텍스트 마이닝을 활용한 시장지표 예측 모델링

Index

1. 관련 논문

2. 절차
   - 데이터 수집
   - 데이터 클렌징
   - 데이터 라벨링
   - 통계 & 딥러닝 모델링

3. 통계기반
   - 전처리
   - Feature Selection
   - NBC
   - Predict

4. 딥러닝기반
   - 전처리
   - Modeling
   - Train
   - Predict

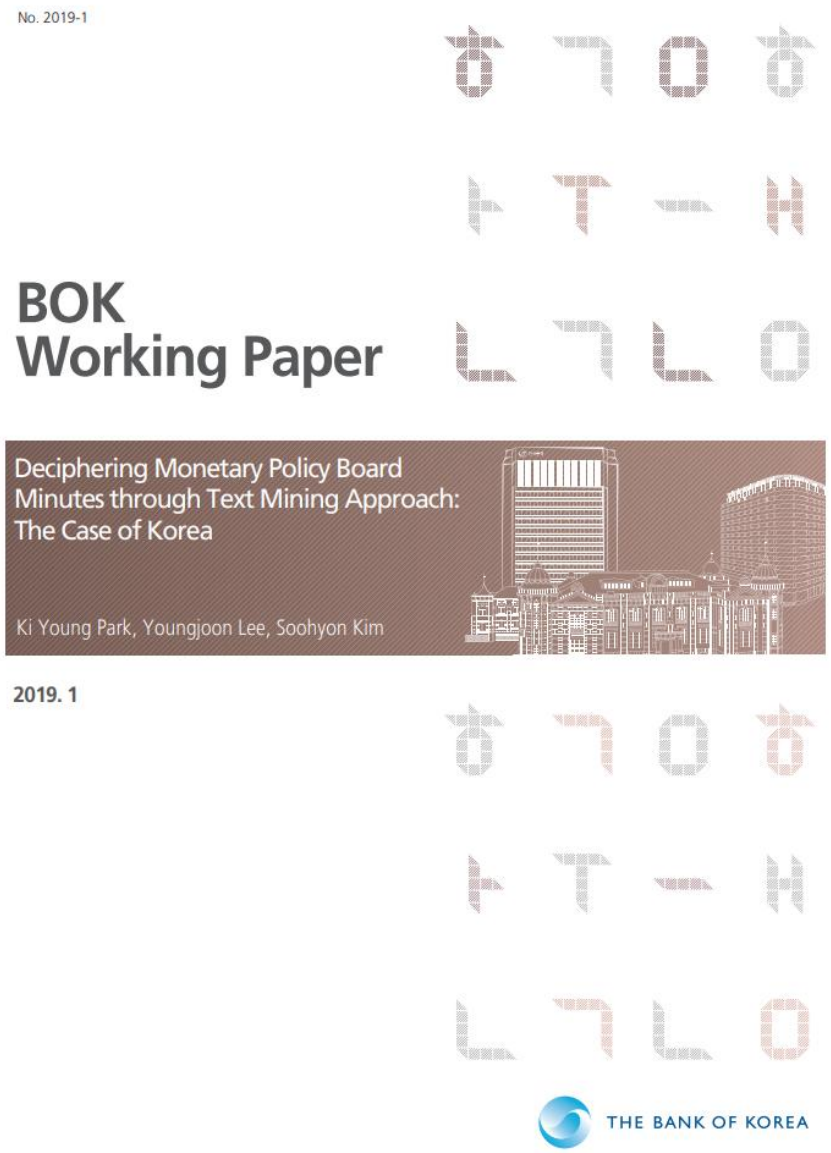
Article

관련 논문

텍스트 마이닝을 활용한 한국은행 기준금리 예측

- 'Deciphering Monetary Policy Board Minutes through Text Mining Approach: The Case of Korea'

- 요약본



Process

절차

데이터 수집

- 기준금리
- 콜금리

- 한국은행 의사록
- 채권보고서
- 뉴스기사

- 권장 수집량 : 뉴스기사 20만 건 이상
- 기술 스택 : Scrapy, 멀티스레딩

데이터 클렌징

- 수집된 데이터에서 분석에 필요 없는 부분을 제거하는 과정
- 수집된 데이터를 눈으로 꼭 확인하고 클렌징 코드를 작성할 것
  (뉴스 언론사별로 다양한 형태의 불필요한 내용이 있을 수 있음)
- 정규표현식 혹은, replace를 활용

뉴욕증권거래소의 트레이더

[AFP 연합뉴스 자료사진. 재판매 및 DB 금지]

삭제

(서울=연합뉴스) 임상수 기자 = 최근 금융시장 혼란 이후 시장에서는 미 연방준비제도(Fed·연준)가 공격적인 금리인하로 경기침체 우려를 잠재우기를 원하고 있다.

데이터 라벨링

- 학습데이터를 구성하는 단계로 목적에 따른 label 전략을 수립
- 목적 : 하루 뒤의 금리가 어떻게 변하는지, 한 달 뒤의 금리가 어떻게 변하는지, 반 년 뒤의 금리가 어떻게 변하는지

기준 금리 변동 분석 (Sentiment Analysis)

- 통계 기반 : 시장 접근법 (NBC)
- ~~사전 접근법 (SO-PMI)~~
- 딥러닝 접근법 : 작은 딥러닝 자연어처리 모델 활용
  - KoBERT, KoELECTRA 등...
- LLM AI : Gemini, GPT API를 활용

Statistical

통계기반 모델링

전처리

- 문장 분리

- 통계기반 텍스트 전처리
- eKoNLPy

Feature Selection

- N-gram

성능 최적화

- 로직 직접 구현
- 파이썬 내장함수 활용
- 멀티프로세싱 활용

모델링

- NBC

예측 & 평가

- 기준금리의 상승, 하락을 맞췄는지
- 앞으로 기준금리가 어떻게 될지 예상

AI

LLM 모델링

프롬프트 작성

- AI가 효과적으로 문서를 분석하여 도출할 수 있도록 분석방향 및 기준, 제약조건, Output format을 정리하여 프롬프트를 작성

Parsing

- Json 포맷, 텍스트 포맷, CSV포맷 등 LLM이 출력으로 반환한 정보를 파이썬 객체로 변환하여 **활용**할 수 있도록 Parsing

예측 & 평가

- 기준금리의 상승, 하락을 맞췄는지
- 앞으로 기준금리가 어떻게 될지 예상

산출물

- Notion (프로젝트 기획, 연구내용 정리)
- Github (코드 관리)
- Google Drive (데이터, 파일 관리)

- 위에서 작성한 내용을 토대로 PPT에 정리