**[ 4월 2주차 프로젝트 수행 일지 ]**

| **프로젝트 타이틀** | 빅데이터 문제해결 프로젝트 | | |
| --- | --- | --- | --- |
| **프로젝트 팀명** | 미안하다 4랑한다 | 프로젝트 팀원 | (김채원)  (윤형석, 전동준, 명재정, 허정윤) |

※ 프로젝트 수행 일지는 매 주 금요일 17:00까지 최종본을 저장해주세요

**1. 프로젝트 수행 계획 및 현황**

| **1. 프로젝트 수행 계획 및 현황** | | |
| --- | --- | --- |
| **이름** | **금주 프로젝트 수행 내용** | **차주 프로젝트 수행 계획** |
| **데이터**  **사이언스** | - 주제 관련 피처데이터 탐색  - 주제 관련 머신러닝/딥러닝 모델 탐색  (어떤 모델을 사용할 수 있을지 탐색 및 회의)  (원하는 결과가 나올 수 있는지 탐색) | - 주제 관련 피처데이터 탐색  - 주제 관련 머신러닝/딥러닝 모델 탐색  - 최종 주제 선정시, 데이터 EDA 및 시각화와  학습용 머신러닝/딥러닝 모델 회의 |
| **데이터**  **엔지니어링** | - 주제 관련 데이터 탐색 및 수집  - 수집한 데이터로 테이블 및 ERD 구상  - 데이터 사이언스의 학습 모델링 진행을 위한  데이터 탐색 및 전처리 관련 회의 | - 주제 관련 데이터 탐색 및 수집  - 수집한 데이터로 테이블 및 ERD 구상  - 최종 주제 선정시, 학습용 머신러닝/딥러닝 모델을 위한 피처 데이터 전처리 및 EDA 회의 |

**2. 강사님 피드백**

| **2. 강사님 피드백 : <외국인들을 위한 한식 식당 추천 서비스> 주제 선정 관련 피드백** | |
| --- | --- |
| **데이터사이언스**  **이경미 강사님** | **- 특정 클라이언트/타겟층(외국인)을 위한 서비스라기에는 폭넓은 서비스 지원이다.**  **- 이미지분류 등 딥러닝 작업 소요 시간을 고려해봐야 한다.**  **- 이미지 검색은 API가 이미 존재한다.** |
| **데이터엔지니어링**  **이동헌 강사님** | **- DE(데이터 엔지니어링)의 파이프라인 구축 과정의 난이도가 다소 쉽게 느껴진다.**  **- 빅데이터의 전처리를 통해 여러개의 entity가 나올 수 있는 과정이 필요하다.**  **- 서비스의 타겟층과 그 타겟층을 위한 기능 제공에 초점을 맞출 필요가 있다.** |

**[ 4월 3주차 프로젝트 수행 일지 ]**

| **프로젝트 타이틀** | 빅데이터 문제해결 프로젝트 | | |
| --- | --- | --- | --- |
| **프로젝트 팀명** | 미안하다 4랑한다 | 프로젝트 팀원 | (김채원)  (이상동, 윤형석, 전동준, 명재정, 허정윤) |

※ 프로젝트 수행 일지는 매 주 금요일 17:00까지 최종본을 저장해주세요

**1. 프로젝트 수행 계획 및 현황**

| **1. 프로젝트 수행 계획 및 현황** | | |
| --- | --- | --- |
| **이름** | **금주 프로젝트 수행 내용** | **차주 프로젝트 수행 계획** |
| **데이터**  **사이언스** | - 주제 찾기  - 주제에 맞는 데이터 탐색 및 수집  - 주제에 관한 기능 구상을 위한 관련 논문 탐색  - 머신러닝/딥러닝 모델 탐색 및 논문 검색 | - KoNLPy, W2V, D2V 관련 조사 및 학습  - 크롤링된 데이터에서 키워드 추출  - 추출된 키워드의 유사도 및 정확도 검증 |
| **데이터**  **엔지니어링** | - 주제 찾기  - 주제에 맞는 데이터 탐색 및 수집  - 주제에 구현할 기능 구상  - 엔지니어링 기술 탐색 | - API 데이터 수집 & 데이터 크롤링 한 후 하둡에 올리고 전처리  - 테이블 정의서 및 ERD 작성  - 하둡에 올린 데이터 Spark로 처리 |

**2. 강사님 피드백**

| **2. 강사님 피드백 : 주제 선정 관련 피드백** | |
| --- | --- |
| **데이터사이언스**  **이경미 강사님** | - **주제에 대해 추상적으로 생각하지 말고 데이터 깊게 분석해보고 보안해 나가면서 하나의 주제를 깊숙히 파야한다.**  **- 벤치마킹을 하면서 비슷한 기능을 제공하는 서비스들을 분석하는 것만으로도 큰 강점을 가질 수 있다.**  **- 벤치마킹을 하려면 차별화되거나 기능적이거나 분석적인 방법에서 더 강점이 있어야 한다.**  **- 시간을 효율적으로 사용하기 위해 팀원들의 역할 분배를 잘 해야한다.** |
| **데이터엔지니어링**  **이동헌 강사님** | **- 엔지니어링 파트가 어떤 기술을 사용할 것인지 잘 구상해야 한다.**  **- 하둡에 데이터를 올리고 스파크에서 어떤 것을 전처리 할 수 있는지 생각해야 한다.**  **- 주제를 정할 때 핵심으로 파고들지 못하고 겉도는 느낌이 있다.**  **- 세세하게 정하고 다음으로 진행해야 한다.** |

**[ 4월 4주차 프로젝트 수행 일지 ]**

| **프로젝트 타이틀** | 빅데이터 문제해결 프로젝트 | | |
| --- | --- | --- | --- |
| **프로젝트 팀명** | 미안하다 4랑한다 | 프로젝트 팀원 | (김채원)  (이상동, 윤형석, 전동준, 명재정, 허정윤) |

※ 프로젝트 수행 일지는 매 주 금요일 17:00까지 최종본을 저장해주세요

**1. 프로젝트 수행 계획 및 현황**

| **1. 프로젝트 수행 계획 및 현황** | | |
| --- | --- | --- |
| **이름** | **금주 프로젝트 수행 내용** | **차주 프로젝트 수행 계획** |
| **데이터**  **사이언스** | - 데이터 크롤링 진행중  - 네이버맵 : 상세정보(주소, 메뉴, 사진 등) + 리뷰  - 다이닝코드 : 1차로 음식점별 page id 수집  - 리뷰 데이터 텍스트분석 모델링 자료 조사  - 리뷰 모델링 (키워드 추출, 군집화)  - 데이터 전처리(KoNLPy(토큰화), 불용어제거, 벡터화)  - 분석 설계서 작성 : 피처데이터 | - 데이터 크롤링 완성 : 네이버맵 / 다이닝코드  - **텍스트 분석 모델링**  - 리뷰 데이터로 텍스트 전처리 진행  (클렌징, 토큰화, stop word 제거 등)  - 전처리된 리뷰 데이터를 벡터화 진행 후 피처데이터 추출 ( Bow->TF-IDF, W2V -> CBOW/SkipGram)  - 벡터화된 피처 데이터들로 머신러닝 모델링 진행  (군집화 진행) |
| **데이터**  **엔지니어링** | - API 데이터 수집 완료  - API로 받은 데이터 하둡에 적재  - 하둡에 적재한 데이터 Spark로 처리  - 상호명, 주소 이상치 처리  - 좌표(lat, lot) 추가  - 데이터 크롤링 진행 중  - 데이터 명세서 작성  - System Architecture 작성  - AWS 환경 구축  - Hadoop, Spark  - MySQL, MongoDB | - Airflow 구축  - ElasticSearch 기능 구현  - 하둡에 올린 데이터 spark로 처리  - 크롤링 데이터 합치기  - django 구축 시작 |

**2. 강사님 피드백**

| **2. 강사님 피드백** | |
| --- | --- |
| **데이터사이언스**    **이경미 강사님** | **- 진행상태가 느리니 크롤링 이후 분석을 빨리 진행해야 할 것**  **- 강사님이 제공한 다양한 참고자료를 보고 텍스트분석쪽 공부를 한 뒤 어떻게 벡터화를 진행할 건지 혹은 토픽모델링을 어떻게 활용할 것인지 결정해야 함** |
| **데이터엔지니어링**    **이동헌 강사님** | **- 크롤링을 빠르게 마무리하고 데이터 파이프라인 구축에 힘을 써야 한다.**  **- 병렬로 일을 처리 해야 한다.** |

**[ 4월 5주차 프로젝트 수행 일지 ]**

| **프로젝트 타이틀** | 빅데이터 문제해결 프로젝트 | | |
| --- | --- | --- | --- |
| **프로젝트 팀명** | 미안하다 4랑한다 | 프로젝트 팀원 | (김채원)  (이상동, 윤형석, 전동준, 명재정, 허정윤) |

※ 프로젝트 수행 일지는 매 주 금요일 17:00까지 최종본을 저장해주세요

**1. 프로젝트 수행 계획 및 현황**

| **1. 프로젝트 수행 계획 및 현황** | | |
| --- | --- | --- |
| **이름** | **금주 프로젝트 수행 내용** | **차주 프로젝트 수행 계획** |
| **데이터**  **사이언스** | - **텍스트 분석 모델링 구현 진행(카카오맵 데이터)**  - 리뷰 데이터로 텍스트 전처리 진행  (클렌징, 토큰화, stop word 제거: 등)  - 전처리된 리뷰 데이터로 LDA(토픽모델링) 진행  - 최적의 파라미터 튜닝 진행  (Perplexity, Topic Coherence)  - pyLDAvis 으로 시각화  - 전처리된 리뷰 데이터로 W2V & LDA 진행 | - **텍스트 분석 모델링 구현 완성 (전체 데이터)**  - 전체 리뷰 데이터(네이버맵,카카오맵,다이닝코드,식신)로  텍스트 전처리 진행  (클렌징, 토큰화, stop word 제거 등)  - 전처리된 전체 리뷰 데이터로 LDA(토픽모델링) 진행  - 최적의 파라미터 튜닝 진행  (Perplexity, Topic Coherence)  - pyLDAvis 으로 시각화  - 유사도 높은 단어사전 생성  - 최종 테마 선정 |
| **데이터**  **엔지니어링** | - ElasticSearch와 MySQL을 연동하여 MySQL의 table의 내용을 ElasticSearch를 이용해 API로 검색할 수 있도록 구축  - Hadoop에 저장한 음식점 API데이터 Spark로 처리  - Spark로 처리한 데이터 MySQL DB에 저장  - Spark로 처리한 데이터 MongoDB에 저장  - MongoDB의 geospatial을 이용하여 특정 주위 반경 1km내에 MongoDB에 저장된 row를 반환하도록 함수 제작  - Django   1. 로그인/로그아웃, 게시물, 댓글, 좋아요 기능 생성 2. 랭킹 페이지 생성 3. 카카오맵 API를 이용한 맛지도 페이지 생성(초안) | - 최종 데이터 정제 후 db에 저장  - 에어플로우 구축 시작  - erd 그리기  - Django   1. 랭킹 페이지 계속 구현 2. 마이 페이지 구현 3. 검색창 만들고 ElasticSearch와 연동 4. 게시물, 댓글 등 CRUD 구현 5. 피드(게시물) 상단에 닉네임과 프로필 사진 나오게 만들기 6. 회원정보 그룹(계급?레벨?) 추가 7. 로그인 시 우측 상단 프로필 사진 변경 8. 음식점 정보 리스트 페이지 구현 9. 그 외 세부 사항들 회의 후 구현 예정 |

**2. 강사님 피드백**

| **2. 강사님 피드백** | |
| --- | --- |
| **데이터사이언스**  **이경미 강사님** | - 크롤링된 4개의 사이트에서 가지고 있는 리뷰데이터들 갯수를 파악하고, 토픽모델링(LDA), Word2Vec를 사용하여 텍스트 분석을 진행하면 됨  - 진행된 부분들을 문서화하고 이를 취합하면서 어디까지 진행되었고 어디를 해야하는지 확인해봐야 함 |
| **데이터엔지니어링**  **이동헌 강사님** | - id를 활용하여 음식점들을 join하는 방법을 모색해봐야 함  - 역할 분담하여 병렬적으로 진행해야 한다. |

**[ 5월 1주차 프로젝트 수행 일지 ]**

| **프로젝트 타이틀** | 빅데이터 문제해결 프로젝트 | | |
| --- | --- | --- | --- |
| **프로젝트 팀명** | 미안하다 4랑한다 | 프로젝트 팀원 | (김채원)  (윤형석, 전동준, 명재정, 허정윤) |

※ 프로젝트 수행 일지는 매 주 금요일 17:00까지 최종본을 저장해주세요

**1. 프로젝트 수행 계획 및 현황**

| **1. 프로젝트 수행 계획 및 현황** | | |
| --- | --- | --- |
| **이름** | **금주 프로젝트 수행 내용** | **차주 프로젝트 수행 계획** |
| **데이터**  **사이언스** | - **텍스트 분석 모델링 구현중 (전체 데이터)**  - 전체 리뷰 데이터(네이버맵,카카오맵,다이닝코드,식신)로  텍스트 전처리 진행  (클렌징, 토큰화, stop word 제거 등)  - 전처리된 전체 리뷰 데이터로 LDA(토픽모델링) 진행  - 최적의 파라미터 튜닝 진행  (Perplexity, Topic Coherence)  - pyLDAvis 으로 시각화 | - **텍스트 분석 모델링 구현완료 (전체 데이터)**  - 전처리된 전체 리뷰 데이터로 LDA(토픽모델링) 진행  - iteration 횟수 결정 (400 or 500)  - 최종 테마 선정  - 테마별 음식점 적용  - **모델링 최종 결과 장고와 연동하기 -> 서비스 배포** |
| **데이터**  **엔지니어링** | **-** 리뷰 데이터 스파크로 처리 후 db에 저장하기  - 에어플로우 구축 시작  - 엘라스틱 서치로 검색기능 구현  - aws에서 데이터 저장, 처리 테스트  - django Backend   1. 리뷰에 좋아요를 받은 순으로 랭킹페이지를 구현 2. 리뷰에 텍스트형(내용) 컬럼 추가 3. 리뷰 작성 시 서비스, 맛, 위생상태, 가격을 매기도록 함. DB컬럼 추가 후 5개 짜리 별점을 받는 form 생성 4. 기존 회원가입 시 User 모델의 gender, birth\_date 위젯을 바꿈 5. MyPage 구현(프로필 수정, 비밀번호 변경, 계정삭제 기능 포함) 6. 리뷰 수정/삭제 기능 구현 | **-** 엘라스틱 서치 장고와 연동 후 자동완성 시키기  - 에어플로우 구축 완료하기  - 서비스 페이지 테스트  - django Backend   1. Store(음식점) App 생성 2. 음식점 리스트, 세부사항 등 모델 생성 후 페이지 구현 3. 검색창과 엘라스틱 서치 연동 4. 현재 위치 데이터 받아오는 기능 탐색 5. 리뷰 업로드 사진에서 위치데이터 찾을 수 있는지 확인하고 해당 음식점 데이터가 맞는지 확인하는 기능 구현 6. 지도에 위치 기반 가까운 음식점 마커 구현 |

**2. 강사님 피드백**

| **2. 강사님 피드백** | |
| --- | --- |
| **데이터사이언스**  **이경미 강사님** | - 리뷰데이터 탐색하는 것이 중요하다. 리뷰 데이터 있는 음식점들이 몇 개인지, 업종별/지역별로 통계도 확인해봐야 한다. |
| **데이터엔지니어링**  **이동헌 강사님** | **-** 엘라스틱 서치를 사용한 추가적인 서비스에 대해 생각해봐야 함  - 프로젝트 진행사항을 보고할 때는 현재까지 진행한 현황을 먼저 보고하고, 정체된 부분의 경우 안되고 있는 부분의 이유를 설명하고 어떻게 해결할 것인지 보고 할 수 있도록 해야한다. |

**[ 5월 2주차 프로젝트 수행 일지 ]**

| **프로젝트 타이틀** | 빅데이터 문제해결 프로젝트 | | |
| --- | --- | --- | --- |
| **프로젝트 팀명** | 미안하다 4랑한다 | 프로젝트 팀원 | (김채원)  (윤형석, 전동준, 명재정, 허정윤) |

※ 프로젝트 수행 일지는 매 주 금요일 17:00까지 최종본을 저장해주세요

**1. 프로젝트 수행 계획 및 현황**

| **1. 프로젝트 수행 계획 및 현황** | | |
| --- | --- | --- |
| **이름** | **금주 프로젝트 수행 내용** | |
| **데이터**  **사이언스** |  | |
| **데이터**  **엔지니어링** |  | |

**2. 강사님 피드백**

| **2. 강사님 피드백** | |
| --- | --- |
| **데이터사이언스**  **이경미 강사님** | - 테마를 보여주는 위치 생각해봐야함.  - 검색시에 결과를 보여주는 순서의 기준 필요  - 테마의 수를 더 늘리는 것도 고려 |
| **데이터엔지니어링**  **이동헌 강사님** | - 엘라스틱 서치의 기능 |