1. SNS 데이터 전부 합쳐서 섹션별로 나눠서 저장

(섹션 : 블로그, 카페, 뉴스)

* CONTENT가 결측값이면 날렸음

colab에서 하면 뭔 IllegalCharacterError 나는데 로컬 주피터 노트북에서 하면 에러 안남,,

-> 그냥 로컬에서 돌려서 파일 올려놓겠슴

blog.xlsx : 블로그

cafe.xlsx : 카페

news.xlsx : 뉴스

2. DATE 변수 날짜형으로 전환

3. TITLE, CONTENT 명사 추출후 빈도별 word cloud

--------------------------------------------------------------------------------------------------------------

회의일 : 2019.07.29

회의자 : 예은, 유경

1. **SNS 데이터**  
     
   **- 미세먼지 관련/비관련 글로 분류**   
    -> ‘‘미세먼지’ 단어의 출현빈도 이용/빈도수가 낮은 하위 50(?)개   
    => 미세먼지 비관련  
     
   ex) 출현빈도가 1~2개인 경우, 주제와 상관없이 여는말(?) 또는 맺음말 정도로 언급되었을 것으로 판단하고, 제외  
     
   **- 미먼감성지수**  
    미세먼지가 없어서 좋아요 -> 긍정  
    미세먼지가 심해서 싫어요 -> 부정  
    미세먼지로 인해서 긍정적인 이야기를 하는지, 부정적인 이야기를 하는지 판단하기.  
    아직 아이디어 단계
2. **유동인구 데이터**  
     
    - 요일별(주중,주말)/시간대별(새벽,아침,점심,저녁) 유동인구 변화  
    => 중요 포인트 : 미세먼지에 의한 변화인지 아닌지 잘 분간해야함 ☆

steady한 검색은 관심사가 아님.

급상승하는 키워드가 중요.

ex) 인기검색어 상위 1~3등은 무조건 ‘은행’ -> 정보의 가치가 없음

분산이 클수록 가산점을 부여하여 인기검색어 동작

seasonality파악 (ARIMA모형)

회의일 : 2019.08.09

회의자 : 예은, 유경

<< 주말동안 할 일 >>

1. **모델링이 가능한 형태로 데이터셋 만들기**- 동별로 데이터셋 생성

<변수목록>

1. index : 일자

2. 환경기상데이터(케이웨더)

1) mise\_med

2) mise\_mode

3) cmise\_mde

4) cmise\_mode

5) 온도

6) 습도

-> 시간대별 추출한 대표값의 **평균**을 일별 대표값으로 사용

(mean 함수 사용시 NaN 값 제외하고 계산됨)

3. 카드매출데이터(신한카드)

1) 각 카테고리 비율 값

4. 유통데이터(GS리테일)

1)

5. 유동인구데이터()

2. **SNS데이터 / 텍마 전에 공부하고 질문할것 정리**