NLP Advanced Assignment

해당 연구 계획서는 한국은행이 발간한 2023-5호 bok이슈노트의 주제를 인용하였으며, 해당 주제를 재현하는 것에 초점을 두었음을 밝힙니다.

1. 연구 주제

증권사 리포트 텍스트 분석을 이용한 산업 모니터링

정보를 주고받는 가장 기본적인 수단이 사람의 언어라는 점에서, 언어분석 기술은 경제 분야에서 활용가치가 매우 높다. 많은 사람의 언어를 종합해서 이해할 수 있다면 경기 방향 등 거시적 예측에 큰 도움이 될 것이다.

특히 증권사 애널리스트 리포트에 나타나는 텍스트 데이터는 가공하는 방식에 따라 다양한 미시적, 거시적 연구가 가능하다. 또한 텍스트 데이터는 발간일 기준으로 취합이 가능하기 때문에 여타 공식 통계보다 신속하게 분석할 수 있는 장점이 있다. 또한 텍스트 데이터는 수치화하기 힘든 여러 주제에 대한 전문가들의 견해를 취합하여 나타낼 수 있다. 이러한 점에서 텍스트를 바탕으로 만들어진 지표들은 기존의 숫자가 반영하지 못하는 새로운 정보를 정량화하여 제공할 여지가 크다.

2. 연구 방법

1) 사용 데이터

기업분석 전문가인 증권사 애널리스트들의 기업평가 보고서를 빅데이터로 구축하여 진행한다. 보고서의 내용 중 수치로 나타나는 지표는 모두 제외하고 오로지 텍스트에 나타나는 정성적(qualitative) 정보만을 이용하여 애널리스트들의 생각을 취합한다.

2) 연구 방법

자연어처리(natural language processing)기술과 데이터 마이닝 기법 등 다양한 통계 기법을 사용한다.

구체적으로 웹 스크래핑 기술을 이용하여 증권사 기업 평가 보고서를 수집한다. 애널리스트 보고서 수집시 산업 및 거시 분석 보고서는 제외하며, 개별 기업 분석 보고서만 수집한다. 텍스트 데이터의 논조 파악을 위해 트랜스포머 기반의 감성분석 모형을 활용하고, 키워드 빈도 분석, 동적인자모형(dynamic factor model) 기반의 시계열 분석, 네트워크 데이터 분석 등을 활용하여 텍스트 데이터로부터 유의미한 경제적 정보를 추출할 수 있다.