Big Data Programming

AI 융합학부
20170403 최혜원

21.04.20

INDEX

1. 주제

2. 알고리즘

3. 진행 상태

4. 개발 계획

5. Q & A

1. 주제

이 주제에 대한 설명

“정치 뉴스와 연예 뉴스의 상관관계 분석”

가설 : 어떤 연예계 사건이 터졌을 때, 정치적, 사회적 문제를 덮고자 하는 것일지도 모른다.

2. 알고리즘
구현 flow

If 특정 정치 키워드가 **이슈가 된 시기**를 파악

If 그 시기 전후로 어떤 **연예 관련 keyword 그래프가 급격히 상승**

Then 두 특정 키워드의 추이를 분석, 결과 도출

2. 알고리즘

구현 flow

If 특정 정치 키워드가 이슈가 된 시기를 파악

1) 키워드를 정한다. (ex. 국정원 해킹, 교과서 국정화 추진 논란, LH) (중간)
2) 정한 키워드를 리스트로 정리한다. (날짜데이터 등) (기말)

2. 알고리즘

구현 flow

If 어떤 연예 관련 keyword 그래프가 급격히 상승

1) 키워드 발생 전후기간에 연예 관련 데이터 추이를 파악한다.(중간)
2) 의미 있는 연예 관련 데이터를 추출한다.(기말)

2. 알고리즘

구현 flow

두 특정 키워드의 추이를 분석, 결과 도출

1) 정치 키워드와 연예 키워드 추이를 시각화한다.
2) 의미를 도출한다.

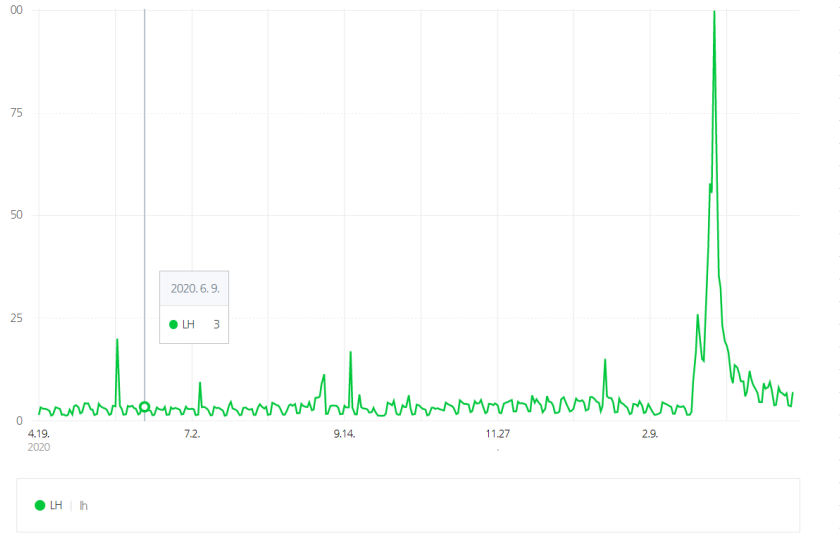
3. 진행 상태

현재 프로젝트 진행 상태 - 데이터 수집

네이버 데이터 랩

	A	B	C	D
1	url	http://datalab.naver.com/keyword/tre		
2	주제	통검		
3	범위	합계		
4	기간	일간 : 2020-04-19 ~ 2021-04-19		
5	성별	전체(여성,남성)		
6	연령대	전체		
7	날짜	LH		
8	2020-04-19	1.30792		
9	2020-04-20	3.16627		
10	2020-04-21	2.80415		
11	2020-04-22	2.8181		
12	2020-04-23	2.67861		
13	2020-04-24	2.30737		
14	2020-04-25	1.15771		
15	2020-04-26	1.38571		
16	2020-04-27	3.18558		
17	2020-04-28	3.03698		
18	2020-04-29	2.7725		
19	2020-04-30	1.35084		
20	2020-05-01	1.41146		
21	2020-05-02	1.14269		
22	2020-05-03	1.31436		
23	2020-05-04	2.61585		
24	2020-05-05	1.42702		
25	2020-05-06	3.44095		
26	2020-05-07	3.70221		

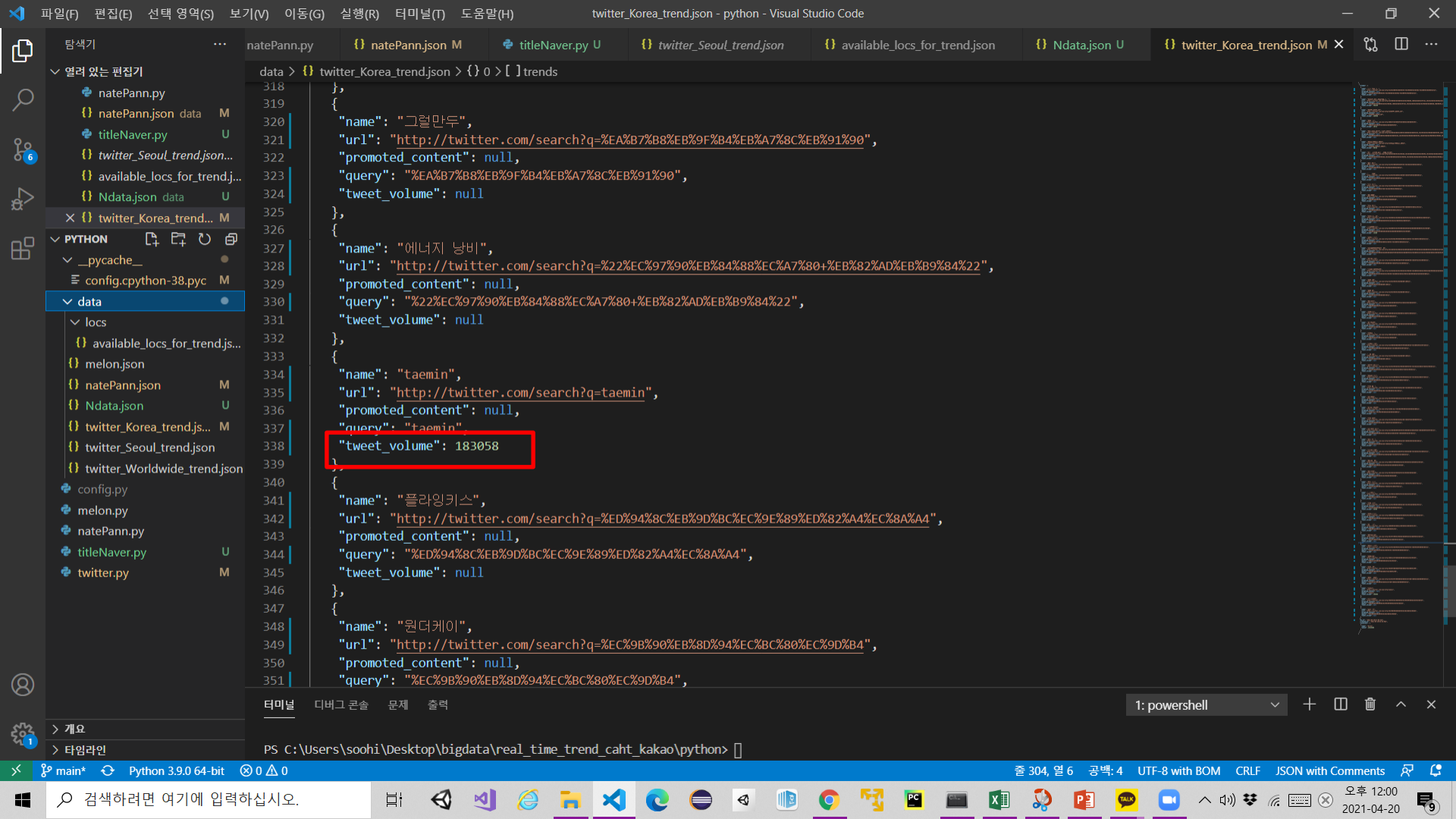
	A	B
322	2021-02-27	1.37069
323	2021-02-28	1.35513
325	2021-03-02	9.76598
326	2021-03-03	16.60175
327	2021-03-04	25.91388
328	2021-03-05	21.25245
329	2021-03-06	14.95799
330	2021-03-07	14.49662
331	2021-03-08	25.51313
332	2021-03-09	41.72165
333	2021-03-10	57.83092
334	2021-03-11	55.53427
335	2021-03-12	100.0
336	2021-03-13	70.38121
337	2021-03-14	35.33277
338	2021-03-15	32.34782
339	2021-03-16	23.11885
340	2021-03-17	19.17093
341	2021-03-18	18.29111
342	2021-03-19	16.55561
343	2021-03-20	10.64205
344	2021-03-21	9.05837
345	2021-03-22	13.48537
346	2021-03-23	12.88022



3. 진행 상태

현재 프로젝트 진행 상태 - 데이터 수집

- 트위터의 파이썬 라이브러리인 tweepy를 활용하여 과거 및 현재 인기 트위터 데이터를 가져와서 가공하였다.
- Tweet volume 이 일정 수 이상 큰 경우의 단어를 모으기 위해 volume 값들을 출력하였다.
- 날짜는 cursor 메서드의 since값을 지정하여 원하는 날짜의 data를 가져올 수 있다.

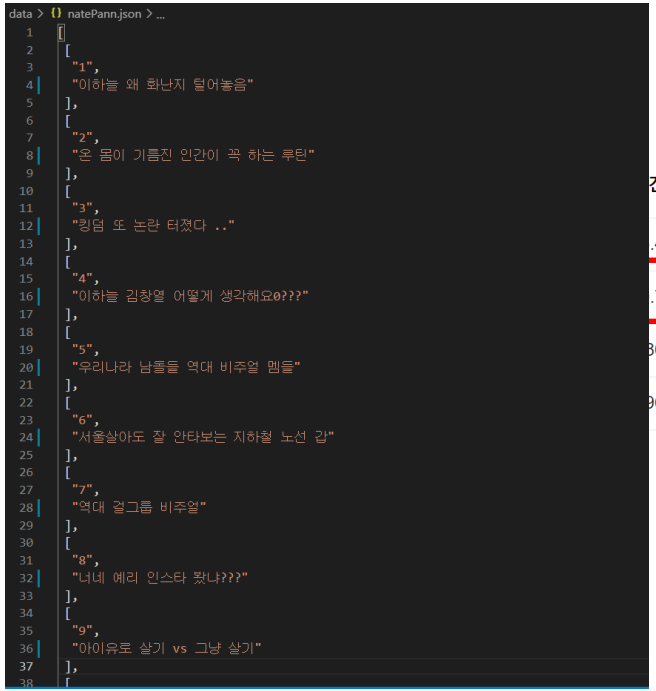

파일(F) 편집(E) 선택 영역(S) 보기(V) 이동(G) 실행(R) 터미널(T) 도움말(H)
twitter_Korea_trend.json - python - Visual Studio Code
탐색기
열려 있는 편집기
natePann.py
natePann.json data M
titleNaver.py U
twitter_Seoul_trend.json...
available_locs_for_trend.j...
Ndata.json data U
twitter_Korea_trend... M
PYTHON
__pycache__
config.cpython-38.pyc M
data
locs
available_locs_for_trend.js...
melon.json
natePann.json M
Ndata.json U
twitter_Korea_trend.js... M
twitter_Seoul_trend.json
twitter_Worldwide_trend.json
config.py
melon.py
natePann.py
titleNaver.py U
twitter.py M
개요
타임라인
natePann.py
natePann.json M
titleNaver.py U
twitter_Seoul_trend.json
available_locs_for_trend.json
Ndata.json U
twitter_Korea_trend.json M
data > twitter_Korea_trend.json > {} 0 > [ ] trends
},
{
"name": "그럴만두",
"url": "http://twitter.com/search?q=%EA%B7%B8%EB%9F%B4%EB%A7%8C%EB%91%90",
"promoted_content": null,
"query": "%EA%B7%B8%EB%9F%B4%EB%A7%8C%EB%91%90",
"tweet_volume": null
},
{
"name": "에너지 낭비",
"url": "http://twitter.com/search?q=%22%EC%97%90%EB%84%88%EC%A7%80+%EB%82%AD%EB%B9%84%22",
"promoted_content": null,
"query": "%22%EC%97%90%EB%84%88%EC%A7%80+%EB%82%AD%EB%B9%84%22",
"tweet_volume": null
},
{
"name": "taemin",
"url": "http://twitter.com/search?q=taemin",
"promoted_content": null,
"query": "taemin",
"tweet_volume": 183058
},
{
"name": "플라잉키스",
"url": "http://twitter.com/search?q=%ED%94%8C%EB%9D%BC%EC%9E%89%ED%82%A4%EC%8A%A4",
"promoted_content": null,
"query": "%ED%94%8C%EB%9D%BC%EC%9E%89%ED%82%A4%EC%8A%A4",
"tweet_volume": null
},
{
"name": "원더케이",
"url": "http://twitter.com/search?q=%EC%9B%90%EB%8D%94%EC%BC%80%EC%9D%B4",
"promoted_content": null,
"query": "%EC%9B%90%EB%8D%94%EC%BC%80%EC%9D%B4",
터미널 디버그 콘솔 문제 출력
1: powershell
PS C:\Users\soohi\Desktop\bigdata\real_time_trend_caht_kakao\python>
main* Python 3.9.0 64-bit 0 0
줄 304, 열 6 공백: 4 UTF-8 with BOM CRLF JSON with Comments
검색하려면 여기에 입력하십시오.
오후 12:00
2021-04-20

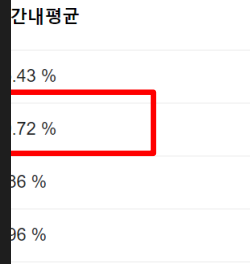
3. 진행 상태

현재 프로젝트 진행 상태 - 데이터 수집

구글 트렌드 api(google-trends-api)를 통해서 원하는 데이터를 가져온다.
(날짜, 키워드 등)

1위 :: 슈퍼리그
2위 :: 비트코인
3위 :: 이현배
4위 :: 버팀목자금플러스
5위 :: 무리뉴 경질
6위 :: 기모란
7위 :: 바르셀로나
8위 :: 아스날
9위 :: F1





4. 개발 계획

앞으로의 개발 계획 - 데이터 가공

1. 정치적으로 굵직한 사건들을 시기별로 정리할 것.
   (선별 방법 : 검색량의 **볼륨 변화량이 급증**했을 경우 정치적 굵직한 사건이라고 판단)
2. 그 시기 전후로 연예관련 데이터를 모을 것.
3. 뉴스, 커뮤니티, 트위터 이외의 연예계 관련 데이터 소스(뉴스 기자, 뉴스 사이트: 피드백 반영)를 찾을 것.
4. 최대한 많은 데이터를 수집하여 의미 있는 데이터만 전처리하여 시각화 할 것.

QnA

Thank you