안녕하세요, 저희는 기업 클러스터링이라는 주제로 프로젝트를 진행하였습니다.

저희는 기업 클러스터링을 위해서, 크게 기업 개요 데이터와 주가 변동 데이터의 2가지 형식의 데이터를 수집하였습니다. 먼저, 기업 개요 데이터 검토 결과를 설명 드리겠습니다.

기업 개요 데이터는 FnGuide에 나와있는 기업정보를 크롤링하여 KOSPI 200 기업들의 기업 개요 텍스트 파일을 얻을 수 있었습니다. 이후, 분석을 위해서 국문 전처리 과정을 거쳐야 했는데요.

간단하게 실제 예시를 통해서 설명 드리면

“2021년 세계 반도체시장 점유율은 D램 28.1%, 낸드플래시 12.7%임”이라는 문장이 있을 때,

첫 번째로, ‘2021년’, ’28.1%’, ’12.7%임’ 등의 숫자가 들어간 단어를 제거한 뒤,

두 번째로, 간이 맞춤법 검사기를 만들어 ‘DRAM’이나 ‘D램’과 같이 여러 형태로 나타날 수 있는 단어들을 하나로 통일시켜 주었습니다.

세 번째로, 낸드플래시와 같이 분석기가 모르는 단어들을 사용자 사전으로 지정해준 뒤에,

마지막 단계로 형태소 분석기 khaiii를 사용해 명사와 영어 단어만을 추출하였습니다.

이렇게 전처리를 한 결과를 WordCloud를 통해 시각화한 결과, 다음 그림과 같이 나타났습니다.

사업, 시장과 같이 많은 기업 개요에서 나타나는 단어와 함께, 자동차, 화학과 같은 특정 분야의 기업들을 나타내는 단어들도 같이 볼 수 있었습니다.

그러나 FnGuide에서 가져온 기업 개요 텍스트 파일만으로는 그 양이 부족하여 이후의 분석에서 어려움이 있다고 판단하였고, dart에서 기업보고서를 추가로 크롤링하여 진행하기로 하였습니다.

다음은 주가 변동 데이터 검토 결과입니다.

2020년 10월부터 2021년 10월까지의 기업 별 주가의 데이터는 좌측 상단 table과 같았으며 이를 시각화한 그림은 좌측 하단 그림입니다.

등락률을 비교하기 위해 주가 데이터를 기반으로, 등락률을 계산하고 시각화 하니 오히려 기업별 주가 변화 특성을 잘 반영하지 못한다는 문제가 있었습니다.

따라서 주가 데이터를 이용하기로 결정했는데, 기업별로 주가의 scale이 다르기 때문에 그대로 거리를 계산한다면 부정확한 클러스터링이 이루어질 수 있다고 판단해 정규화를 거쳤습니다.

각 날짜 주가에서 평균 주가를 빼고 표준편차로 나누어 정규화 후 시각화한 그림은 우측 상단의 그림과 같습니다. 빨간색 그래프는 삼성전자와 SK하이닉스, 파란색 그래프는 카카오와 네이버의 정규화한 주가입니다. 비슷한 종목은 비슷한 변화 추이를 보였습니다.

이후 시계열 데이터에 적합한 거리 계산 방법이 필요하여 DTW를 통해 새로운 거리 matrix를 만들었고 그 결과는 우측 하단과 같습니다.

다음으로 모델 학습에 관해 설명드리겠습니다.

먼저 기업개요 데이터에 대해 벡터화를 진행합니다.

카운트 벡터라이저에 스탑워즈, 엔그램과 민 맥스 디에프 값을 설정하고, 이 결과를 티에프 아이디에프 벡터로 변형시켰습니다.

클러스터링 알고리즘으로 kmeans 를 사용할 것이기 때문에, 적정 군집의 개수를 찾을 필요가 있었습니다. 그래서 Elbow method, 실루엣, 그리고 계층적 클러스터링을 적용시켜 보았는데, 주가 변동 데이터를 정규화 시켰을 때는 Elbow point가 명확했고, 5개에서 7개 정도의 군집 개수가 적절한 것으로 보였습니다.

그러나, 기업개요 데이터에 대해서는 point를 찾기 어려울 정도로 일정하게 떨어지는 그래프가 나왔고, 실루엣 값도 너무 작은 값이 나오는 것을 알 수 있었습니다. 이점을 Dart 에서 얻은 데이터로 보완할 계획입니다.

계층적 클러스터링으로도 개체간 거리값에 따라 선을 그어 군집의 개수를 설정해보려했습니다. 다만, 개체간의 거리를 재는 method에 따라 그 결과가 매우 상이했기 때문에, 과연 어떤 method 가 적절할지에 대한 논의가 더 필요해 보였습니다.

Kmeans 와 같은 분할적 클러스터링 이외에도, 밀도기반의, 군집개수를 정해주지 않아도 되는 알고리즘들을 사용해보았지만, 왼쪽 아래에서 보시는 것과 같이 너무 많은 개체를 이상치로 분류하거나, 기업 대부분을 하나의 클러스터로 묶고 나머지 군집이 개체 하나만으로 이루어지는 결과가 나와서 밀도기반 보다는 분할적 클러스터링으로 진행하기로 정했습니다.

(넘겨주세요)

각각 kmeans 와 time series kmeans 로 군집화를 진행한 결과를 PCA로 차원을 축소시켜 나타낸 것이 왼쪽 그림입니다. 차원이 축소되어 데이터 손실이 있었다는 것을 감안해도, 위쪽 기업개요 군집화 결과를 보시면 개체들간의 거리가 매우 가까워서 클러스터링이 잘 되지 않고 있음을 알 수 있습니다.

앞으로는, 재무지표 데이터의 결측값을 처리하고 나면, 기업개요, 주가변동, 재무지표 총 3번의 클러스터링 결과에서 클러스터 중심간의 거리 Matrix를 도출하고, 그 결과값에 weight를 반영하여 새로운 기업간 거리 Matrix를 만들어 최종 클러스터링을 진행할 계획입니다. 그 과정에서 데이터의 특성에 맞는 거리 산출 방법과 클러스터링 알고리즘 선정할 것이고, 특히 기업개요 데이터는 데이터 자체가 추가되면서 기존의 형태소 사전에 보완이 필요할 것으로 보입니다.

이상으로 발표를 마치겠습니다.