2024 NH투자증권 빅데이터 경진대회 (예선)

팀명	NH방범대				
팀원	성명	생년월일	학교	학과	연락처
	김동영	1999.10.06	한국외국어대학교	영어통번역학과	010-4790-7697
	[redacted]				
주제명	섹션별 군집화와 RAG에 기반한 ETF QA 시스템				

분석 보고서

본 분석은 CRISP-DM 방법론에 기반하며, 분석의 목적은 사용자 맞춤 ETF 추천 시스템을 구현하는데 있다. 제공 데이터와 외부 데이터를 활용하여 필요한 수치형 데이터를 추출한 뒤, ETF의 특성을 파악하고 이를 통하여 사용자는 자신의 투자 성향과 목표 등에 맞는 ETF를 추천받을 수 있다. 수치형 데이터는 단기 수익성, 장기 수익성, 변동성, 유동성, 관심도 총 5가지 섹션을 기반으로 데이터를 분류하고 피처를 결합 및 간소화하여 군집 분석을 통해 ETF의 특성을 파악한다. 자세한 전처리 과정은 다음과 같다.

제공 데이터와 외부 데이터를 활용해 단기 수익성, 장기 수익성, 변동성, 유동성, 관심도의 5가지 섹션으로 나누어 진행하였다. 총 24개의 피처를 추출했으며 특정 섹션으로 분류가 어려운 피처는 상관분석을 통해 상관관계가 높은 피처들이 모여있는 섹션에 포함시켰다. 5개의 섹션은 각각 여러개의 피처를 가지고 있는데 이때 피처가 많아지게 되면 차원의 저주로 인해 군집화의 성능이 저하될 수 있다. 이 문제를 해결하고자 각 섹션별로 피처를 간소화하였다. 두가지 방법으로 간소화를 진행하였다. 첫번째, 피처들을 스케일링하고 같은 섹션 내에서 상관관계가 높은 피처들을 결합하여 하위 피처를 구성하였다. 먼저 단기수익성에서는 '1개월_수익률'과 '3개월_수익률'을 조합해 '단기수익' 피처를, 'MACD', 'RSI'를 결합해 '보조지표' 피처를 새롭게 만들었다. '단기수익'과 '보조지표'는 실제 단기간에 거둔 수익과 매수 타이밍을 평가해줄 수 있다. 장기수익성은 높은 상관관계를 보인 '1년_수익률'과 '누적_수익률(Z)', '정보비율(Z)', '샤프지수(Z)'를 묶어 '장기수익'으로 평가했다. 장기적으로 안정적인 수익을 기대할 수 있는 '1년_배당수익'은 '배당' 피처로 따로 분리하였다. 유동성은 '총 거래 금액'을 통해 '거래 규모'를, '매수매도_차이비율'을 이용해 '매수-매도 간의 균형'을, '유입금액_분산'과 '유출금액_분산'으로 '규모의 불안정성'을 측정했다. 마지막으로 관심도 섹션은 '종목조회수'와 '관심종목등록수'로 'ETF의 인지도'를 계산하고, '계좌수증감율'과 '관심종목대비_매수계좌'로 '매수증가세' 피처를 만들었다. 단순한 조합은 모두 합계로 진행했으며 같은 섹션의 피처들끼리는 상호 영향력이 균등하도록 스케일링을 재차 진행했다. 두번째, 피처가 6개로 가장 많은 변동성은 다른 섹션들과 달리 피처 간의 상관관계가 다양하게 나타났다. 이에 따라 주성분 분석으로 피처의 복잡성을 줄이고자 누적 분산이 90%를 넘기는 3개의 차원으로 축소하였다.

주요 피처들이 정리된 뒤에는 이를 기반으로 카테고리별 K-means 군집화를 수행했다. 군집의 개수는 실루엣 계수로 선정했으며 군집화의 결과를 시각화하고 평가하였다. 이때 차원 축소를 거쳤던 변동성 군집은 별도로 박스플롯을 통해 평가하였다. 최종적으로 각 카테고리별로 해석한 군집의 특성을 ETF에 라벨링하여 ETF별 특성 데이터를 완성하였다.

가공한 데이터를 활용해 ETF 종목 정보를 제공하는 RAG 모델을 구축했다. RAG 모델은 ETF 관련 데이터를 데이터베이스에 저장하고, 사용자의 질문에 맞춰 해당 정보를 검색해 답변을 생성한다. 이를 통해 사용자는 ETF의 특징, 성과, 시장 동향 등에 대해 정확하고 신뢰할 수 있는 정보를 얻을 수 있다. 이러한 도메인 맞춤형 접근 방식은 사용자에게 편리함을 제공하며, 효과적인 투자 결정을 도울 수 있다.

서비스 기획 아이디어 및 발전 방향

정리한 데이터를 기반으로 Langchain RAG 기반의 문서 QA 시스템을 설계한다. RAG는 언어 모델의 성능을 높이기 위해 검색 시스템과 결합한 방식이며, 기존 생성형 AI와 달리 원하는 특정 도메인에 특화된 생성형 AI를 만들 수 있다는 장점을 갖는다. 이에 본 팀의 목표는 ETF 추천에 특화된 생성형 AI 챗봇 서비스를 구현하는 것으로 설정한다. 간단한 프로세스는 다음과 같다.

단기수익성/장기수익성/변동성/유동성/관심도 5개의 섹션에 기반하여 생성한 ETF별 특성 데이터를 RAG에 활용할 수 있는 데이터로 전처리를 진행한다. 그 후, 각 ETF 별로 군집화에 활용한 변수와 군집화 된 결과를 자연어 텍스트로 만들고 이를 하나의 데이터프레임으로 생성한다. 이어서 해당 데이터를 읽어와 여러 개의 청크로 나눈 다음, 벡터 데이터베이스를 구축하여 문서를 벡터로 임베딩하고 파일 시스템에 저장한다. 다음 단계에서는 사용자의 질문이나 주어진 컨텍스트와 가장 관련성이 높은 정보를 검색해주는 retriever 메서드를 사용하여 검색을 하는데, 이 때 코사인 유사도를 기반으로 검색을 수행한다. 마지막으로 이렇게 검색된 정보를 바탕으로 사용자의 질문에 답변을 생성한다. LLM 모델에 검색 결과와 함께 사용자의 입력을 전달하면, 사전 학습된 지식과 검색 결과를 결합해 주어진 질문에 가장 적절한 답변을 생성하는 QA 시스템을 구축한다. 본 팀은 부가적으로 이전 대화기록을 고려하는 생성기를 활용해 사용자의 입력을 지속적으로 분석하여 더 나은 답변을 제공한다.

제공 데이터 외의 여러 데이터를 활용하면 사용자에게 더 나은 서비스를 제공할 수 있다. 본 팀이 제안하는 아이디어는 사용자의 편의성을 향상시키는데 중점을 둔다 .

첫째, 개인 맞춤형 포트폴리오 추천 시스템 구축이다. 제공 데이터 외에도 사용자의 투자 목표, 원하는 투자 기간 등 더 자세하고 추가적인 내용을 담고 있는 정보를 활용한다면 정밀한 분석이 가능할 것이다. 예를 들어 투자 성향이 보수적인 사용자에게는 안정형 상품을 추천하는 등 사용자의 투자 성향에 맞춰 최적의 ETF 포트폴리오를 제공할 수 있다.

둘째, 실시간 질문 및 답변 기능을 제공한다. 사용자가 ETF 투자와 관련하여 궁금한 점을 즉시 해결할 수 있는 QA 기능을 구축한다. 현재 시스템이 미리 정의된 질문에 대한 답변을 내놓는 형식이라면, 사용자가 어떤 질문을 하든 챗봇이 답변을 할 수 있는 형식의 QA 기능을 추가하는 것이다. 더욱 고도화 된 자연어 처리를 활용해 다양하고 복잡한 질문에도 대응할 수 있으며 사용자 또한 원하는 답변을 얻을 수 있다. 추가적으로 자주 묻는 질문(FAQ) 세션을 구축하여 필요한 정보를 쉽게 찾을 수 있도록 하고, 기업 내 새롭게 들어오는 정보를 QA 시스템과 연동하여 실시간으로 데이터베이스를 업데이트하는 인프라를 구축할 수 있다.

이러한 개선점들은 시의성이 중요한 ETF 투자에서 사용자에게 신속한 답변을 제공함으로써 사용자의 편의를 개선한다. 결론적으로 사용자가 효과적인 ETF 투자 결정을 내릴 수 있도록 지원하는 서비스를 제공할 수 있다.