1차 프로젝트

2020년 2월 23일 일요일 오후 6:58

2/23 1일차
하루 100개 기사 내용 크롤링(파이썬 주피터 노트북) 한달 카페 와이파이 기준 대략 1시간 10분 소요
1차 프로젝트중 한국경제 전체 카테고리 하루 100개씩 뽑아내니 문제가... 연예뉴스 겨우 한두줄짜리 무의미 기사가 너무 많다... 예를들어 '한예슬이 레드카펫에 서서 포즈를 취하고 있다' 끝! 이런 기사들이 너무 많은데... 넣어야 하나??
지금 시간에 따라 기사가 나오기 때문에... 자정에 가까운 기사들 100개가 뽑혀짐...
그 시간대 대다수 연예기사들이 많이 나오는거 같다... 아예 시간대를 랜덤으로? 아니면 오전으로? 하는 식의 해결책으로 접근해보자.
자정 넘어서 즉, 초기 기사들(새벽)은 대다수 해외 국제 정세 위주야... 주요 뉴스만을 뽑고싶은데... 방법은?
즉, 초반 데이터 수집 과정에서 예측결과에 영향을 크게 미칠것으로 예상...
양질의 데이터 수집이 매우 중요하다 판단된다.
아쉽지만 일단 경제 카테고리로 진행하자... 수동으로 시간대 새벽으로 할 수는 없을것 같잔아...
랜덤뽑기도 어려워 보인다.(매일 기사 양이 달라 10개링크, 5개링크 등 인덱스 길이 오류 예상...) 연예기사만 제외하기도 어려워 보이고...
또다른 문제 명사 추출에서 여성 임원을 한단어로(붙여야 의미 있음)해야 하지만, 여성과 임원으로 각각 나뉘어 진다... 솔직히 매일 단어로 주가 관련 파악은 무의미해보일지도... 그때 그때 유행하는 단어들이 너무 큰 영향을 미치는거같다... 예를들어 2월달은 코로나의 달... 그러나 1년전이나 1년후의 코로나 단어는 주가와 관련 의미가 거의 없겠지... 아마도 기사 보단 주가의 회사들끼리의 비교나 금리, 환율, 유가 등으로만 작업하는게 더욱 연관성이 높아보인다... 데이터 뽑는데에도 시간 절약되고ㅠ
아... 11월 30일 경제 기사 수 너무 적어;; 신문사 바꿔서 해야되??? ㅅㅂ...ㅜ
데이터... list.append로 한달치 받으면 기사별로 대략 3천개가 쌓이지만...
하루 100개의 기사를 한줄로 저장하고싶어... 한달이면 30row로... 어떻게 할까?
1년치 한번에 카테고리별 뽑는 코드 작성 완료, 주피터 노트북으로 3개의 파이썬 창으로 실행 동시에 3년치 크롤링 1개당 약 10분 소요 1년치 12개월 2시간 정도 걸렸음(집 기준)3개 동시 돌렸으나 큰차이 없이 2시간에 3년치 확보 함.
2/24 2일차
일단 경제, 정치, 국제 20기사씩 3년치 크롤링중 아마도 10년치를 뽑아야 할지도...
오늘의 문제는 3~10년치의 코스닥 외에 여러 회사 일일 마감 주가 데이터 크롤링,
국제 유가, 환율, 금리 등 크롤링...
R로 엑셀 달별 카테고리별 데이터 합치기... 카테고리는 하나의 카테고리로 합치고
로우는 달별 년도별 순서대로 합치기...
교육장 컴퓨터(i5 16g)로 정치 카테고리 3년치 뽑고, 내 노트북(서피스7 i7 16g)로 국제 뽑는중

내 노트북이 빠름 체감, 그러나 큰 차이는 없음.
KRX한국 주식사이트인것 같음
주가 정보 일별로 정보 찾기가 쉽지 않음... 특히 매일 최고가 최저가 정보는 얻기 매우 힘들어 보임
야후 파이낸스 사이트에서 티커?? 검색하여 히스토리칼 데이터로 정보 확인
10년치2010.01.01~2020.01.01 코스피, 삼성전자, cj지주, 롯데 지주, 대한항공, sk하이닉스, 현대차
네이버금융에서 티커? 검색 확인
fred 환율에서 환율 데이터 확보
환율 데이터 빈 곳 많음... 주가데이터도 많음... 같은 해당 날짜 기사 제외 처리?? or 랜덤포레스트 빈칸 채우기??
가장 베스트는 둘다 해보고 더 나은 모델 택...
국제유가 데이터로 검색함 국제유가(WTI)로... itstat??에서 퍼옴 http://www.itstat.go.kr/stat.it?no=1072
데이터 컬럼 로우 바꿔야 할듯...
기사 데이터 + 날짜 데이터 합치기 성공
네이버 기사는 날짜나 데이터 링크 규칙이 자체 함수 화 되어 규칙 찾기 힘듬...
다른싸이트는 예로 1부터 차례대로 커짐 BUT 네이버는 추가 주소가 존재? 무작위 규칙성 숫자 느낌... 크롤링 막은 느낌...
2/26
기사 데이터 카테고리별 1년치 합치기, 10년치 합치기 concat axis=0, rbind
윤년 2월 29일 수동으로 엑셀 기입해줌. 그러나 날짜외에 시간 00:00:00 자동기입;; 원인 모름
기존 데이트 합치는 코드에 윤년 12년 16년 예외처리 해줌
3/8
컬럼명 네이밍 작업, oil데이터 열 행 트랜스포메이션
oil데이터 날짜 2010년1월1일->2010-01-01로 변경 파이썬으로 시도 date 함수 적용 잘안됨.
파이썬으로 for문 일일이 수동으로 넣어줘야 하나?
3/9
r as.date 함수로 쉽게 변경 가능 해결,
결측치 다수 존재. 환율 데이터는 .으로 주가 데이터는 null 값으로
국제유가 데이터는 결측치 존재 x 확인
.정규식으로 제거 시도, but .지우면 숫자 소수점도 사라짐...
as.character로 변경 후 as.numeric을 해주니 자연스레 .과 null값 모두 na처리가 되어버림
3/10
ppt 작업 시작 samsung close 컬럼 up down 진행, 한칸 앞당겨 내일의 income으로 진행
left_join 처리 하여 데이터 합침,
3/11
표준화, 베이지안 필터, 정규화 NN, 상관계수 분석, ppt 작업 완료