**리스크 매니저**

빽투처퓨처

유근태, 윤희재, 조형주

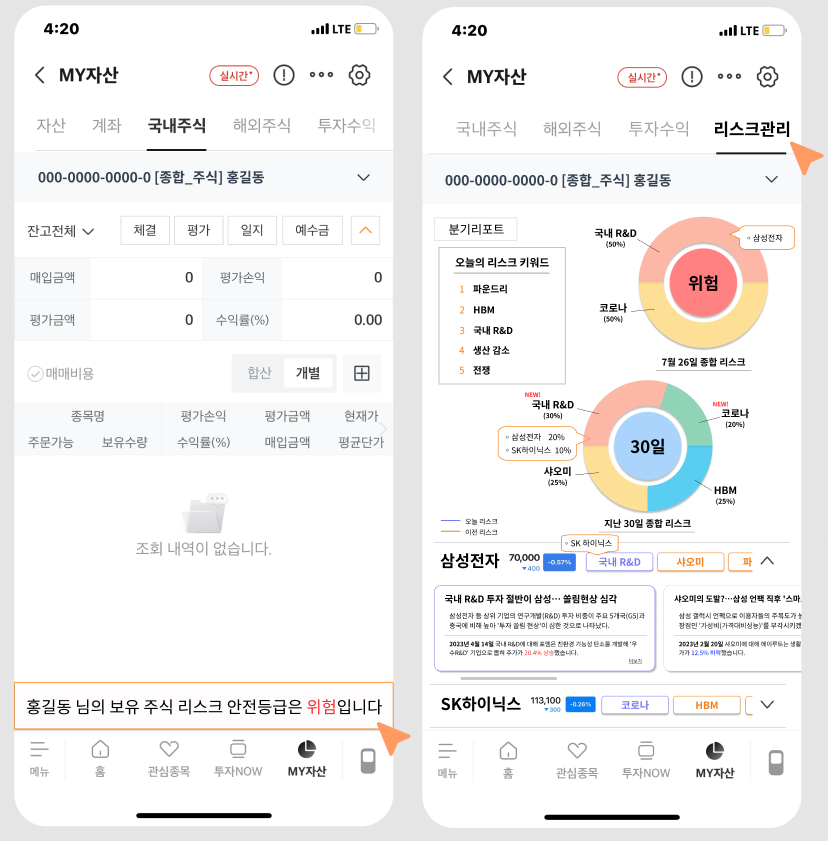
1. 기획 및 분석 배경

다양한 요인들이 복합적으로 상호작용하여 변동하는 주식시장을 개인 투자자가 예측하는 것은 쉽지 않은 일이다. 시장의 규모가 커지고 기업 관계가 복잡해짐에 따라 필요한 정보의 선별이 어려워졌고 리스크가 있는 종목임에도 불구하고 이를 파악하지 못해 손실을 보는 고객들도 많이 존재한다. 그래서 최근에는 분산투자를 통해 보유한 금융 자산의 하락폭을 최소화할 수 있게 하는 리스크 관리의 중요성이 대두되고 있다. 이에 맞추어 증권사들은 전문가들과 발전된 AI 기술을 통해서 다양한 자산 관리 서비스를 제공 중이다. 하지만 리스크만을 관리해주는 서비스는 아직까지는 쉽게 찾아볼 수 없다. 또한 정작 이러한 서비스들은 보유한 금융 자산이 상대적으로 높은 소수의 개인 투자자에게만 제공되고 있으며 아직까지 보통의 개인 투자자들에게는 접근성이 떨어지는 것이 사실이다. 이러한 개인 투자자들을 위해 우리는 뉴스 기사 분석을 통해 보유 종목에 대한 리스크를 키워드 형태로 제공하여 보유 종목들의 리스크를 한눈에 파악할 수 있는 대국민 리스크 관리 서비스를 제공하고자 한다. 우리 서비스의 목적은 고객들이 주식을 거래할 때 더 현명하게 대응하고 주식 관리에 대한 이해도를 높여 능동적인 투자 습관을 가질 수 있도록 하는 것이다.

2. 서비스 설명

현재 미래에셋증권 M STOCK 앱을 포함한 여러 증권사 앱에서 따로 일반 개인 투자자를 위한 리스크 관리만을 위한 서비스는 찾기 어렵다. 미래에셋증권 웹사이트에는 시장 읽어주는 AI, 종목 읽어주는 AI 서비스가 있지만 웹사이트에 들어가서 일일이 찾아봐야 해서 접근성이 떨어진다. 따라서 해당 서비스는 자신의 리스크에 대한 정보를 쉽게 접할 수 있는 개인화 서비스이다.

1. 설명



우리 서비스는 미래에셋증권 앱인 M STOCK에 탑재할 예정이다.

M STOCK 앱의 MY 자산 탭에서 ‘국내 주식’을 누르면 <그림 1>처럼 하단 바에 ‘홍길동 님의 보유 주식 리스크 안전등급은 안전/위험입니다’라는 문구를 고정시켜 사용자가 국내 주식에 대한 리스크 정도를 한 눈에 파악할 수 있도록 한다. 오늘 리스크 키워드가 없을 경우 ‘안전’, 하나라도 존재할 경우 ‘위험’으로 표시한다.

또한 <그림 2>처럼 M STOCK 앱에서 새롭게 리스크 관리 탭을 추가했다.

리스크 관리 탭은 장이 열리는 오전 9시 이전인 오전 7시에 매일 업데이트되며 고객이 미리 뉴스 기사를 기반으로 자신의 보유 주식에 대한 리스크를 파악하여 장이 열리면 바로 대응할 수 있도록 돕는다. 따라서 모든 종목에 대한 뉴스 기사 크롤링을 자동화하여 오전 7시에 리스크관리 서비스를 제공한다.

탭의 왼쪽 상단에 오늘의 리스크 키워드가 있는데 이는 자신이 보유하는 종목 외에도 전체적인 리스크 키워드 TOP5를 보여줌으로써 고객들이 다른 종목에 투자할 때 이 리스크에 주의할 수 있도록 한다. 오른쪽 상단에는 두 개의 도넛형 그래프가 있는데 위쪽에는 오늘의 종합 리스크를, 아래쪽에는 지난 30일의 종합 리스크를 제공한다. 이렇게 두 가지를 제공하는 이유는 리스크가 꼭 당일에만 영향을 미치는 것이 아니라 지속적으로 영향력이 있을 수 있고 며칠 간의 시간 차를 두고 서로 다른 종목에서 같은 리스크가 발생할 수 있기 때문이다. 또한 투자자의 보유 종목들이 리스크가 없는 날에는 추가적인 정보 제공이 없기 때문에 그럴 경우 최근 30일 간 발생한 리스크를 확인하여 잠재적인 리스크를 파악할 수 있게 해준다.

두 도넛형 그래프에서 고객의 보유 종목들의 리스크 키워드를 1) 나의 자산 중 해당 종목의 비율 2) 언급량 3) 기사의 부정도 세 가지 기준을 바탕으로 비율을 책정하여 시각화하고 각 키워드에 마우스 커서를 올리면 해당 리스크 키워드를 보유하는 종목들과 각 종목이 차지하는 비율을 확인할 수 있다. 그래프로 한 눈에 시각화하여 제공해줌으로써 투자자는 보유한 종목들이 동일한 리스크를 공유하고 있을 경우 다른 곳에 분산 투자를 하는 것을 고려해볼 수 있으며 자신이 대응해야 하는 리스크들의 우선순위도 파악이 가능하다.

도넛형 그래프 비율 책정 기준은 다음과 같다.

애플에 대한 키워드가 A, B라 가정하고 LG전자에 대한 키워드가 C, D라고 가정하자. 애플은 30만원, LG 전자는 50만원 가지고 있다.

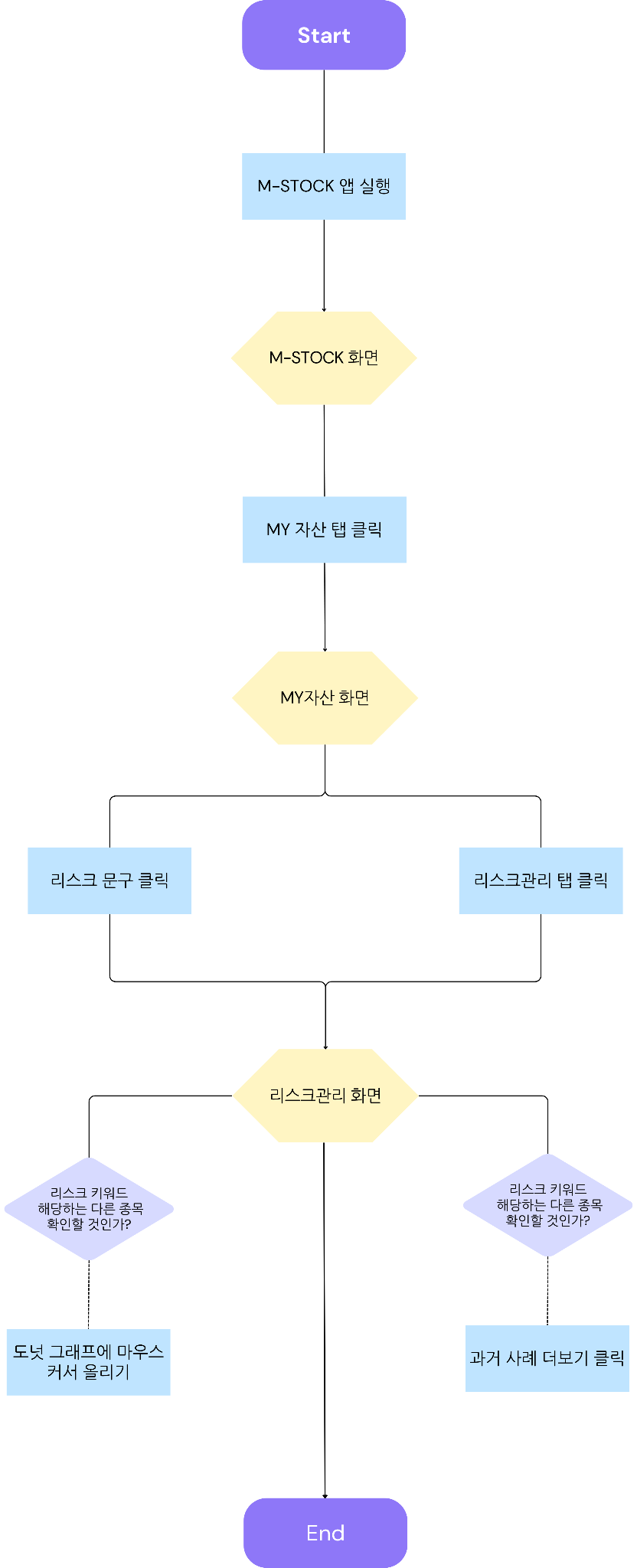
1. 나의 자산 중 해당 종목의 비율 (해당 종목의 자산 / 본인의 전체 자산 , 0 ~ 1)
2. 언급량 (해당 리스크 기사 수 / 전체 기사 수 , 0 ~1)
3. 기사의 부정도의 평균 (0 ~ 1) (해당하는 리스크 기사의 부정도의 합 / 전체 기사)

| 0.6 \* 나의 자산 중 해당 종목의 비율 + (0.4 \* 언급량 \* 기사의 부정도 평균) |
| --- |

모든 키워드에 대해 이 수식을 적용해 값을 구한 다음 값들 간의 비율을 구한 것이 도넛형 그래프의 비율이다. 30일 도넛형 그래프의 경우 새로운 키워드가 오늘 들어올 경우 다시 모든 키워드에 대해 위 수식을 적용해 비율을 구한다. 종목을 매도할 경우 해당 종목에 있던 키워드들은 리스크 관리 탭에서 삭제되고 새로 종목을 매수했을 때는 종목의 최근 30일 간 키워드들이 새로 반영된다.

맨 하단에는 종목별로 리스크 키워드와 그에 해당하는 뉴스 기사 정보 및 해당 리스크가 발생했던 과거 사례의 한 줄 요약과 그 당시 기업의 주가 등락률 정보를 제공한다. 이 때 키워드는 금일 기준으로 30일 전까지의 키워드를 누적으로 보여주고 오늘 등장한 리스크 키워드는 보라색으로, 이전 키워드는 주황색으로 표시하여 구분한다. 특히 해당 키워드에 대한 과거 사례는 실제로 등락률이 어떻게 변했었는지를 보여주며 고객이 현재 주식을 추가매수할지 매도할지 판단할 때 도움을 주는 지표로 작용한다.

1. 서비스 Process Flow Chart



추가해야 할 것들: END, 각 기호가 뭐에 해당되는지 설명하기 (<https://blog.naver.com/ycpiglet/222113989214> 여기에 해당하게끔 양식 맞게 변경하기)

M STOCK 앱에서 MY 자산 탭을 누른 뒤 상위 탭에서 ‘국내 주식’을 누르면 하단 바에 ‘홍길동 님의 보유 주식 리스크 안전등급은 안전/위험입니다’라는 문구가 존재한다. 이 문구를 누르거나 상위 탭을 오른쪽으로 스크롤하면 있는 ‘리스크 관리’ 탭을 누르면 ‘리스크 관리 화면’으로 넘어간다.

본인이 소유한 종목 내에서 같은 리스크 키워드에 해당하는 다른 종목을 보고 싶다면 도넛 그래프에 있는 키워드에 마우스 커서를 올리거나, 종목 별 기사 요약에 있는 리스크 키워드에 마우스 커서를 올리면 확인할 수 있다.

또한 리스크 키워드에 해당하는 과거 사례를 추가적으로 더 보고 싶다면 더보기를 누르면 된다.

3. 사용 데이터 명세 및 수집 방법

1. 미래에셋증권에서 제공하는 6개월치의 뉴스 데이터

23.01.01 ~ 23.06.30 미래에셋증권 MTS/HTS에서 제공됐던 뉴스데이터를 사용한다. 우리 서비스에서 현재 해당하는 리스크와 관련된 과거 사례를 제공해야 하기 때문에 많은 뉴스 기사 데이터가 필요하다. 가장 최우선적으로 접근성이 높은 미래에셋증권의 과거 뉴스 데이터가 적합하다고 판단하여 이를 사용할 생각이다. 6개월치 뉴스 데이터에 대한 요약과 키워드 추출 과정을 진행시킨 뒤 매일매일의 리스크 키워드와 관련된 사례 제공을 진행한다. news\_dataset\_202301.zip부터 news\_dataset\_202306.zip까지 사용할 예정이다.

1. 종목별 실시간 뉴스 데이터

우리는 종목별 실시간 뉴스 데이터를 네이버 증권에서 매일 실시간으로 올라오는 뉴스 데이터를 크롤링하려고 한다. 웹 크롤링이란 웹 상의 정보들을 탐색하고 수집하는 작업을 의미한다. 실시간 뉴스 데이터를 분석해야만 하는 우리 서비스의 특성 상, 웹 크롤링은 꼭 구현해야 하는 기능 중 하나이다. 우리 서비스는 오전 7시에 매일 리포트를 업데이트해야 하기 때문에 그 전날 오전 7시부터 당일날 새벽 4시까지 21시간동안의 국내 주식 기사를 크롤링한다. 모든 뉴스 기사를 크롤링하고 해당 기업이 어디인지 분류하는 것이 시간과 자원 소모가 클 것이라고 판단되어 네이버 증권에 있는 기능인 ‘종목별 뉴스 공시’를 사용하여 종목별 기사 크롤링을 진행한다.

웹 크롤링을 구현하는 모델의 경우, 많은 라이브러리와 다양한 레퍼런스가 이미 존재하고 있기 때문에 우선적으로 가장 대표적이고 성능이 좋다고 판단되는 모델 3가지를 비교 분석해봤다.

| 모델 | BeautifulSoup | Scrapy | Selenium |
| --- | --- | --- | --- |
| 정의 | 웹 스크래핑을 위한 파이썬 라이브러리로, 간단한 웹 페이지에서 사용하기에 적합함 | 웹 크롤링 및 스크래핑을 위한 파이썬 프레임워크로, 대규모 크롤링과 고급 기능이 필요한 경우에 사용함 | 웹 브라우저를 자동화하고. 테스트하는 데 사용되는 라이브러리로, 동적 웹페이지 크롤링과 스크래핑에 적합함 |
| 장점 | 크롤링 자체에 최적화되어 있지 않아서 작업 수행 시 추가적인 기능과 설정이 필요함 | 동적으로 생성된 페이지의 데이터를 크롤링하는 것이 가능함. 또한 웹사이트의 구조를 파악하고 크롤링 규칙을 설정하여 자동화된 크롤러 규칙 설정 가능함 | JavaScript에 의해 동적으로 생성된 페이지의 데이터를 크롤링하는 것이 가능함 |
| 단점 | JavaScript에 의해 동적으로 생성된 페이지의 데이터를 크롤링하는 것이 힘듦 | 간단한 작업에서는 오버헤드 발생할 수 있음 | 웹 브라우저를 실제로 제어하기 때문에 많은 리소스가 필요하고, 페이지 요청과 응답에 있어서 많은 시간 소요할 수 있음 |

우리 서비스는 매일 나오는 수많은 기사를 크롤링하며 분석해야 하기 때문에 데이터 규모가 크다. 또한 매일 서비스를 제공해야 하므로 자동화된 크롤러를 만들어야 할 필요성이 있다. Scrapy 라이브러리는 큰 규모의 크롤링 작업에 적합하며 Spider에 정의된 크롤링 규칙을 설정하여 자동화된 크롤러를 만들기 쉽다는 장점을 가지고 있다. 이러한 장점이 우리 서비스와 가장 적합하다고 판단하여 Scrapy 라이브러리를 사용하여 웹 크롤링을 진행할 예정이다. 최종적인 결과물은 TITLE, COMPANY\_NAME, CONTENT, COUNT 열을 가진 CSV 파일로 변환하여 저장한다.

1. 등락률 데이터

과거 사례를 소개해주는 기능에서 실제로 그 당시에 주가가 어떻게 변화했는지를 명시하기 때문에 해당 날짜의 종목 등락률, 종목명, 날짜 데이터를 네이버 증권에서 크롤링하려고 한다.

1. 개인 보유 주식 데이터

개인 보유 주식 데이터는 미래에셋증권에서 제공하는 고객 데이터인 ‘월말일 기준 평가액 1(2,3)위 국내주식 종목번호’ 컬럼을 사용할 예정이었으나 이 데이터는 월말 기준이기 때문에 다음 달 말일까지 투자자의 새로운 매수 및 매도 이력은 알 수 없다. 매일 업데이트 되는 우리 서비스의 특성 상 월말 데이터만 가지고는 서비스 구현이 제대로 안 되기 때문에 위 데이터는 사용하지 않는다. 따라서 실제 마이데이터는 구하기 어렵기 때문에 직접 고객 데이터를 생성해 서비스를 구현해볼 예정이다. 고객 데이터에는 임의의 고객명, 자산, 보유 종목 데이터가 저장되게끔 구현한다.

4. 데이터 전처리

1. 감정 분석 및 키워드 추출 모델
2. 번역

한국어보다 영어로 번역하여 텍스트분석을 하는 것이 더 효과적이다. 영어의 경우 한국어와는 다르게 형태소 분석을 따로 진행하지 않고 토큰화 작업만 진행시켜도 된다는 특징이 있다. 한국어의 경우 어간에 접사가 붙어 의미가 변하는 교착어의 특성을 가지고 있으며 단어의 순서가 유연하기 때문에 텍스트 분석을 진행하기 어려운 언어에 속한다. 반면 영어는 주로 단어의 형태 변화가 적으며 공백으로 구분되는 단어들이 많기 때문에 토크나이저만으로도 텍스트 분석을 효과적으로 진행할 수 있다.

이러한 이유들로 인해서 앞에서 한국어 기사를 영어로 번역하는 과정을 추가하였다.

여러 번역 API와 라이브러리를 고려해보았지만 DeepL이나 ChatGPT의 API 같은 경우 유료화되어 있고 텍스트 길이가 일정 수준 이상 넘어가면 작동이 제대로 되지 않아 제외하였다. 이에 반해 파이썬에서 import할 수 있는 라이브러리는 googletrans 와 kakaotrans가 있으며 kakaotrans의 경우 한국어 번역 및 한국어를 포함한 아시아 언어들의 번역에 특화되어 있어 한영 번역 시 googletrans에 비해 메리트가 있을 것으로 예상된다.

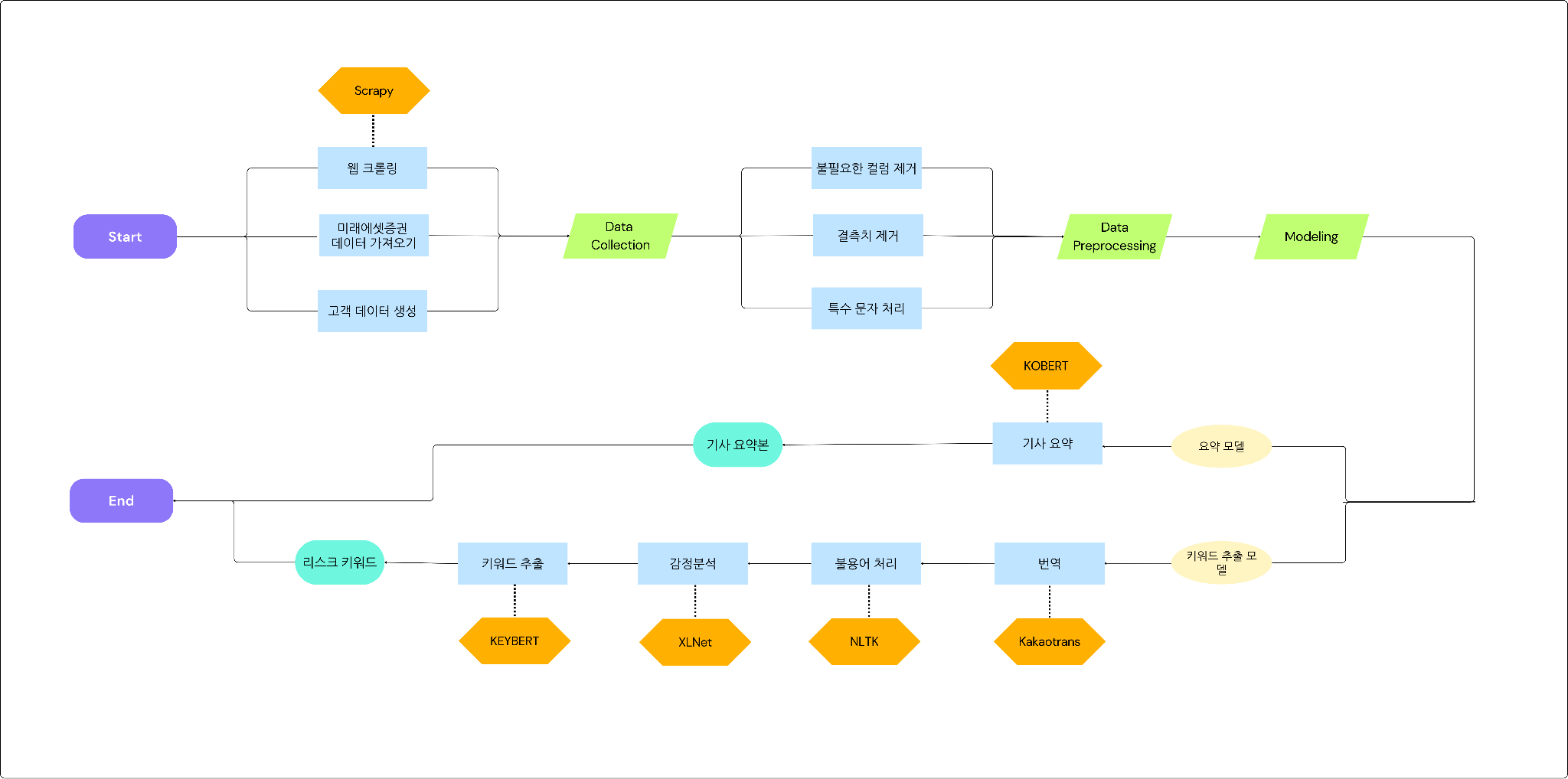
1. 특수 문자 처리: 실시간 뉴스 데이터를 네이버 증권에서 크롤링 해올 때 기사 본문에 있는 ‘/, [ ], $’ 등의 특수문자를 제거할 예정이었으나 뉴스 기사 내에서 유의미하게 쓰일 가능성이 있어 삭제하지 않는다.
2. 6개월치 뉴스 데이터를 살펴보면 총 8개 컬럼이 존재하는데, 우리는 DATE\_TIME, TITLE, CONTENT 세 컬럼을 사용하고자 한다. WRITER와 SUBCATEGORY의 경우, 우리 서비스에서 불필요한 요소라고 판단했고, ITEM\_NAME의 경우 결측치가 너무 많아 사용하지 않을려고 한다. 그리고 TAG\_LIST와 IMPORTANCE의 경우에는 키워드 추출 및 감정분석 단계를 아래 모델링에서 거치기 때문에 미래에셋증권에서 제공하는 값들이 아닌, 우리 모델을 통해 새롭게 정의하고 이를 저장할 예정이다.
3. 종목은 COMPANY\_NAME 열로, 각 기사의 제목과 본문은 TITLE, CONTENT 열에 각각 저장한다. 또한 네이버 증권 종목별 뉴스에서 ‘제목’ 탭 대신 ‘내용’ 탭에 들어가면 이미 유사한 기사명은 군집화가 되어 있어 비슷한 기사가 이미 묶여 있는 것을 확인할 수 있다. 그래서 크롤링을 진행할 때 같은 기사로 묶인 경우에 대해서 기사 개수를 언급량으로 생각하여 COUNT 컬럼에 추가해 저장한다.

5. 모델링

우리 서비스의 AI 모델은 키워드 추출과 텍스트 요약 크게 두가지이다.

첫번째 AI 모델인 키워드 추출은 총 7단계로 구성된다. 우선 한국어의 경우 교착어의 특성상 분석에 어려움이 존재하기 때문에 영어로 번역하여 텍스트 분석을 진행할 계획이다. 그런 다음 불용어 사전을 구축한 뒤 토큰화를 수행하여

우리 서비스에 있어 AI 모델 구현의 전체적인 단계는 7단계이다. (요약 모델이랑 키워드 추출 모델 2가지 케이스이고, 키워드 추출 모델은 크게 6가지 단계, 요약 모델은 크게 2가지 단계를 가지고 있다. )



1. 토큰화 / 벡터화

문자로 이루어진 데이터를 컴퓨터가 처리하기 위해서는 문자를 숫자로 바꾸어주는 벡터화 작업이 필요하다. 또한 벡터화 작업을 진행하기에 앞서 하나의 의미가 있는 단위인 토큰 단위로 나누는 토큰화 과정을 진행해야만 한다. 토큰화를 진행시키기 위한 모델은 BERT로 선정하였다. 최근 가장 유명한 모델인 GPT 모델 대신 BERT 모델을 선택한 이유는 텍스트 속에서 의미를 추출하는 기능은 BERT 모델이 더 강점을 지니고 있는 부분이기 때문이다. GPT는 사용자가 원하는 정보를 정확히 전달하기 위한 일종의 챗봇 서비스이기에 의미 추출보다는 텍스트 생성에 더 강점을 지닌 모델이다. 이러한 이유로 BERT가 GPT보다 더 나을 것이라고 판단하였다.

BERT 모델은 2018년에 구글이 처음 공개한 사전 훈련된 모델이기 때문에 2023년인 지금까지 BERT를 개량한 다양한 모델들이 추가적으로 발표되었다. 다양한 개량된 모델들 중에서 RoBERTa와 ALBERT 모델 2가지를 고려하였다. 두 모델 모두 기존의 BERT 모델보다 더 큰 데이터셋을 사용하여 학습하였으며 학습 시간 또한 더 많이 투자되었다. 뉴스 기사의 경우 다양한 전문 용어들과 신조어들이 자주 나와서 학습 데이터가 더 크면 클수록 좋은 성능을 보일 것이라고 예상하였으며 우선적으로 두 모델을 선정하였다.

| 모델 | ALBERT | RoBERTa |
| --- | --- | --- |
| 공통점 | WordPiece 토큰화 사용 - 단어를 더 작은 단어로 쪼개서 토큰화를 진행하는 것을 의미함. 학습 데이터에 존재하지 않는 새로운 토큰이 등장하여 나타나는 OOV(Out-Of-Vocabulary) 문제를 완화할 수 있음. 이를 통해 새로운 단어나 고유 명사를 더 잘 토큰화할 수 있음. | |
| 특징  및 장점 | 크로스 레이어 변수 공유 - 더 작은 크기의 모델로도 기존 BERT 성능을 낼 수 있음.  BERT는 12개의 인코더가 중첩되어 있어서 학습시 모든 인코더의 변수를 학습해야함. 하지만 ALBERT의 경우 첫번째 레이어에서 계산한 값을 뒤 레이어에서도 공유하므로 효율적인 학습 가능 | 정적 마스킹(bert-based) - 훈련할 때 하나의 토큰을 마스킹시킨 뒤 앞 뒤 단어들을 통해 예측하며 학습 진행  동적 마스킹(RoBERTa) - 문장을 여러 번 복사한뒤 랜덤하게 마스킹할 토큰을 지정함. 이를 통해 다양한 경우에 대한 학습이 가능하며 문맥을 더 잘 이해할 수 있음 |
| 단점 | 1. 직접 조정해야주어야 하는 하이퍼 파라미터는 더 많기 때문에 모델 튜닝 과정에 어려움이 존재함  2. 훈련 데이터 어휘와 파라미터를 공유하기 때문에 훈련 데이터에 의존적임 | 1.대규모의 언어 모델이기 때문에 많은 양의 컴퓨팅 자원이 요구됨  2.모델 자체의 크기가 크기 때문에 토큰화하는 시간이 오래 걸림 |

2) 불용어처리

영어로 번역된 기사에서 유의미한 단어만 대상으로 키워드로 추출하기 위해 불용어(stopword)를 제거하는 과정을 거친다. 불용어는 of, for, I, me 등 문서에서 자주 등장하지만 분석에 있어 유의미하지 않은 단어들을 뜻한다. 대표적으로 영문서에서 불용어를 제거하는 방법은 nltk 라이브러리에서 nltk.download(‘stopwords’)로 불용어 사전을 다운받고 stopwords.words(‘english’)로 설정하여 불용어를 제거할 수 있다. 추가적으로 뉴스 기사의 특성상 동일하게 등장하는 기자명이나 키워드를 추출하였을 때 반복적으로 등장하는 의미없는 단어는 불용어 사전에 추가한다.

3) 감정분석

매일 그 날의 보유 주식에 대한 리스크를 제공하는 서비스이기 때문에 주가에 영향을 미치는 것은 기업의 재무건전성이나 기업가치보다 사람들의 심리가 더 중요하게 작용할 것이라고 판단을 하였다. 따라서 기업과 관련된 부정적인 기사는 주가 하락의 요인이 될 수 있으며 이는 리스크로 정의할 수 있다. 이러한 이유로 뉴스 기사에 대한 감정분석 단계를 추가하였으며 리스크는 키워드 형태로 제공할 것이다.

개별 기사들을 각각 분석하는 것보다 비슷한 주제를 가진 군집으로 분류하고 각 주제에 대한 감정점수를 구해서 진행하는 것이 그 기업의 전체적인 상황과 리스크를 고려하는데 있어서 더 적절할 것이다. 하지만 네이버 뉴스에서 이미 비슷한 주제의 기사를 군집화하는 기능을 제공하고 있어 이를 사용하여 크롤링하였기 때문에 다른 별도의 작업을 진행하지 않았다.

감정점수는 감정분석을 진행할 경우 긍정/부정/중립에 속할 확률로서 결과값이 나오게 되기 때문에 이를 감정점수로 정의한다. 감정점수는 3가지 감정 중 값이 가장 큰 값을 사용할 예정이며 부정인 경우 양의 기호(+)를 붙일 것이며 긍정의 경우 그 반대인(-)를 붙일 예정이다. 긍,부정의 기호를 다르게 함으로써 합산 하였을 때 어느 쪽의 영향력이 더 강한지 확인할 수 있다.

감정분석의 단계는 우선 각 군집에 속한 기사들에 대해서 감정분석을 진행하여 감정 점수를 구할 것이다. 그런 다음 군집별로 기사들의 감정점수를 평균낸다면 그 주제가 긍정인지 부정인지 판단할 수 있을 것이다. 이를 바탕으로 그 기업의 전체적인 감정 점수를 구할 것이다. 기업의 전체적인 감정점수를 기준으로 그 기업의 감정점수가 부정적일 경우에만 즉 양의 실수를 가질 때에만 우리의 리스크 서비스에서 해당 기업에 관한 리스크를 제공할 것이다. 기업에 대한 감정점수를 구하는 이유는 만약 각 개별기사의 감정점수와 언급량을 고려하였을 때 부정인 기사가 존재하여도 긍정인 기사의 영향력이 더 강할 수도 있기 때문이다. 이런 경우 해당 기업은 리스크가 없음에도 불구하고 오히려 잘못된 정보를 제공할 수도 있다. 따라서 기업의 감정점수도 필요하다고 판단하였고, 기업의 감정점수를 구하는 수식은 아래와 같다.

| 수식  기업 A의 감정점수 =  (주제1 감정점수 \* 언급량 + 주제2 감정점수 \* 언급량 + …. + 주제N 감정점수 \* 언급량) / N  (N = 기업A에 속한 주제의 수)  (감정점수의 경우 부정일 경우 양의 기호이며 긍정일 경우 음의 기호를 가짐)  주제 N의 감정점수 =  (기사1 감정점수 + 기사2 감정점수 + …. + 기사N 감정점수) / M  (M = 주제에 속한 기사의 수) |
| --- |

제공되는 리스크 키워드는 부정이라고 판단한 주제속에서만 키워드를 추출해 제공할 것이다.

많은 모델들이 감정분석 기능을 제공하고 있지만 감정분석의 대표적인 모델인 XLNet, ELECTRA 두 모델을 우선적으로 고려하였다.

3-1) XLNET

XLNET은 기존 언어 모델들의 단점을 보완하고 장점만을 흡수한 모델이다. 기존 언어모델은 2가지 방식으로 분석을 진행한다. 첫번째는 자동회귀(Auto-regressive)방식으로서 이전 토큰들을 통해서 다음 토큰을 예측하는 방법이다. 이 방법은 단방향으로만 진행되기 때문에 문맥을 고려하는 데 한계점이 있다. 두번째 방법은 AE(Auto-Encoding)으로서 데이터를 마스킹 해 손상시킨 후 다시 재건하는 것을 목표로 하는 방법이다. BERT는 이 방법을 통해서 문맥을 이해하는 부분이 강화되었지만 재건축을 할 때 다른 마스킹된 단어들과의 조건부 확률을 고려하는 것이 아니기 때문에 독립적으로 예측이 진행된다. 따라서 각 단어들은 독립적이며 이로 인해서 문장 길이가 길거나 복잡할 경우 문제가 발생한다. 이러한 단점들을 보완하여 XLNET은 문장 내 모든 단어들의 토큰을 무작위로 섞어 순서를 맞추는 방식을 도입하였다. 이를 Permutation Language Modeling이라고 한다. 또한 기존 Transformer와 달리 Transformer-XL을 사용하여 고정된 입력 길이를 사용하지 않기 때문에 이전 단어의 정보를 장기적으로 기억할 수 있게 되었다. 이를 통해서 XLNET은 장기의존성 문제를 해결하여 문장의 길이가 긴 문서의 맥락도 더 잘 파악할 수 있게 되었다.

3-2) ELECTRA

ELECTRA는 사전 훈련 방식으로 생성자-판별자 구조를 사용하는 모델이다. 생성자 단계에서는 문장의 일부 단어를 랜덤하게 다른 단어로 대체한 뒤 새로운 문장을 생성한다. 이렇게 생성된 문장에서 생성자는 대체된 단어가 무엇인지 예측하며 학습이 진행된다. 그 다음 판별자에서는 생성자에 의해서 새롭게 만들어진 문장과 기존의 문장을 구분할 수 있도록 훈련한다. 이 학습 방법을 통해서 기존의 BERT가 지니고 있었던 입력 데이터와 학습데이터의 불일치 문제를 해결하였다. 불일치 문제란 기존 BERT의 경우 일부 단어가 마스킹된 문장을 가지고 학습을 진행하지만 fine-tuning을 위해 새롭게 입력되는 데이터의 경우 마스킹이 되어 있지 않다는 차이점이 존재한다. 이로 인해서 오히려 마스킹 된 문장을 더 잘 이해하며 실제로는 성능이 달라지는 문제점이 바로 불일치 문제이다. 앞서 언급하였듯이 ELECTRA는 마스킹이 아닌 새로운 단어로 대체하는 방식으로 학습을 진행하여 완전한 문장을 가지고 학습하기 때문에 기존 BERT보다 성능 면에서 우수하다.

SOTA에서 진행한 감정분석 대회 결과를 보면 두 모델 모두 감정분석에서 우수한

성능을 보이지만 XLNET이 더 좋은 성능을 보였기에 우선적으로 XLNET을 사용하기로

한다.

4) 키워드 추출

텍스트 요약

보유한 주식의 리스크를 키워드 형식으로 전달해주는 것이 서비스의 목적이기 때문에 뉴스 기사에서 키워드를 얼마나 잘 뽑는지에 따라서 서비스의 만족도가 달라질 것이라고 예상된다.

하지만 기업에 관한 뉴스의 수가 적은 경우가 존재하기 때문에 하나의 뉴스기사로만 키워드 추출하는 것이 올바르게 키워드를 뽑지 못할 것이라고 판단된다. 그렇기 때문에 요약문을 여러 개 생성하여 키워드 추출의 정확성을 높이기로 하였다. 텍스트 요약모델로는 TextRank를 선택하였다. TextRank를 선택한 이유는 이 모델의 경우 해당 방법이 문장들 간의 유사도를 기반으로 문장을 선택한 뒤 요약문을 생성할 수 있는 모델이기 때문에 다양한 유사도 방법을 선택할 수 있고, 문서의 문장 그대로 사용하기 때문에 전문 용어나 기업명과 같은 단어들을 못 잡을 가능성을 배제할 수 있다. 그래서 우리는 코사인, 유클리드, 자카드 유사도 3가지를 사용하여 서로 다른 3가지 요약문을 생성할 것이다. 이렇게 생성된 요약문들을 사용하여 키워드를 추출한다.

키워드 추출 모델

키워드 추출 모델로는 KEYBERT와 TF-IDF를 고려하였으며 둘 중 하나를 선택하여

진행할 것이다

4-1) KEYBERT

KeyBERT는 BERT-based 모델이기 때문에 문맥의 의미를 파악하고 키워드를 잘 추출할 수 있다는 장점이 있다. 또한 코사인 유사도를 기준으로 문서와 키워드 간의 유사도를 판단하고 문서 맥락과 가장 연관성이 높은 키워드만을 추출한다.

KeyBERT로 키워드를 추출하는 방법으로는 두가지가 존재하는데 첫번째 방법은 MMR(Max Marginal Relevance)이다. 이 방법은 키워드를 최대한 다양하게 뽑아야 하는 경우 사용한다. 가장 먼저 키워드와 문서와의 연관성을 계산한 다음 추출된 키워드들 중에서 중복된 것을 최소화하여 다양성이 유지될 수 있도록 해준다.

두번째 방법은 Max Sum Similarity이다. 이 방법은 하나의 문서에서 키워드를 추출할 때 키워드들 간의 유사도 합을 최대화하는 방법이다. 첫번째 방법이 키워드의 다양성을 중요시했다면 두번째 방법은 서로 연관도가 높은 키워드들이 뽑힐 수 있도록 하는 방법이다.

4-2) TF-IDF

TF-IDF는 전체 문서에서 공통적으로 등장하는 단어의 경우 중요도가 낮다고 판단하며 특정 문서에서만 등장하는 단어가 중요한 단어라고 판단한다.

TF-IDF에서 TF는 단순빈도를 의미하며 하나의 문서에서 사용된 단어를 카운팅한 뒤 벡터로 표현하는 DTM을 생성한다. DF의 경우 여러 문서들 속에서 특정 단어가 몇개의 문서에 존재하는지를 의미한다. 이때 하나의 문서에서 몇 번 언급이 되었는지는 중요하지 않고 등장의 유무를 중시한다. 따라서 5개의 문서에서 하나의 문서에만 등장한 단어가 해당 문서속에서 1000번이 언급이 되었든 100번이 언급이 되었든 DF값은 1이 된다. 그런 다음 IDF 값은 log(N/DF(w)) 를 통해서 구할 수 있다. N은 전체 문서의 수를 의미하며 DF(w)는특정 단어 w가 등장하는 문서의 수를 나타내는 빈도를 의미한다. 이때 로그를 취해주는 이유는 총 문서의 수가 많아질 경우 값이 기하급수적으로 커지기 때문이다. 이와 같이 TF-IDF는 특정 문서에서만 등장하는 단어를 중요한 키워드로 인식하기 때문에 뉴스 기사에서 키워드를 뽑는 방식으로 적절하다고 판단을 하였다.

키워드 정제

중복된 키워드로 인해 제대로 된 리스크 전달이 안 되는 경우를 배제하기 위해 중복된 키워드를 제거하려고 한다. 그런 다음 제목과 본문을 가중합하여 정규화한 벡터와 추출된 키워드 벡터간의 코사인 유사도를 계산하여 최종적으로 상위 2~3개의 리스크 키워드를 선정할 것이다. 이때 사용되는 벡터화 모델로는 1)에서 사용한 모델을 사용할 것이다.

(2) 요약 모델

한국어는 형태소 분석 자체가 어려워 키워드 추출 과정에서 어색한 단어가 나올

가능성이 높다. 따라서 키워드 추출을 할 때 한국어를 영어로 번역하여 진행한다면 더

나은 성능을 보일 것이라고 판단하였다. 하지만 요약 모델의 경우 한국어 요약 모델을

사용해도 성능의 큰 차이가 없기 때문에 CONTENT 열에 존재하는 한국어 기사를

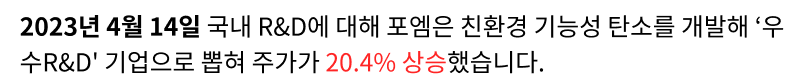
사용하여 요약을 진행한다. 요약 모델의 과정은 아래와 같다.

1) 해당 기업에 대한 가장 부정적인 뉴스 기사 요약

BERT의 경우, 양방향 Transformer 기반으로 사전훈련된 자연어 처리 모델로, 문맥을 파악하기 위해 양방향 정보를 사용하기 때문에 문장의 의미를 이해하는 데에 뛰어나며, 요약과 같은 작업에서 좋은 성과를 보인다. 이러한 이유 때문에 많은 논문에서 한국어 기사를 요약할 때 BERT 계열의 모델을 많이 사용하는 것을 확인할 수 있다. 그 중에서도 KOBERT는 대규모의 한국어 텍스트 데이터로 사전 훈련되어 있고 풍부한 언어 지식을 보유하고 있어 요약 작업에서 높은 성능을 보인다. 또한 한국어의 특성을 고려해 구축되어 있으며, 문장 구조와 단어 표현에 있어서 최적의 성능을 가지고 있기 때문에 한국어 기사를 요약 제공하기 위해서 KOBERT 모델을 사용하고자 한다. 우리가 현재 주식 관련 기사를 다루어서 그 도메인에 맞게끔 요약을 진행해야 하는데, KOBERT 모델은 사전 훈련된 모델을 다른 작업에 맞게 Fine-tuning하는 것이 용이하기 때문에 Fine-tuning 작업을 진행하여 기사 요약을 진행하고자 한다.

2) 부정 뉴스와 유사한 키워드를 가진 과거 사례 기사 요약

리스크 키워드에 해당하는 과거 사례와 실제 등락률에 대한 정보를 제공해줌으로써 동일한 키워드에 대한 다른 종목에서의 이슈들과 주가 변화를 파악할 수 있고, 고객들의 주식거래에 대한 의사결정을 도와줄 수 있다. 우리가 제공하고자 하는 예시는 아래와 같다.



이전 단계들에서 가장 부정한 뉴스 키워드 추출 및 기사 요약을 진행했기 때문에 가장 부정하다고 판단되는 뉴스의 리스크 키워드가 뽑혀 있는 상태이다. 과거 뉴스 데이터 중에서 이와 동일한 사례로 판단되는 것들을 뽑아야 하기 때문에 엄청난 양의 뉴스 데이터를 살펴야 한다. 하지만 모든 기사를 다 크롤링하고 비교 분석하면 연산량이 너무 많아지고, 효율적이지 못하게 된다. 그렇기 때문에 우선 과거의 뉴스 데이터는 TITLE만 확인하고 TITLE에 키워드가 포함되어 있는 경우 비슷한 사례라고 판단하여 그 기사에 대해서만 1번 단계와 동일하게 벡터화를 진행한다. 제목으로 먼저 기사를 분류한 뒤, 본문을 확인하는 방식을 취함으로써 연산량을 줄이고자 한다. 이후 본문에 대해서 실제로 유사한 사례인지 비교하기 위해 코사인 유사도나 임베딩 기반 유사도 측정 방법을 통해 기사들을 비교하고, 특정 임계값 이상의 유사도를 가지는 기사를 선택한다. 이렇게 해서 뽑아낸 기사를 요약하여 과거 사례로서 제공하려고 한다. 이 때 미래에셋증권에서 제공한 6개월치의 뉴스 데이터를 사용하고자 하며, 한글 기사로 되어 있기 때문에 (2)-1) 단계와 동일하게 KOBERT 모델 사용해서 요약을 진행한다.

그리고 그 사례가 발생한 DATA\_TIME을 수집하여 해당일의 주가 등락률을 제공한다.

6. 결론 및 확장 방향

결론 및 기대효과

리스크 관리 서비스의 주 목적은 매일 뉴스 데이터를 기반으로 고객들의 보유 종목들에 대한 리스크 정보를 실시간으로 제공함으로써 주식을 거래할 때 더 현명하게 대응하고 주식 관리에 대한 이해도를 높일 수 있다는 것이다. 고객들에게 종목을 추천하는 등 인사이트를 제공하는 것이 아니라 유의미한 정보를 바탕으로 스스로 판단할 수 있게끔 유용한 정보를 제공하는 것이 우리 서비스의 핵심이다.

아무리 사람이 분산투자를 한다고 하더라도 자신의 보유 종목들이 비슷한 리스크를 공유하고 있다면 이를 매도한 뒤 다른 종목에 투자해 손실을 막을 수 있다. 또한 오늘의 리스크 키워드를 확인함으로써 내가 지닌 리스크가 전체 리스크 키워드들 중 심각성이 어느 정도인지, 다른 종목에 새로 투자할 때 해당 종목이 어떤 리스크를 보유하고 있는지 확인할 수 있어 손실을 미리 방지할 수 있다는 것이 우리 서비스의 차별점이다. 타 증권사에서도 아직 대국민 리스크 관리 서비스는 찾아보기 어렵기 때문에 이는 신규 고객을 끌어들이는 효과도 기대해볼 수 있다.

최근들어 다양한 AI 서비스와 개인 자산 관리 서비스가 등장하였지만 아직까지는 보통의 개인 투자자들을 위한 서비스는 접근성이 떨어진다. 이러한 개인 투자자들을 위해서 우리는 매일 뉴스 기사에 대한 감정분석을 진행하여 보유한 각 종목에 대한 그 날의 부정적인 기사들 중 가장 영향력이 클 것으로 예상되는 리스크 키워드를 제공하여 리스크 관리를 실현하고자 한다. 각 기업들에 대한 리스크가 키워드 형태로 제공되기 때문에 보유한 종목들이 오늘은 어떤 리스크를 가지고 있는지 파악하기 쉬우며 만약 보유한 주식들에서 공통적으로 동일한 키워드가 반복된다면 제대로된 분산 투자를 하지 못하고 있는 것으로 간주하여 동일한 리스크를 공유하지 않는 새로운 주식을 탐색해볼 수 있는 기회를 제공할 수 있을 것이라고 전망한다.

주식 투자를 하면서 수익을 얻는 것도 중요하지만 보유하고 있는 자산을 잃지 않도록 하는 것도 그만큼 중요하다. 리스크관리 서비스를 통해 투자자들이

확장방향

1. 해외 주식으로의 서비스 확장

현재 우리 서비스는 국내 주식을 대상으로만 리스크 관리 서비스를 제공하고 있다. 국내 주식뿐만 아니라 해외 주식을 거래하는 고객들도 많은데 서비스를 개발하는 입장에서 서비스의 필요성을 파악하기 위해 우선적으로 국내 주식으로만 서비스를 기획했다. 이후 필요성이 있다고 판단되면 해외 주식 관련해서 리스크 관리 서비스를 진행할 수 있다. 그러면 고객이 보유한 국내 및 해외 뉴스를 바탕으로 해외 종목들까지 리스크를 확인하면 더 다채로운 정보를 확인할 수 있으며, 국내 주식과 해외 주식이 공통적으로 가지고 있는 리스크나 경제 상황에 대해서 분석을 추가적으로 진행할 수 있을 것이라고 판단한다.

1. 더 많은 과거 사례 뉴스 데이터 분석

현재 우리 서비스는 미래에셋증권에서 제공하는 6개월의 뉴스 데이터를 활용하여 리스크에 연관된 과거 사례를 제공하고 있다. 6개월보다 더 이전의 기간 뉴스 데이터는 수집하지 못한 상태로, 그 부분까지 크롤링을 진행한다면 너무 많은 데이터와 모델을 돌려야 한다고 판단했다.

1. 서비스 성능 확인

현재 우리 서비스는 팀 내에서 자체적으로 구현한 가상 고객 데이터로만 구현을 진행한다. 추후에 미래에셋증권 본선을 진행하는 과정에서 추가적으로 대용량 가상 고객 데이터나 민감한 고객의 개인 정보가 포함되지 않은 암호화된 파일을 받을 수 있다면 그 데이터를 바탕으로 더 나은 서비스가 될 것이라고 생각한다.

**참고 문헌**

<https://velog.io/@ssokeem/data-news-project-final>

[버트(BERT) 개념 간단히 이해하기 (tistory.com)](https://moondol-ai.tistory.com/463)

[BERT와 BERT 파생모델 비교(BERT, ALBERT, RoBERTa, ELECTRA, SpanBERT) - 코딩 일기 (tistory.com)](https://chanmuzi.tistory.com/163)

[[NLP] KeyBERT - 키워드 추출(Keyword extraction) (velog.io)](https://velog.io/@jochedda/KeyBERT-%ED%82%A4%EC%9B%8C%EB%93%9C-%EC%B6%94%EC%B6%9CKeyword-extraction)

[TF-IDF (Term Frequency - Inverse Document Frequency) (tistory.com)](https://yeong-jin-data-blog.tistory.com/entry/TF-IDF-Term-Frequency-Inverse-Document-Frequency)