통해 분석 할 데이터를 축소 시켰습니다.

3. FastText를 통해 임베딩 벡터로 변환 시켰습니다.
- 자연어에 가까운 데이터 형태를 반영하여 Distributed Representation 방식을 이용했습니다.

4. 문장에 등장 한 단어 벡터의 평균을 군집화하여 라벨링을 진행했습니다.

5. FastText Supervised 방식을 통해 모델을 개발했습니다.

6. Flask를 이용해 로컬 서버에서 서비스 작동을 검증했습니다.

5. 프로젝트 설계

프로젝트 과정

0. 프로젝트의 전체 프로세스를 설계 및 구체화 했습니다.

1. 데이터 수집 (Data Setting)
- Python을 이용한 크롤링

2. 데이터 전처리 (Filtering Usable Data)
- 불필요한 데이터 제거
- 형태소 분석

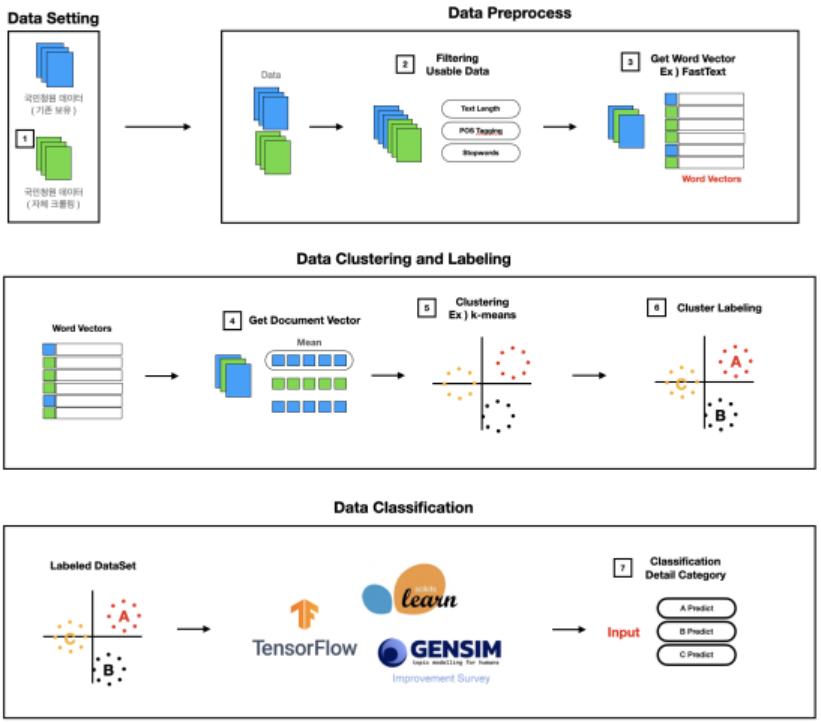
3. FastText를 이용한 단어를 임베딩 벡터로 표현

4. 임베딩 벡터를 이용하여 문장 벡터를 추정
- 열단위 평균으로 문장 벡터를 대체

5. FastText로 얻은 단어 벡터를 이용한 문장 군집화

6. 군집 별 라벨링 진행

7. 분류 모델 개발



프로젝트에 기여 한 점

1. 프로젝트의 전체 프로세스를 설계 및 구체화 단계
2. FastText로 얻은 단어 벡터를 이용한 문장 군집화 구현
3. 군집 별 라벨링의 편의성을 위한 Frequency 기반 라벨링 프로그램 개발
4. Tensorflow를 통한 분류 모델 작성 후 문제점을 인지
- FastText Supervised 를 사용하여 분류기 개발

모델 성능 평가

```
pd.DataFrame(tmp, columns=['1','2','3','4','5','평균'])
```

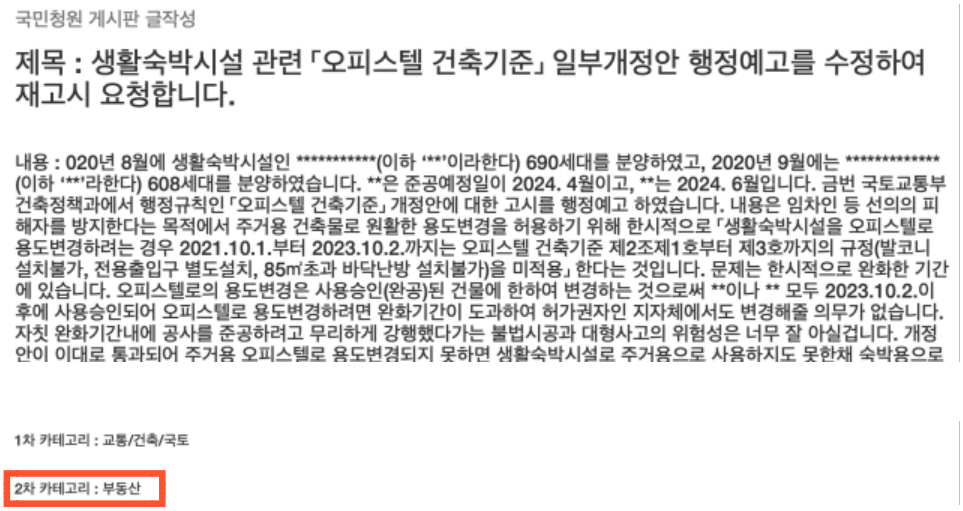
	1	2	3	4	5	평균
0	0.738401	0.735718	0.738708	0.727415	0.736136	0.735276
1	0.779919	0.779513	0.775862	0.784178	0.783570	0.780609
2	0.688217	0.694512	0.693638	0.688751	0.690631	0.691150
3	0.753931	0.726415	0.734277	0.750000	0.732704	0.739465
4	0.756814	0.769972	0.753994	0.764333	0.775141	0.764051
5	0.911950	0.911129	0.905934	0.911403	0.911676	0.910418
6	0.682353	0.670103	0.670103	0.687776	0.668630	0.675793
7	0.697341	0.705235	0.699211	0.695472	0.708913	0.701234
8	0.708967	0.699903	0.702169	0.705406	0.711349	0.705559
9	0.766078	0.768744	0.771743	0.773000	0.770000	0.769913
10	0.724348	0.714348	0.715870	0.719287	0.718200	0.718410
11	0.777819	0.782877	0.781379	0.776695	0.793517	0.782458
12	0.646075	0.669241	0.644788	0.646075	0.662371	0.653710
13	0.443804	0.504323	0.498559	0.476879	0.427746	0.470262
14	0.739710	0.764288	0.748297	0.751555	0.733116	0.747393
15	0.734589	0.725782	0.723656	0.722053	0.710814	0.723379
16	0.854101	0.858062	0.854407	0.856388	0.855563	0.855704

```
pd.DataFrame(tmp, columns=['1','2','3','4','5','평균'])
print('전체 카테고리 기준 평균은 : {}'.format(sum(list(map(lambda x:x[5] , tmp))) / 17 ))
```

```
전체 카테고리 기준 평균은 : 0.7308696221612786
```

6. 사용 예시

Flask 웹 서비스를 통한 (로컬)서비스 페이지 개발



프로젝트 소개

개인 프로젝트

1. 프로젝트 제목 :

K-means 및 유사 문장 탐색 알고리즘 구현

2. 프로젝트 개발 배경 :

패스트캠퍼스의 "텍스트 분석을 위한 머신러닝" 강의를 듣고 텍스트 군집화에 관심을 갖게 되었습니다.
대표적인 군집화 방식인 K-means 알고리즘을 직접 구현해보고 싶은 생각으로 프로젝트를 시작했습니다.

K-means를 구현하던 중 벡터간의 거리를 이용 한 분석에 대해 관심을 갖게 되어, 단어의 벡터를 독립적으로 이용하여 유사 문장을 찾는 알고리즘을 구현하게 되었습니다.

3. 요구 사항 :

1. K-means 알고리즘을 구현합니다.
- 거리를 계산하는 공식은 K-means의 성능을 측정 한 논문을 참고하여 [Jaccard, Cosine, Euclidian, Correletion, KL-Diverse] 를 이용하여 구현합니다.

2. 입력 된 문장과 비슷한 문장을 찾는 알고리즘을 구현합니다.
- 문장 벡터 및 단어 벡터를 기반으로 거리 공식을 이용하여 유사 문장을 탐색합니다.

4. Github

https://github.com/Lion2me/Lionlp

AI - 개별 프로젝트 K-means 및 유사 문장 탐색 알고리즘 구현

5-1 . K-means 개발 과정 :

1. 문장의 전처리 과정을 설계했습니다. (Lionlp Github의 preprocess 내 소스)
- 정규표현식을 이용 한 전처리 및 Branching 를 이용 한 부분 단어 등장 횟수 탐색, N-gram 추출 등을 구현했습니다.

2. 클러스터 클래스를 작성했습니다.
- ID 를 구분하고 getter와 setter 등을 구현 한 뒤 벡터를 표현하는 centroid_ 를 초기화 했습니다.

3. Distance 관련 함수를 구현했습니다.
- scipy를 사용하여 가능한 간략하게 구현했습니다.
- KL-Diver는 zero-division 문제로 인해서 구현하지 못했습니다.

4. 클러스터의 초기 벡터에 초기값을 지정해주는 함수를 구현했습니다.
- 추가적으로 모든 클러스터를 랜덤 초기화 한 후 Entropy를 계산하여 가능 한 높은 Entropy의 분포를 가진 클러스터를 선택하는 알고리즘을 구현해보았습니다.

5-2 . K-means 실행 과정 :

AI - 개별 프로젝트 K-means 및 유사 문장 탐색 알고리즘 구현

6-1 . 유사 문장 탐색 알고리즘 개발 과정 :

1. K-means 구현 시 사용했던 전처리 함수를 공유했습니다.

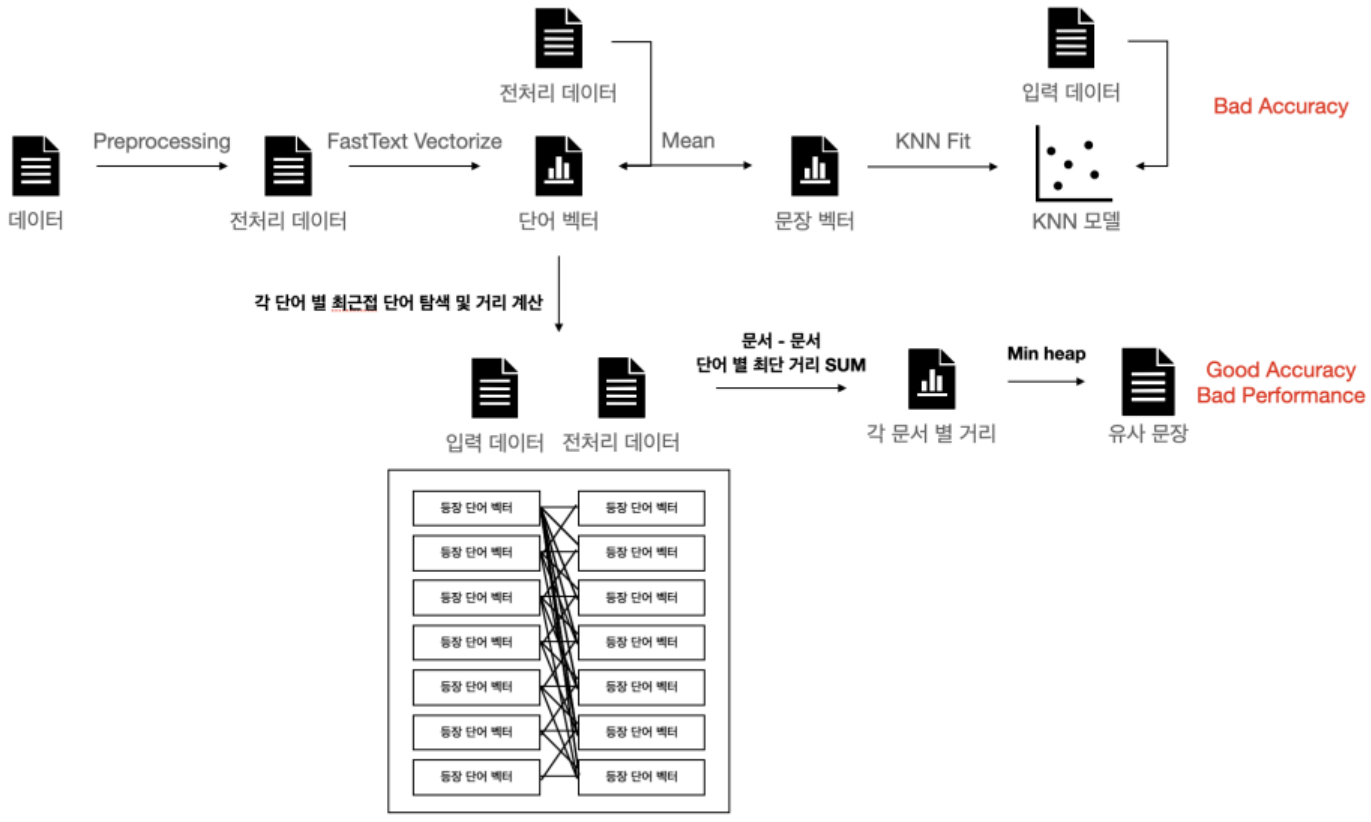
2. pretrained 된 Distributed Representation 모델을 이용하여 단어를 벡터화 시킨 후 평균을 이용하여 문장의 벡터를 유추했습니다.
- 각 컬럼의 값을 평균 연산하였습니다.

3. 문장의 벡터을 이용하여 KNN 알고리즘을 통해 가장 유사한 단어를 탐색해보았습니다.

[KNN을 통한 유사 문장이 좋은 결과를 얻지 못해 다른 방식으로 문제에 접근했습니다.]

4. 문장의 벡터가 아닌 단어 벡터를 이용하여 두 문장의 등장 단어 전체를 거리 연산하여 가장 가까운 단어 셋을 Heap에 저장하는 형태로 문제에 접근했습니다.
- 가장 유사한 단어가 함께 사용 되었다면 Input 데이터의 단어 셋과 해당 문서의 단어 셋의 거리가 짧아지고 min_heap을 통해 유사한 문장이 추출 될 것이라고 예상했습니다.
- 연산 시간이 오래 걸리는 문제가 발생했습니다.

6-2 . 유사 문장 탐색 알고리즘 실행 과정 :



교육 관련 증명 서류

1. 카카오 엔지니어와 데이터 엔지니어링 입문 on cloud 3기

러닝스푼즈 주관

2022-09-17 ~ 2022-10-22

교육 소개

데이터 엔지니어링의 기본 지식에 대한 강의
ELK 파이프라인에 대한 실습

교육을 통해 배운 점

- 람다 아키텍처와 카파 아키텍처에 대한 지식 습득
- 배치 처리 파이프라인과 스트림 처리 파이프라인에 대한 지식 습득
- ELK 스택의 기본적인 실습 제공

CERTIFICATE
of Completion

This Is To Certify That
노경주
Has Successfully Completed The Course Of
카카오 엔지니어와 데이터 엔지니어링 입문 on Cloud 3기
Course Duration
2022. 09. 17 – 2022. 10. 22 (총 6회, 총 18시간)
Course Cost
640,000원

위 사람은 러닝스푼즈에서 실시한
교육 과정을 성실히 수료하였으므로
이 증서를 수여합니다.

Changmin LEE
The CEO of Learning Spoons

2. 텍스트 분석을 위한 머신러닝

패스트 캠퍼스 주관

2019-12-07 ~ 2020-02-22

교육 소개

텍스트 분석에 필요한 지식과 기술을 배울 수 있었던 교육

교육을 통해 배운 점

- 텍스트 데이터를 벡터화 시키는 임베딩 기법의 종류와 특징
- 텍스트 데이터 분류를 위한 기법의 종류와 특징
- TextRank를 통한 랭킹 알고리즘

fast campus Life Changing Education fastcampus.co.kr

패스트캠퍼스
교육과정
수료 증명서

문서번호 FC-200302-CA0001
발행일자 2020. 03. 02

성명
노경주

수료정보

수료 과정	교육 기간	이수 시간	수료일
텍스트 분석을 위한 머신러닝 3기	2019. 12. 07 ~ 2020. 02. 22	30시간	2020. 02. 22

위 사람은 패스트캠퍼스의 상기 과정에서
성실하게 수료하였음을 증명합니다.
2020. 03. 02

패스트캠퍼스(주)

교육 관련 증명 서류

3. Statistic Learning : 기계학습의 기초부터 응용까지

한국통신학회 주관

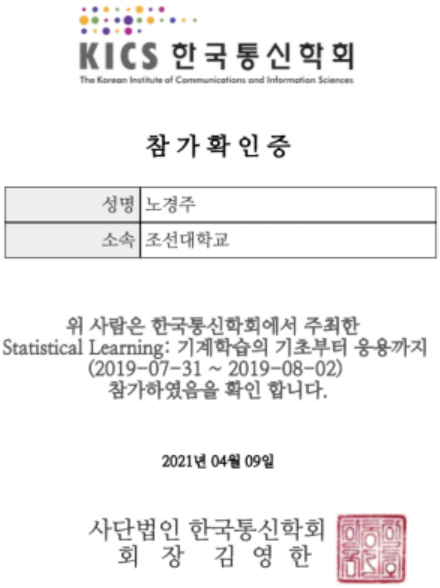
2019-07-31 ~ 2019-08-02

교육 소개

기계학습의 기초 지식과 모델 학습 방법론 및 주의 할 점에 대한 세미나

교육을 통해 배운 점

- 회귀를 통한 데이터 분석의 이해
- F-Fold와 같은 모델 테스트 방법론 및 검증 시 주의 할 점
- 데이터에 따른 분석 기법과 문제 해결 방안



KICS 한국통신학회
The Korean Institute of Communications and Information Sciences

참 가 확 인 증

성명	노경주
소속	조선대학교

위 사람은 한국통신학회에서 주최한
Statistical Learning: 기계학습의 기초부터 응용까지
(2019-07-31 ~ 2019-08-02)
참가하였음을 확인 합니다.

2021년 04월 09일

사단법인 한국통신학회
회 장 김 영 한

4. SNS 기반 빅데이터 분석 교육

스마트인재개발원 주관

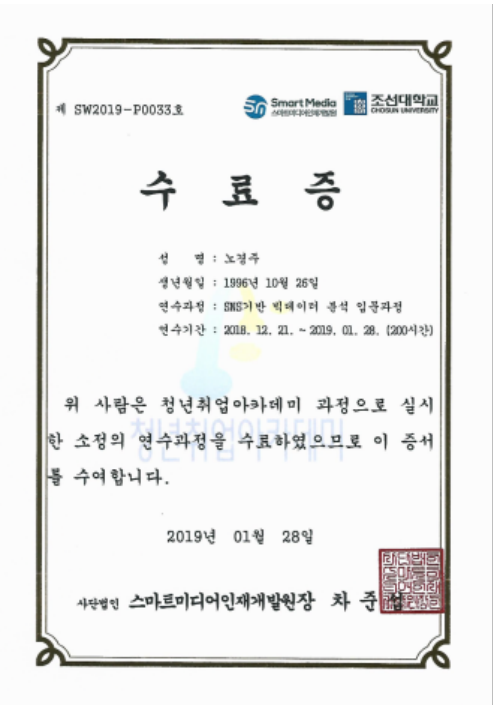
2018-12-21 ~ 2019-01-28

교육 소개

빅데이터 분석을 위한 데이터 수집 및 정제 / 분석 과정에 대한 교육

교육을 통해 배운 점

- 크롤링을 통한 데이터 수집 방법
- API를 통한 공공데이터포털 데이터 활용



제 SW2019-P0033호

Smart Media 조선대학교 CHOSUN UNIVERSITY

수 료 증

성 명 : 노경주
생년월일 : 1996년 10월 26일
연수과정 : SNS기반 빅데이터 분석 입문과정
연수기간 : 2018. 12. 21. ~ 2019. 01. 28. (200시간)

위 사람은 청년취업아카데미 과정으로 실시한 소정의 연수과정을 수료하였으므로 이 증서를 수여합니다.

2019년 01월 28일

사단법인 스마트미디어인재개발원장 차 준

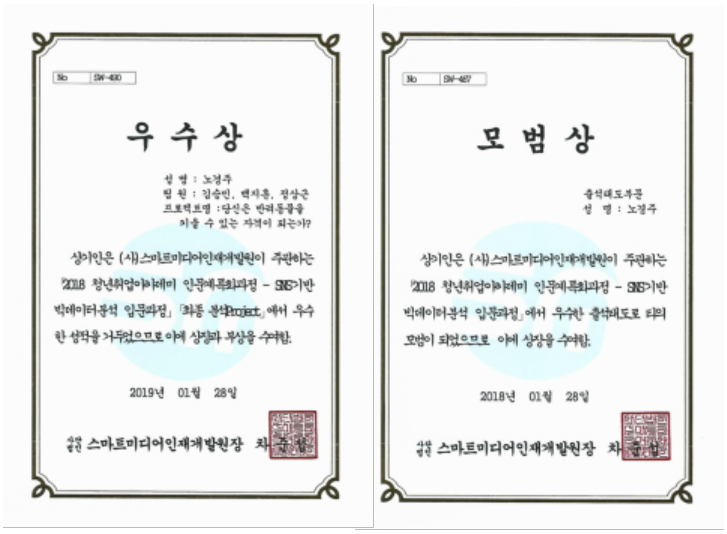
수상 관련 증명 서류

1. SNS 기반 빅데이터 분석 최종 프로젝트

스마트인재개발원

2019-01-28

우수상/모범상



우 수 상

성 명 : 노경주
팀 원 : 김승민, 백시홍, 정상준
프로젝트명 : 당신은 반려동물을 이해 할 수 있는 자격이 되는가?

상기인은 (사)스마트미디어인재개발원이 주관하는 '2018 청년취업아카데미 인공예측최과정 - SNS기반 빅데이터분석 입문과정」 최종 분석Project」에서 우수한 성적을 거두었으므로 이에 상장과 부상을 수여함.

2019년 01월 28일

(사)스마트미디어인재개발원장 차 [redacted]

모 범 상

공식대도자문
성 명 : 노경주

상기인은 (사)스마트미디어인재개발원이 주관하는 '2018 청년취업아카데미 인공예측최과정 - SNS기반 빅데이터분석 입문과정」에서 우수한 출석태도로 타의 모범이 되었으므로 이에 상장을 수여함.

2018년 01월 28일

(사)스마트미디어인재개발원장 차 [redacted]

프로젝트 소개

유기동물 입양 과정에 대한 데이터 분석

기여 한 점

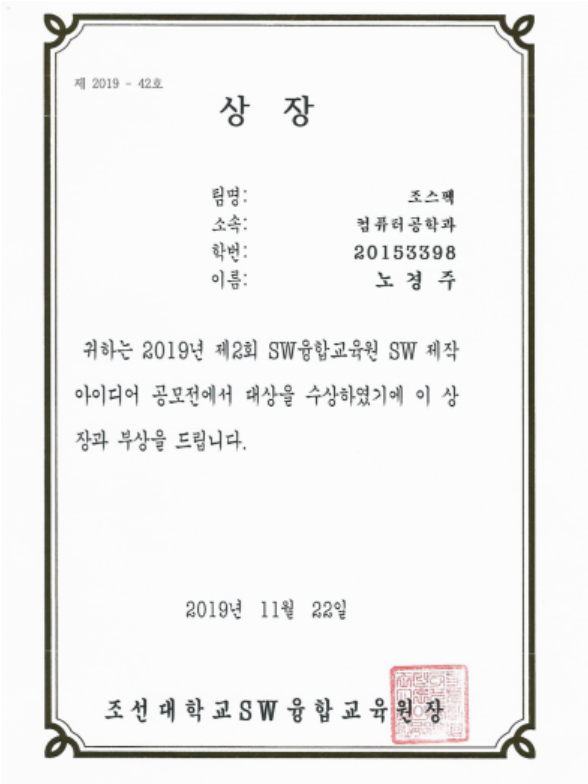
- 프로젝트의 팀장으로서 SNS기반 빅데이터 프로젝트를 진행하며 데이터 분석을 맡아 활동했습니다.
- 팀원 중 장애로 의사소통이 어려운 팀원이 있었지만 각 팀원에게 업무를 분배하여 프로젝트를 진행했습니다.

2. SW 제작 아이디어 공모전 대상

조선대학교 SW융합교육원

2019-11-22

대상



제 2019 - 42호

상 장

팀명: 조스팩
소속: 컴퓨터공학과
학번: 20153398
이름: 노 경 주

귀하는 2019년 제2회 SW융합교육원 SW 제작 아이디어 공모전에서 대상을 수상하였기에 이 상장과 부상을 드립니다.

2019년 11월 22일

조선대학교SW융합교육원장

프로젝트 소개

스펙을 쌓을 수 있는 방법을 공유하거나 졸업한 학생들과 함께 취업을 준비할 수 있는 소프트웨어 제안

기여 한 점

- SW 아이디어 공모전에서 개발하는 과정의 기술적인 설계를 맡아서 활동했습니다.
- 프로그램 예시를 만드는 과정에서 UX/UI에 대한 설계를 맡았습니다.

수상 관련 증명 서류

3. 교내 프로그래밍 대회 동상

조선대학교 SW융합교육원

2019-10-17

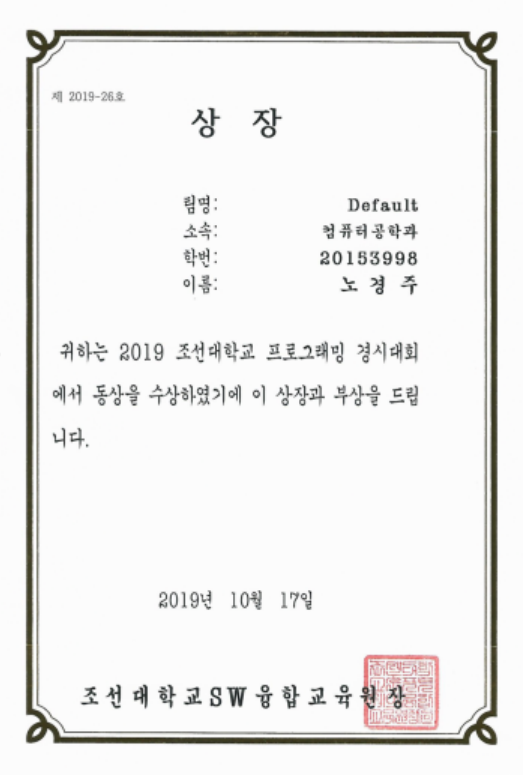
동상

대회 소개

교내에서 진행한 프로그래밍 대회입니다.

기여 한 점

- 팀장으로서 출제 된 알고리즘 문제를 팀원과 함께 문제를 해결했습니다.



제 2019-26호

상 장

팀명: Default
소속: 컴퓨터공학과
학번: 20153998
이름: 노 경 주

귀하는 2019 조선대학교 프로그래밍 경시대회에서 동상을 수상하였기에 이 상장과 부상을 드립니다.

2019년 10월 17일

조선대학교SW융합교육원장

추가 정보

GitHub / 블로그 운영

2020-08-03 ~

개인 공부

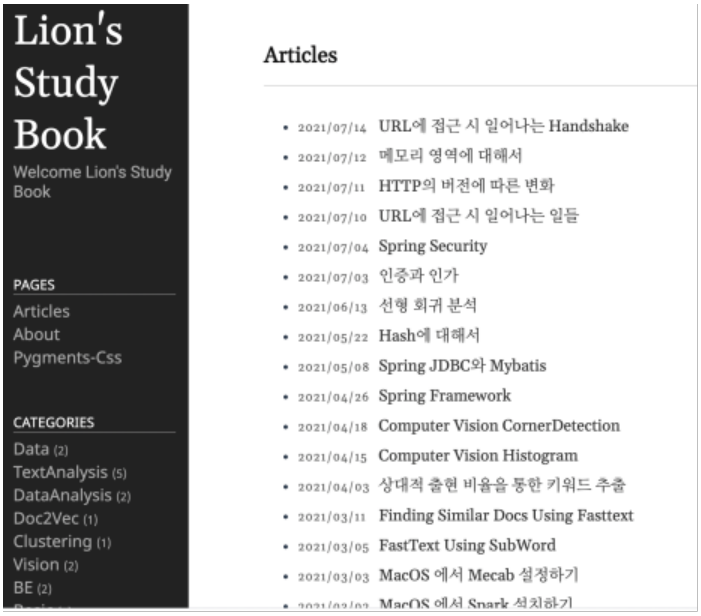
블로그 소개

기본 CS 지식 / BE / 데이터 분석 등 공부한 내용을 포스팅

URL

https://lion2me.github.io/

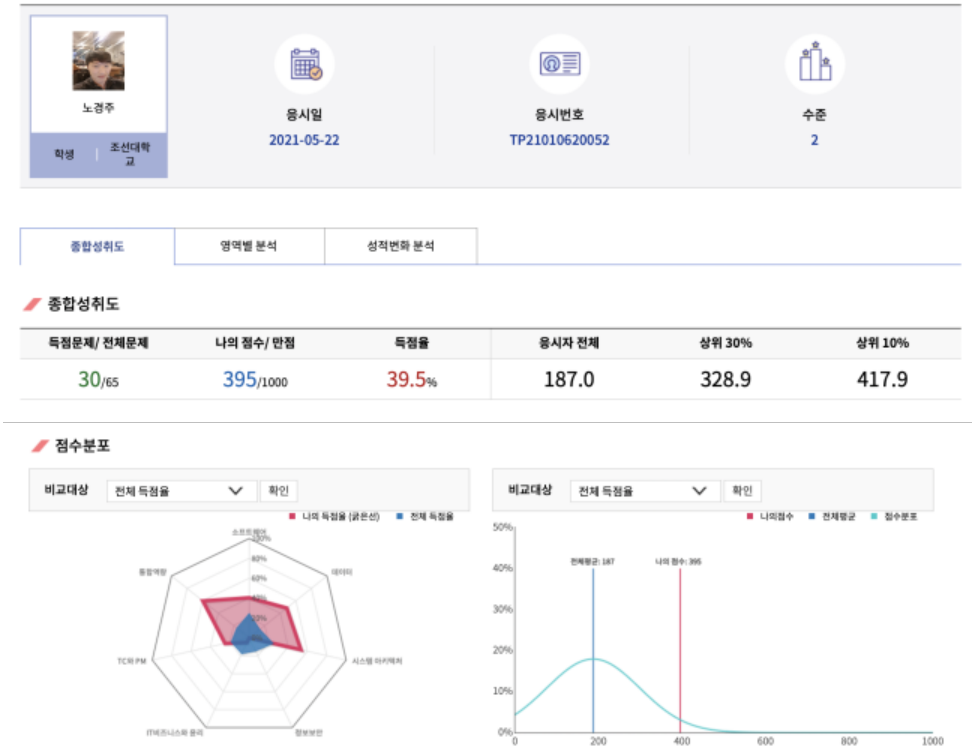
https://github.com/Lion2me



Topcit 395

2021-05-22

개인 공부



추가 정보

Notion

상세 경력 정리

URL

https://flax-flat-3cf.notion.site/afb7a61c6ad44c71a9a3e5d7d48787b0?v=45c16188d9a14824a8fc375c6778070a&pvs=4