

# AI파이썬빅데이터수집 기말고사 프로젝트 보고서

2021481054 박태희

## 1 ) 중간고사 목표 :

2022 월별 인천 코로나 확진자 수, 관련 뉴스 취합 후 연간 여론 파악  
지역은 서구, 동구, 미추홀구 / 2020 ~ 2022년까지 조사할 예정

## 2-1 ) 결과 :

2022년 상반기까지의 확진자 수와 코로나 관련 뉴스 데이터 수집 후 동향 파악  
2022년 월별 뉴스 데이터 수집 완료 (1.2만개)

### - 목표 차이 이유 :

1. 확진자 수 데이터를 찾는 데 성공했으나 날짜별 입력에 많은 어려움이 있었음
2. 교수와의 상의하에 지나친 분량을 줄이고 2022년의 데이터를 수집하는 것으로 결정함

## 2-2 ) 프로젝트를 위해 중간고사 이후 소비한 시간 : 약 5일

## 2-3 ) 이번 프로젝트를 통해 보충해야 할 것 :

1. 웹드라이버를 통해 어떤 동작이 발생하는지에 대해 이해
2. 주기적인 크롤링 실습과 통계 도출
3. 크롤링 효율 향상을 위해 코드를 개선

## 2-4 ) 어려웠던 점

1. 뉴스 내 일부 html 구조 변화로 인해 잘못된 정보를 가져와서 시간적인 손해를 봄
2. 최근까지 뉴스 데이터를 수집해야 하기에 분량과 시간이 상당히 많았음
3. 포털 사이트에서 상세 날짜까지 보여주지 않았기 때문에 불가피하게 링크에 접속해서 추가 데이터를 수집해야 했음
4. 종합적으로 크롤링에 수많은 시간이 요구되었음

## 변화하는 코로나

학번/이름 2021481054 박태희

### 데이터 수집 방법

- Visual Code Studio (Python 3.10)

브라우저 : 크롬 (Chrome 102)

드라이버 : ChromeDriver 102.0.5005.61

데이터 : 네이버 포털 사이트

(비고 : 언론사 3개, 2022-01-01 이후 뉴스 데이터)

1월

분석 결과





- 2022년 1월에 요점이 된 키워드는,  
지난해 + 신규 + 오미크론 + 확산

- 지난해 신종 코로나바이러스 '오미크론'이  
떠들썩해지면서 전체 확진자 수 급증 및  
근로시간 단축으로 인한 임금 감소가  
진행되고 있다.

2월

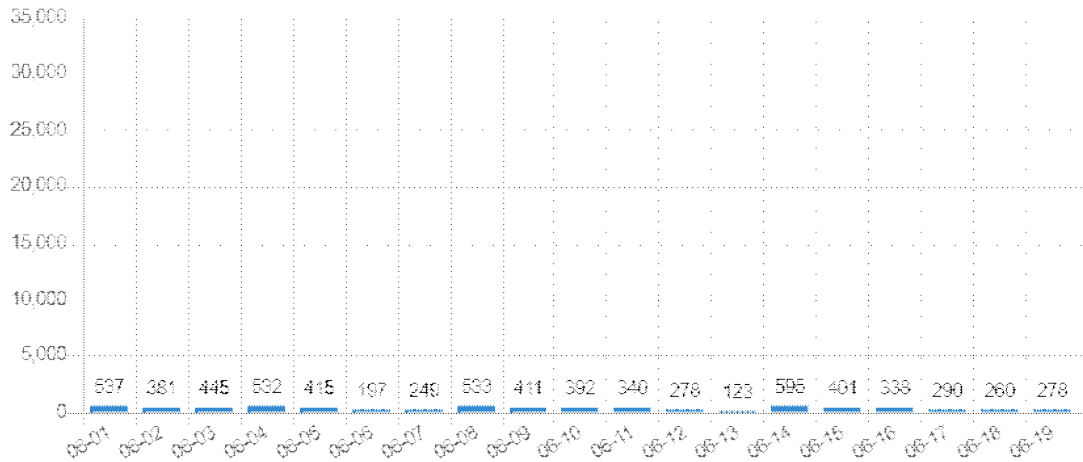
분석 결과

	<ul style="list-style-type: none"> <li>- 2022년 2월에 요점이 된 키워드는, <b>오미크론 이슈 외</b>, 재택 + 자가 (근무) 후보 + 대표 + 지역 + 대선</li> <li>- 근로시간 급감으로 인한 노동 정책의 변화가 발생하게 되었고, 근로시간 유연화를 위해 재택근무를 시행한 업체가 나타나는 것을 알 수 있다.</li> <li>- 3월부터 대선 투표가 시작되기 때문에 후보자 이름과 후보, 대선 및 관련된 단어들이 나타나고 있다.</li> </ul>
<p style="text-align: center;"><b>3월</b></p> 	<p style="text-align: center;"><b>분석 결과</b></p> <ul style="list-style-type: none"> <li>- 2022년 3월에 요점이 된 키워드는, <b>오미크론 이슈 외</b>, 방역 + 상황 + 격리 + 검사 + 정부, (선거관리)위원회 + 투표 + 당선인 + 대통령</li> <li>- 종식을 기대하기 힘든 코로나19 확산을 막기 위한 사회적 거리 두기가 18일부터 실내외 마스크 착용을 제외하고 전면 해제되는 소식이 나타났다. 이때부터 확진자 수가 급증하기 시작한다.</li> <li>- 대통령 선거일로 인해 관련된 단어들이 언급되는 것으로 확인된다.</li> </ul> <p style="text-align: center;"><b>4월</b></p>

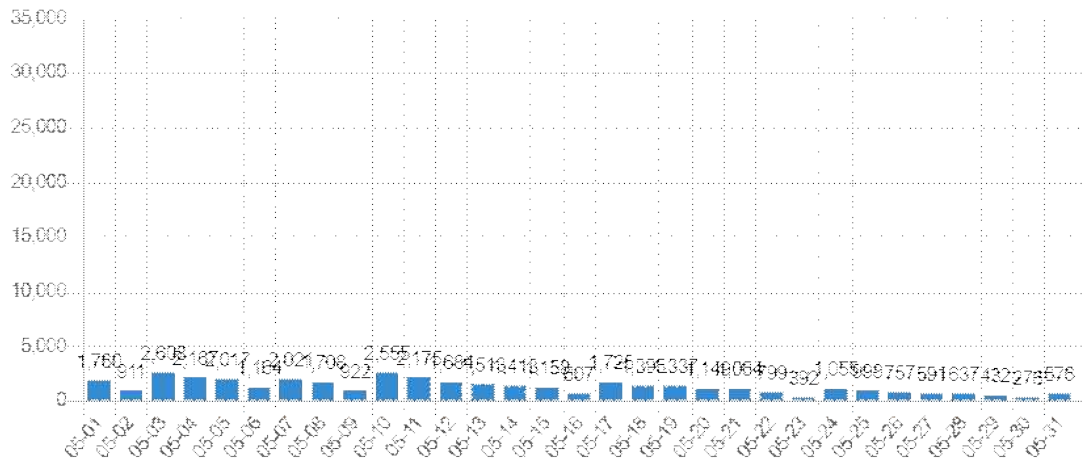
	<ul style="list-style-type: none"> <li>- 2022년 4월에 요점이 된 키워드는, <b>오미크론 이슈 외</b>, 올해 + (거리 + 두기) + 해제 중국 + 상하이 + 봉쇄 우크라이나 + 러시아 + 전쟁</li> <li>- 정부 측의 거리 두기 해제 및 방역 규정 완화, 중국에 있는 상하이 및 도시 봉쇄 명령, 우크라이나와 러시아 간의 전쟁 발발</li> </ul>
<p>5월</p>	<p>분석 결과</p> <ul style="list-style-type: none"> <li>- 2022년 5월에 요점이 된 키워드는, <b>오미크론 이슈 외</b>, <b>팬데믹</b> 북한 + 김정은</li> <li>- 북한에서 코로나 확산세로 인한 사망자 수에 대해 일절 없다는 것으로 의심을 사고 있고, 지속되는 핵 도발로 인해 남북 양측간의 마찰이 일어나면서 지원에 대해 갈등을 겪고 있는 상황이다.</li> </ul>
<p>6월</p>	<p>분석 결과</p> <ul style="list-style-type: none"> <li>- 2022년 6월에 요점이 된 키워드는, <b>오미크론 이슈 외</b>, <b>팬데믹</b> 중국 + 봉쇄 + 해제 정부 + 투표 + 진행 (원숭이 + 두창)</li> <li>- 사람들이 팬데믹을 겪으면서 인공지능 채용 여지와 로컬의 가치가 급부상하는 것을 느끼고 있다. 그 외에도 피해에 관한 내용을 전체적으로 다루는 때인 것으로 보인다.</li> </ul>

- 중국은 7월부터 봉쇄 해제를 하겠다는 것을 알 수 있었다. 상하이 봉쇄 이후 공급망 교란과 소비 침체로 인해 실업률과 경제가 안 좋아진 것으로 보인다.
- 또 다른 감염병인 원숭이 두창이 새롭게 발생하게 되면서 확산하는 것에 대해 두려움이 나타나고 있다.
- 6월 지방선거 투표가 있는 날이므로 관련 단어가 드러난 것으로 확인된다.

## 일별 확진자 추이

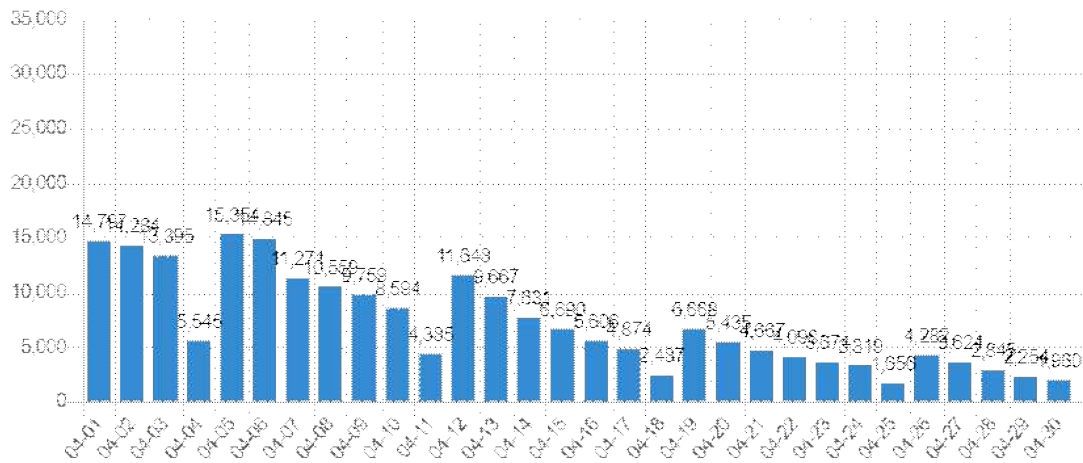


## 일별 확진자 추이

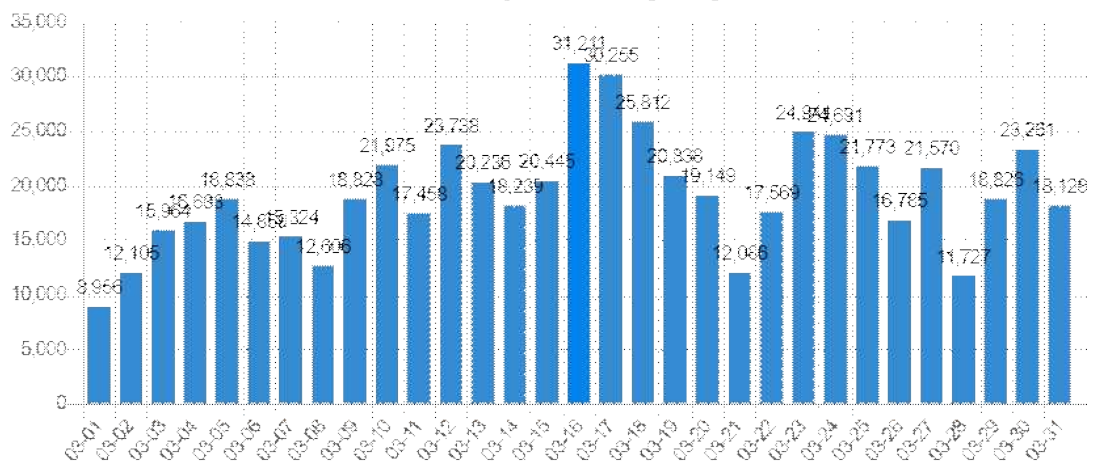




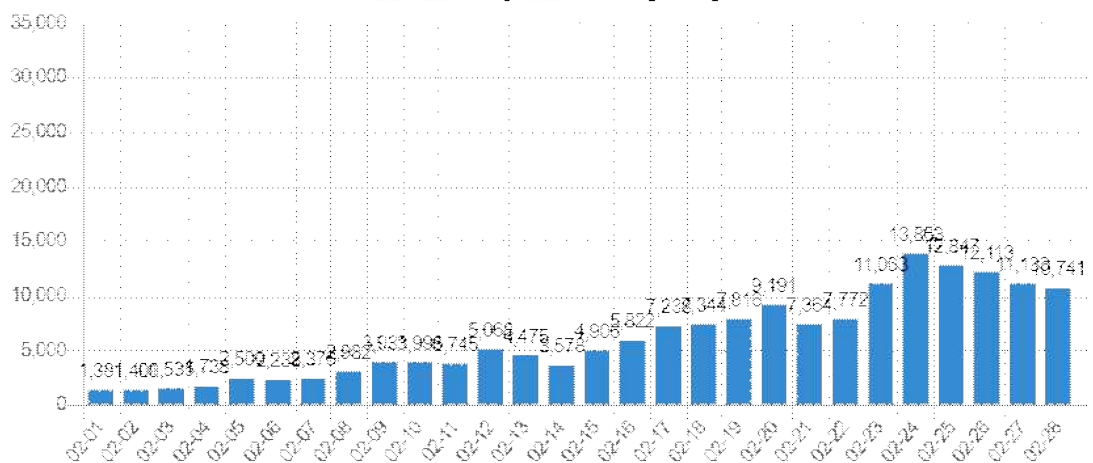
## 일별 확진자 추이



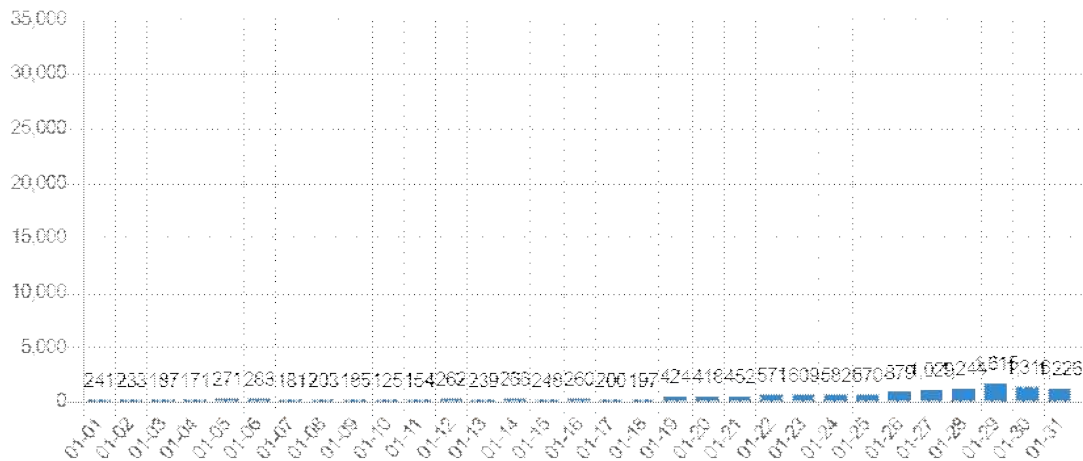
## 일별 확진자 추이



## 일별 확진자 추이



## 일별 확진자 추이



### 결론

- 코로나는 아직도 끝날 기미가 보이지 않고 있다.
- 감염병으로 인해 손해를 보고 있었으나 사람들은 적응해 나가고 있다.
- 새로운 감염병인 '원숭이 두창'은 전파력이 강하지 않으나 확산에 주의가 필요하다.
- 북한의 핵 도발로 인해 전과 달리 한·미 대 북·중으로 대치 우려가 일어날 수 있다.

### 프로젝트 주제 선정 이유

코로나로 인해 변화하는 세상에서 뉴스를 통해 올해 동안 어떤 키워드가 등장했고, 흐름이 어떻게 변화했는지 알아보기 위해 조사해 보기로 했다.

초기에는 공공 데이터와 네이버 API를 사용하기로 했지만, 생각했던 것과는 다른 데이터가 있었기 때문에 통계청이나 네이버 포털 사이트를 통해 가능한 조사를 하고 얻은 데이터로 시각화를 나타낸 것이 위의 워드 클라우드들이다.

제대로 얻지 못한 것은 '2022년 상반기 코로나 확진자 수' 데이터인데, 본래 파이썬을 이용해서 크롤링 한다는 취지와는 맞지 않게 수집했기 때문에 많이 아쉽다고 생각한다.

출처 : <https://www.incheon.go.kr/covid19/index>