데이터마이닝 프로젝트 제안서

1. 프로젝트 개요 및 주제

1.1 프로젝트 개요

- 개요 : MLLAB 연구실에서 함께 작성 중인 주식 추천 논문 바탕으로 프로젝트 진행
- 프로젝트 목표 : 논문 투고(CIKM24 or AAAI25 or ICDM)
- 참여자 : 정가연

1.2 주제 정의

- 주제 : High-Order Mean-variance Neural Utility Function for Stock Recommendation
- 주제 선정 배경 : 기존의 주식 추천 관련 연구에서는 주가 예측 모델 기반으로 기대 수익률이 높은 주식을 추천해주는 식으로 진행되어 왔음. 이러한 접근법에서는 주식의 변동성이 반영하지 못했다는 한계가 존재함. 이를 해결하고자 주식의 기대 수익률과 변동성을 함께 고려한 효용 함수를 활용해서 투자자 선호도 기반의 주식 추천 모델을 제안하고자 함.

1.3 연구 문제 정의

- 목표 : Markowitz의 The Modern Protfolio Theory에서 제시한 mean-variance utility function을 사용하여 return-risk의 공간에서 investor의 preference를 바탕으로 주식을 추천
- 세부 목표 1 : Hypergraph Neural Network(HNN)의 Laplacian matrix를 사용해서 mean-variance utility function을 approximation함, 이를 활용해서 개별 주식에 대한 utility을 구하고 utility 기반 ranking을 진행.
    - Laplacian matrix의 semi-positive/negative 성질을 활용해서 quasi-convex/concave한 objective fuction 디자인
- 세부 목표 2 : N개 주식에 대한 다음 거래일의 return과 risk를 예측하는 HNN-based prediction model 개발
    - 현 시점에서 불확실성이 존재하는 미래 주가를 고려하기 위해서 expected return과 expected risk 두 값을 예측하는 모델 활용
- 세부 목표 3 : 개별 주식에 대한 text data를 기반으로 text-encoded vector를 만들어서 주가 예측에 활용.
    - S&P500 Stock에 대한 tweet data와 Chinese Stock에 대한 News data 활용

---

2. 데이터 소개

2.1 데이터 출처

- S&P500 Stock Data : Yahoo Finance (yfinance library)
- Chinese Stock Data : Yahoo Finance (yfinance library)
- S&P500 Tweet Data : Twitter
- Chinese News Data : Wind-Financial Terminal, sina(중국 온라인 뉴스 사이트), hexun(중국 온라인 뉴스 사이트)

2.2 데이터 명세

- S&P 500 Stock Data
  - 미국 상위 88개 S&P500 주식
  - 기간 : 2014.01.01~2015.12.31
  - Conglomerates 산업군의 전종목(8 종목)
  - 그 외 산업군(Basic Materials, Consumer Goods, Healthcare, Services, Utilities, Financial, Industrial Goods, Technology)에서 자본금 규모 상위 10개 종목(8개 산업군 * 상위 10종목 = 80 종목)

- Chinese Stock Data
  - Shanghai, Shezhen, Hong Kong 주식 시장에서 가장 많이 거래되는 80개 중국 A주 주식
  - 기간 : 2014.01.01~2015.12.31

- S&P500 Tweet Data
  - 미국 상위 88개 S&P500 주식과 관련된 트위터 텍스트 데이터
  - 기간 : 2014.01.01~2015.12.31
  - 개별 주식에 대한 종목별 일자별 트윗
  - 각 일자에 대한 트윗들의 단어들이 value값으로 저장되어 있음

- Chinese New Data
  - 중국 상위 80개 A주 주식과 관련된 90,361개의 Chinese의 금융 뉴스 헤드라인
  - 기간 : 2014.01.01~2015.12.31
  - 각 종목(ticker) 별로 2년 간의 뉴스 게시일, 뉴스 헤드라인이 기록되어 있음

---

3. 진행 계획

- 선행 논문 정리→완료
- 데이터 수집→완료
- 선행 논문 코드 구현 및 성능 확인→진행 중
- 모델 아키텍처 구체화 및 개발→진행 예정
- 실험→진행 예정
- 논문 작성→진행 중