|  |
| --- |
| **해커톤2 개발계획서** |

***※ 본문 10페이지 내외(개요 및 Reference 제외)로 작성, ‘파란색 안내 문구’는 삭제하고 검정색 글씨로 작성하여 제출, 필요시 사진(이미지) 또는 표 추가 가능***

|  |  |  |  |
| --- | --- | --- | --- |
| **개발아이템명** | 뉴스와 공시에서 주가 영향 키워드 학습을 통한 종목의 주가 상승구간 예측 | | |
| **소속** | 양재 아이펠 양재 | | |
| **신청자 성명** | 박설현 | **담당퍼실** |  |

□ 아이템 개요(요약)

|  |  |
| --- | --- |
| **개발 목적** | 주식 투자 중 재료매매 기법이 있습니다. 주가에 영향을 줄 기사나 공시가 뜬 경우 짧은 시간 안에 주가 상승 혹은 하락 가능성을 판단하고 트레이딩을 해야 하는데 지금까지는 감에 의지하거나 증권사의 별점 키워드로 판단을 해 왔음.  개인이 기관을 이기기 위해서는 신속히 재료의 강도를 판단할 수 있는 시스템이 필요함. |
| **아이템 소개** | - 핵심기능은   1. 주식투자에 맞게 pre-train된 word embedding (stock-BERT) 2. 특정 뉴스나 공시 텍스트를 입력할 경우 주가의 상승, 하락율 예측 3. 특정 뉴스나 공시 재료가 주가에 미칠 영향의 기간   -1차 사용자층은 전업 투자자, 개인투자자이고, 2차 사용자는 증권사 투자팀  -개발은 웹과 하이브리트앱으로 진행할 예정 |
| **아이템의**  **차별성** | 기존 솔루션은 증권사의 키워드 강도자료임. 예를 들어 신약개발이라는 키워드가 기사 등에 나온 경우 주가반영 강도 2라는 정도의 서비스만 제공됨. 증권사에서는 -3 ~ +3까지의 강도 표시만 되어 있어서 주가가 얼마나 변동할지, 얼마 동안 영향을 받을지 등에 대해 알 수가 없음. 그래서 실제 투자에 약한 참조역할만 됨.  이번에 만들어서 제공할 시스템의 차이점은   1. 특정 기사나 공시의 대상이 되는 회사 주가의 예상 상승률, 하락률을 알 수 있고. 2. 특정 재료가 어느 기간 동안 주가에 영향을 미치는지를 알 수 있고 3. 주식에 특화된 word embedding을 통해 관련 키워드를 찾을 수 있음. |
| **이미지** | 아직 없음. 서비스 계획도는 작성 중 |
| **레퍼런스** | ☞ 참고 논문 (계속 Update 중)  - 개별 기업에 대한 네이버 검색량과 주식 수익률의 관계  - A Robust Predictive Model for Stock Price Prediction Using Deep Learning and Natural Language Processing  - NLP for Stock Market Prediction with Reddit Data  - Sentence-BERT: Sentence embeddings using Siamese BERT-networks.  - Stock Prices Prediction using the Title of Newspaper Articles with Korean Natural Language Processing |
| **데이터 수집** | 1차 데이터 수집 대상 기간 : 2022.01.01 – 03.31 (최근 3개월)  2차 데이터 수집 대상 기간: 2021.04.01 – 2022.03.31 (최근 1년)  - 주가 상승 10% 이상 상승 or 하락하는 회사의 당일 기사와 공시 텍스트 수집 (직접)  - 주가 상승 10% 이상 상승 or 하락하는 회사의 일별 주가, 거래량 수집 (직접)  - 수집 대상 기간의 코스피와 코스닥 주가, 거래량 일별 자료 (직접) |

|  |  |
| --- | --- |
| **1. 문제인식 (Problem)** |  |

**1-1. 개발 목적**

한국의 경우 개인과 기관투자자간의 정보량의 갭이 점점 커지고 있다. 기관투자자는 방대한

정보 네크워크와 인공지능을 통해 시장 대비 초과수익률을 높이고 있어서 돈을 가진 사람

점점 더 돈을 벌고 있다. 이에 개인이 기관을 이기기 위해서는 재료 투자에서 인공지능을 통

해 뉴스나 공시에서 주가에 영향을 주는 단어를 빨리 캐치하고 그 강도에 따라 투자할 수 있

는 시스템을 갖춘다면 기관에 필적하는 수익률을 올릴 수 있을 것이다.

**1-2. 관련 기술 동향**

NLP를 투자에 활용하기 시작한 것은 그리 오래되지 않았다. 지금까지 연구된 기술들을 보면

1. Reddit이라는 카페에 언급된 단어를 가지고 감성분석을 해 주가 예측
2. 주가에 영향을 주는 키워드를 뽑아 주가 강도를 부여

NLP가 아닌 경우 주가 예측은 시계열 분석, LSTM 등 시계열 자료를 다루는 모델을 사용하는 경우가 많았다.

**1-3 개발아이템의 독창성**

이 아이템의 특이한 것은 NLP와 Data2vec를 결합한 것이다. 주가를 예측할 때 정량 자료로만 예측하거나 정성 자료를 부가적으로 사용한 적은 있으나 본 개발과 같이 word embedding과 정량 정보를 같이 학습시켜서 주가를 예측한 연구결과는 지금까지는 없었다.

|  |  |
| --- | --- |
| **2. 개발 및 연구 내용** |  |

**2-1. 개발 내용 상세**

세부적으로 개발하고자 하는 것은

1단계: 일일 주가 10% 이상 상승 주식들의 관련 기사 및 공시를 Transformer와 BART를 통해

학습시켜서 word embedding하기 (2개의 NLP 모델 만들기)

2단계: 학습된 word embedding과 거래량 데이터를 Transformer에 넣어서 학습시키기

(target : 일일 주가 상승률(%))

3단계: 증권사에서 기업 및 공시 관련 기사 스크래핑 Bot 만들기

4단계 : 스크래핑한 기사 및 공시를 실시간으로 학습하는 Online 모델 만들기

5단계: 학습 모델에 따라 일일 예상 상승률 20% 이상인 주식에 자동 투자하는 증권사 trading 모델 만들기

**2-2. 개발 목표**

최종 목표는 뉴스나 공시를 활용해 실시간으로 주가 10% 이상 상승할 주식을 찾고 자동으로 투자해 일일 수익률 3% 이상되는 투자 알고리즘을 만드는 것

단계별로 개발 목표는

1단계: 주가 상승 10% 이상인 기업들의 기사 및 공시 word embedding

2단계: 주가 예측 모델

3단계: 뉴스 및 공시 scraping Bot

4단계: Online live 학습 모델

5단계: 자동 주식 Trading 알고리즘

|  |  |
| --- | --- |
| **3. 실현가능성 (Solution)** |  |

**3-1. 개발 전략**

개발 전략은 자체 개발과 외주의 병행

* 자체 개발 : word embedding과 학습 모델, 기사 Scraping
* 외주 : 개발된 학습 모델과 자동 트레이딩 연결

**3-2. 해커톤 수행 일정**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **주**  **내용** | | 5월  1주 | 2주 | 3주 | 4주 | 6월 1주 |
| **1단계** | 데이터 수집 |  |  |  |  |  |
| 세부 계획 수립 |  |  |  |  |  |
| 기본 모델 만들기 |  |  |  |  |  |
| **2단계** | 학습 모델 만들기 |  |  |  |  |  |
| 거래량 모델 만들기 |  |  |  |  |  |
| **3단계** | 스크래핑 Bot 제작 |  |  |  |  |  |
| 4단계 | 온라인 Live 모델 |  |  |  |  |  |
| 5단계 | 자동 Trading 알고리즘 |  |  |  |  |  |

|  |  |
| --- | --- |
| **4. 팀 구성 (Team)** |  |

**4-1. 팀장 및 팀원의 역할 분배**

◦ **팀장** 프로필 **및 역량**

* 투자 전문가 : CFA 자격증 및 미국 회계사 시험 all pass
* 주식 투자 경력 10년

|  |  |
| --- | --- |
| **5. Reference** |  |

**■ 논문**

- 개별 기업에 대한 네이버 검색량과 주식 수익률의 관계

- A Robust Predictive Model for Stock Price Prediction Using Deep Learning and Natural Language Processing

- NLP for Stock Market Prediction with Reddit Data

- Sentence-BERT: Sentence embeddings using Siamese BERT-networks.

- Stock Prices Prediction using the Title of Newspaper Articles with Korean Natural