# 멀티캠퍼스 금융/마케팅 데이터 분석가 과정

: 정형/비정형 데이터를 이용한 연방준비은행 기준금리 예측 모델

전체 기간: 2024. 4. ~ 10.
프로젝트 기간: 2024. 7. ~ 10.

프로젝트 팀원 수: 4명

***Why choose this topic?*** *금융 topic 안에서 팀원들의 공통된 관심사였기 때문*

**정형 데이터**
- 기준금리
1. 한국은행
2. ECB(유럽중앙은행)
- 삼성전자 주식데이터 (일봉)
- 미국 소비자 물가 지수

**비정형 데이터 (긍/부정 -> 1/0)**
- 커뮤니티
1. 네이버 삼성전자/sk하이닉스 종목토론방
2. “DC인사이드” 삼성전자 토론방
3. “Reddit” 주식 토론방
- 주식 애널리스트 보고서
- Beige Book

**데이터 수집 기간**
2014-2024(10년)


**연준 기준 금리**

## 최종 목표?

1. 실시간 데이터를 통한 연준 기준금리 예측 모델
2. 연준 기준금리 예측에 관련성이 높았던 비정형 데이터를 중심으로,
   연준 기준금리 전망에 대한 요약 보고서 모델

# Content

# Crawling

## - crawling data

비정형 데이터(커뮤니티)

1. 네이버 삼성전자/sk하이닉스 종목토론방
2. DC인사이드 삼성전자 토론방
3. “Reddit” 주식 토론방

## - crawling의 이유

영향력

1&2. 높은 국내 시가 총액 및 커뮤니티 활성화 정도
3. 국외 가장 활성화 수 많은 커뮤니티

## - 문제점

네이버 삼성전자 종목 토론방 데이터 100만개,
100만개 crawling 예상 시간 2일.
crawling 오류가 발생했고,
100만개->50만개->30만개 조정하면서,
예상치 못한 3주간의 시간 허비.

# 텍스트 데이터 긍/부정 정확도 측정

## - 측정 목적

연준 기준금리 예측 정확도를 높이기 위해서

## - 텍스트 데이터 라벨링 모델(긍/부정 -> 1/0): ChatGPT 4o

## - 측정

측정 방법: 긍/부정 정확도 비율(단순 수학적 매칭 비교)
측정 모델: VADER, TextBlob, Hugging Face, Bert(KoBert 등), **ChatGPT 4o(채택)**

## - 문제점

ChatGPT를 포함한 모델들의 복합적인 맥락 이해 X(반어법, 긍/부정 섞임)
ex1) “하 참 좋네 ;;” - 1(긍정)
ex2) “그래 갈때까지 가자, 가보는 거야, 버틴다” - ?
-> ChatGPT 프롬프트 수정(좀 더 복합적인 이해 포함)
-> 효과 미미

**ChatGPT prompt**
“주식 애널리스트 관점에서, 해당 텍스트가 긍정은 1, 부정은 0로 대응하시오.”

## Process

1. 정형/비정형 데이터(비정형데이터는 정형데이터로 가공)와 연준 기준금리와의 상관성 확인.
2. 그 중 상관성이 높은 데이터(Beige Book, ECB 기준 금리 등)를  통해 연준 기준 금리 예측 예정.
   - 예측 선택지: 실제 연준 기준금리 지표 or 상승/하락/동결
3. 상관성이 높은 데이터들의 API를 받아, 실시간 연준 기준금리 예측 모델 구현 구상.

## Insight

ChatGPT를 제외한 감정분석 모델들은 한국어 텍스트 감정분석 정확도가 떨어진다고 느낌(2024 기준)
ex) 그래그래 올라가자 – 0(부정)

## Review

- 실패한 이유
Bottom-up 방법으로 초기 목표를 연준 기준금리를 예측하는 어떤 모델로 추상적으로 설정하고,
데이터와 process를 지켜보면서 점점 목표를 구체화하는 방식으로 계획했지만,
예상치 못했던 작업(crawling)에서 시간이 흐르고 실제 모델 구현 방법에 대한 준비와 합의가 이루어지지 못했음.
- 회고
프로젝트의 완성은 실패하였지만 충분히 완성난이도가 높은 프로젝트였기에, 실패를 경험으로 여길 것. 멘토님께 도움을 적극적으로 요청하였다면, 진행 속도가 더 빨랐을 것 같음.

**python 문법, crawling – 1개월**

**Machine Learning – 2개월**

**Deep Learning – 1개월**
RNN/CNN, LSTM, transformer