금융 시계열데이터 분석

팀 원 김영훈 지도교수 이슬

개발 동기 및 목적

주식차트에 대해 과거의 데이터를 토대로 미래를 예측한다는 것은 주식을 하는 사람들에게는 언제나 큰 관심사이다. 주식 시장에서 사용할 수 있는 투자분석 기법은 크게 두 가지로 구별되는데, 기업의 내재적 가치를 평가하여 투자하는 기본적분석과 오직 주식의 변동가격 움직임 자체만을 연구하는 기술적 분석이 있다. 본 연구에서는 주식의 종가 데이터를 이용하여 기술적 분석을 통해 해당 주식의 과거 주가 움직임 중 특이점을 찾아내어 그 특이점(Anomaly)을 기반으로 분석함으로써 과거를 분석하고, 현재를 진단하여 미래에 대한 대비에 목적을 가진다. 기계학습과 딥러닝으로 비슷한 패턴의 차트 움직임을 예상하는 것이 아닌, 뉴스/금융 정보를 활용해 해당 주식이 어떠한 키워드에 민감하게 반응하는지를 파악하는 것에 목적을 둔다.

개발내용

1. Data Crawling & Data Preparation

Matrix Profiling에 필요한 시계열 데이터를 수집하기위해 Python 패키지 중 하나인 Selenium을 이용해 Yahoo finance에서 Raw 데이터를 수집한다. 원하는 주식의 이름을 검색하면 자동적으로 해당 주식의 code와 featur를 가진 timestamp 데이터를 수집한 후 Pandas DataFrame을 이용해 수집한 데이터를 기간 이동평균계산(rolling)을 진행한 후 평균(mean)값을 내어 원하는 기간(window)의 주가평균을 구하여 종가 데이터들을 시계열 데이터로 만들어주는 데이터 전처리 작업을 진행하였다.

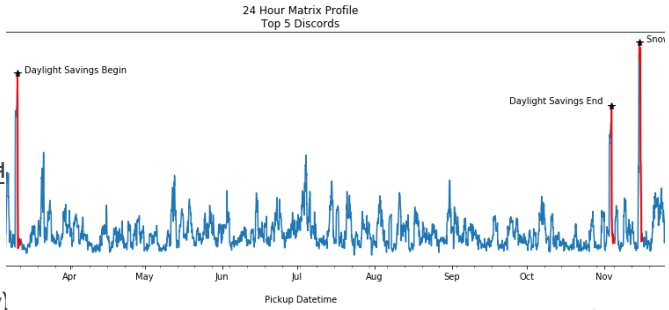
2. Matrix Profile

구한 시계열 데이터 T의 부분열을 m만큼의 윈도우(window)를 움직여(sliding) 얻어낸 모든 가능한 연속물의 순서집합과 모든 부분열 집합에 존재하는 랜덤 부분열 D와 유클리드 거리(Euclidean Distance)을 얻어내어 모든 부분열에 대해서 짝간 유클리드 거리를 얻어 거리행렬을 만들어낸다. 이러한 거리행렬을 토대로 모티프(적어도 두번 이상 비슷한 패턴으로 나타남)과 디스코드(가장 거리가 큰 값 = 유사한 부분열이 없거나 매우 다름)를 얻어 이상점 탐지(anomaly)를 진행한다.

3. TF-IDF

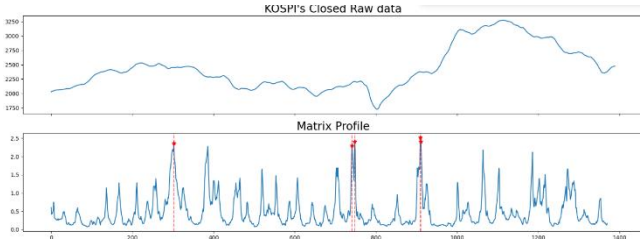
Matrix Profile 을 통해 이상점을 감지하여 감지된 시기에서의 뉴스/금융 정보를 스크래핑 하여 TF-IDF분석을 통해 궁극적으로 원하는 주식의 이상에 대한 분석을 진행한다.

주요기술



Matrix Profile이란 주어진 쿼리에서 가장 가까운 이웃을 찾는 방법인 전체 유사성 검색을 거짓 음성 없이 찾는 방법으로, 이를 활용해 금융 시계열 데이터에 적용하여 지난 과거로부터 가장 상이한 패턴(Anomaly)인 디스코드(Discord)를 얻을 수 있다. 이때 디스코드가 생긴 시점을 수집하여, 해당 시점에서의 뉴스와 금융레포트 내용들을 텍스트 마이닝하여 TF-IDF 분석을 한다. TF-IDF란, 특정 문서 내에서 어떤 단어가 얼마나 중요한지를 나타내는 통계적 수치를 말하는데, 이러한 기술을 이용하여 이상탐지가 발생된 시기의 금융데이터를 분석하여 과거 주식시장을 해석하는데 도움을 줄 것이다.

결과 및 분석



위쪽 그래프는 데이터 전처리를 통한 종가 시계열 데이터를 Pyplot을 통해 시각화한 모습이고, 아래쪽 그래프는 KOSPI를 검색하였을 때, Matrix Profile을 통한 Anomaly를 발견한 그래프이다. 실제로 해당 Anomaly의 datetime을 얻어 뉴스 데이터를 크롤링 한 후 TF-IDF를 통해 의미한 데이터들을 얻어내었다. 실제로 KOSPI의 Top 디스코드의 경우엔 미중 무역전쟁이라는 키워드를 얻었고, 그 다음 디스코드에 대해선 코로나19라는 키워드를 얻을 수 있었다.

하지만, 공매도나 내부적/외적/소비자심리지수 등 여러 지수를 고려하였을 때, 모든 주식에 있어 유의미한 결과를 얻을 수는 없었다.