<크롤링 부분 대본>

(참고: 단락마다 한 페이지)

이제 뉴스기사를 크롤링 한 과정과 그렇게 나온 데이터에 대한 EDA 결과 몇 가지를 소개해드리겠습니다.

우선, 단일 페이지를 크롤링할 때는 URL 분석이 필요 없을 수도 있지만, 저희는 여러페이지에서 뉴스 데이터를 받아와야 하는 구조이기 때문에, URL 분석을 필수적으로 해야했습니다. 그래서 간단하게 URL분석의 과정을 먼저 보여드리겠습니다.

URL을 보면 이렇게 중간중간 하이라이트 친 부분들이 핵심어들인데, 이 부분을 하나하나 보겠습니다.

먼저, query를 시작으로 나와있는 이 부분은 네이버 검색창에서 입력한 검색어가 무엇인지 표기해주는 부분인데, 예를 들어서 비트코인이라고 검색을 하면 query=비트코인 이라고 URL에 표기됩니다. 참고로, 현재 이 URL은 한글을 유니코드로 표기했기때문에 이런 알수없는 문자들로 나오고 있습니다. 파이썬에서 구동시 무시하고 한글로 입력해도 정상적으로 실행이 됩니다.

Sort는 정렬방식을 표기하는데, 네이버 기사는 관련도순, 최신순, 오래된순 3가지 방식으로 선택하여 정렬할 수 있게 만들어져 있습니다. 이때 sort값은 관련도순 설정시 0 최신순은 1 오랜된순은 2 이런식으로 URL에 들어가게 됩니다.

Ds와 de는 검색 날짜 범위 설정시 시작일과 종료일을 표기합니다. 주의할 점은 형식이 연도.월.일 의 형식으로 넣어야 한다는 점입니다. 저희는 2013년4월1일부터 2021년 4월1일의 기사를 대상으로 했으므로, 다음과 같이 나타나는 것을 알 수 있습니다.

From과 to역시 시작일과 종료일을 표기합니다.

주의할 점은, ds/de와 from/to의 날짜는 동일하지만 입력하는 형식이 다르다는 것입니다. 보시는 것과 같이 ds/de는 점을 넣고, from/to는 점을 넣지 않습니다.

마지막으로, start는 현재 몇 페이지에 있는 지를 나타내는데, 이 때 주의할 점이 페이지는 1,2,3,4 이런식으로 나가지만 이에 해당하는 start값은 1, 11, 21, 31 이런식으로 나간다는 점입니다.

예를들어, 400페이지는 start값이 3991에 해당하는 것을 볼 수 있습니다.

이렇게 살펴본 것과 같이, URL분석을 통해 다음과 같은 정보들을 알 수 있고, 이는 크롤링 과정에서 중요한 역할을 하게 됩니다.

이제, 본격적으로 크롤링 과정에 대해 알아보겠습니다.

우선 간단하게 워크 플로우를 말씀드리면 다음과 같습니다(피피티 읽어주면 될듯)

그런데, 이 과정에서 한 가지 문제가 있었습니다.

바로 네이버가 데이터를 최대 400페이지, 즉 한 번에 4000개만 제공한다는 점이었습니다.

그래서 또다른 워크플로우를 생각했는데, 그것은 다음과 같습니다(이것도 피피티 읽어주면 될듯)

그래서 저희는 날짜, 언론사, 제목, 뉴스 요약본, 뉴스 링크 이렇게 5개의 변수를 대상으로 크롤링 했고, 그 결과 다음과 같은 데이터 프레임을 얻었습니다. 그 결과 총 357957개의 데이터를 얻었습니다.

이제 이 데이터를 가지고 어떻게 EDA를 시행했는지 소개해드리겠습니다.

저희는 크롤링해서 얻은 데이터와 비트코인 가격 데이터를 가지고 EDA를 진행했는데, 주요 사안은 다음과 같았습니다. (오른쪽 박스 읽기)

먼저, 일일 뉴스 개수와 코인 가격 간의 관계에 대해서 살펴보겠습니다.

그에대한 플로우는 다음과 같았는데, (피피티 읽어주면 될듯)

즉, 다음과 같이 value\_counts()로 파생변수를 생성하고 종가 데이터를 숫자만 나오게 전처리를한후 , merge()함수를 써서 새로운 데이터 프레임을 만들었습니다.

해당 데이터 프레임을 시계열 그래프로 시각화 한 결과는 다음과 같은데, 육안으로는 상관관계가 없어 보이고, 따라서 상관계수를 구해봤습니다.

그 결과, 상관계수가 작아 뉴스 데이터 개수와 비트코인 가격 간의 상관관계가 없다고 말할 수 있을 것 같습니다.

이제, 일일 뉴스 개수와 일일 수익률 간에는 어떤 관계가 있을지에 대해 살펴보겠습니다. (피피티 읽어주세요 )

이러한 과정을 통해 다음과 같이 5개의 코인에 대한 일일 수익률을 구했는데, 이 그래프들 역시 육안으로는 뉴스 개수와 상관관계가 없어보입니다.

일일수익률은 그래프가 모두 비슷한 형태라 비트코인 그래프만 뉴스 개수와의 상관계수를 구했는데, 거의 0에 수렴하여 상관관계를 갖고 있지 않다고 말할 수 있을 것 같습니다.