# 추천시스템 & 데이터 스크래핑 과제

[DGB대구은행] iM DiGital Banker Academy(데이터분석 전문가 양성 과정)

### **배경**

최근 금리가 상승하고 있는 추세에도 불구하고, 부동산에 대한 투자 관심은 여전히 높아지고 있습니다. 특히 대환대출 및 주택담보대출 갈아타기와 같은 금융 상품에 대한 관심이 증가하면서, 많은 사람들이 부동산 시장에 재진입하고 있습니다. 이러한 상황 속에서, 주택금융연구원이 발표한 2024년 상반기 주택시장 및 향후 전망 보고서에 따르면, 서울 지역 아파트 가격은 상승세를 보이고 있으며, 비수도권 지역의 가격은 하락세가 지속되고 있습니다. 또한, 서울 지역의 거래량은 증가하고 있어, 투자 수요가 서울 아파트로 집중되고 있는 것으로 추정됩니다. 이러한 추세는 향후 부동산 투자 결정에 중요한 요인으로 작용할 것입니다.

따라서, 본 과제에서는 부동산 매물 데이터를 스크랩하여 투자에 적합한 아파트를 추천하기 위한 시스템을 구축하고자 합니다. 데이터 스크래핑을 통해 실시간으로 변동하는 부동산 시장 정보를 수집하고, 이를 분석하여 매매자에게 유용한 추천 정보를 제공할 것입니다. 특히, 금리 상승에도 불구하고 투자 기회를 찾고 있는 고객들에게 맞춤형 조건 아파트 추천을 통해 부동산 투자 결정을 지원할 계획입니다

### **데이터 설명**

대구광역시 수성구 범어동, 만천동 3가지 아파트들의 대한 매물 데이터

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **아파트 이름** | **유형** | **가격** | **평수** | **층** | **방향** | **지역** |
| 대구범어아이파크 103동 | 월세 | 5,000/150 | 109B/84m² | 6/23층 | 남서향 | 대구광역시 수성구 범어동 |
| 대구범어아이파크 101동 | 전세 | 7억 5,000 | 148/121m² | 고/23층 | 남향 | 대구광역시 수성구 범어동 |
| 대구범어아이파크 101동 | 월세 | 3억/250 | 148/121m² | 중/23층 | 남향 | 대구광역시 수성구 범어동 |

### **방법론**

##### 데이터 수집

부동산 매물 정보를 웹 크롤링하여 아파트 이름, 가격, 평수, 층수, 방향 등 다양한 속성을 포함한 데이터셋을 생성합니다. 이 데이터는 사용자의 아파트 선택에 기반한 추천 시스템의 기초 자료로 활용됩니다.

##### 데이터 전처리

크롤링한 데이터에서 문자열 형태의 데이터를 수치형 데이터로 변환합니다.

##### 행렬 인수분해

(SVD)행렬 인수분해 기법을 사용하여 데이터를 분해합니다. 이를 통해 각 아파트의 특성과 사용자의 선호도를 잠재 요인으로 변환합니다.

##### 코사인 유사도 계산

추천하고자 하는 아파트와 다른 아파트 간의 코사인 유사도를 계산합니다. 코사인 유사도는 두 벡터 간의 각도를 기반으로 유사성을 측정하며, 1에 가까울수록 더 유사한 것으로 간주됩니다.

##### 추천 시스템 구축

사용자가 선호하는 아파트 이름을 입력하면, 해당 아파트와 코사인 유사도가 높은 다른 아파트를 추천합니다. 추천된 아파트 목록에는 각 아파트의 가격, 평수, 층수, 방향 등의 정보를 포함하여 사용자에게 제공합니다.

##### 결과 제공

추천된 아파트 리스트를 사용자에게 보여주고, 각 항목에 대한 상세 정보를 제공하여 사용자가 더 나은 결정을 내릴 수 있도록 지원합니다.

### **분석 과정**

##### 부동산 매물 크롤링

* 1. extract\_listings : 함수를 정의하여 매물 정보를 수집합니다.

이 함수는 스크롤하여 페이지의 데이터를 로드하고, BeautifulSoup을 사용하여 매물 정보를 추출합니다.

텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

* 1. execute\_script 메서드를 사용하여 JavaScript로 스크롤을 구현합니다.
  2. 각 매물의 제목, 유형, 가격, 평수, 층, 방향 등의 정보를 추출하여 사전 형태로 저장하고 리스트에 추가합니다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

* 1. 같은 방식으로 다른 아파트 매물 페이지에서 데이터를 수집합니다. 정보를 수집한 후 지역 정보를 추가 및 데이터 결합 및 저장합니다.

텍스트, 스크린샷, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

##### 부동산 추천 시스템

* 1. 데이터 준비 및 아파트 선택

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

* 1. 아파트 특징 결합 및 벡터화
* 특징 결합 : 아파트의 여러 정보를 하나의 문자열로 결합하여 각 아파트의 특징을 표현합니다. 이 정보는 아파트 이름, 유형, 가격, 평수, 층, 방향, 지역을 포함합니다.
* TF-IDF 벡터화 : 결합된 문자열 데이터를 TF-IDF 행렬로 변환합니다. 이는 텍스트 데이터의 중요도를 수치적으로 나타내어 유사도 계산을 가능하게 합니다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

* 1. 선택한 아파트의 TF – IDF 벡터 추출 및 유사도 계산
* TF-IDF 벡터 추출: 선택한 아파트의 인덱스를 사용하여 TF-IDF 벡터를 추출합니다.
* 코사인 유사도 계산: 선택한 아파트의 TF-IDF 벡터와 모든 아파트의 TF-IDF 벡터 간의 코사인 유사도를 계산합니다. 이 유사도는 선택한 아파트와 다른 아파트 간의 유사성을 나타냅니다.
* 추천 아파트 선정: 유사도가 높은 상위 10개의 아파트를 추천 목록으로 정렬하여 추출합니다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

* 1. 가상의 평점 생성 및 SVD 모델 학습
* 가상의 평점 생성: 1부터 5까지의 랜덤한 평점을 생성하여 APT 데이터프레임에 추가합니다.
* SVD 모델 학습

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

* 1. 아파트 입력 및 추천 아파트 출력

텍스트, 스크린샷, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

자동 생성된 설명

### **결론**

본 과제에서는 금리가 상승하는 상황에서도 부동산 투자에 대한 관심이 높아지고 있음을 반영하여 아파트 매물 추천 시스템을 구축했습니다. 투자자들은 대환대출과 같은 금융 상품을 활용해 부동산 시장에 재진입하고 있습니다. 이를 바탕으로 사용자에게 맞춤형 아파트 추천 서비스를 제공하고, 더 나은 투자 결정을 지원하고자 했습니다.

### **고찰**

대구 아파트 매물 정보를 직접 크롤링하여 유용하게 사용하여 좋았다.

추천 시스템을 만들 때 코사인 유사도만 사용했을 때 보다 SVD를 혼합하여 사용했을 때 유사도가 높아지고 추천하는 항목이 많아져 혼합하여 사용하면 괜찮다는 걸 알았다.

* 한계점

아파트 매물의 링크기 좌표마다 달라져 규칙을 찾아 적용하여 대구 아파트의 매물을 전체 정보를 크롤링하지 못해 아쉬웠다.

SVD 모델을 사용할 때 문자열 인코딩과 가격의 전세, 매매 등의 문자열을 수치화하지 않고 훈련을하여 모델의 성능을 더 높이지 못해 아쉬웠다.

데이터가 부족하여 아쉬웠다.