DS INVESTMENT & SECURITIES

퀀트

숨겨진 알파를 여는 비밀 코드

:토픽 모델링으로 보는 국내 주식시장

신민섭 퀀트

02-709-2667
sms200uk@ds-sec.co.kr

2024.12.19

Con-tents

텍스트데이터 속에서 가능성을 찾다

모두가 손쉽게 시장을 관찰하는 시대

숨겨진 수익률을 찾기 위한 연속된 선택의 게임

빠르게 정보가 반영되는 주식시장

예전보다 편해진 세상이다. 실시간으로 투자 정보가 업데이트되는 시대 속에서 투자자들은 계속해서 기회를 탐색하고 있다. 시장은 복합적인 요소들이 혼재하는 시장 속 매수와 매도가 경합하면서 매 순간 움직인다. 주식시장을 움직이는 주체는 투자자들의 매매이다. 이들을 움직이게 만드는 촉매는 시장에 관한 정보이다. 기업들의 공시와 언론 보도 등이 대표적이다. 증권가의 전망도 시장을 움직이는 촉매를 제공한다. 이처럼 끊임없이 나오는 정보에 투자자들은 반응하며 주식시장이 유기체와 같이 역동적으로 반응한다.

불확실한 미래 속에서의 투자는 선택의 연속

투자자들은 각자 다른 시각으로 시장을 해석한다. 향후 시장은 어떻게 흘러갈까? 자신이 투자한 종목은 미래에 어떤 수익률을 실현할까? 최근 나타난 이슈로 인해 시장에 미칠 영향은 무엇일까? 같은 숫자와 텍스트 데이터들을 보더라도 다른 생각을 가질 수 있다. 때에 따라서는 다양한 투자자들의 관점이 한쪽으로 쏠릴 수도 있다. 반대로 투자자들의 관심이 분산되어 시장 지수가 역동적으로 움직이지 않을 수도 있다. 불확실한 미래 속에서 투자는 선택의 연속과 다르지 않다.

매일 쏟아지는 많은 정보에 압도될 정도이다. 연속된 선택의 게임 속에서 무엇에 투자해야 할지 고민이 들 정도이다. 그럼에도 투자하기 위해서는 정보의 홍수 속에서 압축적인 의사 결정이 필요하다. 주식시장 참가자들은 자신만의 투자 전략과 투자 철학이 있다. 각자가 쌓은 경험치가 제각기 다르기에 투자자 자신들이 특화하고 있는 분야도 다르다. 불확실성이 누군가에게 기회일 수도 있고 위험일 수도 있다. 미래를 그 누구도 정확하게 예측할 수 없기에 투자에서 정답을 찾는 것은 무모한 도전에 가깝다.

투자를 하는 이유는 미래의 결과를 실현하기 위해서 존재한다. 투자자에게는 미래에는 더 큰 자산으로 경영진에게는 더 큰 성과를 실현할 것이라는 믿음이 있기에 자본이 움직이는 것이다. 물론 투자에 있어서 다른 이유도 존재할 수 있지만 미래에 성과를 기대하거나 이루기 위해서 투자한다는 점에 있어서는 부정할 수 없을 것이다.

구슬도 진주도 꿰어야 보배다

언론은 중요한 정보 제공자

날마다 보도되는 국내 경제 그리고 기업에 대한 소식을 통해 어떻게 세상이 움직이는지 언론을 통해 알 수 있다. 기업들의 실적부터 한국 경제에 대한 전망까지 언론을 통해 대중들에게 공개된다. 언론을 통해 다양한 분야의 전문가들의 해석과 전망을 볼 수 있다. 이처럼 언론은 투자에서 중요한 정보 제공자 역할을 담당하고 있다.

투자 정보를 찾는 과정은 선택에 대한 정당성을 부여하는 과정

투자 정보를 찾아보는 것은 확신을 가질 수 있는 작은 조각들을 모으는 과정이다. 자신이 투자한 대상에 대한 정보들은 변화가 나타날만한 각각의 가치를 지니고 있다. 기업을 포함하여 국가 내 경제 주체들의 흔적을 언론을 통해 파악할 수 있으며 어떻게 흘러갈지 판단할 수 있다.

때로는 지금 시점에서 왜 이 정보가 나타났는지 고민해야 할 때도 있다. 하지만 대개 시간이 지나 자연스럽게 드러나는 경우가 많다. 진실인지 거짓인지 확실하지 않은 정보는 시간이 지나면 본질이 드러난다. 빙산의 일각이 드러나기 전에 주가는 먼저 움직일 수 있다. 그전에 실마리를 찾아야 한다. 구슬도 진주도 꿰어야 보배가 되기 때문이다.

압축적인 정보 전달을 효율적으로 수행하는 모델의 필요성

투자의 실마리를 찾는 과정은 적지 않은 정신적인 노동과 스트레스를 동반한다. 방대한 정보 속에서 정보들을 찾다 보면 시장의 큰 그림을 잊어버릴 수도 있다. 같은 정보를 보더라도 각자가 다르게 꿰기 때문에 자신의 생각대로 흘러가지 않는다. 움직이는 이유는 저마다 다르다. 긍정적인 정보와 부정적인 정보가 오락가락하는 흐름 속에서 중심을 잡기가 쉽지 않다. 연속된 정보들의 주요 흐름을 몇 페이지의 도표 또는 압축된 키워드로 요약할 필요가 있다.

이러한 고민들을 반영하여 한국 경제와 증시에 대한 정보를 압축적으로 전달할 수 있는 모델을 제안하고자 한다. 토픽모델링(Topic Modeling)이다. 투자의 기회를 자동으로 찾기보다는 투자의 기회를 찾기 위한 지도를 제공하는 모델에 가깝다. 다만 세부적인 사항까지 파악하려면 원문 출처를 찾아서 추가로 확인하는 과정이 필요하다. 여기서 제안하는 모형의 주요 목적은 국내 시장과 관련한 큰 그림을 먼저 파악하는 것에 중점을 두었기 때문이다.

제안하는 모델은 넓은 선택지를 제공하는 역할

투자에 정답은 없다. 다만 자신이 흥미를 가지고 있는 분야에서 돈의 흐름을 따라가다 보면 자신만의 답이 있다. 당장 돈의 흐름이 많이 나타나는 곳에 주목할지 미래에 많이 나타날 곳에 주목할지는 선택에 달려있다. 제안하고 있는 모델을 통해 어떻게 데이터를 취합하여 결과물을 도출했는지 다음 절부터 설명하고자 한다. 모델의 구조와 기타 수식 설명은 글 후반부에 확인할 수 있다.

한국 증시와 경제의 흐름을 보여주는 직관적인 그림

텍스트 데이터에서 투자의 실마리를 체계적으로 찾고자 하는 시도의 필요성

어떠한 상황에서도 위험을 최소화하고 수익률을 극대화하는 전략을 구성하는 것은 매우 어려운 일이다. 하지만 수익률로 이어질 수 있는 촉매를 찾아서 수익률을 극대화할 수 있는 환경을 마련하는 것은 시도해 볼 만하다. 투자 전략은 여러 가지가 있고 모델도 다양하지만 투자하는데 대다수가 공감하는 대상은 경제와 기업의 행태이다. 누구나 쉽게 직관적으로 다가갈 수 있는 방법은 키워드와 토픽이라는 생각을 가지게 되었다.

투자에 있어서 단순히 키워드로 접근하는 거에 대해서는 사실 조심스럽다. 그럼에도 퀀트 모델에서 보완할 수 있는 부분이 있다면 시도할 가치가 있다고 생각한다. 불확실한 미래에 가장 필요한 것은 전망이다. 어떤 정보를 이용하고자 하는 사람들에게 단순히 정보를 얻는 것에 그치지 않아야 할 것이다. 어떤 전략이 유효할지 어떤 것에 집중해야 하는지 선택하기 위한 의견을 제시해야 한다.

모델에서 의견까지 자동으로 도출되지 않는다는 점이 제안하는 모델이 가진 한계이기도 하다. 효율적인 추론을 위한 필요한 정보들을 요약하는 것이 이 모델의 주된 목표이다. 특정 이슈에 대한 의견은 아직까지는 인간의 해석과 추론이 담당하는 영역이기에 이 부분은 자체적인 코멘트로 보완할 예정이다.

국내 증시뿐만 아니라 한국의 경제가 어떠한 방향으로 흘러가고 있는지 그림을 직관적으로 확인한다면 모델이 줄 수 있는 통찰은 다양할 것이다. 미래에 대해 자신이 가지고 있는 전망과 시장에서 내재된 다수의 전망 간의 큰 차이는 미래에 큰 성과의 차이를 가져올 수 있다. 이러한 괴리를 확인할 수 있다면 투자에 큰 도움이 될 것이라고 확신한다.

그림1 토픽모델링의 결과는 시장 속 군중의 시각과 자신의 시각을 복기할 수 있는 중요한 정보

자료: ChatGPT, DS투자증권 리서치센터
주: ChatGPT DALL-E를 활용하여 출력된 이미지

비정형데이터 속에서 기회를 찾다

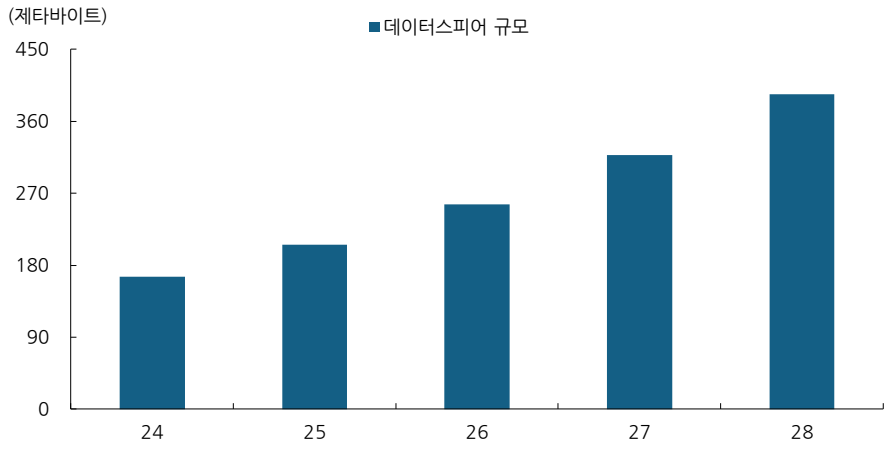
새로운 금맥

데이터의 양은 상상을 초월할 정도로 방대하다. IDC의 보고서와 Seagate의 2024년 사업보고서(10-K)에 따르면 2028년에 전 세계 데이터스피어(Datasphere)가 약 394 제타바이트에 이를 것으로 전망했다. 이를 기가바이트(Gigabyte, GB)로 환산할 경우 약 433조 2,076억 기가바이트(433,207,581,343,740GB)이다.

방대한 양의 데이터 증가량의 원천은 다양한 하드웨어 기기에 있음

데이터스피어는 전세계 다양한 하드웨어 기기에서 생성 및 복제되는 모든 유형의 새로운 데이터를 측정한 수치를 의미한다. 스마트폰, 노트북, PC, 촬영기기, 의료기기 등 다양한 전자기기로부터 이미지 또는 음성 및 영상까지 다양한 데이터들이 생산된다. IDC와 Seagate는 디지털 컨텐츠가 생성되는 원천을 코어(Core), 엣지(Edge), 엔드포인트(Endpoint) 3가지로 분류했다. 과거에 비해 데이터들이 디지털 컨텐츠로 전환되고 있으며 디지털화된 데이터들이 생성되는 경로가 다양해졌다.

그림2 전세계 연도별 데이터스피어 규모 추정



자료: IDC, Seagate 10-K (2024), DS투자증권 리서치센터

일상생활에서 접하는 데이터들은 상당 부분이 비정형 데이터

실시간으로 데이터와 상호작용하는 시대인 만큼 최근 이슈에 대한 해석도 짧은 시간 내에 볼 수 있는 시대이다. 새로운 정보에 대한 다양한 사람들의 의견을 영상 또는 텍스트로 확인할 수 있다. 예를 들면 미연방공개시장위원회(FOMC)가 진행된 다음 날 기자의 의견도 확인할 수 있고 증권가의 전망도 확인할 수 있다. 신제품에 대한 리뷰도 신제품이 출시된 다음날에 짧은 시간 안에 볼 수 있다. 이러한 데이터들은 정형화되어 있지 않다. 대부분 비정형화된 데이터들이다.

전체 데이터 중 80% 이상이 이미 비정형데이터

전체 데이터의 80% 이상의 비중을 차지하는 비정형데이터는 새로운 금맥이다. IDC에 따르면 향후 비정형데이터의 용량은 매년 빠른 속도로 증가할 것으로 전망했다. 또한 IDC는 전체 데이터에서 비정형데이터가 차지하는 비중은 90% 이상에 도달할 것으로 전망했다. 90%의 기회에서 비정형데이터를 정제하여 투자에 도움을 주는 데이터를 제공할 수 있다면 그 유용성은 무궁무진할 것이다.

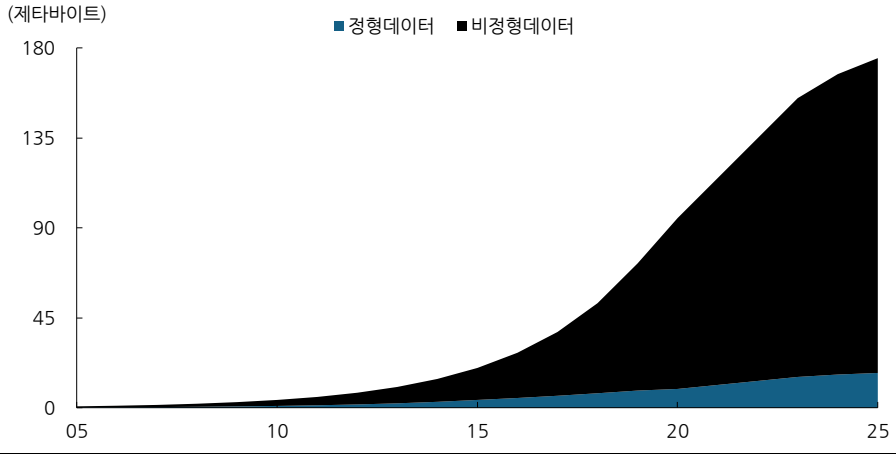
비정형 데이터에서 문맥과 디테일을 찾는다면 숨어있는 알파를 찾을 수도 있을 것

정형데이터는 이미 대다수 투자자가 공통으로 보는 데이터이다. 정형데이터 속에서 미래에 어떤 수치가 입력될지 예측한다. 미래에 특정 숫자가 기록될지 세부적인 이유와 과정을 찾기 위해서 치열한 수싸움이 주식시장에 녹아있다. 다만 수치로 입력된 정형데이터는 해당 수치가 입력된 결과만 이야기할 뿐이다. 왜 이 숫자가 기록되었는지 세부적인 이유와 과정을 찾기 위해서는 문맥(Context)이 필요하다. 다시 말해서 비정형데이터가 필요하다. 이처럼 정형데이터에서 정당화된 이유를 강화할 또 다른 증거를 비정형데이터에서 찾을 수 있다면 투자의사결정에서 좀 더 확신을 부여할 것이다.

투자에서 의사결정의 최종 단계는 1) 매수, 2) 매도, 3) 매수 또는 매도하지 않고 보유와 같이 3가지로 압축된다. 투자에서 왜 이런 선택을 하는지 명확한 근거를 확보해야 한다. 정형데이터만으로 의사결정을 하기에 어려운 부분을 비정형데이터가 상당 부분 보완할 것이다.

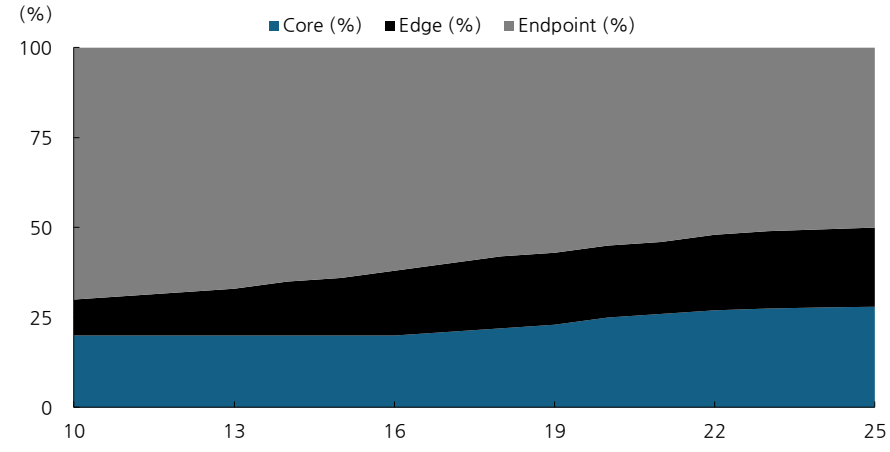
과거와는 다르게 정보의 이용성이 다양해졌고 합리적인 의사결정의 과정이 주식시장에 빠르게 반영되는 시대에 살고 있다. 수치데이터는 과거와 달리 빠르게 실시간으로 계산이 가능해졌다. 이제는 비정형데이터에서 숨겨진 알파를 찾아야 할 수도 있다. 비정형데이터를 찾아서 체계적으로 수집할 방법이 있다면 정형데이터만 사용했을 때보다 큰 부가가치를 창출할 수 있을 것으로 기대한다.

그림3 비정형데이터의 압도적인 성장세



자료: IDC, DS투자증권 리서치센터
주: 2017년에 전망한 수치에 기반하고 있으며 현재는 이보다 더 높은 성장률로 증가할 것으로 추정

그림4 어디에서 데이터가 생산되는가



자료: IDC, DS투자증권 리서치센터

표1 저장 형식에 따른 데이터 분류

특징	정형 데이터 (Structured Data)	비정형 데이터 (Unstructured Data)	반정형 데이터 (Semi-structured Data)
정의	사전에 정의된 형식에 따라 정의된 데이터	사전에 정의된 형식이 없는 데이터	데이터 구조에 대한 사항이 일부 있지만 엄격하지 않아서 형식과 구조가 변경될 수 있는 데이터
데이터 특성	양적 데이터	질적 데이터	양적/질적 데이터 모두 포함 데이터 구조에 대한 설명이 함께 존재
예시	관계형데이터베이스, 스프레드시트, CSV 데이터	NoSQL(비관계형데이터베이스), 동영상, 이미지, 텍스트 등	HTML, XML, RDF, JSON, 웹 로그 등

자료: Pure Storage, DS투자증권 리서치센터

때로는 간단한 요약이 필요하다

압축적 정보를 제공하는 텍스트마이닝 모델을 고안

세 줄 요약이라는 단어를 심심찮게 들어본 적이 있을 것이다. 현대인들은 수없이 많은 정보들 속에서 빠르게 핵심적인 정보를 파악하고 싶어 한다. 때로는 너무 많은 정보들이 한꺼번에 나타나 방대한 데이터로 인해 스크롤 압박감을 느낄 수도 있다. 일상이 바쁘고 핵심적인 아이디어만 필요한 사람들에게 정보를 압축하여 시각화하는 모델을 제시할 필요성을 느꼈다. 이를 해결하기 위해 텍스트마이닝(Text Mining)에 기반한 방법론을 고안했다. 바쁜 현대인들의 고민을 해결하는데 기여할 수 있을 것이다.

텍스트마이닝의 정의

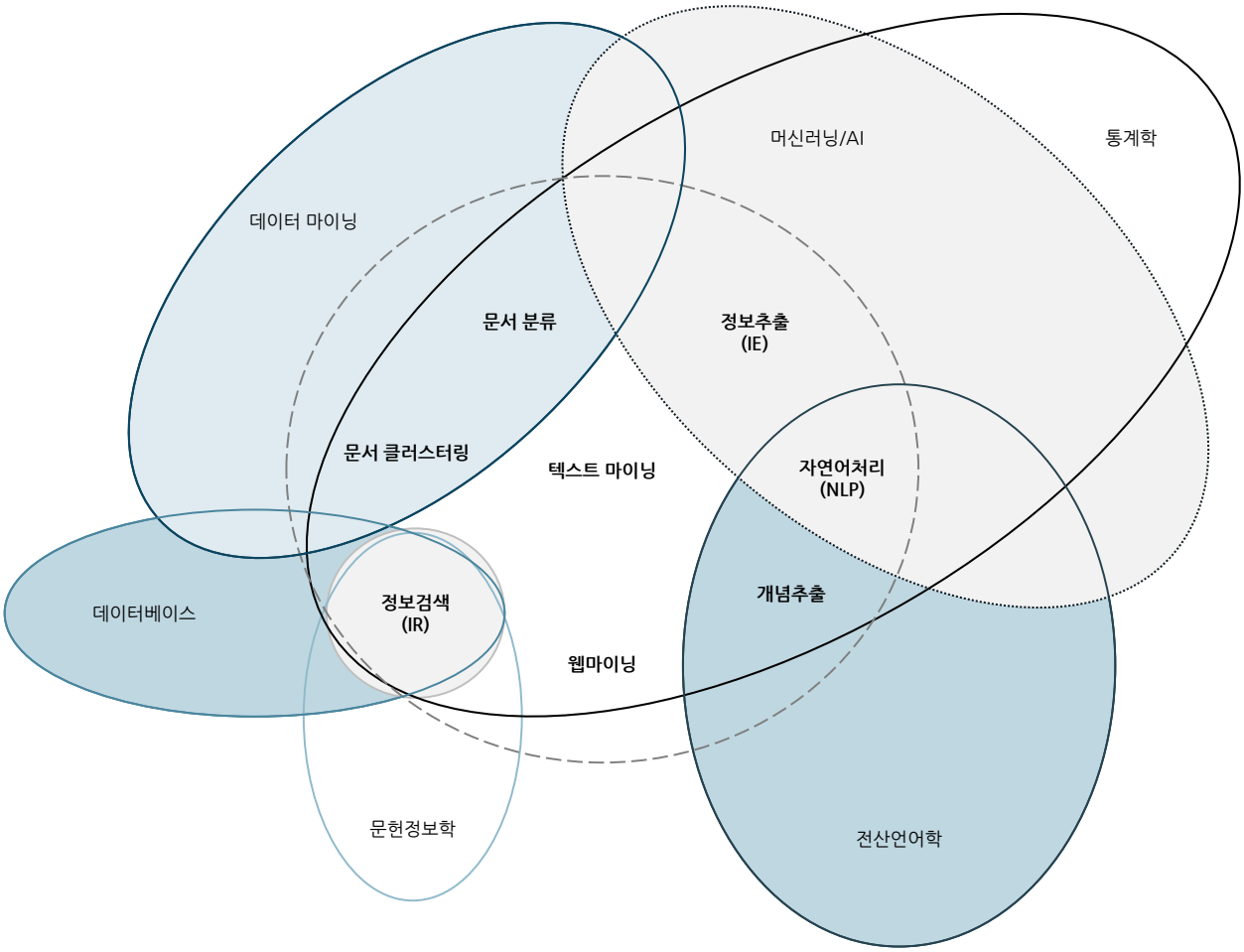
텍스트마이닝은 텍스트 데이터에서 고품질 정보를 추출하는 과정을 일컫는다. 뉴스 기사, 웹사이트, 이메일, 블로그, 댓글 등에 있는 문자들 모두 텍스트 데이터에 해당한다. 고품질 정보는 텍스트 데이터들 속에서 패턴과 트렌드를 통해 얻어진다. 제안하는 모델을 통해 달성하고자 하는 목표는 국내 증시의 주된 주제를 파악하는 것이다. 토픽을 구성하는 단어들을 통해 증시의 큰 그림을 신속하게 파악하는 것이 목표이다. 뉴스들의 흐름 속에서 명확한 주제가 있다면 그에 부합하는 투자자들의 동향이 나타날 것이다.

텍스트마이닝의 종류는 다양하며 때에 따라 다양한 텍스트마이닝 방법론들이 적용됨

텍스트마이닝의 종류는 다양하다. 텍스트마이닝을 적용하기 위해서 다양한 분야의 지식들이 동원된다. 때에 따라서 특정 프로젝트에 다양한 텍스트마이닝 분야들이 적용되는 경우도 있다. Miner et al. (2012)에 따르면 정보 검색, 정보 추출, 문서 분류 등 8개의 분야로 나눴다. 인간의 언어를 컴퓨터가 이해할 수 있도록 데이터를 처리하는 분야인 자연어처리(Natural Language Processing, NLP)도 텍스트마이닝의 한 분야에 속한다.

텍스트마이닝에서 가장 중요한 것은 문제의 정의이다. 텍스트마이닝을 통해 어떤 정보를 추출하여 어떤 문제를 해결할지 목적을 명확하게 설정하는 것이 중요하다. 문제의 유형에 따라 텍스트마이닝을 적용하는 방법론이 좁혀진다. 또한 텍스트마이닝은 다른 데이터와 결합했을 때 좀 더 유용성이 나타난다. 경제/주식 뉴스데이터의 경우 경제지표나 상장되어 있는 기업들의 주가와 결합되면 투자에서 활용도가 높아질 것이다.

그림5 다양한 분야의 지식들이 응용되는 텍스트마이닝의 분야



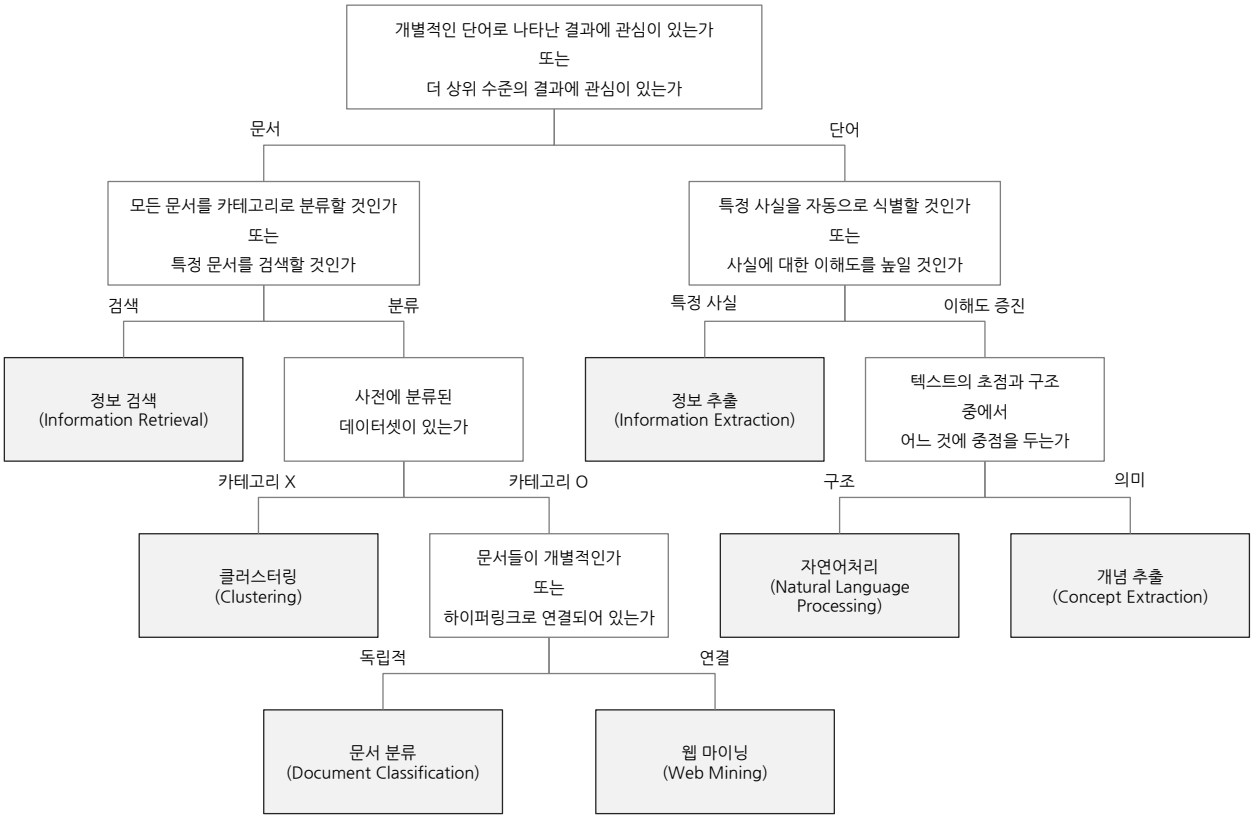
자료: Miner et al. (2012), DS투자증권 리서치센터

표2 주요 목적에 따라 텍스트마이닝은 분류가 다양하며 겹쳐진 영역이 존재

텍스트마이닝 분류	주요 목적
정보 검색 (Information Retrieval, IR)	텍스트 데이터 속 사용자가 필요로 하는 정보를 효율적으로 검색
웹 마이닝 (Web Mining)	웹페이지, SNS, 로그 데이터 등에서 데이터를 수집하고 분석하여 유의미한 패턴 또는 관계 발견
개념 추출 (Concept Extraction)	텍스트 데이터에서 추상적인 개념이나 의미를 파악하여 구조화
자연어 처리 (Natural Language Processing, NLP)	인간의 언어(자연어)를 컴퓨터가 이해할 수 있도록 처리
정보 추출 (Information Extraction, IE)	텍스트 데이터에서 키워드, 인물, 기관, 객체 등을 추출
문서 클러스터링 (Document Clustering)	문서를 그룹화
문서 분류 (Document Classification)	문서를 미리 정의된 범주에 따라 분류

자료: Miner et al. (2012), DS투자증권 리서치센터

그림6 텍스트마이닝 방법론 적용을 위한 의사결정 나무



자료: Miner et al. (2012), DS투자증권 리서치센터

토픽모델링은 상대적으로 접근 가능한 텍스트마이닝 방법론

핵심 정보를 축약할 수 있는 적합한 텍스트마이닝 모델 중에서 토픽모델링(Topic Modeling)이 가장 적합하다고 판단했다. 물론 ChatGPT와 Claude같이 생성형AI를 활용하여 맞춤형으로 학습하여 활용하는 방법이 존재한다. 딥러닝 모델을 직접 구축하는 방법도 있지만 시간과 비용이 적지 않게 든다. 데이터도 충분히 확보되어야 하며 필요에 따라 고성능 컴퓨팅이 가능한 인프라가 갖춰져야 한다. 데이터 과학에 대한 지식도 요구되기에 모델을 개발하기에는 쉽지 않을 것이다.

시장과 밀착된 모델일수록 가장 필요한 요소는 시의성이다. 비교적 짧은 주기에 나타난 정보들을 처리해서 필요한 의사결정을 할 수 있도록 도움을 주는 체계적인 절차가 필요하다. 데이터 수집부터 결과물을 도출할 때까지 반복적인 과정을 안정적으로 할 수 있는 구조를 구성하는 것은 필수이다. 체계적인 절차가 마련된다면 시장을 해석하는데 적지 않은 효용을 줄 수 있을 것으로 기대한다.

요약을 위한 주춧돌

토픽모델링은 시장에 잠재된 주제의 구조를 보여주는 역할을 수행

국내 주식과 관련한 정보를 효율적으로 요약하는 도구를 만들기 위해서는 많은 시간과 자원이 소모된다. 모델도 무거울 것이기에 부담이 크다. 최적화된 하나의 해답이 존재하지 않고 사용자가 요약을 잘했다고 평가할 만한 기준이 명확하지 않다.

주의해야 할 점은 토픽모델링은 텍스트 데이터에서 정보를 요약하는 목적으로 활용되지 않는다는 것이다. 토픽모델링은 결과물에서 어느 정도 요약의 역할을 담당하고 있지만 근본적으로 텍스트 데이터에 내재된 주제 구조를 도출하는 것에 주된 포커스를 두고 있다. 국내 증시를 몇 문장으로 요약하도록 결과를 도출하기 위해서는 고도의 사고 과정이 필요할 것이다. 토픽모델링은 해당 부분을 담당하고 있지 않기에 다른 모형을 활용해야 한다.

제시하는 토픽모델링은 머신러닝 모델 범주에 해당한다. 딥러닝 모델에 비해 가벼운 편이다. 컴퓨팅 자원이 한정적인 환경에서도 구현이 가능하며 적재적소에 활용할 수 있다. 제한된 자원 하에서 가공할만한 결과를 제공하고 요약을 위한 주춧돌을 제시한다는 점에서 의미 있다고 판단한다. 제안하는 토픽모델링 모형을 통해 투자자들에게 다양한 지도를 제시할 것으로 기대한다.

이 모델의 주된 취지는 사고의 과정을 단축할 수 있도록 기초적인 정보를 제시하는 것이다. 여러 가지의 토픽 모델 중에서 최선의 모델을 제시하는 것은 아니기에 얼마든지 개선하고 확장할 여지는 많다. 구현할 수 있는 모델 중 효율적인 것을 고안했기 때문에 추후에 다른 모델들과의 성능 비교는 필요하다. 성능보다도 중요한 것은 시장 그 자체를 반영하는 정보를 보여주는 것이다. 직관과 아이디어가 잘 나타나도록 지속적으로 모델을 검증하고 다듬는 것이 목표이다.

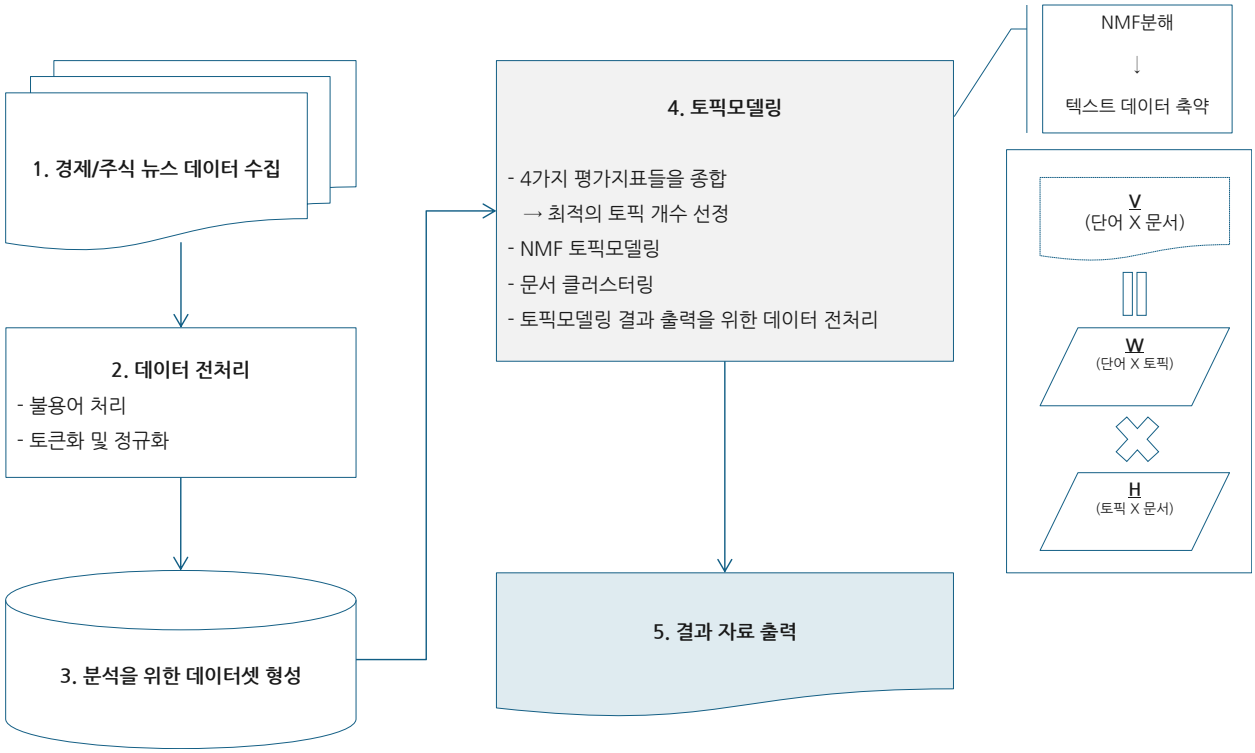
표3 토픽모델링 파이프라인

단계	절차 설명
1. 설정	사전에 필요한 Python 라이브러리 불러오기 및 데이터 입출력 경로 설정
2. 데이터 수집 & 종목 매칭	데이터 수집 유사도에 기반하여 국내주식 종목 매칭
3. 텍스트 전처리	수집한 데이터 중 키워드 등을 데이터를 전처리 및 토큰화
4. TF-IDF 생성	텍스트 데이터를 벡터화
5. 최적의 토픽개수 탐색	다양한 토픽 수에 대해 토픽모델 학습 및 평가 지표 계산
6. 최종 모델학습	최종적으로 선정된 토픽모델 학습 추가적으로 최적화 수행 행렬을 2 개의 행렬(W, H)로 분해
7. W 행렬 분석	문서-토픽 행렬에서 나타난 가중치들을 고려하여 주요 토픽 식별 문서 클러스터링 수행
8. H 행렬 분석	토픽별로 가중치가 높은 단어 10 개 선정
9. 토픽 분석	토픽별로 주요 단어 및 관련된 기사 분석
10. 평가지표 계산	**2 가지 평가지표 계산** 1) 토픽 일관성(Coherence Score), 2) 재구성 오차(Reconstruction Error) 사전에 설정한 가중치를 적용하여 토픽모델 성능 평가 토픽의 일관성을 중점적으로 가중치를 적용하여 반영
11. 결과 저장	최종적으로 선정된 NMF 토픽모델의 결과를 파일로 출력하여 저장

자료: DS투자증권 리서치센터

주: 문서 클러스터링을 수행하는 이유는 본 문서에서 사용하는 토픽모델링에서 토픽마다 배정된 문서와 문서클러스터는 유사한 면이 있기 때문. 데이터 시각화 목적으로 수행됨

그림7 토픽모델링 파이프라인 개요도



자료: DS투자증권 리서치센터

텍스트 속 숨겨진 알파를 찾는 비밀코드

빅카인즈(BigKinds) 소개 및 활용하기

빅카인즈 플랫폼 소개

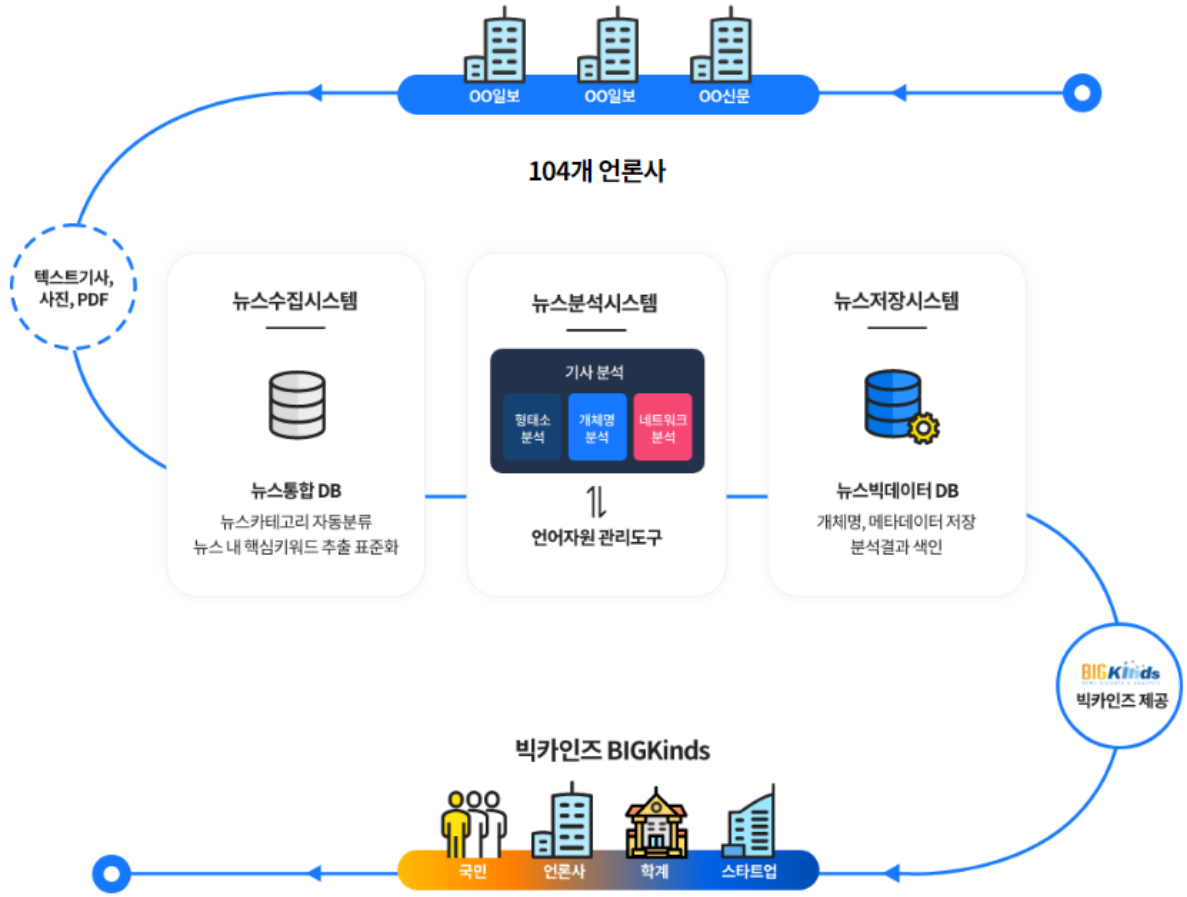
빅카인즈(BigKinds)는 한국언론진흥재단이 운영하는 뉴스 검색 및 뉴스 데이터플랫폼이다. 2016년 4월 19일부터 서비스를 시작하였으며 2024년 기준으로 104개의 여러 언론사와 협약을 맺은 상태이다. 서울, 경기, 경북, 경남, 강원 등 7개 지역별 10개 언론사의 뉴스를 빅카인즈를 통해 검색할 수 있다. 빅카인즈 플랫폼을 통해 검색할 수 있도록 협약을 맺은 언론사들의 수가 증가하는 추세이기에 데이터의 다양성이 강화될 것이다.

빅카인즈에서 제공되는 뉴스들은 실시간으로 수집된다. 작성일자 기준으로 약 1억 700만의 기사들이 수집되었으며 지속적으로 빅카인즈와 협약을 맺은 뉴스들이 빅카인즈 데이터베이스에 저장되고 있다. 수집된 기사들이 국내 언론 전체를 대표하기에는 어려운 측면이 있지만 뉴스데이터를 합법적으로 활용할 수 있는 몇 안 되는 플랫폼이다. 개인이 연구 목적으로 사용하는 것은 자유롭게 사용가능하다. 다만 상업적 목적으로 이용할 경우 뉴스 콘텐츠를 이용할 수 있는 상품을 뉴스토어라는 웹사이트에서 구매해야 한다.

빅카인즈의 장점은 전처리된 뉴스데이터를 이용할 수 있음

빅카인즈의 가장 큰 장점은 빅카인즈 자체 알고리즘으로 전처리한 텍스트 데이터를 사용자들에게 공급한다는 것이다. 데이터 전처리(Data Preprocessing)단계에서는 처리에 많은 시간과 노력이 든다. 빅카인즈에서는 데이터 전처리에 대한 수고를 크게 덜어준다. 빅카인즈 자체에서 뉴스 본문들을 분석하여 키워드, 객체, 인물, 기관 등 1차적으로 가공된 데이터를 제공한다.

그림8 빅카인즈 서비스 개념도



자료: 빅카인즈, DS투자증권 리서치센터

표4 빅카인즈에서 검색가능한 언론사

분류	개수	언론사
전국일간지	12	경향신문, 국민일보, 내일신문, 동아일보, 문화일보, 서울신문, 세계일보, 아시아투데이, 조선일보, 중앙일보, 한겨레, 한국일보
경제일간지	13	대한경제, 매일경제, 머니투데이, 메트로경제, 브릿지경제, 서울경제, 아시아경제, 아주경제, 이데일리, 이투데이, 파이낸셜뉴스, 한국경제, 헤럴드경제
지역일간지	45	강원도민일보, 강원일보, 경기신문, 경기일보, 경남도민일보, 경남신문, 경남일보, 경북도민일보, 경북매일신문, 경북일보, 경상일보, 경인일보, 광남일보, 광주매일신문, 광주일보, 국제신문, 금강일보, 기호일보, 남도일보, 대구신문, 대구일보, 대전일보, 동양일보, 매일신문, 무등일보, 부산일보, 새전북신문, 영남일보, 울산매일, 울산신문, 인천일보, 전남일보, 전라일보, 전북도민일보, 전북일보, 제민일보, 제주일보, 중도일보, 중부매일, 중부일보, 충북일보, 충청일보, 충청타임즈, 충청투데이, 한라일보
지역주간지	5	당진시대, 설악신문, 영주시민신문, 평택시민신문, 홍성신문
방송사	5	KBS, MBC, OBS, SBS, YTN
전문지	10	기자협회보, 디지털타임스, 미디어오늘, 소년한국일보, 시사 IN, 일요신문, 전자신문, 주간한국, 한겨레 21, 환경일보
스포츠신문	3	스포츠서울, 스포츠월드, 스포츠한국
인터넷신문	11	EBN, PD-저널, 노컷뉴스, 뉴스펭귄, 뉴스핌, 데일리안, 브레이크뉴스, 비즈워치, 쿠키뉴스, 프레시안, 헬로디디

자료: 빅카인즈, DS투자증권 리서치센터

저작권 존중하기

데이터를 이용하기 위해서는 사전에 저작권에 대한 검토가 필요

많은 언론사들을 한꺼번에 이용할 수 있는 만큼 빅카인즈를 활용하기 위해서는 저작권에 대한 존중이 필요하다. 다양한 언론사들의 데이터를 이용할 수 있는 만큼 저작권 침해될 수 있는 부분을 생각해야 한다. 언론사의 허락을 구하지 않고 무단으로 기사를 블로그나 SNS 등에 공유하는 경우가 흔하다. 뉴스를 타인에게 공유하는 것은 저작권 불법에 해당될 수 있다. 복제권과 공중송신권이 저작자에게 있기 때문이다.

저작권 보호대상의 범위

모든 기사들이 저작권 보호 대상은 아니다. 저작권법에 따르면 저작물은 저작자의 사상과 감정을 표현한 결과물을 일컫는다. 뉴스에 있는 기사들 중 의견이 담겨있는 기사들의 경우 저작자의 생각이 담겨있으므로 저작권의 보호를 받는다. 다만 단순 사실을 전달하기 위한 보도는 저작권의 보호 대상에 포함되지 않는다. 인사 동정, 부고 등과 같은 기사도 저작권의 보호 대상에 포함되지 않는다.

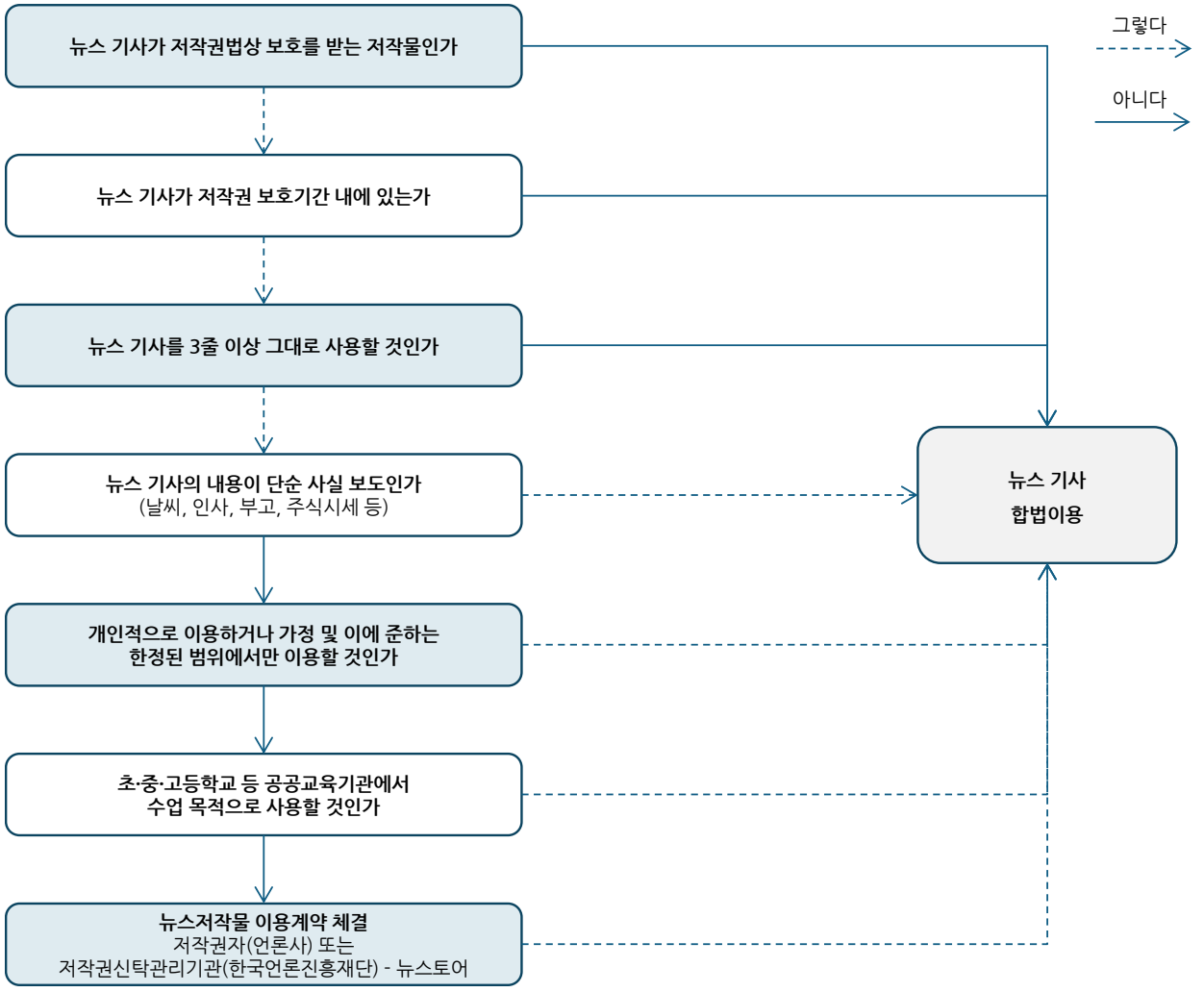
증권분야 기사는 사실을 전달하거나 사건을 육하원칙에 따라 정리하여 전달하는 경우가 많다. 사실을 전달하는 과정에서도 저작자의 의견이 포함될 수 있으므로 세부적인 내용 확인이 필요하다. 저작권의 보호를 받을 수 있는 내용이 있을 수도 있기 때문이다. 기사에서 촬영한 사진과 영상은 저작권법 보호 대상이다. 사건을 보도하는 과정에서 기자의 추측과 판단이 포함되어 있을 경우 마찬가지로 저작권법의 보호를 받을 수 있다. 이처럼 인간의 사상 또는 감정을 독자적으로 담고 있다면 저작권 보호 대상에 포함된다.

표5 뉴스 저작물 저작권 침해 대표유형

유형	내용
무단전재	언론사의 뉴스를 사전 동의나 허락 또는 계약 없이 무단으로 자신의 사이트에 게재 신문기사를 홈페이지, 블로그, 내부 인트라넷, SNS에 전재 업무상 목적으로 뉴스를 스크랩하여 다수의 사람들에게 배포
데이터베이스화	뉴스 저작물을 허락 또는 계약 없이 무단으로 수집하여 사내 데이터베이스를 구축 그리고 제3자에게 제공하는 경우
프레임 링크	자신의 사이트 프레임 내 언론사의 뉴스페이지나 목록페이지를 가두어두고 서비스하는 방식 다른 사이트의 내용을 자사 홈페이지 내용처럼 보이도록 연결
상업적 활용	뉴스 콘텐츠를 동의 없이 광고, 유료 서비스 등에 사용하여 수익을 창출하는 경우

자료: 한국언론진흥재단, 한국저작권위원회, DS투자증권 리서치센터

그림9 뉴스데이터 저작권 침해 여부 체크리스트



자료: 한국언론진흥재단, DS투자증권 리서치센터

뉴스토어는 뉴스 저작권을 손쉽게 구매할 수 있는 플랫폼

뉴스 데이터를 이용함에 있어서 저작권 문제를 간단하게 해결하는 방법은 뉴스토어(Newstore) 웹사이트를 활용하는 것이다. 뉴스토어를 통해 뉴스 저작권을 합법적으로 구매하여 이용할 수 있다. 뉴스 데이터 하나하나가 언론사에게는 무형자산이다. 빅카인즈에서는 전처리된 데이터를 이용자들에게 제공하고 있지만 뉴스 전체 내용이 담긴 데이터를 제공하지는 않는다. 뉴스 전문을 활용하기 위해서는 뉴스토어를 통해 저작권을 구입해야 한다.

뉴스토어를 통해 뉴스 저작권을 구입하여 활용할 수 있는 것은 한국언론진흥재단이 저작권을 수탁하기 때문이다. 저작권자에 해당하는 언론사는 한국언론진흥재단에 뉴스 데이터를 공급함과 동시에 저작권을 신탁한다. 뉴스이용고객이 뉴스토어를 통해 상품을 구매하면 그에 대한 저작권료를 뉴스 저작권 신탁관리업자인 한국언론진흥재단이 언론사에게 배분하는 구조로 이뤄져 있다.

뉴스 본문 전체를 사용할 경우 상품 구매가 필요할 것으로 예상

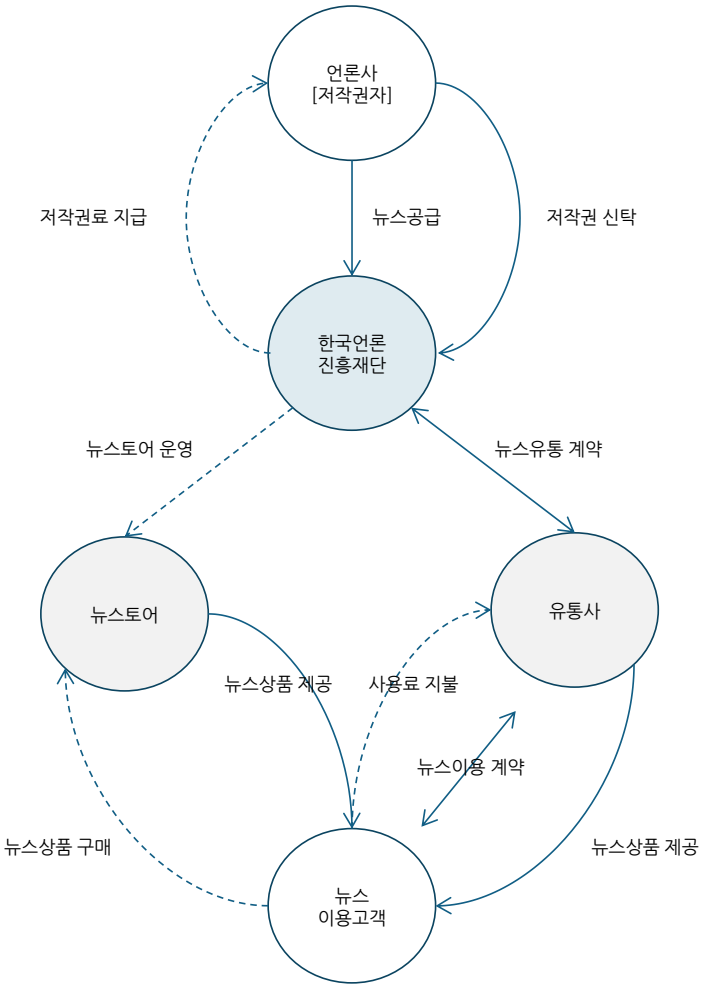
제안하는 모델은 빅카인즈에서 전처리한 데이터만을 바탕으로 추가적으로 데이터 가공작업을 거친 결과물이다. 뉴스 전문을 사용하지 않았다. 뉴스 본문 데이터 전부를 활용한다면 세부적인 문맥까지 반영한 모델로 고도화할 수 있을 것이다. 연관어 분석, 키워드랭킹, 오늘의 키워드 등과 같이 빅카인즈 플랫폼에서 정형화 작업을 거친 데이터를 이용하기 위해서는 해당되는 상품 구매가 필요하다.

표6 뉴스토어 제공 상품

상품 유형	설명
기사 단건	출판, 광고, 온라인게시 목적으로 뉴스를 활용하기 위한 목적의 상품 인쇄부수에 따른 가격정책 적용
보도사진	뉴스에 포함된 보도사진을 판매하는 상품
크리에이터	유튜버 또는 인플루언서와 같은 디지털 크리에이터를 위한 상품 구독자수에 따라 차등적인 가격정책 적용
라이선스	기업과 기관 관련 뉴스 전문을 사내외에 게시 및 발송할 수 있는 저작권 라이선스 해당 상품 구입시 사내 게시판이나 매일, 홈페이지에 뉴스게시 가능
뉴스데이터	뉴스 데이터베이스를 월단위로 구매하여 이용하는 상품
뉴스분석 API	한국언론진흥재단의 빅데이터 분석시스템을 이용 비정형데이터인 뉴스를 정형데이터로 만들어 API로 제공하는 상품
주문형	위 사항 중 고객이 필요로 하는 조건에 맞춰 제공하는 상품

자료: 뉴스토어, DS투자증권 리서치센터

그림10 뉴스 저작권을 둘러싼 생태계



자료: 한국언론진흥재단, DS투자증권 리서치센터

주: 뉴스이용고객은 국가기관 · 지방자치단체 · 공공기관 · 기업체 · 출판사 · 기타 단체 · 개인 등을 포함

빅카인즈 활용하기

빅카인즈의 자체 알고리즘으로 전처리한 텍스트 데이터를 활용할 차례이다. 빅카인즈 웹사이트에서 검색기능을 활용하여 원하는 데이터를 구하면 텍스트마이닝을 수행할 수 있다. 좀 더 정형화되고 풍부한 데이터가 필요할 경우 뉴스토어에서 데이터를 구입하거나 API를 구독하여 활용해도 좋다.

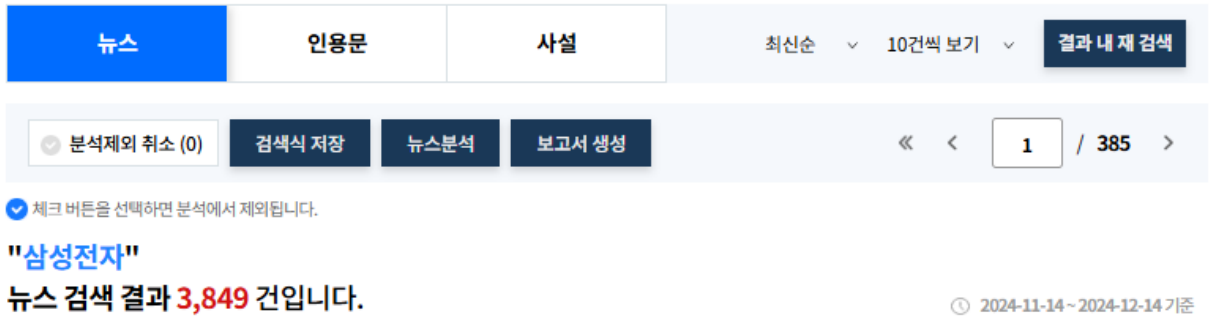
빅카인즈 기능 소개

"뉴스 분석"이라는 이름의 탭에서 "뉴스검색 · 분석"은 원하는 키워드에 따라 기사를 검색할 수 있는 기능을 제공하고 있다. 예를 들면 삼성전자를 최근 1개월의 기간으로 설정하여 검색하면 아래 그림과 같이 나타난다. 뉴스 검색 결과 아래에는 다양한 언론사들의 제목과 본문이 있다. 필요에 따라 언론사나 기사 분류 별로 범위를 설정하여 검색할 수 있다.

검색 결과 화면 밑에는 "분석 결과 및 시각화"라는 탭을 확인할 수 있다. 해당 탭을 확인해 보면 검색된 뉴스들을 빅카인즈에서 전처리한 데이터를 다운로드할 수 있다. 삼성전자라는 종목을 검색했을 때 나타난 기사마다 빅카인즈의 자체 알고리즘에 의해 기사 분류, 인물, 위치, 기관, 키워드 등과 같이 전처리되어 있음을 확인할 수 있다. 스프레드시트 형태로 된 데이터를 다운로드하여서 텍스트마이닝을 활용한 분석에 활용하면 될 것이다.

여기서 키워드로도 충분히 토픽모델링 결과를 낼 수 있지만 전처리 과정을 조금 더 거친다면 이전보다 정제된 결과를 기대할 수 있다. 키워드 이외에도 제목, 기사 분류, 인물, 위치, 기관 등과 같은 항목을 활용한다면 마찬가지로 토픽모델링의 성능개선에 도움 될 것이다.

그림11 빅카인즈 웹사이트 뉴스 검색결과 스크린샷



자료: 빅카인즈, DS투자증권 리서치센터

그림12 빅카인즈 뉴스 데이터 다운로드 화면

제목	통합 분류1	통합 분류2	통합 분류3	사건/사고 분류1	사건/사고 분류2	사건/
반도체·배터리·항공 업계, 美 견제·환율 급등에도 기댈 곳 없다 [尹 대통령 탄핵 가결]	경제>반도체	경제>무역	경제>외환			
"보조금 지연에 중국 추격까지"...대통령 리스크에 반도체?전자 업계 '먹구름' [탄핵가결]	경제>반도체	정치>정치일반	경제>외환			
잡스가 뿌린 AI 혁명의 씨앗 [AI오답노트]	IT_과학>모바일	IT_과학>보안	경제>외환	사고>산업사고>폭발		
[논현로] 민주적 의사결정 존중하는 사회	경제>경제일반	IT_과학>보안	경제>외환			
용인 서천동 'The 테라스 프라이빗 43' 분양	경제>부동산	경제>자동차	경제>외환			
국평 4억원대 반세권 새아파트 '용인 둔전역 에피트' 16일 무순위 청약 진행	경제>부동산	경제>취업_창업	경제>외환			
당근과 채찍으로 중국 견제하는 美 산업 정책[별별법]	경제>산업_기업	국제>중국	경제>무역	범죄>기업범죄>거래제한		
[VC's Pick]오픈AI가 픽한 유니콘...'스픽' 1100억 투자 유치	IT_과학>과학	IT_과학>모바일	경제>증권_증시			
대만 계엄 정국서 탄생한 TSMC, 지속성장 가능했던 이유 [기업&이슈]	경제>반도체	경제>증권_증시	경제>산업_기업			
사람·자원 넘쳐나는 '기회의 땅'에 몰린 반도체 기업들 "지금부터 공들여야" [헬로인디아]	경제>반도체	경제>취업_창업	경제>유통			
매경이 전하는 세상의 지식 (매-세-지, 12월 14일)	경제>무역	경제>증권_증시	경제>금융_재테크			
'게임 쇼크'에도 이 종목은 샀다...외국인들 쓸어 담았다는데 [한경우의 케이스스터디]	경제>증권_증시	경제>금융_재테크	경제>유통			
한 달 남은 삼성 갤럭시 언팩, 깜짝 공개 제품은	IT_과학>모바일	IT_과학>콘텐츠	경제>유통			
"한국미래가 암울하다는 증거"...유학인재 급감이 韓경제에 던지는 위기음 [★★글로벌]	국제>국제일반	경제>경제일반	경제>유통	사고>산업사고>화재		
북미펀드로 빠져 나간돈 10兆 넘었다	경제>증권_증시	경제>국제경제	경제>금융_재테크	사회>사회갈등>시위		
삼성전자 '더 프리미어 8K', 업계 최초 8K협회 표준 인증 획득	IT_과학>모바일	IT_과학>콘텐츠	IT_과학>보안			
공모주 투자는 필패?...상장 이후 주가 '주르륵'?...?증권사만 배불리나 [MONEY톡]	경제>증권_증시	경제>금융_재테크	경제>유통	범죄>기업범죄>내부자거래		
"음식 상할뻔 했는데 다행"...삼성전자 '오늘보장' 소비자 호평	IT_과학>모바일	경제>서비스_쇼핑	경제>유통			
[NNA] 글로벌 스마트폰 생산대수, 3Q 7%↑	경제>유통	IT_과학>모바일	경제>반도체			
후계자도 구매부문 전진배치...원가절감 힘주는 철강사	경제>산업_기업	경제>유통	경제>자동차			

자료: 빅카인즈, DS투자증권 리서치센터
주: 삼성전자를 예시로 검색하여 나타난 결과 화면 중 일부

정제된 결과를 위한 데이터 사전 작업

텍스트 전처리

오탈자와 필요 없는 데이터 제거는 전처리 과정에서 필수적인 과정

텍스트 전처리는 텍스트마이닝 성능에서 중요한 부분을 차지한다. 데이터를 수집했으면 데이터 전처리할 부분이 남아있는지 검토하고 추가적으로 정제하는 과정이 필요하다. 데이터 분석에 용이하도록 오탈자를 수정하거나 필요 없는 부분을 제거해야 한다. 토픽모델링에서 중요도가 떨어지는 텍스트 데이터들이 주된 제거대상이다.

텍스트데이터 분석에 불필요한 데이터를 제거하는 작업을 불용어 제거(Stopword removal)라고 한다. 불용어를 제거한 이후에도 데이터 클리닝(Data Cleaning) 작업은 계속된다. 텍스트마이닝에서의 데이터 전처리는 1) 토큰화(Tokenization), 2) 정규화(Normalization), 3) 불용어 처리(Stopping Words) 3가지가 대표적이다.

토큰화는 텍스트 데이터를 문장 또는 단어 단위로 나누는 작업

토큰화는 텍스트 데이터들을 토큰 단위로 나누는 작업이다. 문장 단위로 나눌 경우에는 문장 토큰화(Sentence tokenization)라고 하며 단어 단위로 나눌 경우 단어 토큰화(Word tokenization)라고 한다. 토큰화를 하고자 하는 언어마다 단어들을 구분하는 기준이 다르다. 각 언어에 적절한 토큰화 및 형태소 분석기를 활용한다. 또한 정규표현식(Regular Expression)을 활용한 토큰화도 있다. 텍스트 데이터에 내재된 패턴을 정규표현식으로 사전에 설정하여 단어 토큰을 추출할 수 있다.

정규화는 단어들을 통일되고 표준화된 단어로 변환하는 작업

정규화는 동일한 의미를 가졌지만 다른 형태로 표현된 단어들을 통일되고 표준화된 단어로 변환하는 작업이다. 여기서 정규화는 다시 어간 추출(Stemming)과 표제어 추출(Lemmatization) 2가지로 나뉜다.

어간 추출
: 단어에서 어형변화가 나타나지 않는 부분을 추출

어간 추출은 단어의 핵심부분을 추출하는 작업이다. 단어로부터 어미나 접사 등을 제거하여 분리한다. 여기서 핵심부분이라는 것은 단어가 변형되어도 핵심 의미가 변하지 않는 부분을 뜻한다. 한국어에서 “가다”를 예시로 들 경우 어간은 “가”에 해당한다. “가다”의 어형변화가 “간다”, “갔다”와 같이 시간에 따라 바뀔 수 있는데 “가”는 변하지 않았기 때문이다. “가”를 제외하고 나머지 바뀌는 부분을 어미라고 한다.

표제어 추출
: 사전에 등재된 단어의 기본형으로 변환

표제어 추출은 사전에 등재된 단어의 기본형(Lemma)으로 변환하는 작업을 말한다. 표제어 추출은 각 언어의 사전에서 의미적으로 독립적으로 정의된 기본형으로 변환한다. 단어의 어미를 단순하게 잘라내는 것과 구별된다. 어간 추출과 다르게 표제어 추출의 경우 문법적 의미를 고려하기 때문이다. “가다”를 다시 예로 들면 “갔다”, “가고 있다”, “가세요” 등과 같은 단어는 표제어 추출에서는 모두 기본형인 “가다”로 변환된다. 사전에 등제된 기본형은 “가다”이기 때문이다.

불용어 설정은 토픽모델링 성능에 도움되지 않는 단어들을 제거하는 작업

불용어(Stop words) 설정은 토픽모델 성능에 중요한 영향을 미친다. 불용어는 토픽모델링 분석에 도움 되지 않으며 문서의 주제나 의미 파악에 가치가 낮은 단어들을 뜻한다. 사전에 불용어를 합리적으로 설정하고 제거한다면 토픽모델의 성능이 개선될 수 있다. 불용어를 설정하지 않으면 토픽모델 결과물의 해석이 불명확해질 수 있으며 효율성이 떨어질 수 있다.

불용어를 설정하는 방식은 다양하다. 토픽모델링 결과물의 희소성(Sparsity)을 증가시키는 가장 간단한 방법은 빈도 기반 불용어 제거이다. 문서에서 지나치게 반복적으로 등장하거나 드물게 등장하는 단어들을 불용어로 설정할 수 있다. 수집한 문서에서 공통적으로 높은 빈도를 보인 단어들은 토픽모델링에서 큰 의미를 가지지 않을 가능성이 높다. 반대로 희소한 단어의 경우에도 분석에 토픽모델 해석에 기여하지 않는다면 불용어 목록에 추가하여 제거할 수 있다.

도메인 분야에 적합한 불용어 설정의 필요성

경우에 따라서는 특정 분야에서의 불용어를 설정하는 것도 중요하다. 기업, 사업, 거래와 같은 단어는 경제주식 분야의 뉴스에서 자주 등장하는 단어들이다. 하지만 이 단어들은 주제 파악에는 도움 되지 않을 수 있다. 기업과 사업의 경우에는 이미 구체적인 사명과 사업부가 언급되어 있을 수도 있기 때문에 토픽모델 결과에서 제거해도 큰 영향을 미치지 않을 수도 있다. 거래의 경우 주식거래를 의미하는 것인지 또는 고객과의 거래를 의미하는 것인지 해석이 모호한 경우가 있으므로 토픽모델 결과에서 제거를 고려할 수도 있다.

불용어를 설정할 때 중요한 단어가 불용어로 설정되지 않도록 신중하게 고려해야 한다. 지속적으로 불용어 목록을 업데이트하여 토픽모델의 성능을 높이는 것도 중요하다. 분석 과정에서 발견된 불필요한 단어들을 지속적으로 추가한다면 토픽모델의 성능은 개선될 것이다.

표7 대표적인 불용어(Stopwords) 예시

불용어 선정 이유	불용어 추가 방식	불용어 목록
단위를 나타내는 단어	불용어 목록에 추가	하나, 둘, 셋, 넷, 다섯, 여섯, 일곱, 여덟, 아홉, 열, 일, 이, 삼, 사, 오, 육, 일곱, 여덟, 아홉, 열, 십, 백, 천, 만, 억, 조, 월, 년, 평, 명, 개, 주, 위, 원, 시, 분, 초, 건, 호, 종, 심, 곳, 원어치, 차, 척, 대, 약, 선, 보, 좌, 우, 점, 피, 프로, 포인트, 배, 배수, 대략, 어치, 약, 분기, 반기, 연간, 연초, 연말
토픽모델링에 연관 없는 단어	불용어 목록에 추가	document, write, td, tr, nbsp, txt, html, gif, png, alt, ad, object, var, ent, embed, banner, head, index
토픽모델링에 중요도가 떨어짐	정규표현식	re.sub('₩b₩w*(자산운용|투자연구소|스탁론|stockloan|상담센터|투자증권|증권|리서치|센터장|전략팀장|부문장|본부장|기자|신문|품질보증제)₩b','',text)
주식/경제와 연관성이 떨어짐	불용어 목록에 추가	상담, 무료, 유료, 무료공개, 유료공개, 상담원, 전문상담원, 운세, 사주, 연말정산, 취재, 인터뷰, 발언, 언급, 쿠폰, 증정, 사은품, 경품, 추첨, 상품권, 검색, 기념, 할인, 공짜
토픽모델링에 연관 없는 대명사, 부사	불용어 목록에 추가	매우, 정말, 아주, 너무, 더, 가장, 즉, 따라서, 굉장히, 등, 때문, 위해, 통해, 대해, 의해, 따라, 대한, 같은, 이런, 저런, 그동안, 이번, 저번, 다음, 그녀, 무엇, 그, 이렇게, 저렇게, 이날, 어제, 그저께, 사흘, 나흘, 이전에, 은, 는, 도, 만, 까지, 조차, 나, 는, 은, 던, 곧

자료: DS투자증권 리서치센터

정보 축약하기

행렬분해는 단순하면서 강력한 방식

행렬분해 방식의 토픽모델링

핵심 정보만을 추려내는 방법 중에서 행렬분해 기반의 방식(Matrix Factorization Approach)이 있다. 행렬분해 기반의 토픽모델링은 행렬 연산을 통해 뉴스 기사와 단어 간의 관계를 학습하는 방식을 취한다. 문서-단어 행렬을 낮은 차원으로 분해하여 잠재 요인을 찾아내는 과정을 거친다. 행렬분해 기반 방식은 텍스트 데이터에 명확한 분포나 패턴이 보이지 않더라도 유연하게 적용할 수 있다.

비음수행렬분해 : 정보를 축약하기 위해 2개의 행렬로 분해

여기서는 비음수행렬분해(Non-negative Matrix Factorization, NMF)에 기반한 방식을 적용했다. NMF는 문서-단어 행렬의 모든 요소들이 음수(-)가 아니라는 (Non-negative) 제약조건 하에서 문서-단어 행렬을 문서-토픽 행렬과 토픽-단어 행렬로 분해하는 과정이다. 문서-토픽 행렬에서 각각의 요소는 각 문서가 토픽들과 어떤 관계를 가지는지 양(+)의 값으로 표현된다. 토픽-단어 행렬에서도 마찬가지이다. 수집한 뉴스 기사들을 모은 데이터를 2개의 행렬로 분해하는 과정에서 정보의 축약이 나타난다. 정보가 축약되면서 특정 토픽과 강한 연관성을 보인 단어만 남게 된다.

그림13 행렬분해 기반 토픽모델링 도식화

문서1 & 단어1	문서1 & 단어2	문서1 & 단어3	문서1 & 단어4	문서1 & 단어5
문서2 & 단어1	문서2 & 단어2	문서2 & 단어3	문서2 & 단어4	문서2 & 단어5
문서3 & 단어1	문서3 & 단어2	문서3 & 단어3	문서3 & 단어4	문서3 & 단어5
문서4 & 단어1	문서4 & 단어2	문서4 & 단어3	문서4 & 단어4	문서4 & 단어5
문서5 & 단어1	문서5 & 단어2	문서5 & 단어3	문서5 & 단어4	문서5 & 단어5

=

문서1 & 토픽1	문서1 & 토픽2	문서1 & 토픽3
문서2 & 토픽1	문서2 & 토픽2	문서2 & 토픽3
문서3 & 토픽1	문서3 & 토픽2	문서3 & 토픽3
문서4 & 토픽1	문서4 & 토픽2	문서4 & 토픽3
문서5 & 토픽1	문서5 & 토픽2	문서5 & 토픽3

×

토픽1 & 단어1	토픽1 & 단어2	토픽1 & 단어3	토픽1 & 단어4	토픽1 & 단어5
토픽2 & 단어1	토픽2 & 단어2	토픽2 & 단어3	토픽2 & 단어4	토픽2 & 단어5
토픽3 & 단어1	토픽3 & 단어2	토픽3 & 단어3	토픽3 & 단어4	토픽3 & 단어5

자료: DS투자증권 리서치센터

주: 5개의 문서와 5개의 단어집합으로 이루어진 데이터를 예시로 들어 간단하게 도식화

NMF 토픽모델링 소개

NMF 토픽모델링 소개

NMF(Non-negative Matrix Factorization) 토픽모델링은 비음수행렬분해를 통해 여러 문서들 속에 잠재된 주요 주제를 도출하는 방법이다. 아래 간단한 수식으로 표현할 수 있다.

$$\mathbf{V} \approx \mathbf{W} \times \mathbf{H}$$

행렬 **V**는 "뉴스 기사 × 단어"로 구성된 데이터로 문서-단어 행렬을 의미한다. 행렬 **W**는 "뉴스 기사 × 주제"로 구성된 데이터로 문서-토픽 행렬을 나타내며 각각의 문서가 특정 토픽에 얼마나 연관되어 있는지 의미한다. 행렬 **H**는 "주제 × 단어"로 구성된 데이터로 토픽-단어 행렬을 의미하며 각 단어가 특정 토픽에 얼마나 연관되었는지 수치로 나타낸다.

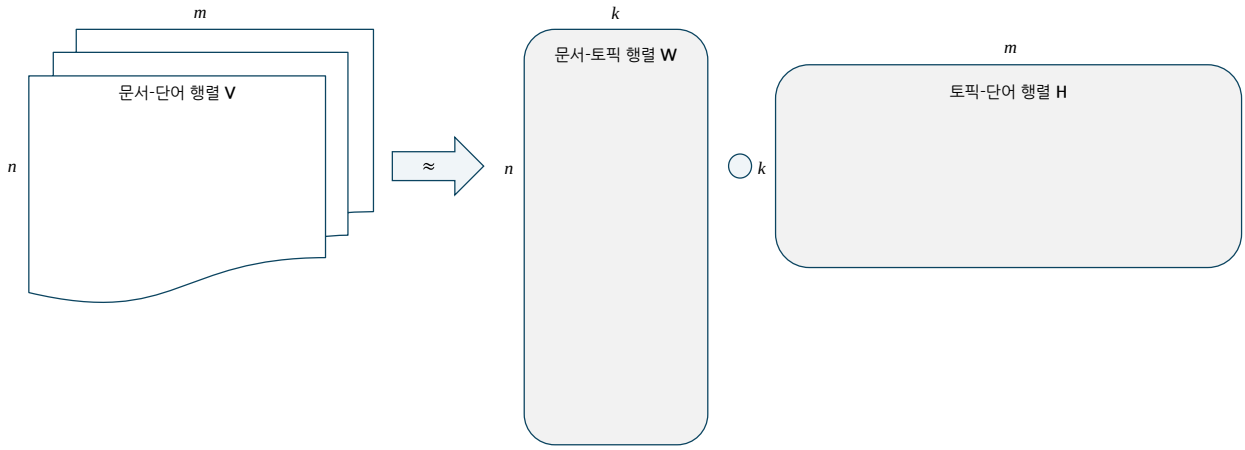
NMF 토픽모델링에서는 뉴스 기사 하나를 단어들이 모인 하나의 거대한 행렬이라고 가정한다. 이후 2개의 행렬로 분해하는 과정을 거친다. 이때 비음수분해의 제약조건이 적용되는데 각 행렬요소가 0 이상의 값을 지니도록 규칙이 적용된다. 이는 단어의 빈도수나 특정 토픽과의 관련도를 직관적으로 해석할 수 있도록 돕는다.

NMF의 주요 목적은 원래 행렬의 정보를 가장 잘 보존하면서 축약하는 것

분해된 2개의 행렬 **W**와 **H**를 서로 곱한 수치와 원래 행렬을 비교하면 완벽하게 일치하지 않는다. 2개의 행렬로 분해되는 과정에서 일부 정보는 손실되어 정보가 축약된다. 문서와 단어의 관계가 중요한 부분만 남게 되므로 데이터를 압축적으로 요약하는 효과가 발생한다.

곱셈과정은 원래 행렬 **V**의 정보를 가장 잘 근사(Approximate)하기 위한 과정으로 볼 수 있다. NMF 토픽모델 상에서 최종적인 결과가 도출될 때까지 2개 행렬의 곱셈연산이 수학적인 알고리즘을 통해 반복적으로 수행된다. 오차($||\mathbf{V} - \mathbf{WH}||$)를 최소화하는 방향으로 학습한다.

그림14 NMF 토픽모델 도식화



자료: Egger, R., & Yu, J. (2022), DS투자증권 리서치센터

주: 여기서는 토픽의 개수가 k개로 불명확함. 사전에 정하거나 교차검증 등을 통해 최적화된 토픽 개수인 k를 선택하는 과정이 필요

NMF 토픽모델링의 장점

NMF 토픽모델링의 장점
1) 빠른 계산 및 학습
2) 해석의 용이성

첫 번째 장점은 빠른 계산과 학습이다. 구현난이도가 토픽모델링 중에서는 어렵지 않은 편에 속한다. Python에서 sklearn 라이브러리 내에서 구현이 가능하다. 투입시간 대비하여 효율적인 결과를 가져다줄 수 있다. 또한 행렬 기반의 연산이기에 복잡한 추론 과정을 거치지 않는다. 대용량의 데이터를 학습하기 위해 GPU가 반드시 필요하지는 않다.

두 번째는 해석하기 용이하다는 점이다. 비음수행렬분해 연산과정에서 가중치들이 모두 0 이상의 수치를 지니기 때문에 토픽모델링 결과가 상대적으로 명확하게 해석된다. 추가적으로 이미지나 오디오 데이터 등에도 활용할 수 있다. 컴퓨터 입장에서 행렬로 표현할 수 있는 데이터라면 다른 분야에도 적용할 수 있다.

종합적으로 계산복잡도가 낮은 행렬연산을 통해 효율성과 해석의 용이성 2가지가 NMF 토픽모델의 주요 장점이다. 20 Newsgroups 데이터셋을 대상으로 여러 토픽 모델 성능을 비교한 연구에서 NMF 토픽모델링은 효율성 측면에서 우수한 성능을 보인 바 있다.

표8 NMF 토픽모델은 20 Newsgroup 데이터셋 성능비교에서 효율적인 성능을 보임

토픽 모델	모델링 방법 및 주요 특징	일관성	다양성	안정성	실행 시간(초)
LSI (LSA, Latent Semantic Indexing/Analysis)	행렬분해 중 특이값분해(SVD)를 사용하여 데이터를 단순화하여 단어 간의 숨겨진 관계 파악	0.013	0.481	1.000	3.252
NMF (Non-negative Matrix Factorization)	비음수행렬분해(NMF)를 통해 데이터의 모든 값을 양수로 취급하여 해석을 쉽게 함	0.082	0.695	0.816	4.880
LDA (Latent Dirichlet Allocation)	디리클레 분포를 사용하여 확률 모델링. 각 문서의 주제 비율을 계산	0.058	0.711	0.726	8.057
HDP (Hierarchical Dirichlet Process)	사전에 주제 개수를 정하지 않음 계층적 구조로 새로운 주제를 유동적으로 생성	-0.160	0.626	0.894	24.472
ETM (Embedded Topic Model)	인공신경망 구성하여 단어들 간의 복잡한 관계를 탐색함	0.040	0.378	0.894	45.977
CTM (Correlated Topic Model)	토픽 간의 관계를 고려하여 주제들이 서로 어떤 상관관계에 있는지를 파악함	0.075	0.890	0.697	23.981
ProdLDA (LDA with Product of Experts)	LDA 와 유사하지만 좀 더 복잡한 주제 구조를 표현 신경망 구성하여 다양한 주제를 탐색	0.055	0.879	0.627	27.308

자료: Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., & Hassan, A. (2023), DS투자증권 리서치센터
주: 소수점 넷째 자리에서 반올림한 수치

유사도에 기반한 종목 매칭

코사인 유사도

코사인 유사도
: 데이터 간의 유사성을 측정하기 위한 지표

코사인 유사도(Cosine Similarity)는 2개의 벡터가 얼마나 유사한지 측정하는 지표이다. 코사인 유사도는 문서 간의 텍스트 유사도를 분석할 때 주로 활용된다. 각도를 나타내는 지표가 유사성을 측정하는 지표가 될 수 있는 이유는 벡터가 방향성을 가지고 있기 때문이다.

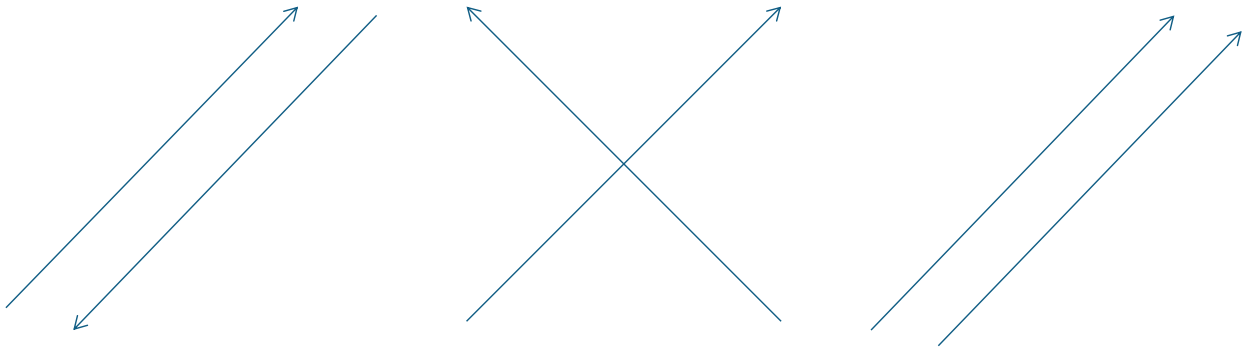
방향성이 서로 반대라면 코사인 유사도는 -1의 값을 지닐 것이다. 반대로 같은 방성이라면 코사인 유사도는 1의 값을 가질 것이다. 2개의 벡터가 서로 직각 방향으로 진행될 경우(Orthogonal) 코사인 유사도는 0의 값을 가질 것이다. 0의 코사인 유사도는 서로 관련이 없다는 의미로 해석할 수 있다. 코사인 유사도를 계산하는 수식은 아래와 같다.

$$\cos\theta = \frac{\mathbf{A}\cdot\mathbf{B}}{||\mathbf{A}||\,||\mathbf{B}||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

코사인 유사도는 각도의 개념이다. 코사인(Cosine)을 나타내는 기호에 각도를 나타내는 θ(Theta)을 확인할 수 있다. 직각삼각형을 기준으로 빗변과 밑변의 비율로 계산된다. 코사인을 측정하는 방법은 여러 가지 존재하기 때문에 코사인 유사도는 코사인을 측정하는 것과 다르지 않다. 텍스트마이닝 분야에 활용되면서 코사인을 측정하는 것을 유사도로 해석한 것으로 추정된다.

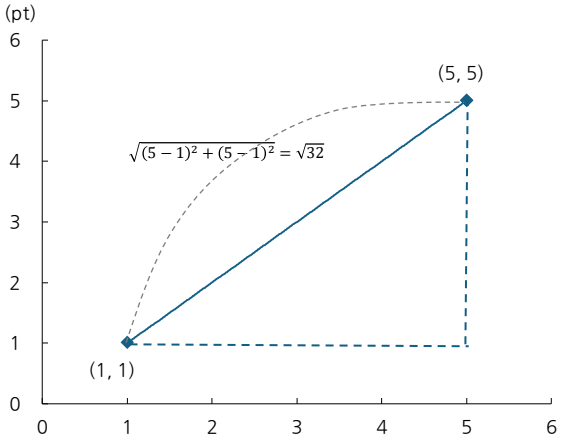
A와 **B**는 각각 벡터를 나타낸다. **A**와 **B**는 두 벡터의 내적(Dot product)을 의미한다. 내적(Dot product)은 벡터를 구성하는 각각의 요소들끼리 서로 곱한 후 모두 더한 결과이다. ||**A**||와 ||**B**||는 각각 벡터의 유클리드 노름(Euclidian Norm)을 나타낸다. 2차원 평면에서 한 점에서 다른 점까지의 거리는 2차원 유클리디안 거리(Euclidian Distance)라고 정의한다. 2차원 평면에서 거리를 계산하는 방법은 피타고라스 정리를 활용하는 것이다. 유클리드 노름은 이를 일반화한 개념이라고 볼 수 있다. 2차원 평면에서 두 점 사이의 거리를 계산했던 체계를 N차원 평면으로 확장하면 유클리드 노름이 계산된다.

그림15 코사인 유사도에 따른 2개 벡터 간의 방향성



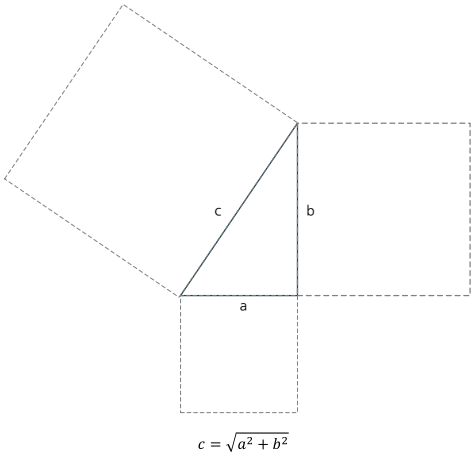
자료: 딥러닝을 이용한 자연어 처리 입문(2024), DS투자증권 리서치센터

그림16 2차원 유클리디안 거리 예시



자료: DS투자증권 리서치센터

그림17 2차원 유클리디안 거리는 피타고라스 정리와 유사



자료: DS투자증권 리서치센터

종목 매칭하기

매매 의사결정에 도움되기 위해 기사와 종목 매칭을 수행

텍스트마이닝의 결과물이 요약으로만 그치면 아쉽다. 결국에는 무엇을 매매하면 좋을지 결론까지 도출할 수 있는 정보가 필요하다. 이를 위해 필요한 작업은 텍스트데이터들과 종목에 대한 정보를 서로 비교했을 때 유사도가 가장 높은 투자후보를 추정하는 것이다. 여기서는 국내주식이라는 투자자산을 주로 다룰 것이기에 투자 후보군을 종목으로 용어를 통일하여 서술할 예정이다.

주식·경제 분야의 뉴스데이터들을 전반적으로 봤을 때 나타나는 데이터들의 특성이 하나 있다. 투자대상을 직접적으로 언급한다는 것이다. 뉴스에서 사건을 서술할 때 어떤 회사인지 직접적으로 언급하지 않는다면 정보 전달이 어려울 것이다. 사건이 발생했을 때 주체와 객체가 명확하기 때문에 기사에 있는 내용과 종목 이름을 서로 매칭하는 것으로 충분하다고 판단했다. 관련된 키워드까지 포함하여 종목을 매칭시킨다면 좀 더 정교한 결과가 나타날 것으로 기대한다.

뉴스 하나를 볼 때도 여러 개의 종목들이 직관적으로 생각나는 경우가 있다. 뉴스 기사 1개가 반드시 1개의 종목만 매칭되어야 하는 법은 없다. 뉴스 기사 별로 최대 5개의 연관된 종목이 유사도가 높은 순으로 매칭되도록 설정했다. 시가총액 1,000억 미만의 종목의 경우에는 매칭되지 않도록 추가했다.

축약어, 별칭, 브랜드명과 종목을 매칭시키기 위해 추가적인 리스트 필요

몇몇 종목의 경우에는 제목에 해당 종목을 나타내는 별칭 또는 축약어를 쓰는 경우가 있다. 키워드에서도 특정 종목을 나타내는 별칭과 축약어가 있기에 관련된 단어가 나타나면 종목명을 일치시킬 수 있도록 리스트를 추가했다. 아래 표는 그에 대한 예시를 일부 발췌한 것이다.

표9 축약어 또는 별칭에 따른 종목명 매칭 추가 규칙

별칭 또는 축약어	종목명
청정원, 쉐프원, 미원	대상
아프리카티비, 아프리카TV	SOOP
네이버	NAVER
국민은행, KB은행	KB금융
삼전	삼성전자
하닉	SK하이닉스
iM뱅크, iM증권	DGB금융지주
부산은행, BNK투자증권	BNK금융지주
동부화재	DB손해보험
신행은행, 신한증권, 신한금융투자, 신금투, 신한투자증권	신한지주
하나증권, KEB하나은행, 하나금융, 하금투, 하나은행	하나금융지주
LGES, LG엔솔, 엘지엔솔	LG에너지솔루션
SKIET	SK아이이테크놀로지
KAKAO, kakao	카카오
포스코	POSCO홀딩스

자료: 빅카인즈, DS투자증권 리서치센터

압축적 의사결정을 위한 국내 증시 우회도로

토픽모델 평가

토픽 개수의 범위는 너무 적거나 많지 않게 설정

해석의 용이성 및 계산 시간을 고려하여 토픽 개수 범위 설정

정기적으로 토픽모델의 결과가 잘 나타나고 있는지 확인하기 위해서는 토픽모델의 성능을 평가할 필요가 있다. 모델의 실행시간을 고려하여 토픽이 나타날 수 있는 범위는 5개에서 35개까지 설정했다. 1주일 동안 나타난 이벤트를 너무 적은 개수로 주제를 분류할 경우에는 각각의 토픽 내 배정된 기사들이 어떤 주제를 가지고 있는지 해석하기 어려워지는 단점이 있다.

토픽모델링 결과에서 1개의 토픽 내에 다양한 유형의 기사가 포함될 확률이 있다. 실제로 약 30개의 토픽을 고정시켜 토픽모델을 실행했을 때도 여러 섹터의 기사들이 모여 하나의 토픽으로 형성된 경우가 있었다. 국내 증시를 너무 단순하게 해석하고 적은 수의 토픽을 고집하기에는 투자 기회를 포착할 수 있는 범위가 좁아지는 한계점이 존재할 것이라고 판단했다.

반대로 너무 많을 경우에는 어떤 주제에 집중해야 하는지 어려움이 존재할 수 있다. 또한 종목별로 상대적으로 같은 숫자의 기사가 관찰되더라도 영향력이 다르다. 이를 해결하는 방법은 뉴스 데이터와 연관된 종목들의 거래대금을 같이 보면 해당 문제는 완화될 것으로 판단했다.

평가지표 소개 및 최적의 토픽개수 선정

토픽모델링의 결과를 평가하기 위해서는 정량적인 평가와 정성적인 평가 모두 고려해야 한다. 정량적인 평가의 경우 토픽의 일관성 지표를 우선적으로 고려했다.

토픽의 일관성
: 상위 키워드들이 얼마나 의미 있게 일관성을 유지하는지 평가

토픽의 일관성(Coherence Score)은 각 토픽 내에서 상위 10개에서 20개의 키워드들이 얼마나 의미 있게 일관성을 유지하는지 평가하는 지표이다. 0부터 1까지의 범위를 지니며 1에 가까울수록 토픽 내 단어들이 서로 의미적으로 일관적으로 평가된다. 토픽모델링의 품질을 평가하는 대표적인 지표이다. 측정하는 방법론은 다양하다. UMass Coherence, C_v Coherence, UCI Coherence 등 토픽의 일관성을 측정지표들이 있다. 여기서는 C_v Coherence 측정지표를 사용했다.

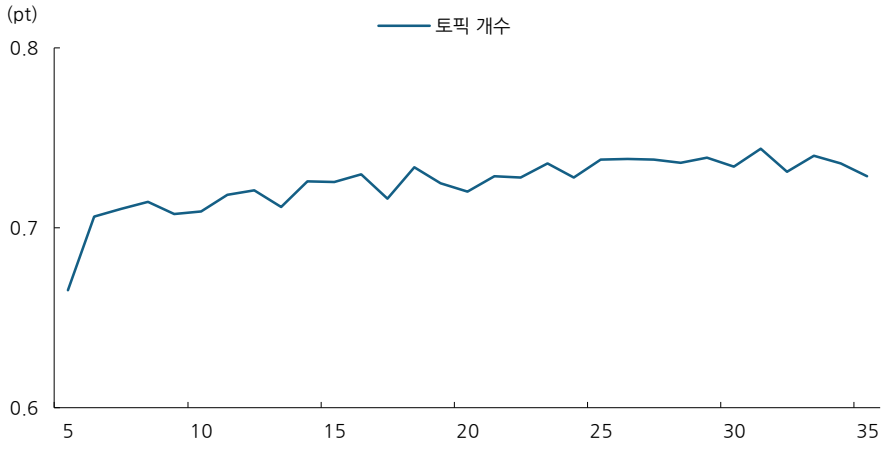
재구성 오차
: 원본 데이터 행렬을 얼마나 재구성을 가깝게 했는지 측정

그 외에 재구성오차 (Reconstruction Error)지표도 고려했다. 재구성 오차는 NMF 토픽모델이 원본 데이터 행렬을 얼마나 재구성을 가깝게 했는지 측정하는 지표이다. 모델의 데이터를 얼마나 잘 설명하고 있는지 평가한다. 재구성오차가 낮을수록 모델이 원본 데이터의 핵심적인 부분을 보존하면서 설명하고 있다고 평가된다.

금융 분야에서 지속적으로 활용되기 위해서 중점적으로 평가해야 하는 특성은 일관성이다. 2가지의 지표를 중점적으로 고려했으며 일관성 점수에 가장 높은 가중치를 부여했다. 정량적인 지표만으로는 토픽모델이 실용적으로 활용할 수 있는지 평가하지 어려운 면이 있다. 전문가나 일반인들의 정성적인 평가를 통해 토픽모델링의 결과물을 검토해야 한다.

최적의 토픽 개수를 선정하기 위해 토픽 5개부터 35개까지 순차적으로 NMF 토픽모델을 학습시켜서 일관성 지표(Coherence Score)를 추정했다. 31개의 토픽의 일관성이 0.744이 계산되었고 가장 높은 수치를 기록했다. 따라서 최적의 토픽 개수를 31개로 선정하고 최종적으로 NMF 토픽모델링을 수행했다.

그림18 일관성 점수에 기반한 최적의 토픽 개수 선정



자료: DS투자증권 리서치센터

표10 참고: 토픽 일관성 측정지표 종류

종류	값의 범위	공식	특징	어울리는 데이터 유형
UMass	음(-)의 무한대부터 0 까지	$C_{UMass}(W,D) = \sum_{i<j} log \frac{P(w_i,w_j)+\epsilon}{P(w_i)}$	공출현도 기반으로 직접적인 단어관계를 평가 0 에 가까울 수록 일관성이 좋다고 여겨짐	특정 단어들이 직접적으로 언급되는 짧은 문서들에 적합
C_v	0 부터 1 까지	$C_v = \frac{\sum_{i<j} log\left(\frac{P(w_i,w_j)}{P(w_i)}\right)}{\sum_{i<j} \log P(w_i,w_j)}$	공출현 및 의미적 유사성까지 평가 1 에 가까울수록 일관성이 우수	뉴스 기사, 서적, 논문 등 긴 문서 데이터에 적합
UCI	음(-)의 무한대부터 양(+)의 무한대까지	$C_{UCI} = \sum_{i<j} log\, P(w_i,w_j)$	슬라이딩 윈도우 기반으로 공출현도 계산 값이 클수록 공출현이 많음을 의미	긴 문서 데이터에 적합

자료: Röder, M., Both, A., & Hinneburg, A. (2015), Towards Data Science, DS투자증권 리서치센터

군중의 관심도 파악

거래대금 속에는 군중의 관심이 스며들어 있다

종목과 관련된 기사의 개수 뿐만 아니라 거래대금도 고려할 필요가 있음

종목과 관련된 기사의 개수를 단순하게 매칭시켜 쓰기에는 아쉬운 점이 있다. 대형주라면 당연히 언론의 입장에서 중요할 수밖에 없으니 기사 수가 많을 수밖에 없다. 대형주를 분석하는 증권사의 입장에서도 중요하게 다룰 수밖에 없다. 그렇다면 언론의 관심뿐만 아니라 투자하는 사람들의 관심도를 반영하는 방법을 찾아야 한다.

그에 대한 해답은 거래대금에 있다. 시장에서 거래된 종목의 가격과 수량을 모두 고려한 지표는 거래대금이다. 이것을 활용하면 종목에 대한 관심도를 반영할 수 있다.

종목의 거래대금이 기록되기까지의 과정은 다양한 유형의 투자자들의 매매에서 나타난다. 거시경제에 대한 정보와 국가/산업에 대한 정보를 기반으로 판단하는 Top-down과 기업에 대한 정보에 주목하는 Bottom-up의 경로로 나타난다. 물론 상관없이 투자하는 사람들도 존재할 수 있다.

기업의 실적은 매 분기별로 공시된다. 실적과 상관없이 투자하는 사람들도 존재할 수 있지만 기업의 실적을 어느 정도 반영하는 투자자들은 기업의 실적 발표 전후로 이용 가능한 정보들을 취합하여 판단한다. 기업의 실적이 좋고 업황이 긍정적이라면 주가는 우상향하는 특성을 가질 수는 있다. 하지만 투자자들에게 증명되기 전에 변동성이 발생하는 경우가 종종 발생한다. 변동성의 과거 흔적은 가격과 거래대금에 녹아있다. 주식시장에서 거래되는 순간 고스란히 기록되기에 가장 빠르게 확인할 수 있는 데이터이다. 반대로 한국 경제에 대해 긍정적인 경제지표가 발표되거나 한국 경제의 미래에 긍정적이라는 시각을 가진다면 투자하는 과정에서 특정 종목들에서 거래가 나타날 수 있다.

주간누계거래대금 대비 일평균거래대금 비율로 공정하게 반영

시가총액과 상관없이 적용할 수 있는 상대거래강도 지표 소개

거래대금도 그대로 활용하면 문제가 있다. 대형주와 중소형주에 나타난 관심도를 공정하게 평가할 수가 없다. 대형주는 당연히 대형주에 부합하게 거래대금이 매우 큰 규모로 나타날 것이다. 소형주의 경우 대형주에 비해 작은 시가총액 규모로 인해 거래대금이 작게 나타날 확률이 높다. 거래대금이 많이 발생하더라도 대형주와 소형주에서 느끼는 체감이 다르다.

예를 들면 1조원의 거래대금이 300조원의 시가총액을 기록하고 있는 종목에서 발생하는 것과 900억원의 시가총액을 기록하고 있는 종목이 체감하는 것은 매우 상이할 것이다. 대형주와 소형주 가리지 않고 공평하게 비교하는 방법은 거래대금에 있다. 이를 가중치로 반영한다면 기업별로 관심도를 국내에 상장된 기업들의 시가총액 규모와 상관없이 비교할 수 있을 것이다.

5일누계거래대금을 일평균거래대금으로 나누어서 일종의 거래강도 지표를 활용하면 된다. 이를 반영한 수식은 아래와 같다. 상대거래강도라는 명칭은 공식적으로 있는 용어가 아니기에 해석에 유의해야 한다.

$$상대거래강도 = \frac{\sum_{i=1}^{5} 거래대금_{T-i}}{일평균거래대금_{YTD}}$$

5일누계거래대금을 나타내는 $\sum_{i=1}^{5} 거래대금_{T-i}$는 최근 5거래일 동안 누적되어 거래된 금액을 나타낸다. 일평균거래대금$_{YTD}$은 연초부터 최근 날짜까지 거래대금 누적금액을 누적거래일수로 나눈 값이다. 종목의 유동성과 거래 활성도를 종합적으로 평가하는데 사용된다. 상대거래강도를 활용하면 종목의 시가총액 크기와 상관없이 해당 종목이 과거 평균적인 상황에 비해 거래가 많이 발생했는지 확인할 수 있다.

비교를 위해 5일누계거래대금 가장 많이 발생한 종목 20개와 상대거래강도 순으로 나열한 종목 20개를 정리하면 다음과 같다. 2024년 12월 16일 종가를 기준으로 나타난 결과이다.

표11 5일누계거래대금 기준 상위 20개 종목

Code	Name	5 일누계거래대금(억원)	일평균거래대금(억원)	시가총액(억원)	TTM P/E	12MF P/E	TTM P/B	WICS 업종명(대)
A005930	삼성전자	54,820	15,625	3,235,622	11.5	10.3	1.0	IT
A000660	SK 하이닉스	30,020	7,901	1,339,524	12.8	5.2	1.9	IT
A035420	NAVER	14,125	1,551	331,926	20.2	18.9	1.3	커뮤니케이션서비스
A196170	알테오젠	11,799	2,823	160,756	-5488.5	680.3	90.3	건강관리
A328130	루닛	11,488	453	24,254	-108.3	-51.4	10.4	건강관리
A105560	KB 금융	8,598	1,254	332,138	7.3	6.2	0.6	금융
A007660	이수페타시스	8,364	1,234	16,887	27.1	13.0	5.4	IT
A068270	셀트리온	7,989	1,368	419,396	207.0	33.6	2.4	건강관리
A035720	카카오	7,528	723	190,067	-21.2	42.2	1.9	커뮤니케이션서비스
A000100	유한양행	7,125	1,961	90,476	54.7	49.1	4.0	건강관리
A373220	LG 에너지솔루션	6,919	1,009	895,050	-246.8	76.4	4.4	IT
A005490	POSCO 홀딩스	6,685	1,597	218,128	16.9	9.9	0.4	소재
A267260	HD 현대일렉트릭	6,643	1,244	140,043	27.4	20.3	10.1	산업재
A087010	펩트론	6,527	648	24,555	-107.4		104.7	건강관리
A012450	한화에어로스페이스	6,481	1,485	139,251	25.0	14.7	4.6	산업재
A001570	금양	6,379	542	16,440	-9.7		18.3	소재
A065350	신성델타테크	6,151	996	28,254	-633.6		13.7	경기관련소비재
A010130	고려아연	5,891	641	230,221	36.8		2.5	소재
A034020	두산에너빌리티	5,878	1,318	112,034	136.3	25.2	1.5	산업재
A005380	현대차	5,633	2,350	432,444	4.5	4.2	0.5	경기관련소비재

자료: QuantiWise, DS투자증권 리서치센터

주: 시가총액 1,000억 미만의 종목 제외

표12 상대거래강도 기준 상위 20개 종목

Code	Name	상대거래강도	일평균거래대금(억원)	시가총액(억원)	TTM P/E	12MF P/E	TTM P/B	WICS 업종명(대)
A053800	안랩	46.7	84.2	8,166.90	21.87		1.99	IT
A293490	카카오게임즈	43.8	58.72	15,697.20	-6.26	28.21	1.05	커뮤니케이션서비스
A017800	현대엘리베이	26.1	40.1	22,165.40	33.28		1.74	산업재
A328130	루닛	25.3	453.2	24,254.20	-108.33	-51.42	10.38	건강관리
A950160	코오롱티슈진	22.1	63.83	20,740.70	-103.73		13.93	건강관리
A001530	DI 동일	21.6	30.05	10,888.10	200.44	59.75	1.69	경기관련소비재
A002840	미원상사	20.7	4.66	9,143.80	17.41		2.29	소재
A377300	카카오페이	20.6	152.96	39,580.50	-201.19	176.75	2.13	IT
A000240	한국앤컴퍼니	19.6	23.53	18,977.60	5.12	4.18	0.44	경기관련소비재
A381970	케이카	18.4	12.42	6,933.40	17.57	13.7	3.02	경기관련소비재
A052020	에스티큐브	17	22.39	5,552.80	-18.23		22.52	IT
A263750	펄어비스	15.4	128.27	18,182.10	190.86	13.79	2.33	커뮤니케이션서비스
A004800	효성	14.9	23.69	8,738.50	3.22	13.74	0.44	산업재
A005250	녹십자홀딩스	14.3	11.52	8,347.50	-12.93		0.88	건강관리
A084110	휴온스글로벌	14.2	15.68	5,219.00	14.3		0.97	건강관리
A302440	SK 바이오사이언스	14	64.48	41,680.80	-94.44	-38.03	2.38	건강관리
A014620	성광벤드	12.5	90.41	6,563.20	17.27	11.74	1.21	산업재
A376300	디어유	12.4	53.24	9,210.50		22.32	4.87	커뮤니케이션서비스
A241560	두산밥캣	12.2	206.61	41,854.00	6.38	7.11	0.66	산업재
A001570	금양	11.8	541.61	16,439.60	-9.74		18.25	소재

자료: QuantiWise, DS투자증권 리서치센터

주: 시가총액 1,000억 미만의 종목 제외. 상대거래강도 = 5일누계거래대금 / 일평균거래대금

표13 5일누계거래대금 & 상대거래강도 기준 상위 종목 기간 수익률

종목명	1 주전대비 수익률 (%)	1 개월전대비 수익률 (%)	3 개월전대비 수익률 (%)	종목명	1 주전대비 수익률 (%)	1 개월전대비 수익률 (%)	3 개월전대비 수익률 (%)
삼성전자	0.4	1.3	-15.8	안랩	-3.7	16.9	38.2
SK 하이닉스	8.0	3.3	13.0	카카오게임즈	8.4	22.4	10.1
NAVER	0.2	10.3	30.9	현대엘리베이	16.1	9.7	31.4
알테오젠	-5.2	-31.1	-5.6	루닛	7.4	73.9	103.2
루닛	7.4	73.9	103.2	코오롱티슈진	23.7	56.6	90.9
KB 금융	1.3	-5.8	3.2	DI 동일	19.5	24.1	55.6
이수페타시스	17.1	19.2	-27.4	미원상사	-1.1	-0.8	-7.5
셀트리온	6.1	16.8	-1.0	카카오페이	-3.0	34.3	20.7
카카오	-2.3	25.8	20.4	한국앤컴퍼니	23.9	19.6	12.3
유한양행	-4.2	-6.4	-9.8	케이카	15.8	16.9	9.6
LG 에너지솔루션	-0.7	3.1	-4.3	에스티큐브	63.6	61.1	75.4
POSCO 홀딩스	-2.6	-4.9	-28.7	펄어비스	-19.4	-25.4	-22.8
HD 현대일렉트릭	8.2	7.2	37.0	효성	14.2	7.0	4.6
펩트론	26.5	-19.5	105.6	녹십자홀딩스	16.2	23.2	12.2
한화에어로스페이스	5.0	-24.9	5.3	휴온스글로벌	7.5	37.4	52.9
금양	21.5	-17.4	-47.0	SK 바이오사이언스	7.7	19.0	-4.1
신성델타테크	29.2	145.6	126.7	성광벤드	10.1	52.9	62.6
고려아연	-27.4	7.7	67.0	디어유	9.0	1.4	108.7
두산에너빌리티	1.8	-20.0	-3.7	두산밥캣	-3.4	10.3	1.0
현대차	-1.9	0.2	-12.9	금양	21.5	-17.4	-47.0

자료: QuantiWise, DS투자증권 리서치센터

주: 좌측 청색열은 5일누계거래대금 기준 상위 20개 종목의 기간수익률을 나타냄. 우측 회색열은 상대거래강도 기준 상위 20개 종목의 기간수익률을 나타냄.

국내 주식시장을 다른 방식으로 조망하는 우회로

2024년 12월 3주차 DS Weekly 탑픽(Top-PicK)모니터링

2024년 12월 9일부터 2024년 12월 16일까지 국내 경제 일간지를 대상으로 뉴스 데이터를 종합했을 때 국내 증시의 토픽은 31개가 형성되었다. 표를 보면 토픽번호마다 10개의 상위 키워드를 확인할 수 있다. 전반적으로 최근 한국의 정세가 어지러운 점을 반영한 토픽모델링 결과가 나타났다.

토픽 번호 중에서 주요 사항들을 브리핑하면 아래와 같다.

1번 토픽: 탄핵정국을 반영

1번 토픽에 있는 상위 단어 10개는 탄핵정국을 반영하고 있다. 관련된 주제의 토픽 번호는 26번으로 중복되어 잡힌 모습이 보인다. 토픽 번호 X 문서 클러스터 히트맵을 봤을 때는 약 100여개에 달하는 일부 문서들이 다른 클러스터로 분류된 모습이 나타난다. 배정된 기사들 중에서 분류하는 알고리즘이 조금씩 다르기에 나타난 결과이다.

2번 토픽: 외환시장 및 물가와 관련된 주제

2번 토픽은 외환시장과 물가와 관련된 단어들로 구성되어 있다. 탄핵이 불발된 소식이 나타난 이후 환율이 1,430원/달러를 넘어서면서 원화약세가 외환시장에 반영되었다. 올해 김밥, 삼겹살 등의 외식 메뉴 가격 인상을 다룬 기사도 2번 토픽으로 분류되었다. 한국소비자원 가격정보포털에 업데이트된 데이터를 참고하여 생활물가가 올랐음을 최근에 문제 삼고 있는 것으로 추정된다.

3번 토픽: AI를 적용하는 기업들의 사례 그리고 규제 완화

3번 토픽은 AI와 관련된 주제를 나타낸다. 글로벌 기업을 포함한 국내 대기업들이 AI를 활용한 제품과 서비스를 출시하는 소식들이 있었다. 금융업종에서도 AI를 접목한 서비스들을 대중들에게 선보일 것으로 기대된다. 금융위원회에서 AI 은행원, AI 보험설계사, AI 증권정보 제공 등을 비롯하여 AI를 활용한 혁신금융서비스를 제공할 수 있도록 규제를 완화한 결과이다. 망 분리 규제가 완화되었기에 AI 서비스를 개발하기에 이전보다 더 수월해질 것으로 전망된다. AI를 활용하여 기업의 비용 절감 및 운영 효율화를 통한 실적 개선을 기대할 수 있는 부분이다.

4번 토픽: 미중 무역분쟁과 국내 기업의 해외진출 관련

4번 토픽은 3가지의 이슈가 혼재되어 있다. 중국의 엔비디아 반독점법 위반조사에 대한 소식과 미국의 중국에 대한 추가관세를 예고하는 등 중국과 미국이 서로 견제하는 모습이 보인다. 예상치를 밑도는 중국 11월 소매판매 3.0% 발표 등으로 부양책에 대한 기대감이 꺾이는 모습이 중국 지수에 반영되었다. 그 외 삼양식품의 중국 공장 설립, 현대차 아이오닉 5N의 중국 올해의 고성능차 수상 등이 4번 토픽에 잡혔다.

7번 토픽: 금리와 관련된 주제

7번 토픽에는 금리와 관련된 기사들로 구성되었다. 미국 11월 CPI가 2.7% 발표됨에 따라 금리인하 기대감이 유효하다는 기사들이 있었다. 한국은행과 금융위원회에서는 내년부터 지표금리를 CD금리에서 KOFR로 점진적으로 전환할 계획을 발표했다.

13번 토픽: 국내 기업들의 해외 수출 실적관련 소식

13번 토픽에는 국내 기업들의 해외 수출 소식 등을 다루고 있다. 식품업을 비롯하여 국내 몇몇 산업부문에서 해외 수출이 늘어나고 있고 있다는 기사들이 분포되어 있다. 베트남이라는 단어의 경우 예전보다 더 한국과의 교역규모가 증가했으며 한국 기업들의 베트남 수출 소식이 몇몇 존재함에 따라 상위 키워드 중 하나로 포함된 것으로 추정된다.

16번 토픽: 메모리 반도체 및 TSMC 관련 소식

16번 토픽은 반도체 산업과 관련된 주제를 나타낸다. TSMC의 소식을 중점적으로 다루면서 반도체 밸류체인에 대한 전반적인 소식들이 골고루 분포되었다. 그 외에 중국의 DRAM 가격 반값 공세로 인해 DRAM 가격 하락을 다룬 소식들도 존재했다.

24번 토픽: 지주사 관련소식 종합

24번 토픽에서는 국민연금이 두산에너빌리티와 두산로보틱스 합병과 관련된 소식들이 주된 키워드로 잡히는 모습이 나타났다. 주식매수청구권보다 높은 가격으로 대해서만 찬성한다는 조건을 제시했으나 양사의 주가가 매수 예정가보다 한참 하회하는 수준이라서 사실상 기권 의견으로 해석되었다. 롯데쇼핑이 자회사 롯데인천타운을 흡수합병할 예정이라는 소식과 한미약품 경영권 분쟁에 대한 기사도 일부 포함되었지만 상위 키워드로는 잡히지 않았다.

29번 토픽: 테슬라와 양자컴퓨터 관련 주제

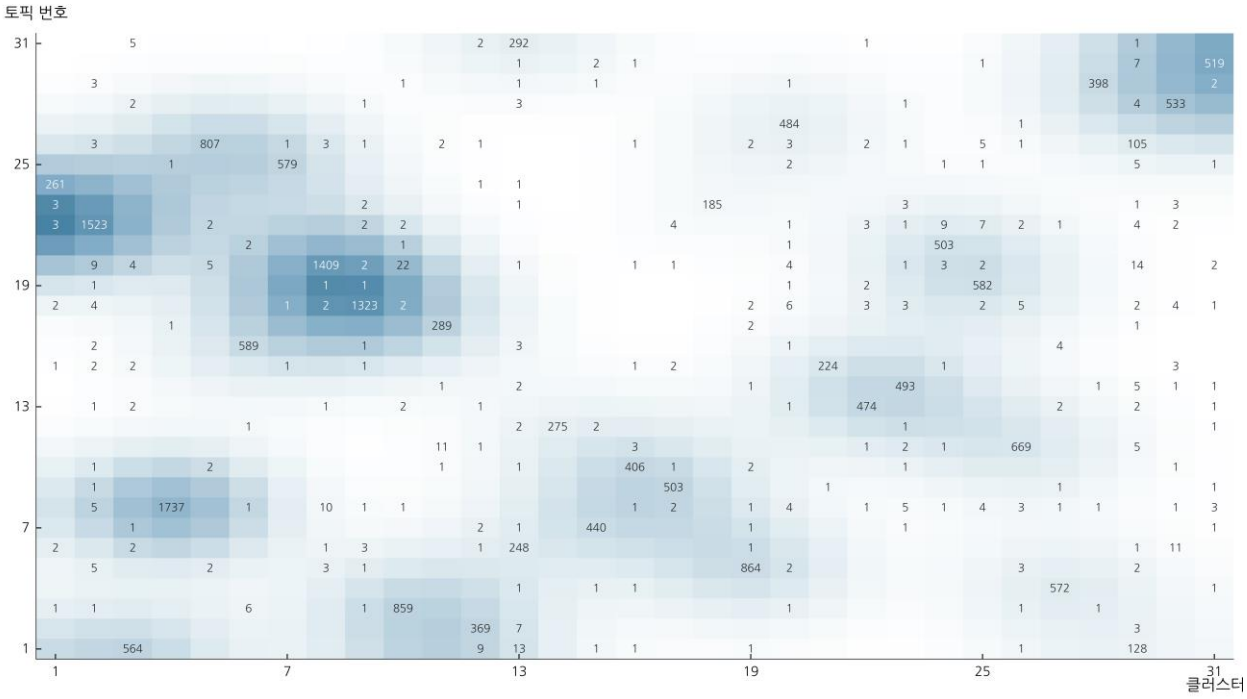
29번 토픽에서는 테슬라와 양자컴퓨터 2개의 주제가 하나의 토픽으로 형성되었다. GM은 로보택시 사업을 철수할 예정이라고 발표했다. 스페이스X가 기업가치 500조로 평가됨에 따라 테슬라에게 긍정적인 소식들이 나타났다. 한편으로 구글은 양자컴퓨팅 칩인 '윌로우'를 공개함에 따라 주가가 상승했다. 10조 7000억 년 걸리는 복잡한 계산을 5분 만에 수행하는 뛰어난 성능을 보여 투자자들의 이목을 끌었다.

표14 배정된 토픽 상위 키워드 10개

토픽 번호	TOP 10 키워드
1	탄핵, 대통령, 정국, 정치, 한국, 국회, 가결, 사태, 계엄, 비상계엄
2	환율, 외환, 외환시장, 원화, 서울, 전거래일, 물가, 거래일, 기준, 수입
3	AI, 생성, 데이터, 모델, 활용, 에이전트, 인공지능, 학습, 영상, 지능
4	중국, 전기차, 엔비디아, 배터리, 부양책, 경기, 현대차, 정부, 정책, 미국
5	아파트, 서울, 청약, 분양, 부동산, 단지, 분양가, 가격, 주택, 전용
6	코스닥, 종목, 알테오젠, 거래일, 순매도, 리가켐바이오, 클래시스, 에코프로비엠, 엔켐, 총액
7	금리, 인하, 연준, 물가, CPI, FOMC, Fed, 미국, 채권, 국채
8	브랜드, 제품, 상품, 매장, 판매, 크리스마스, 선물, 구매, 출시, 인기
9	인증, CCM, 소비자, 경영, 획득, 소비자중심경영, 중심, 한국소비자원, 공정거래위원회, 평가
10	대출, 가계, 은행, 연체, 신용, DSR, 주담대, 소득, 증가, 규제
11	감소, 증가, 취업자, 전년, 고용, 제조업, 일자리, 통계청, 소상공인, 건설업
12	ETF, 미국, TIGER, ACE, KODEX, 주식, 운용, 종목, 상품, 돌파
13	수출, 해외, 무역, 방산, 수출액, 반도체, 베트남, 수상, 증가, 정부
14	연금, 퇴직, 펀드, 계좌, IRP, 상품, 가입, 운용, 고객, 금액
15	대한항공, 아시아나항공, 항공사, 통합, 마일리지, 노선, 인수, 항공, 아시아나, 공정위
16	반도체, 삼성전자, TSMC, SK 하이닉스, 파운드리, 삼성, 공정, HBM, 엔비디아, 메모리
17	가구, 주택, 소득, 부채, 공공, 착공, LH, 비중, 물량, 평균
18	치료제, 임상, 계약, 바이오, 치료, 허가, 글로벌, 의약품, 질환, 제품
19	스타트업, 창업, 벤처, 글로벌, 컴업, 프로그램, 혁신, 성장, 협력, 중기부
20	서비스, 제공, 고객, 정보, 이용, 플랫폼, 네이버, 디지털, 데이터, 이용자
21	게임, 오징어, 콘텐츠, 넷플릭스, 출시, IP, 공개, 캐릭터, 시즌, 크래프톤
22	수상, ESG, 경영, 에너지, 대상, 선정, 산업, 우수, 사회, 디지털
23	고려아연, MBK, 영풍, MBK 파트너스, 주주, 소각, 금감원, 주장, 처분, 자사
24	두산에너빌리티, 두산, 합병, 두산로보틱스, 분할, 주가, 주주, 두산밥캣, 두산그룹, 임시
25	승진, 현대차, 인사, 임원, 현대차그룹, 기아, 선임, 조직, 신임, 미래
26	정부, 안정, 부총리, 회의, 대응, 상황, 간담회, 점검, 금감원, 추진
27	CES, LG, LG 전자, 현대모비스, 모빌리티, 전시, 운전자, 혁신, 전자, 차량
28	배당, 현금, 결정, 주식, 공시, 주주, 밸류업, 소각, 자사, 취득
29	테슬라, 머스크, 주가, 구글, 양자, 전기차, 스페이스 X, 나스닥, 사상, 엔비디아
30	트럼프, 미국, 관세, 행정부, 당선인, 비트코인, 한국, 부과, 정부, 정책
31	코스피, 거래일, 순매도, 종목, SK 하이닉스, 삼성전자, 삼성바이오로직스, LG 에너지솔루션, 셀트리온, 매수세

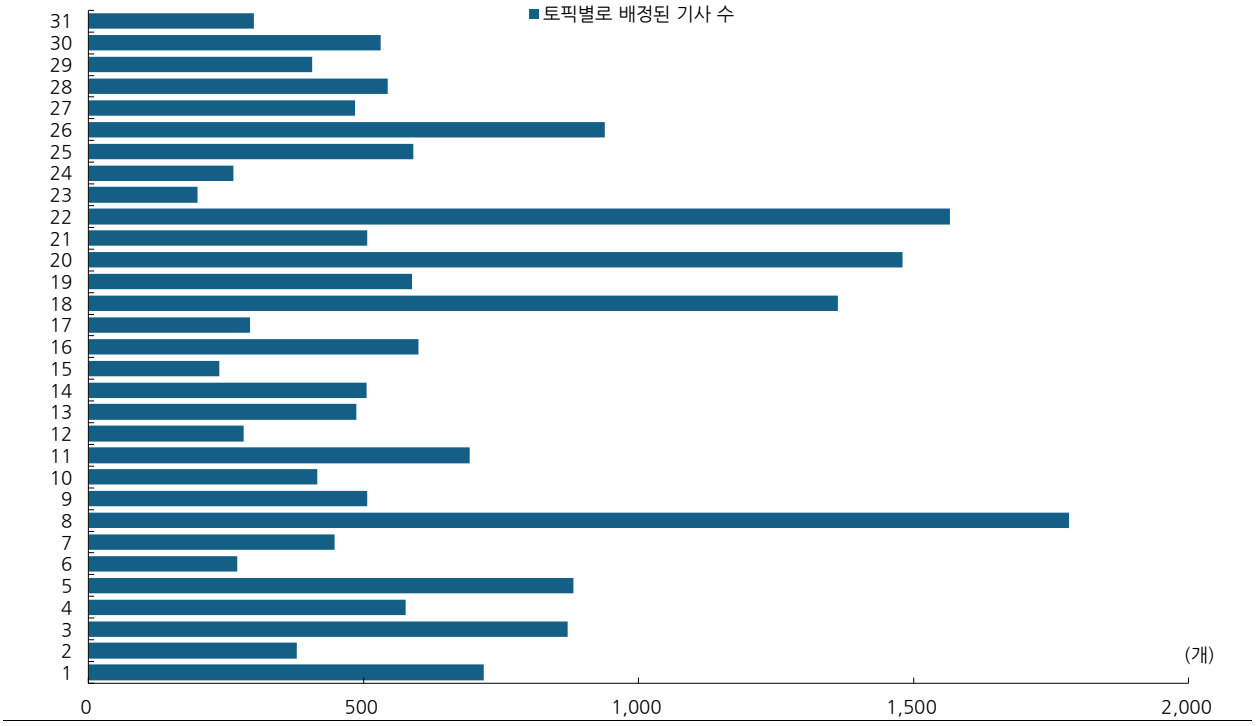
자료: 빅카인즈, DS투자증권 리서치센터

그림19 토픽 번호 X 문서 클러스터 히트맵



자료: 빅카인즈, DS투자증권 리서치센터

그림20 토픽번호별로 배정된 기사 분포



자료: 빅카인즈, DS투자증권 리서치센터

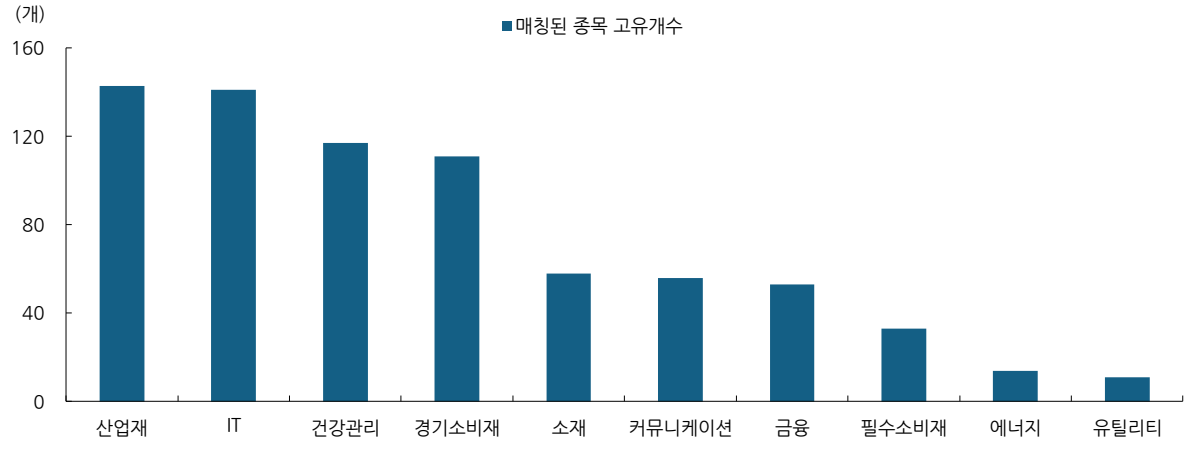
표15 기사 매칭 상위 TOP 50

티커	종목명	5일누계거래대금(억원)	일평균거래대금(억원)	시가총액(억원)	TTM P/E	12MF P/E	TTM P/B	WICS업종명(대)
A065500	오리엔트정공	7,372	37	1,857	-27.4		4.8	경기관련소비재
A035420	NAVER	14,125	1,551	331,926	20.2	18.9	1.3	커뮤니케이션서비스
A086790	하나금융지주	4,003	713	167,461	4.6	4.3	0.4	금융
A055550	신한지주	4,618	885	244,926	5.4	4.9	0.4	금융
A105560	KB금융	8,598	1,254	332,138	7.3	6.2	0.6	금융
A082800	비보존 제약	1,995	35	4,660	-167.6		4.2	건강관리
A035720	카카오	7,528	723	190,067	-21.2	42.2	1.9	커뮤니케이션서비스
A010130	고려아연	5,891	641	230,221	36.8		2.5	소재
A003550	LG	741	196	117,976	12.6	8.6	0.4	산업재
A005930	삼성전자	54,820	15,625	3,235,622	11.5	10.3	1.0	IT
A205500	액션스퀘어	140	2	1,015	-9.6		4.7	커뮤니케이션서비스
A002630	오리엔트바이오	696	6	1,397	42.0		1.8	건강관리
A025950	동신건설	5,685	91	4,120	62.2		4.1	산업재
A053800	안랩	3,932	84	8,167	21.9		2.0	IT
A045660	에이텍	5,378	97	2,416	16.7		2.2	IT
A002020	코오롱	829	12	1,900	1.0		0.2	산업재
A115500	케이씨에스	4,192	53	1,306	67.9		7.2	IT
A013360	일성건설	2,934	35	1,788	73.0		1.4	산업재
A356680	엑스게이트	9,503	131	2,991	76.0		7.6	IT
A293490	카카오게임즈	2,572	59	15,697	-6.3	28.2	1.1	커뮤니케이션서비스
A003490	대한항공	1,322	210	88,925	8.2	6.0	0.9	산업재
A000150	두산	2,458	292	43,788	-17.6	24.0	3.0	산업재
A034730	SK	1,060	369	101,649	-43.9	6.3	0.4	에너지
A005490	POSCO홀딩스	6,685	1,597	218,128	16.9	9.9	0.4	소재
A241560	두산밥캣	2,524	207	41,854	6.4	7.1	0.7	산업재
A030200	KT	1,799	270	115,300	9.9	7.6	0.7	커뮤니케이션서비스
A139130	DGB금융지주	323	51	14,158	6.6	3.3	0.2	금융
A032640	LG유플러스	786	100	47,853	9.0	7.8	0.5	커뮤니케이션서비스
A050960	수산아이앤티	773	9	1,074	18.6		1.2	IT
A373220	LG에너지솔루션	6,919	1,009	895,050	-246.8	76.4	4.4	IT
A377300	카카오페이	3,158	153	39,580	-201.2	176.8	2.1	IT
A005380	현대차	5,633	2,350	432,444	4.5	4.2	0.5	경기관련소비재
A000660	SK하이닉스	30,020	7,901	1,339,524	12.8	5.2	1.9	IT
A003540	대신증권	94	14	8,164	14.4	8.2	0.4	금융
A053580	웹케시	1,253	9	1,478		16.2	1.4	IT
A017670	SK텔레콤	1,678	274	122,430	10.9	9.8	1.0	커뮤니케이션서비스
A034020	두산에너빌리티	5,878	1,318	112,034	136.3	25.2	1.5	산업재
A328130	루닛	11,488	453	24,254	-108.3	-51.4	10.4	건강관리
A000880	한화	476	100	20,501	-42.3	3.3	0.3	산업재
A084690	대상홀딩스	4,736	112	3,755	14.9		0.5	필수소비재
A298000	효성화학	183	12	1,775	-0.6		5.4	소재
A138930	BNK금융지주	894	102	34,415	5.0	4.1	0.3	금융
A263750	펄어비스	1,981	128	18,182	190.9	13.8	2.3	커뮤니케이션서비스
A005940	NH투자증권	322	81	45,161	7.2	6.5	0.6	금융
A068270	셀트리온	7,989	1,368	419,396	207.0	33.6	2.4	건강관리
A020560	아시아나항공	82	24	7,456	8.4		1.4	산업재
A012330	현대모비스	2,466	518	222,723	6.5	5.4	0.5	경기관련소비재
A000270	기아	5,177	1,692	380,573	4.0	3.7	0.7	경기관련소비재
A089860	롯데렌탈	159	17	11,595	13.6	7.1	0.8	산업재
A039490	키움증권	408	80	30,683	6.9	4.1	0.6	금융

자료: 빅카인즈, QuantiWise, DS투자증권 리서치센터

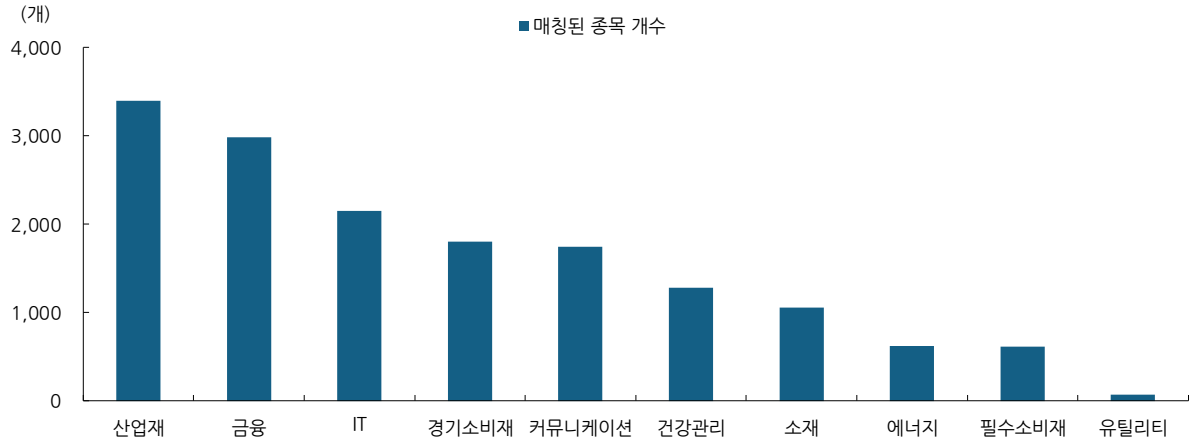
주: 시가총액 5,000억원 미만 종목의 경우 청색 음영표시

그림21 WICS 업종 대분류별 기사와 매칭된 종목 고유개수



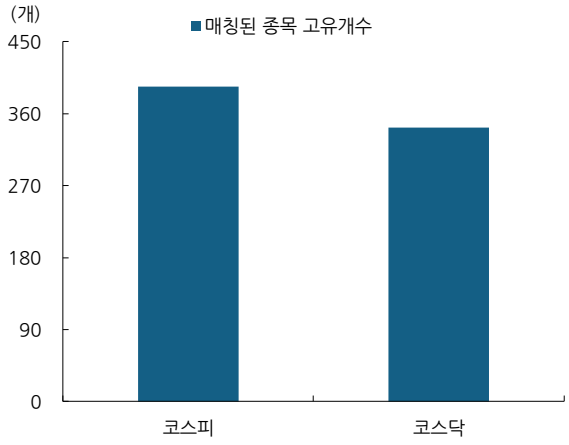
자료: 빅카인즈, QuantiWise, DS투자증권 리서치센터

그림22 WICS 업종 대분류별 기사와 매칭된 종목 개수



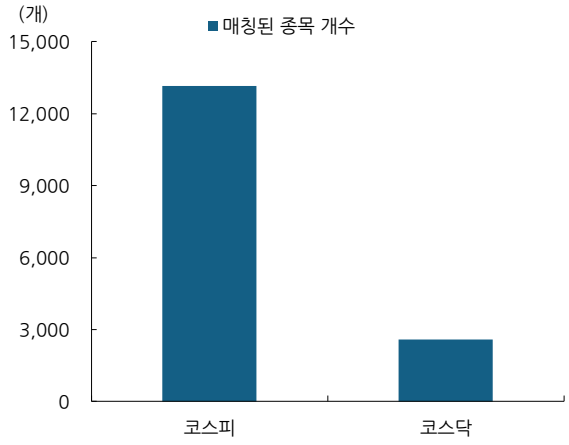
자료: 빅카인즈, QuantiWise, DS투자증권 리서치센터

그림23 시장별 매칭된 종목 고유개수



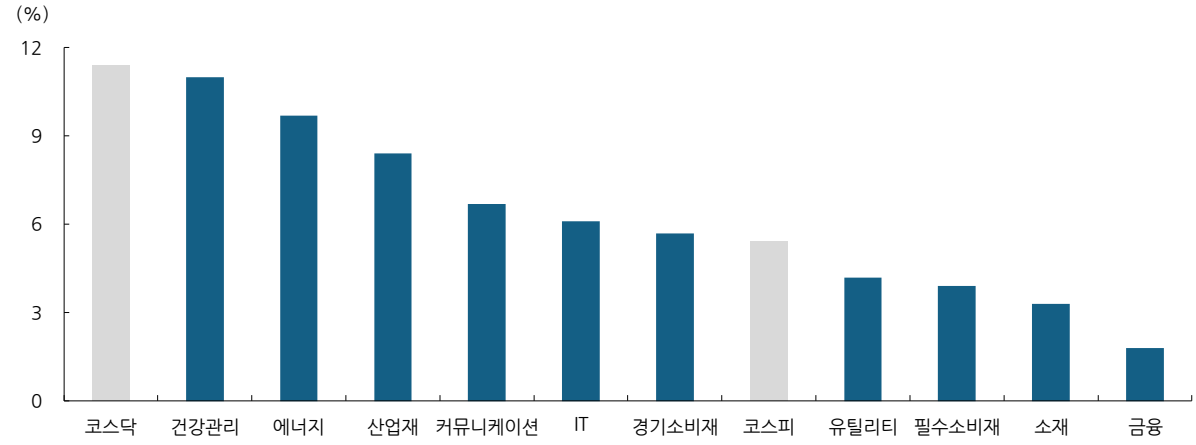
자료: 빅카인즈, QuantiWise, DS투자증권 리서치센터
주: 2개 이상의 기사에 대해 중복으로 매칭된 종목들을 제외하여 고유한 개수로 합산

그림24 시장별 매칭된 종목 개수



자료: 빅카인즈, QuantiWise, DS투자증권 리서치센터
주: 기사에 대해 2개 이상 매칭된 종목들도 중복하여 합산

그림25 2024년 12월 9일 ~ 2024년 12월 16일 WICS 업종별 기간수익률



자료: QuantiWise, DS투자증권 리서치센터

전반적 평가 및 개선사항

정치테마주가 상당 부분 포함된 결과

최근 토픽들의 전반적인 특성은 탄핵정국을 반영

1주일간 토픽 31개를 종합적으로 고려했을 때 국내 증시를 정치테마주, 비트코인, 수출 3개로 요약할 수 있다. 추가로 중국과 미국 간의 무역전쟁 속에서 관련된 사항들도 포함될 수 있다. 다만 무역전쟁 속에서 수혜를 받기 위해서는 미국 또는 중국과 우호적인 관계를 형성하여 실적 개선으로 이어져야 할 것이다.

거래대금을 고려했을 때 최근에 급등이 나타났던 소형주들이 상위 매칭 50개에 상당부분 포함되었다. 탄핵정국 속에서 특정 인물과 연관되어 있는 종목들에서 거액의 거래대금이 발생했기 때문이다. 단기적인 효과로 해석하고 있으며 이것이 중장기간 지속되기에는 어려울 것으로 판단된다. 정치테마주의 경우 중·소형주에 해당되는 경우가 많기 때문에 지수에 미치는 영향은 제한적이다.

국내에도 여전히 AI에 대해 주목하고 있는 중

AI와 관련한 토픽이 주요 토픽으로 지속적으로 유지되고 있음

연말이 다가오는 12월 3주차에도 토픽모델링의 결과는 AI에 대해 주목하고 있음을 알 수 있다. 과거에는 AI 인프라와 관련한 하드웨어 영역에서 주목도가 높았다면 이제는 소프트웨어와 규제 부문에서 이목이 쏠리고 있고 앞으로도 더 해당 부분의 관심도가 높아질 것이다.

AI에 노출되어 있는 기업에 주목하는 것은 여전히 유효

국내의 경우 금융업과 소프트웨어 산업 쪽에서 AI 규제 완화를 통해 비용효율화 또는 실적 개선이 이뤄지는지 지켜보면 좋을 것으로 기대된다. 부가적으로 내년 1월 초에 예정된 CES라는 단어가 토픽 내 상위 단어로 보이는 모습이 나타난다. AI를 적용하여 어떤 부분에서 혁신이 이뤄질지 지켜보고 여전히 IT산업 중 소프트웨어에 대한 관심을 높이고 선별하면 좋을 것이다.

불확실성만 해소된다면 다시 긍정적인 분위기로 전환할 여지는 남아있어

정치불안과 대외 변수들이 부정적으로 작용하고 있음

대외 변수들이 여럿 있지만 어려움 속에 혁신을 이뤄내는 기업에 주목하면 좋을 것이다. 최근 국내 정치불안과 미중무역 분쟁 속에서 한국 경제에 대한 전망이 우호적인 상황이 아니지만 기회는 찾으면 항상 있을 것이다.

최근 토픽모델에 반영된 결과는 아니지만 국내 증시의 방향성이 명확해질 요소들이 아직 발현되지 않아서 지수가 아래 방향으로 눌려있는 상황이다. 밸류업 프로그램, 금융투자소득세, 상법개정 등 불확실하지만 긍정적인 방향으로 결정된다면 얼마든지 국내 증시가 긍정적인 분위기로 전환할 여력은 존재한다.

실제로 통하는 전략일지 추가적인 검증 필요

체계적인 투자 전략으로 구성하여 성과를 검증할 필요가 있음

텍스트 데이터라는 방식으로 효율적인 국내 증시의 정보를 압축하는 방법을 지금까지 고찰해 봤다. 하나씩 검색하여 의사결정까지 도달하는 과정을 줄이면 투자에 도움 되는 부분이 있을 것이다. 하지만 어느 부분을 핵심적으로 먼저 정보 이용자들에게 보여줄지는 아직 보완작업이 필요할 것으로 판단된다.

또한 해당 정보를 참고하여 실제 투자 전략으로 옮겼을 때 성과가 어떻게 나타나는지도 검증이 필요하다. 스크리닝 기준을 설정하여 정기적으로 리밸런싱을 수행했을 때 성과가 어떻게 나타나는지는 데이터가 누적되는 대로 추가적인 검증을 해볼 예정이다.

Compliance Notice

- 동 자료는 기관투자가 등 제 3자에게 사전 제공한 사실이 없습니다.
- 동 자료에 게시된 내용들은 본인의 의견을 정확하게 반영하고 있으며, 외부의 부당한 압력이나 간섭 없이 작성되었음을 확인합니다. (작성자: 신민섭)
- 동 조사자료는 고객의 투자에 참고가 될 수 있는 각종 정보제공을 목적으로 제작되었습니다. 이 조사자료는 당사의 리서치센터가 신뢰할 수 있는 자료 및 정보로부터 얻어진 것이나, 당사가 그 정확성이나 완전성을 보장할 수 없으므로 투자자 자신의 판단과 책임하에 종목 선택이나 투자시기에 대한 최종 결정을 하시기 바랍니다. 따라서 이 조사자료는 어떠한 경우에도 고객의 증권투자 결과에 대한 법적 책임소재의 증빙자료로 사용될 수 없습니다.
-