PROJECT.3

텍스트 마이닝
네이버 환경 뉴스 분석

03

ABOUT PROJECT

1) 데이터 수집 및 전처리

2) 형태소 분석: 빈도분석 / 연관분석

3) 시각화 : 워드클라우드

데이터 수집

자료 입수처 : 네이버 뉴스 > 사회 > 환경

(https://news.naver.com/main/list.naver?mode=LS2D&mid=shm&sid1=102&sid2=252)

수집 데이터 내용 :

pd.date_range로 지정한 1년간의 환경 뉴스 헤드라인 추출

데이터 기간 : 2021.08.01 ~ 2022.07.31

사용 : PYTHON


```
dt_index = pd.date_range(start='20210801', end='20220801')
dt_list = dt_index.strftime('%Y%m%d').tolist()
print(dt_list)
all_data_frame=[]
append = all_data_frame.append
for i in dt_list:
    url = 'https://news.naver.com/main/list.naver?mode=LS2D&mid=shm&sid1=102&sid2=252&date=' + i
    headers = {
        "user-agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_5) AppleWebKit/537.36 (KHTML,
```

데이터 수집

```
#뉴스 헤드라인이 아닌 텍스트 제외
df_concat = df_concat[
    (df_concat['headline'].str.contains('안내헤드라인')) | (df_concat['headline'].str.contains('더보기')) | (
            df_concat['headline'].str.contains('오늘의 인사 종합') | df_concat['headline'].str.contains(
        '동영상기사')) == False]
#공백이 있을 시 drop
df_concat['headline'].replace('', np.nan, inplace=True)
new_df = df_concat.dropna(how='any')


print(new_df.isnull().sum())
# 엑셀파일로 저장하기
writer = pd.ExcelWriter('../Users/wldms/Downloads/environment_news/1year_environment_issue.xlsx')
new_df.to_excel(writer, sheet_name='Sheet1', index=False, header=False, na_rep=' ',
                encoding='utf-16')  # 엑셀로 저장
writer.save()
```

크롤링한 비정형 데이터 정제

결측값 (ex 공백) /이상값 (ex 더보기, 동영상기사 등) 처리하여 수집하고자 한 값만 엑셀 저장

수집된 데이터의 결과물 - xlsx 파일

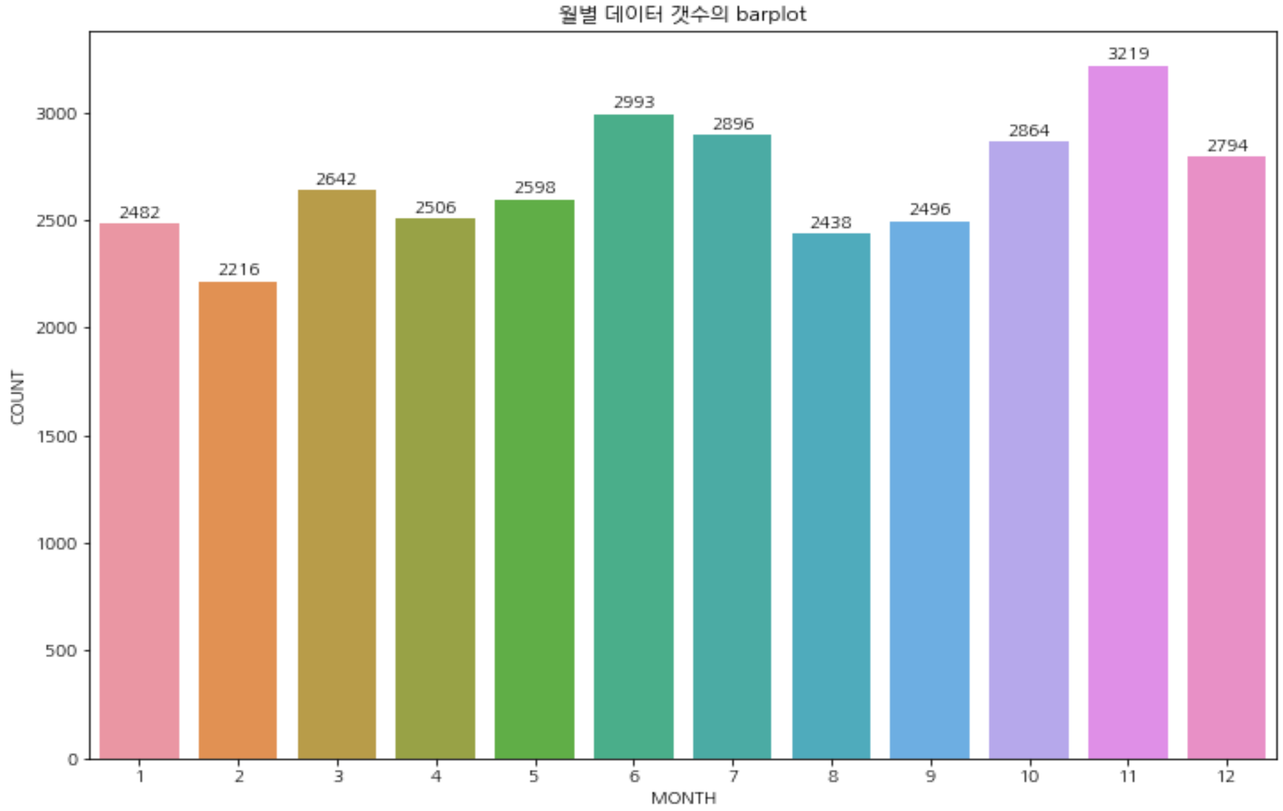
	A	B
32058	대우건설, 발전사업 연계 스마트팜 실증사업 MOU	20220729
32059	대전·충남지역 가뭄 여전…일부 저수지 저수율 30% 안팎	20220729
32060	홈앤쇼핑, MZ세대 환경 서포터즈 '홈앤그리너' 수료식	20220729
32061	올들어 국내 말라리아 환자 193명 발생…작년과 비슷	20220729
32062	고창군, 전북 서해안권 국가지질공원 재인증…내년에 유네스코 도전	20220729
32063	국립수목원, 전나무 숲길 첫 휴식년제 실시	20220729
32064	[내일날씨] 낮 최고 35도 무더위 계속…태풍 간접 영향 강풍	20220729
32065	국립수목원, 전나무 숲길 조성 이후 첫 '휴식년제'	20220729
32066	[내일날씨] 오후 흐려져 곳곳 소나기…서울 낮 최고 35도	20220729
32067	전국이 잠 설쳤다…열대야, 수도권에서 전국으로 확대	20220729
32068	[출근길 인터뷰] '씨낵' 캠페인…"바다 쓰레기 주워오면 과자 제공"	20220729
32069	울산시, 수소 이동수단 규제자유특구 등 혁신 우수사례 6건 선정	20220729
32070	[오늘은] '어흥!' 백두산 호랑이는 언제 사라졌나?	20220729
32071	전국 찜통더위 계속…오후부터 일부 지역 소나기	20220729
32072	민주콩고 '지구허파' 열대우림 석유개발 입찰 공고	20220729
32073	'장마 끝 폭염 시작' 출하 앞둔 복숭아·배추 병충해 심각	20220729
32074	"지구는 스스로 진화하는 생명체" 가이아 이론 만든 러브록 별세	20220729
32075	[내일 날씨]전국 많은 비…태풍 '송다' 영향	20220730
32076	[날씨] 태풍 '송다' 영향으로 전국 많은 비…남부 50~100㎜	20220730
32077	펩시 제로 '겨드랑이 암내' 이유는 이것 때문이었다	20220730
32078	청주 폐기물처리업체 불…경산 자동차 도장 공장 화재	20220730
32079	태풍 '송다' 영향 제주 비…수도권은 불볕더위	20220730
32080	태풍 '송다' 쫓아오는 몬순…찜통 수증기가 한반도 에워싼다	20220730
32081	태풍 '송다' 8월 1일까지 영향…"최대 300㎜ 비·5.3m 파도"	20220730
32082	'탈 플라스틱' 제주플러스 환경포럼에 국제기구도 관심	20220730

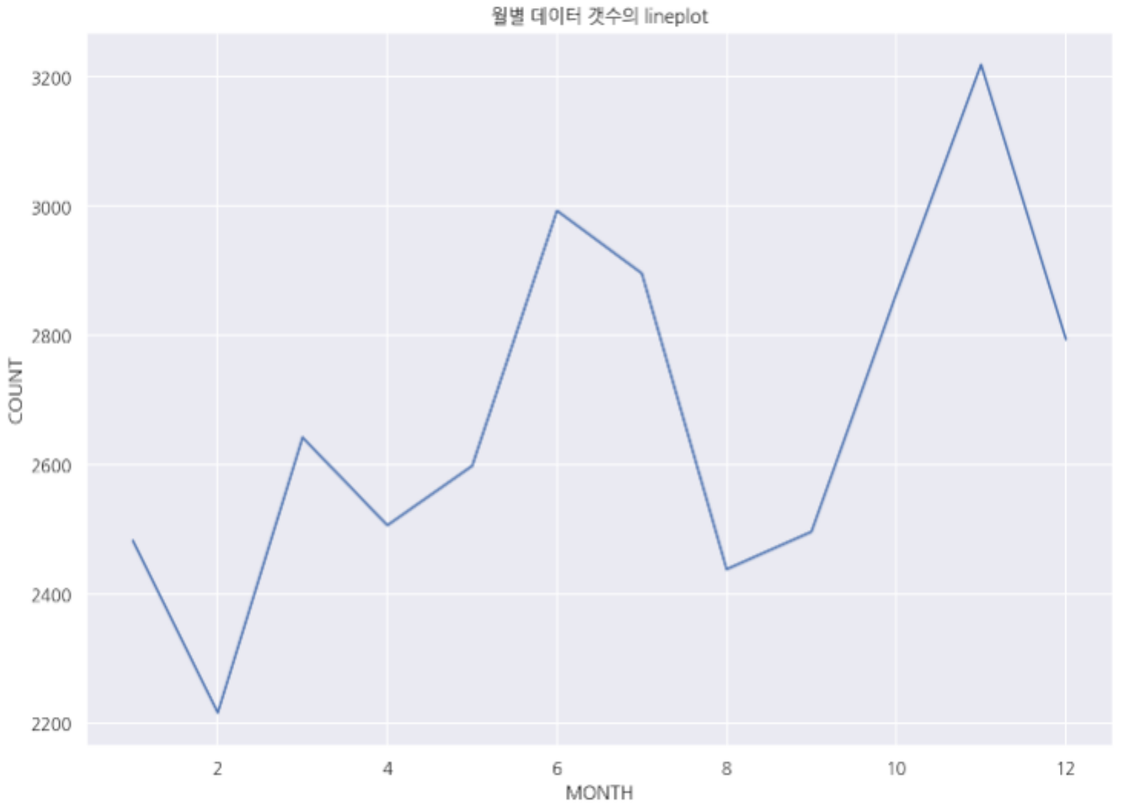
Sheet1

데이터 분석

```
#뉴스의 총갯수 :32144건
df.shape
```

```
(32144, 2)
```





2월의 기사수가 2216건으로 가장 낮고

11월의 기사수는 3219건으로 가장 높다.

데이터 분석

Okt함수이용하여 형태소 분석

```
#올해의 키워드 TOP5
from konlpy.tag import Okt
from collections import Counter
news_list=df['News'].to_list()

okt = Okt()
news_data=[]
extend = news_data.extend
line =[]
for num in news_list:
  line = okt.pos(num)
  n_adj =[]
  # 명사 또는 형용사인 단어만 n_adj에 넣어주기
  for word, tag in line:
      if tag in ['Noun','Adjective']:
          n_adj.append(word)

  #제외할 단어 추가
  stop_words = "하자 곳 도 관 환경 등 명 개 낮 위 첫 곳곳 제 올해 종합 감 날 중 회 종 진 중립 환경부 장관 전국 사업"
  stop_words = set(stop_words.split(' ')) #추가할 때 띄어쓰기로 추가해주기
  # 불용어를 제외한 단어만 남기기
  n_adj = [word for word in n_adj if not word in stop_words]
  #print(n_adj)
  #리스트끝에 리스트항목들을 추가
  extend(n_adj)
#가장 많이 나온 단어 100개 저장
counts = Counter(news_data)
tags = counts.most_common(5)
```

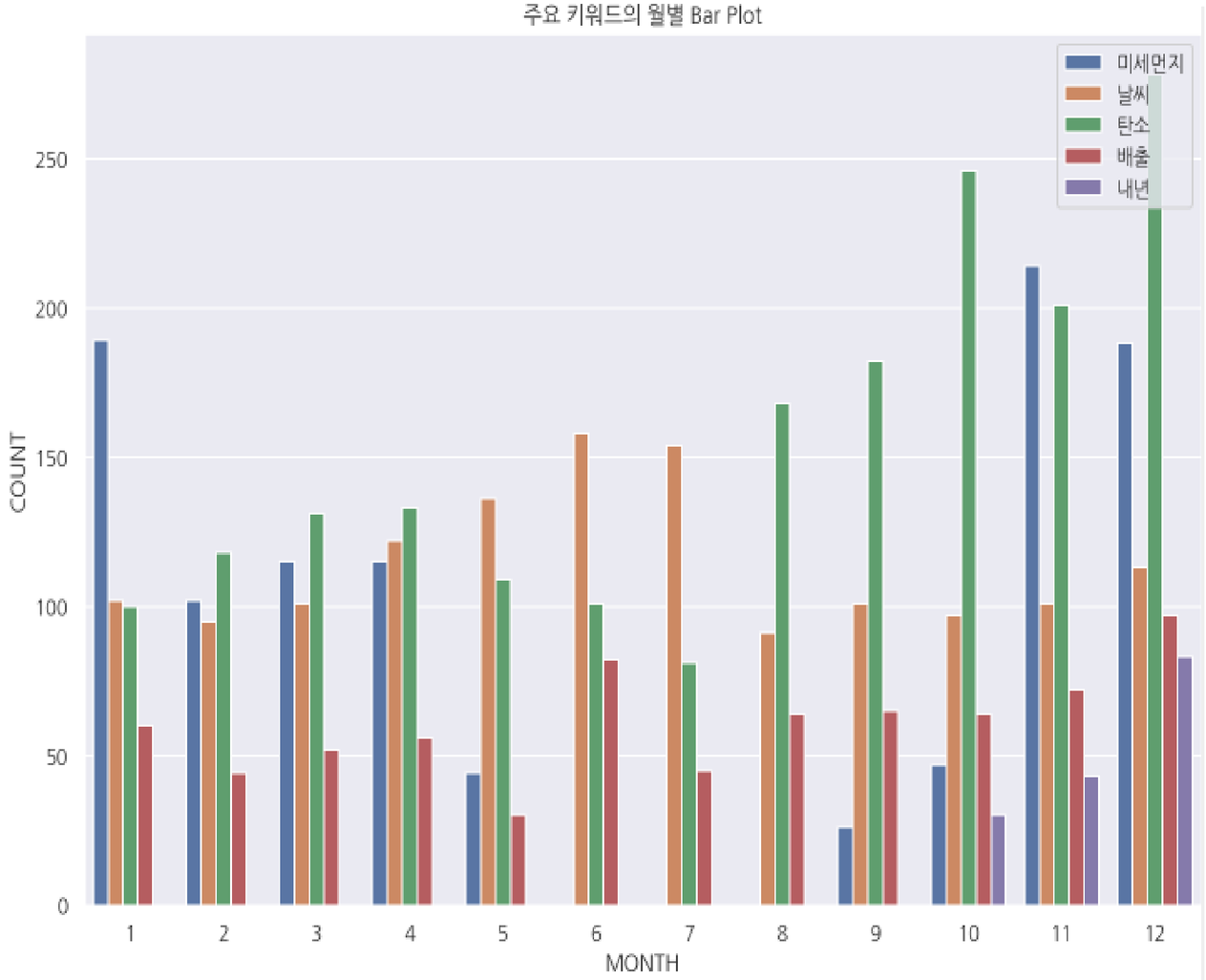
```
#tags(리스트내의튜플) 데이터를 쪼개서 글자와 빈도수로 나누어 데이터프레임에저장
Word= [x[0] for x in main_keyword['tags']]
Wordcount =[x[1] for x in main_keyword['tags']]

main_keyword['Word']=Word
main_keyword['Wordcount']=Wordcount
#기존 tags컬럼 삭제
main_keyword=main_keyword.drop('tags',axis=1)
main_keyword.head()
```

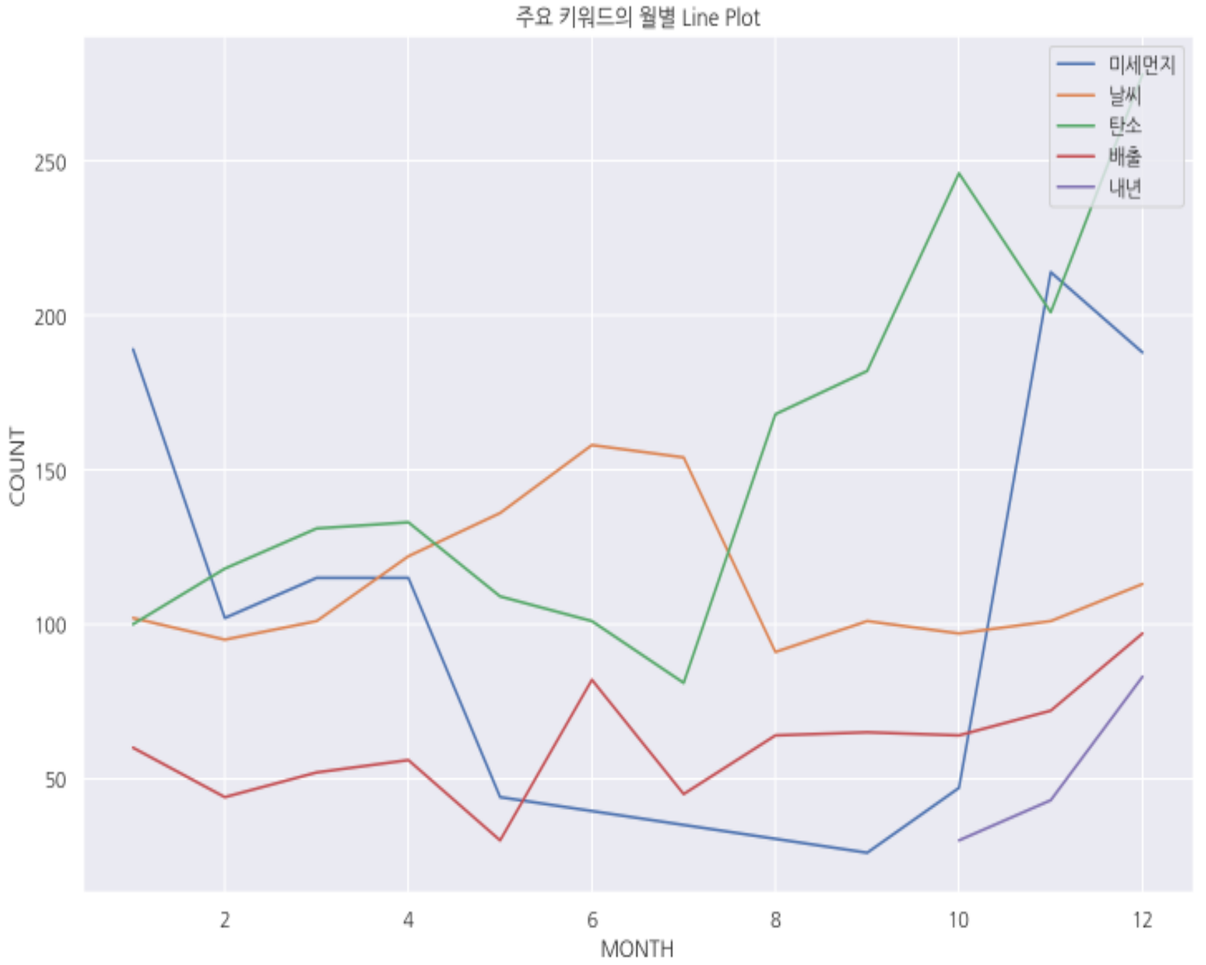
	Word	Wordcount
0	탄소	1848
1	날씨	1371
2	기후	1124
3	미세먼지	1093
4	비	873

빈도 분석 :

올해 환경 뉴스에 가장많이 쓰인단어 5가지

탄소/ 날씨 /기후/미세먼지 / 비





- 겨울인 11,12,1월에 미세먼지관련 기사가 많았다.
- 탄소는 일년내내 꾸준한 기사가 났지만 12월에 가장 두드러지게 나타났다.
- 여름기간에 비 관련 기사가 많았던것을 확인할수 있다.

워드 클라우드

21.08~22.07
1년간의 환경 주요 키워드

탄소중립선언과 행동에 미온적 태도를 보이던 주요국이 앞다퉈 탄소 배출량 '제로'를 선언하면서

최근 1년간 환경뉴스엔
탄소 중립, 탄소 배출 ,녹색,친환경등의 키워드가 가장 많이 쓰였던것으로 보인다.

그외에
22년 호랑이 해를 맞아 호랑이 ,한정애 환경부장관 등 매년 또는 일정 기간 마다 비뀌는 키워드와

미세먼지/초미세먼지/영하/한파/추위/날씨/눈/비 등의 날씨관련 키워드들이 많이 쓰인것으로 보인다.

봄(3,4,5월)의 환경 주요 키워드

봄의 가장 큰 키워드는 날씨

일회용품규제로인해 다회용 컵 사용 뉴스 多

여름(6,7,8월)의 환경 주요 키워드

여름의 가장 큰 키워드는 비

21.08월 카자흐스탄 환경장관과 회담 & 에코스쿨 업무협약식&낙동강 보 개방 및 4대강 회복 방안 논의 & 무공해택 관련기사多 -> 한정애장관님 언급 多

제주 -유해물질 검출 ,멸종위기종 남방큰돌고래 사체 발견

가을(9,10월)의 환경 주요 키워드

가을의 가장 큰 키워드 '기후 '

21.10 기후변화 및 재난대비 취수시설 개선 협약식 , COP26 관련 기사 多

'감축' : 21.10 온실가스 감축목표 급상향 관련 기사 多

겨울(11,12,1,2월)의 환경 주요 키워드

겨울의 가장 큰 키워드는 미세먼지

'미세먼지' : 계절특성에 따른 키워드

'내년' : 연말 특성에 따른 키워드

연관 규칙 분석

support(지지도)

P(A∩B):A와 B가 동시에 일어난 횟수 / 전체 거래 횟수

-> 전체 거래에서 특정 물품 A와 B가 동시에 거래되는 비중
해당 규칙이 얼마나 의미있는지 보여줌.

Confiddence(신뢰도)

P(A∩B) / P(A):A와 B가 동시에 일어난 횟수 / A가 일어난 횟수

-> A를 포함하는 거래 중 A와 B가 동시에 거래되는 비중

lift(향상도)

P(A∩B) / P(A)*P(B) = P (B|A) / P (B)
: A와 B가 동시에 일어난 횟수 / A, B가 독립된 사건일 때 A,B가 동시에 일어날 확률

-> A라는 상품에서 신뢰도가 동일한 상품 B와 C가 존재할 때, 어떤 상품을 더 추천해야 좋을지 판단.
A와 B가 동시에 거래된 비중을 A와 B가 서로 독립된 사건일 때 동시에 거래된 비중으로 나눈 값

연관 규칙 분석

Mlxtend의 Apriori 알고리즘을 적용

1. 데이터셋 샘플 만들기

```
[['탄소','중립','환경부','기후','비']]
```

2. TransactionEncoder를 이용한 학습 시작

```
from mlxtend.frequent_patterns import association_rules
from mlxtend.preprocessing import TransactionEncoder
te = TransactionEncoder()
te_ary = te.fit(dataset).transform(dataset)
```

3. 보기좋게 데이터 프레임으로 생성

```
df= pd.DataFrame(te_ary,columns=te.columns_)
df[['탄소','중립','환경부','기후','비']].head(10)
```

	탄소	중립	환경부	기후	비
0	False	False	False	False	False
1	False	False	False	False	True
2	True	False	False	False	False
3	False	False	False	False	False
4	False	False	False	False	False
5	False	False	False	False	False
6	False	False	False	True	False
7	False	False	False	False	False
8	True	False	False	True	False
9	False	False	False	False	False

연관 규칙 분석

항목 개수가 2개이고
지지도(support)가 0.01 이상인 항목집합만 추려낸 결과

	support	itemsets	length
109	0.055288	(탄소, 중립)	2
102	0.023638	(기후, 위기)	2
108	0.022035	(추석, 연휴)	2
104	0.017628	(비, 날씨)	2
110	0.016026	(찬투, 태풍)	2
106	0.014423	(비, 전국)	2
103	0.012821	(기후, 행동)	2
101	0.011218	(날씨, 교차)	2
105	0.010417	(전국, 날씨)	2
100	0.010016	(온실가스, 감축)	2

** 지지도는 절대 신뢰도보다 높을 수 없다.
같거나 낮을 수 밖에 없는 수치이다.

지지도로 인해서, 일정 이상의 데이터만 가져오고,
특정 이상값의 신뢰도를 추천한다

```
frequent_itemsets = apriori(df, min_support=0.01, use_colnames=True)
frequent_itemsets['length']=frequent_itemsets['itemsets'].apply(lambda x: len(x))
frequent_itemsets[(frequent_itemsets['length'] == 2 ) &
                  (frequent_itemsets['support'] >=0.01)].sort_values(by='support',ascending=False).head(10)
```

연관 규칙 분석

신뢰도(confidence)가 0.7이상인 연관규칙나열

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
2	(순환)	(자원)	0.010016	0.019231	0.010016	1.000000	52.000000	0.009823	inf
7	(찬투)	(태풍)	0.017628	0.021234	0.016026	0.909091	42.813036	0.015651	10.766426
8	(태풍)	(찬투)	0.021234	0.017628	0.016026	0.754717	42.813036	0.015651	4.005054
3	(추석)	(연휴)	0.029647	0.026042	0.022035	0.743243	28.540541	0.021263	3.793311
4	(연휴)	(추석)	0.026042	0.029647	0.022035	0.846154	28.540541	0.021263	6.307292
0	(위기)	(기후)	0.027644	0.048478	0.023638	0.855072	17.638520	0.022298	6.565505
1	(행동)	(기후)	0.015224	0.048478	0.012821	0.842105	17.371031	0.012082	6.026309
5	(탄소)	(중립)	0.071715	0.055288	0.055288	0.770950	13.944134	0.051323	4.124472
6	(중립)	(탄소)	0.055288	0.071715	0.055288	1.000000	13.944134	0.051323	inf

antecedent(전항)과 consequent(후항)컬럼확인

```
from mlxtend.frequent_patterns import association_rules

association_rules(frequent_itemsets,metric='confidence',

                  min_threshold=0.7).sort_values(by='lift',ascending=False).head(10)
```

연관 규칙 분석

4번개체 (연휴) -> (추석) 신뢰도와 향상도를 직접계산

(연휴) -> (추석)규칙의 지지도: ' + str(format(0.022035,".6f")

(연휴) -> (추석)규칙의 신뢰도 : ' + str(format(0.022035/0.029647,".6f")

(연휴) -> (추석)규칙의 향상도 : ' + str(format((0.022035/0.029647)/0.026042,".6f")

```
(연휴) -> (추석)규칙의 지지도: 0.022035
(연휴) -> (추석)규칙의 신뢰도 : 0.743246
(연휴) -> (추석)규칙의 향상도 : 28.540263
```

지지도와 신뢰도는 확률의 개념이므로 1에 가까울수록 연관성이 높다

향상도를 평가도구로 사용하여 판단해볼때

향상도가 1보다 크기에 (연휴)->(추석)은 양의 상관관계가 있다고 말할 수있다.