**기업요구사항 기반의 문제해결 프로젝트 기획안**

**2021년 08월 27일**

| **프로젝트 조** | **6조** | |
| --- | --- | --- |
| **프로젝트 팀원** | 팀명: 업빛투  팀장: (정) 정길종, (부) 김형림  팀원: (사이언스) 윤보람, 채길호 (엔지니어링) 인태우 | |
| **프로젝트 주제** | **뉴스 매체와 유튜버의 주식가격 등락 예측 정확도 분석** | |
| **프로젝트 목적** | **미디어 키워드 분석으로 업종별 이슈가 한국 주식에 미치는 영향을 파악하고, 웹 페이지 대시보드로 구현** | |
| **개요** | 1. 데이터 수집  ① 크롤링으로 유튜브 스크립트, 언론 보도 이슈 수집  - 유튜브의 경우 ( 구독자수 기준 ) : 금융 유튜브 채널 상위 4개  - 언론보도 자료 ( 유료부수 구독자 수 기준 ) : 매일경제, 머니투데이  - 매일 업로드 되는 자료에 대한 스크립트 추출 예정  ② FinanceDataReader 라이브러리로 주식가격 데이터 수집 ( 한국거래소 자료 )  - 매일 실시간 종가 데이터 수집  - 추후 업종 및 종목을 선택 예정  2. 데이터 전처리  ① 유튜브 스크립트 & 언론 보도 이슈  - 텍스트 마이닝  ② 주식가격  - 종목마다의 가격 차이를 고려하여 MinMaxScaler 적용  3. 데이터 분석  ① 유튜브 스크립트 & 언론 보도 이슈  - 주가 등락 예측을 위한 감성 분석(목적성)    ② 주식 가격  - LSTM으로 주식 가격 시계열 예측(분석 방법 비교)  ③ Label  - ①+② 분석을 종합하여 주식가격의 상승/정체/하락 예측  4. 데이터 시각화  - 일별 이슈 키워드를 나타낸 워드 클라우드로  - 키워드 간의 연관성(긍정/부정)  - 주가의 시계열 그래프  5. 웹 페이지 구현  - 이슈로 예측한 주식가격 예측 결과를 한 눈에 볼 수 있도록 웹 페이지 구현  - 해당 회사의 주식 그래프(전날 종가) 및 예측 등락  - 유튜브 이슈 워드 클라우드  - 보도 이슈 워드 클라우드 | |
| **프로젝트 수행 방향**  **(주요 기능 설명)** | **사이언스** | **엔지니어링** |
| (DE로부터 데이터를 전달 받아)  - 텍스트 데이터 EDA 및 텍스트 벡터화  - 텍스트 감성 분석  - LSTM을 사용한 시계열 모델링  - LSTM 파라미터 튜닝  - 감성분석 결과와 LSTM 모델 결과 결합 | - 데이터 전처리  - 데이터 수집  (DS에 데이터를 전달)  - Tableau-웹 페이지 연동  - 백엔드 구현 |
| **프로젝트 수행 도구** | **사이언스** | **엔지니어링** |
| - Python  - TensorFlow  - Tableau  - R | - Python  - JavaScript  - Django  - MySQL  - Spark  - MongoDB |

| **필수 기능** | **사이언스** | **엔지니어링** |  |
| --- | --- | --- | --- |
| * 미디어 키워드 빈도 분석 시각화 (텍스트, 동영상 텍스트 등) * 키워드 감성 분석 결과 시각화 * LSTM 예측 주가 그래프 * 예측 결과 표출 | * 데이터 크롤링 * 시각화 결과 표출 * 대시보드 * 웹 페이지 구현 |  |
| **포함 기술** | **사이언스** | **엔지니어링** |  |
| * LSTM * 워드 클라우드 * 텍스트 마이닝 * 감성분석 | * 크롤링 데이터 DB 저장 * 비정형 데이터 처리 기술 * 정형 데이터 변환 * 데이터 파이프라인 구축 |  |

**## 질문 사항 ##**

1. **주제 선정이 적절했는지에 대한 의견**
2. **데이터 수집 & 분석 범위** 
   1. **코스닥은 다른 영향이 많을 것 같아서 뺏는데 괜찮을지 대한 의견**
   2. **분석 업종이나 종목 선택  
      → 분석에 유리한 업종이나 종목을 선택**

**→ 주가 변동성이 작은 종목/업종을 선택해야 하는지**

**→ 거래량이 많은 종목으로 해야하는지**

**→ 수집 연도를 코로나의 영향이 적은 기간으로 하는 것이 좋은지**

* 1. **매체 보도 자료, 유튜브 데이터 수집에 대한 의견**

1. **프로젝트 필요 지식**
   1. **현 분석 목표에 맞는 효율적이거나, 필요한 기술 지식에 대한 설명 필요**
   2. **단어사전(세종사전)**
   3. **(https://kbig.kr/portal/kbig/datacube/niadict.page)**
   4. **AWS 스크래핑 -> 몽고 -> PySpark -> MySQL**
   5. **모델 업데이트 방법, 트레이닝 셋 Train, Test 고려할 점**
2. **프로젝트 일정**
   1. **WBS를 각 단계별 난이도를 고려한 일정 조율**