**[ 9월 1주차 프로젝트 수행 일지 ]**

| **프로젝트 타이틀** | **뉴스기사와 유튜브 영상 스크립트 감성분석과 LSTM을 이용한 주식 등락 예측** | | |
| --- | --- | --- | --- |
| **프로젝트 팀명** | 업빛투 | 프로젝트 팀원 | (팀장) (정) 정길종, (부) 김형림  (팀원) (사이언스) 윤보람, 채길호  (팀원) (엔지니어링) 인태우 |

※ 프로젝트 수행 일지는 매주 금요일 23:59 전까지 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** | **□ LSTM**  **- FinanceDataReader로 주식 데이터 로드 후, 테스트 수행**  **□ KoNLPy**  **- 수집한 2만여 개 뉴스 데이터 분류 작업**  **- 텍스트 분석 학습**  **□ 감성 어휘 사전**  **- 수집한 종목 뉴스 데이터로 테스트 수행**  **- VADER 감성사전이 종목 뉴스 기사를 대부분 긍정으로 분류하는 문제점 발견**  **- 주식가격 변동으로 자체 감성 어휘 사전을 구축 후 감성점수 확인** | **□ EDA**  **- 데이터 이상치 제거**  **- 데이터 속성 관계 분석**  **□ Py-Hanspell**  **- 네이버 한글 맞춤법 검사기 전처리 패키지(YouTube 스크립트)**  **□ 감성 어휘 사전**  **- 주식가격 분석에 적합한 감성 사전 서칭** |
| **데이터**  **엔지니어링** | **□ 뉴스 크롤링**  **- 4개 종목 2020~2021년 뉴스 데이터 수집(매일경제)**  **- 타 언론사 크롤링에 문제(머니투데이)**  **- YouTube 스크립트 크롤링 실습 및 결과 확인**  **□ AWS 내 작업환경 구축**  **- 필요 라이브러리 설치**  **□ MySQL DB 구축**  **- 연결 환경** | **□ 뉴스 크롤링**  **- 필요 종목에 대한 2018~2021년 뉴스 데이터 수집**  **- 타 언론사(머니투데이) 새 크롤링 코드 작성 및 데이터 수집**  **□ YouTube 스크립트 크롤링**  **- 주식 채널 크롤링 데이터 수집**  **□ 데이터 전처리**  **- 데이터 전처리**  **- 데이터 정규화** |

**2. 강사님 피드백**

| **2. 강사님 피드백** | |
| --- | --- |
| **데이터사이언스**  **이경미 강사님** | - 텍스트 데이터 활용한 주가 예측 참고 문헌 추가 조사  - 주식과 관련된 감성사전 조사  - 서비스 타겟과 내용에 대한 고민 필요 |
| **데이터엔지니어링**  **김성환 강사님** | - 뉴스 데이터 수집 기간을 객관적 기준 정하기  - 크롤링 데이터 전처리 후 DB 적재 |

**[ 9월 2주차 프로젝트 수행 일지 ]**

| **프로젝트 타이틀** | **뉴스기사와 유튜브 영상 스크립트 감성분석과 LSTM을 이용한 주식 등락 예측** | | |
| --- | --- | --- | --- |
| **프로젝트 팀명** | 업빛투 | 프로젝트 팀원 | (팀장) (정) 정길종, (부) 김형림  (팀원) (사이언스) 윤보람, 채길호  (팀원) (엔지니어링) 인태우 |

※ 프로젝트 수행 일지는 매주 금요일 23:59 전까지 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** | **□ 감성 어휘 사전**  **- 재무분석에 사용하는 KOSELF 감성사전 확보**  **- KOSELF 감성사전은 애널리스트 리포트로부터 만들어졌기 때문에, 뉴스 기사의 감성 분석에 완벽하게 적합하지는 않다는 문제점 발견 → KOSELF 감성사전 보완 방안 논의**  **□ TF-IDF**  **- 주가 변동률을 Label로 하여 텍스트를 TF-IDF 벡터화 후 Logistic Regression을 적용한 결과 예측 정확도는 50% 정도**  **□ LSTM**  **- 주식 데이터 추출 라이브러리를 FinanceDataReader로 통일**  **- 재무정보로 재무지표를 계산하여 주식데이터에 추가**  **- LSTM 수행 결과 MSE 확인**  **□ Tableau EDA**  **- 주가 데이터 시각화** | **□ 감성 어휘 사전**  **- 뉴스 기사를 통해 자체 제작한 감성사전을 KOSELF에 추가한 후 감성 분석 진행**  **□ TF-IDF**  **- 모델의 예측 정확도 향상을 위한 하이퍼파라미터 튜닝**  **□ 유튜브 데이터 분석**  **- 벡터화된 단어 간 유사도 확인**  **□ 주식 데이터를 통한 종가(Close) 예측**  **- 머신러닝 기법과 강화학습을 통한 종가 예측**  **- 등락 정확도 파악**  **□ Tableau 시각화**  **- 분석 결과 시각화** |
| **데이터**  **엔지니어링** | **□ 크롤링**  **- 4개 종목 2018~2021년 뉴스 데이터 수집 완료(매일경제)**  **- 4개 종목 2018~2021년 뉴스 데이터 수집 완료(아시아경제)**  **- YouTube 3개 채널 스크립트 크롤링 데이터 수집 완료**  **□ AWS 내 Django**  **- 서버 구축 완료**  **- 태블로 시각화 자료 Django 연결 테스트 완료**  **□ 데이터 수집 스케줄링**  **- 데이터 DB에 적재되는 것까지 확인** | **□ Django**  **- Django-MySQL 연동**  **- 태블로 연동**  **- DB 데이터 장고 대시보드 연결**  **□ 유튜브**  **- 유튜브 전체에서 4개 종목 일별 데이터 수집**  **□ 데이터 수집 스케줄링**  **- 주식, 뉴스, 유튜브 데이터 수집 스케줄링** |

**2. 강사님 피드백**

| **2. 강사님 피드백** | |
| --- | --- |
| **데이터사이언스**  **이경미 강사님** | [9/8 수]  - 서비스 구체화  - 범용적인 감성사전 제작 (종목별이 아닌)  - KOSELF 감성사전과 자체 제작 감성사전 단어들 간의 유사도 확인  - 유튜브 스크립트 Raw Data 그대로 분석 (요약 전처리 X)  - 시계열 다양한 분석 비교해볼 것 (ARIMA, Prophet, LSTM, RL)  [9/10 금]  - 수집 데이터 기간 설정의 객관적인 이유   (현재 삼성전자 액면 분할 이후 시점인 2018년 1월 1일부터 데이터 수집)  - 감성 사전을 위한 종목 추가 필요  - PPT 데이터 구축 및 모델 요약 장표 ***Good-!!*** |
| **데이터엔지니어링**  **김성환 강사님** | [9/8 수]  - interactive한 웹페이지 구현  [9/10 금]  - 등락 정확도 파악 |

**[ 9월 3주차 프로젝트 수행 일지 ]**

| **프로젝트 타이틀** | **뉴스기사와 유튜브 영상 스크립트 감성분석과 LSTM을 이용한 주식 등락 예측** | | |
| --- | --- | --- | --- |
| **프로젝트 팀명** | 업빛투 | **프로젝트 팀원** | (팀장) (정) 정길종, (부) 김형림  (팀원) (사이언스) 윤보람, 채길호  (팀원) (엔지니어링) 인태우 |

※ 프로젝트 수행 일지는 매주 금요일 23:59 전까지 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** | **□ 감성 어휘 사전**  **- FastText의 Facebook 한글 사전을 이용한 유사도를 통해서 KOSELF 단어와 유사한 단어 추가, 뉴스 데이터에서 긍부정 단어 추가를 통한 정확도 확인**  **① FastText 한글 모델 안의 단어를 KOSELF에 추가**  **② 크롤링 뉴스 데이터 안의 단어를 KOSELF에 추가**  **- 종목별+연도별 감성사전 구축 中**  **□ ML**  **- 업종별로 ARIMA와 Prophet을 사용해서 파라미터 수정 및 RMSE 측정**  **□ RL**  **- 강화학습 업종별로 시행 및 정확도 측정**  **□ LSTM**  **- 업종별로 LSTM 예측, RMSE 측정**  **□ Tableau EDA**  **- 주가 데이터 및 워드 클라우드 시각화** | **□ 감성 어휘 사전**  **- 뉴스 기사를 통해 자체 제작한 감성사전을 KOSELF에 추가한 후 감성 분석 진행**  **□ 유튜브 데이터 분석**  **- 스크립트 텍스트 전처리 (불용어 처리 고도화)**  **- TF-IDF 벡터화한 Logistic Regression을 유튜브 데이터에 적용**  **- 자체 제작한 감성사전을 유튜브 스크립트에 적용**  **□ 앙상블(Ensemble)**  **- 머신러닝 기법과 강화학습을 통한 주가 예측**  **- 텍스트 분석(TF-IDF & 감성 분석)을 통한 주가 등락 예측**  **- Logistic Regression으로 최종 주가 등락 예측**  **□ Tableau 시각화**  **- 분석 결과 시각화**  **- 웹 페이지와 연동** |
| **데이터**  **엔지니어링** | **□ 셀트리온 종목 추가 크롤링**  **- 뉴스 매체 셀트리온 기사 크롤링 (완료)**  **- 유튜브 3개 채널 셀트리온 스크립트 크롤링 (완료)**  **□ Django**  **- MySQL 연동**  **- css 적용한 html 페이지 구성**  **- DB데이터 연동 후 시각화 작업**  **□ AWS 내 자동 데이터 수집**  **- 주식데이터, 코스피, 뉴스. 유튜브 자동 수집 저장(완료)** | **□ AWS에 Django 서버 이전**  **- 5개 업종에 대한 그래프, 시가 종가, 거래량을 종목별 페이지에서 상세 그래프 시각화**  **- 워드클라우드 종목 별 페이지에 연동 작업**  **- 태블로 분석 결과 시각화 Django 서버에 연동 작업**  **- 일별 뉴스 데이터 변화 가능하게 하도록 웹페이지에 적용** |

**2. 강사님 피드백**

| **2. 강사님 피드백** | |
| --- | --- |
| **데이터사이언스**  **이경미 강사님** | - 감성사전 진행 상황 : 순조롭게 진행 중인 듯  - 주가 등락 기준을 10% → 조절하면서 결과를 확인하는 것도 좋을 듯  - 셀트리온 추가 이유 : 기존 종목과 변동률 면에서 차별화 되어 분석에 좀 더 유의한 결과를 얻을 수 있을 것으로 기대  - 발표하면서 생각한 이슈 : Feature가 12개여서 LSTM Shift를 12일로 진행 |
| **데이터엔지니어링**  **김성환 강사님** | - 웹 페이지 프로토타입 구축  - 유튜브 크롤링 데이터 개수가 현저히 적은 문제 추후 개선 |

**[ 9월 4주차 프로젝트 수행 일지 ]**

| **프로젝트 타이틀** | **뉴스기사와 유튜브 영상 스크립트 감성분석과 LSTM을 이용한 주식 등락 예측** | | |
| --- | --- | --- | --- |
| **프로젝트 팀명** | 업빛투 | **프로젝트 팀원** | (팀장) (정) 정길종, (부) 김형림  (팀원) (사이언스) 윤보람, 채길호  (팀원) (엔지니어링) 인태우 |

※ 프로젝트 수행 일지는 매주 금요일 23:59 전까지 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** | **□ 주식 데이터**  **- 상승/하락/보합 기준으로 원 change 데이터 사용**  **- 보합 기준 설정(Kospi 지수 변동률과 비교)**  **□ 감성 어휘 사전**  **- 뉴스 기사를 통해 자체 제작한 감성사전을 KOSELF에 추가한 후 감성 분석 진행**  **□ 유튜브 데이터 분석**  **- 스크립트 텍스트 전처리 (불용어 처리 고도화)**  **- 워드 임베딩 DNN**  **- TF-IDF 벡터화한 Logistic Regression을 유튜브 데이터에 적용**  **- 자체 제작한 감성사전을 유튜브 스크립트에 적용**  **□ 앙상블(Ensemble)**  **- 머신러닝 기법과 강화학습을 통한 주가 예측**  **- 텍스트 분석(TF-IDF & 감성 분석)을 통한 주가 등락 예측**  **- Logistic Regression으로 최종 주가 등락 예측**  **□ Tableau 시각화**  **- 분석 결과 시각화**  **- 웹 페이지와 연동** | **□ 주식 데이터**  **- 각 모델 결과 확인 및 결과 취합**  **□ 감성 어휘 사전**  **- 어휘 사전 성능 확인 및 정확도 상승을 위한 방안 모색**  **□ 유튜브 데이터 분석**  **- 토큰화된 텍스트를 감성분석에 활용 및 정확도 확인**  **□ 앙상블(Ensemble)**  **- 모델 별 결과 확인 테이블 작성** |
| **데이터**  **엔지니어링** | **□ AWS에 Django 웹페이지 구현**  **- 5개 업종에 대한 종가, 거래량을 종목별 표출 및 종가 데이터 그래프 시각화**  **- 일별 뉴스 데이터 변화 가능하게 하도록 웹페이지에 적용**  **- 종목별 페이지 데이터 모두 DB와 연동하여 구축**  **- 정확도 데이터 메인 및 종목별 DB 데이터 표출**  **- 단말기 별로 데이터 표출 (윈도우, 안드로이드, ios 일부) 해결**  **□ Tableau 시각화**  **- 분석 결과 시각화**  **- 웹 페이지와 연동** | **□ AWS에 Django 웹페이지 구현**  **- 종목별 워드클라우드**  **- 정확도가 언론 데이터나 최근 데이터 불러오도록 코드 수정**  **□ Github 내 Django 서버 구축**  **- github 내에 작업 내용 업로드**  **-** |

**2. 강사님 피드백**

| **2. 강사님 피드백** | |
| --- | --- |
| **데이터사이언스**  **이경미 강사님** | [9월 23일]  홈페이지 구성 추가  구축된 사전을 통해서 확률을 계산→ 사용자가 판단  LSTM, 머신러닝 window-size = 10으로 통일  [9월 24일]  변동률(change) 값 그대로 반영  긍정, 부정, (보합) → 이부분의 구분점 확립  최근 뉴스들의 키워드를 통한 긍부정 판단  LSTM 오차 줄이기(window\_size 조절, layer 조절)  분석결과를 적절한 서비스로 구현하는 방법을 생각 |
| **데이터엔지니어링**  **김성환 강사님** | [9월 24일]  - 장고 서버 github에 옮기기  - 분석한 데이터는 db에 저장 후 db에서 데이터 로드하도록 프로세스 구성  - |

**[ 10월 1주차 프로젝트 수행 일지 ]**

| **프로젝트 타이틀** | **뉴스기사와 유튜브 영상 스크립트 감성분석과 LSTM을 이용한 주식 등락 예측** | | |
| --- | --- | --- | --- |
| **프로젝트 팀명** | 업빛투 | **프로젝트 팀원** | (팀장) (정) 정길종, (부) 김형림  (팀원) (사이언스) 윤보람, 채길호  (팀원) (엔지니어링) 인태우 |

※ 프로젝트 수행 일지는 매주 금요일 23:59 전까지 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** | **□ 주식 데이터**  **- 각 모델 결과 확인 및 결과 취합**  **□ 감성 어휘 사전**  **- 어휘 사전 성능 확인 및 정확도 상승을 위한 방안 모색**  **□ 유튜브 데이터 분석**  **- 토큰화된 텍스트를 감성분석에 활용 및 정확도 확인**  **□ 앙상블(Ensemble)**  **- 모델 별 결과 확인 테이블 작성**  **□ Tableau 시각화**  **- 분석 결과 시각화**  **- 웹 페이지와 연동** | **□ 분석 소스코드 정리**  **- 분석 코드 정리**  **□ 최종 산출물 정리**  **- 소스코드, 화면설계서, ppt 정리**  **□ 앙상블(Ensemble)**  **- 최종 결과 확인**  **□ 분석 결과 DB 업로드**  **- 웹 페이지와 연동 되도록 분석결과 DB 업로드**  **□ 프로젝트 최종 발표** |
| **데이터**  **엔지니어링** | **□ AWS에 Django 웹페이지 구현**  **- 5개 업종에 대한 종가, 거래량을 종목별 표출 및 종가 데이터 그래프 시각화**  **- 일별 뉴스 데이터 변화 가능하게 하도록 웹페이지에 적용**  **- 종목별 페이지 데이터 모두 DB와 연동하여 구축**  **- 정확도 데이터 메인 및 종목별 DB 데이터 표출**  **- 단말기 별로 데이터 표출(윈도우, 안드로이드, iOS 일부) 해결**  **□ Tableau 시각화**  **- 분석 결과 시각화**  **- 웹 페이지와 연동** | **□ AWS에 Django 웹페이지 구현**  **- 종목별 워드 클라우드**  **- 정확도가 언론 데이터나 최근 데이터 불러오도록 코드 수정**  **- 웹 페이지 필요 기능 추가 구현**  **□ 코드 자동실행**  **- 코드 자동 실행 수정 및 정리** |

**2. 강사님 피드백**

| **2. 강사님 피드백** | |
| --- | --- |
| **데이터사이언스**  **이경미 강사님** | [9월 28일]  - 제목 긍부정 라벨링  - 서비스 구축 신속!  - 텍스트 긍부정 점수화(긍부정 비율을 앙상블 모델의 최종 변수로 제시)  [9월 30일]  - 분석 결과를 웹 페이지에 더 추가 |
| **데이터엔지니어링**  **김성환 강사님** | [10월 1일]  -웹 페이지 개발 관련 산출물 정리 필요 |