2020 년도 동계 **g-step SDSA** 활동 보고서

학과 : 시스템경영공학부
전공 : 기술서비스공학과
학번 : 201711906
이름 : 박민규

1) SDSA 교육과정 정리

1 주차에는 데이터 사이언스란 무엇인가와 중요한 이유와 같은 기본원리에 대한 수업이 이루어졌습니다.

비디오: 산업이 직면한 '빅데이터' 문제는 무엇입니까?
비디오: '데이터 사이언스의 기본 원리(테크니컬)'에 오신 것을 환영합니다.
우리는 수많은 데이터에 둘러 싸여 있습니다. 우리가 사용하는 모든 도구, 우리가 취하는 모든 행동이 데이터를 생성하고 사용합니다. 조직, 사람, 세계의 모든 사람은 우리 주변의 모든 데이터로부터 통찰력과 지식을 얻습니다. 그렇기 때문에 데이터를 잘 다루는 것이 중요합니다.

데이터 사이언스는 여러 분야의 주제와 다양한 방법이 혼합된 것입니다. 주로 컴퓨터 및 머신 러닝과 결합된 통계 분석과 주제 관련 전문 지식, 즉 도메인과 관련이 있는 올바른 질문을 하는 능력과 지식과 통찰력을 얻을 수 있는 능력이 결합된 것입니다. 데이터에서 얻은 것을 행동으로 옮기고, 더 나은 의사결정을 내릴 수 있게끔 하며, 프로세스를 개선하고, 기존에 하고 있는 모든 것을 향상시키고 더 나은 삶을 창조합니다.

데이터 사이언스 프로젝트의 간략한 프로세스는 데이터를 수집 및 처리하고 수집한 데이터를 사용하기 위해 클렌징 작업을 한 후 통계 및 머신 러닝을 통하여 데이터 간 분석을 실시하여 패턴을 발견하고 발견한 패턴으로부터 창출한 유의미한 효과를 스토리텔링을 활용한 시각화를 통해 보고하는 순으로 이루어집니다.

SDSA 과정에서 데이터 사이언스 실습을 하는데 기반으로 사용한 언어는 Python 입니다. 저는 학교 1,2 학년 수업시간에 Java 만을 배워 Python 을 다뤄본 적은 없었으나 1,2 주차 수업에선 그런 사람들을 위해 Jupyter 사용환경과 변수, 코드블럭, 불린 오퍼레이터, 리스트 자료형, 함수, 딕셔너리 자료형, 객체와 클래스, 메소드, 상속과 같은 Python 기초 실습을 통해 기초능력을 기를 수 있었습니다.

2 주차에는 간략한 데이터 사이언스, 데이터 사이언스 프로세스, 데이터 사이언스 툴킷, 데이터 유형, 응용 프로그램 예, 세부적인 판독과 같은 데이터 사이언스의 핵심 개념과 기술에 대한 수업이 이루어졌습니다.

토론 활동 : 데이터 사이언스가 다른 분야와 어떻게 다른가요?

빅데이터는 IT 또는 컴퓨터 과학 및 엔지니어링에 중점을 두며 특허를 식별하고 고객과 조직에 유용한 것을 발견하기 위해 데이터를 수집하고 사용하며 이를 사용함에 있어 볼륨, 다양성, 속도, 정확성이라는 4V 측면의 처리에 중점을 두지만, 데이터 사이언스는 데이터를 분석하고 통계를 작성하기 위할 뿐만 아니라, 그 숫자들을 의미 있는 것으로 바꾸어 프로세스와 삶을 개선하고 자신과 조직 및 고객의 유익을 창출하기 위해 다양한 분야의 방법과 관점의 혼합에 중점을 둔다는 점에서 차이가 있습니다. 또한, 데이터 사이언스는 통계 및 컴퓨터 공학, 수학, 머신 러닝, 프로그래밍 및 계산도구 분야를 통합하는 학문입니다. 데이터 사이언스 프로세스를 수행하기 위해서는 데이터를 다양한 형식으로 저장하고 대용량 데이터를 관리하며 다양한 소스에서 데이터를 가져와 집계하고 통합할 수 있어야 합니다. 그렇기 때문에 각 분야에 대한 이해를 모두 갖추고 있어야 합니다. 이러한 데이터 사이언스 프로젝트를 통해 창출한 정보를 상황에 맞게 적용하고 비즈니스 관리자나 고객이 이해할 수 있도록 스토리텔링을 하는 것 까지가 데이터 사이언티스트의 역량입니다.

비디오: 데이터 사이언스 프로세스

데이터는 다양한 형태를 가지고 있으므로 작업할 수 있는 형식으로 만들기 위해 많은 처리작업을 수행해야 합니다. 데이터 사이언스 프로세스의 첫 단계는 데이터

수집입니다. 온라인 데이터를 수집한다고 하면 해당 사이트에서 크롤링을 통해 전체 데이터를 수집합니다. 그러나 수집한 데이터를 그대로 사용할 수 없기 때문에 적절한 샘플링 작업을 진행하여야 합니다. 이 과정에서 누락되거나 불완전한 정보가 있는 경우, 혹은 누락시키고 수집해야 하는 데이터 포인트가 있는 경우에 대비하여 데이터 큐레이션, 데이터 클렌징 및 탐색 데이터 분석과 같은 몇가지 전처리 작업을 진행하여야 합니다. 이를 위해 몇가지 표준 테스트를 실행하고 데이터에 대한 표준 시각화를 그려 일반 패턴이나 데이터 품질 문제를 확인합니다. 전처리 작업 후에 실제 분석을 시작할 준비를 합니다. 통계 분석이나 머신 러닝 기술을 활용하여 목적에 맞는 결과를 얻습니다. 분석을 통해 얻은 결과를 바탕으로 보고를 진행해야 합니다. 이 과정에서는 어떤 결과가 잠재 고객에게 가장 흥미로울지 실제로 청중은 누가 될 것인지를 결정하고 그 결과의 형태와 순서를 정하되, 스토리를 만들어 호소력 있고 설득력 있도록 해야 합니다. 여기 까지가 데이터 사이언스 프로세스입니다. 프로세스 진행 과정에서 분석으로 돌아가서 다시 분석을 실시하거나 업데이트된 데이터 세트에서 분석을 실행하는 등의 과정이 발생할 수도 있습니다.

비디오: 데이터 사이언티스트의 역할

데이터를 분석하는 것은 데이터를 더욱 간결한 형식으로 표현하는 모델을 설계한다는 것입니다. 따라서 통계적 방법에 기반하여 데이터의 특정 기능을 알아야 하며, 사전 정의된 카테고리에 따라 해당 데이터를 특정 요소별로 분류할 수 있어야 합니다. 예를 들면 이메일의 스팸여부의 분류 등이 있습니다. 또한 보유한 데이터를 기반으로 무엇이 일어날지, 미래에 데이터가 어떻게 표시될지 예측할 수 있어야 합니다. 위와 같은 것들을 수행하기 위한 다양한 기법들이 있습니다. 최종적으로 데이터를 통해 유의미한 결과를 얻었으면 이를 상황에 맞게 적용할 수 있어야 합니다. 그리고 스토리 또는 보고서를 비즈니스 용어로 작성할 수 있어야 하며 여기에는 인포그래픽과 시각화 요소가 포함이 됩니다. 보고서를 작성할 때에는 흥미로운 결과를 일관성 있게 스토리를 구성하여 청중에게 흥미롭고 매력적으로 전달할 수 있도록 구성할 수 있어야 합니다.

토론 활동 : 데이터 사이언티스트에게 필요한 기술은 무엇이라고 생각하십니까?

데이터 사이언스는 데이터 수집, 데이터 큐레이션, 데이터 클렌징 및 탐색데이터분석, 데이터 분석, 시각화를 통한 보고 및 결정의 프로세스로 이루어 집니다. 이러한 프로세스를 수행하기 위해 데이터 사이언티스트는 어떤 데이터를 발견하여 수집, 탐구하며, 데이터가 대답 할 수 있는 질문의 종류를 파악하고, 이야기를 전달하는 데 관련된 질문을 계획하고, 이해하려는 방식 및 알고리즘 유형을 계획, 선정하며 분석을 수행할 모델을 만든 다음, 배운 내용을 보고하기 위해 시각화를 생성하는 등의 활동을 하여야 한다. 이 과정에서 적절한 데이터 샘플링기술, 크롤링 기술, 데이터를 전처리 하기 위한 관리 기술, 갖고 있는 데이터를 다른 외부소스와 결합하는 기술, 얻은 결과를 적절하게 시각화 하고 보고할 수 있는 기술 등이 필요합니다.

비디오: 데이터 사이언스 도구 키트

요즘 데이터 사이언스 분야에서 일자리를 찾고 있다면 자바스크립트, 파이썬, 자바 뿐만 아니라 Scala 및 자체 프로그래밍 언어를 갖고 있는 R 이라는 다양한 프로그래밍 언어를 접하게 됩니다. SQL 또는 관계형 데이터베이스와 NoSQL 데이터 베이스와 MongoDB, 대부분의 빅 데이터 및 데이터 사이언스 프로그래밍 프레임 워크에서 주로 사용되는 Cassandra, 데이터 시각화를 위한 D3 같은 도구들도 있으며 데이터로 시각적인 보고서를 만드는 포괄적인 도구인 Tableau, 유명한 도구인 Excel 과 전문 시각화 도구인 Gapminder, Hadoop, HDFS, MapReduce, Spark, Storm 등과 같은 빅데이터 처리 도구들도 있습니다. 마지막으로, 통계 분석으로 유명한 R 에서 데이터를 분석하는 상당 수의 소프트웨어 라이브러리와 머신 러닝 라이브러리 Weka, 데이터 마이닝 및 데이터 분석을 위한 RapidMiner 등도 있습니다. 텍스트 형식의 문서를 분석하고 텍스트에서 통찰력을 추출하는 도구인 Open Calasis 와 같이 특정 유형의 데이터를 처리하는 전문적인 도구들도 있습니다. 위와 같이 매우 다양한 도구들을 활용하여 데이터 사이언스 프로젝트를 진행해야 합니다. 도구를 선택할 때에는 데이터 유형에 따라, 데이터 규모에 따라 데이터의 업데이트 속도에 따라 적절한 도구를 사용할 수 있어야 합니다.

비디오: 데이터 사이언스의 '데이터'

데이터는 어디에나 있고 다양한 형태와 종류를 갖고 있습니다. 데이터는 정성적, 정량적, 구조화, 비구조화, 공개, 비공개로 크게 6 가지 유형으로 분류할 수

있습니다. 정량적 데이터는 숫자로 나타나며 이산형데이터와 연속형데이터로 나뉩니다. 정량적 데이터의 예시는 보유하고 있는 차량의 수, 고객연령, 국가의 GDP, 온도 등이 있습니다. 정성적 데이터는 주로 프리텍스트로 제공되며 카테고리 또는 자연언어로도 나타납니다. 이항 데이터와 명목데이터 같은 특별한 유형으로도 나뉩니다. 정성적 데이터의 예시로는 예 아니오, 일주일의 요일, 크다 중간이다 작다 등이 있습니다. 구조화데이터는 예측가능한 데이터이며 알고리즘을 사용하는 컴퓨터가 처리하기에 용이한 데이터입니다. 비구조화데이터는 소셜미디어의 게시물과 같은 텍스트 스트림 또는 태그 등이 있으며 이러한 비구조화 데이터를 사용하려면 NoSQL 데이터베이스와 같은 이에 특화된 언어를 통해야 합니다. 공개데이터는 데이터를 활용할 수 있도록 데이터 공급자가 공개해 두는 데이터입니다. 정부 혹은 페이스북과 같은 조직에서 공개 데이터를 주로 제공합니다. 공개 데이터를 사용할 때엔 해당 데이터가 손상되었거나 신뢰할 수 있는지의 유무를 파악하는 데이터 선별력과 통찰력이 필요합니다. 비공개데이터는 타인이 활용할 수 없도록 공개하지 않는 데이터이며, 회사의 기밀데이터, 소매업체의 고객거래내역 등이 있습니다.

토론 활동 : 이미 사용하고 있는 데이터는 어떤 유형의 데이터입니까? 이 데이터를 사용하거나 분석하려고 할 때 어떤 유형의 문제점을 예상할 수 있습니까?

이미 사용하고 있는 데이터의 유형에는 정량적데이터, 정성적데이터, 구조화데이터, 비구조화데이터, 공개데이터, 비공개데이터 등이 있습니다. 다양한 유형의 데이터들을 분석하고 활용하기 위해서는 데이터 속성의 복잡성과 다양성을 이해해야 하며 각 차원을 잘 파악하는 것이 중요합니다. 차원을 잘 파악하면 각 차원에 맞는 데이터를 적절한 유형의 모델, 알고리즘, 기술을 선택하여 데이터 사이언스 프로젝트를 진행할 수 있기 때문입니다.

비디오: 데이터 사이언스의 다양한 응용

데이터 사이언스의 응용 사례 중 유용한 영역은 크게 건강관리, 개인정보, 스포츠 분석, 소셜미디어 분석으로 4 가지 영역이 있습니다.

건강관리 분석은 의료 데이터가 대량으로 제공되기 때문에 제약회사나 의료 사이언티스트가 임상연구에서 활용하여 분석해야 할 데이터의 양을 예상 할 수 있습니다. 애플리케이션의 시나리오는 매우 다양하며 임상시험이나 의학연구부터 기존 시험 데이터 및 환자 보고서를 통한 관찰을 통해 조사하는 방법 등의 소스를 통해 분석하면 의학 및 제약 연구에 대한 새로운 통찰력을 얻고 연구주기를 단축하여 신약을 테스트하고 시장에 더 빨리 출하할 수 있게 됩니다. 예시로는 IBM Watson 의 데이터 분석이 있습니다.

개인 정보 분석은 일정, 이메일, 일상활동에서의 데이터를 관리하여 우리가 일상생활에서 필요한 정보를 가지고 있는지 확인할 수 있습니다. 이를 위해 Goggle Now 라는 검색엔진이 유명합니다. 고객에게 보다 관련성 있는 정보를 제공하여 그 정보를 바탕으로 고객의 일정 선호도 및 매일 하는 일에 대해 해당 정보가 보유한 지식을 바탕으로 고객에게 특정 활동의 수행을 제안합니다.

스포츠 분석은 운동선수들의 경기장에 카메라와 센서를 두고 팀 전원의 모든 움직임을 분석하고 통계를 내어 누가 부상을 입었는지 또는 퍼포먼스를 향상시키기 위해 어떤 방법으로 훈련 받았는지 등을 확인할 수 있습니다.

이러한 분석을 통해 얻은 통찰력을 바탕으로 운동 선수의 훈련이나 영양의 변화 측면에서 퍼포먼스에 어떤 측면의 영향을 미치는지를 결정하는 한계이익이라는 이론이 새롭게 나타났습니다. 스포츠 분석의 예시로는 Opta 가 있습니다.

마지막으로 소셜미디어 분석에서는 트위터에서 나온 '당신이 자리를 비운 사이'라는 도구가 관심을 받고 있습니다. 트위터에서는 팔로우한 사람들의 게시물을 가장 최근 게시물부터 가장 오래된 게시물까지 시간순으로 정렬하여 볼 수 있는데 '당신이 자리를 비운 사이'라는 도구로 인해 플랫폼에 로그인하지 않은 상태에서 수백만 건의 게시물 중에 관련이 있는 것만 선택하여 필터링을 하여 그동안 올라온 수천, 수만 건의 게시물 중에서 이용자와 가장 관련이 있다고 생각하는 5 개를 선정하여 게시합니다.

2 주차 실습 **'데이터 사이언스를 위한 Python'**에서는 Python 을 활용하여 영국의 오픈소스 데이터 중 교통 흐름 데이터의 csv 파일을 읽어 들여 list 자료구조와

Dictionary 자료구조, for 함수, 그리고 magic.py 에서 함수들을 상속하여서 원하는 정보를 추출하였습니다. 또한

Matplot lib 의 일부인 pyplot 을 사용하여 산점도를 통해 시각화 하였고, 반응변수와 예측변수를 지정하여 회귀방정식을 만들어 원하는 정보를 가져오는 방법을 실습하였습니다.

3 주차 강의에서는 데이터 수집, 저장 및 관리에 중점을 두고 데이터 사이언스의 체험적 실습을 통해 다양한 데이터 소스를 결합하는 방법에 대해 배웠습니다.

비디오: 데이터 수집, 데이터 소스

데이터에는 정적 데이터와 동적(스트리밍) 데이터가 있습니다. 정적 데이터는 주로 데이터베이스에서 얻으며, 딱히 변하지 않습니다. 예시로는 웹상의 데이터가 있습니다. 반면, 동적 데이터는 실시간으로 처리해야 하는 데이터이며 예시로는 금융데이터 또는 GPS 데이터가 있습니다.

웹에서 데이터를 가져오기 위한 가장 좋은 방법은 RESTful API 를 사용하는 것입니다. 이를 통해 회사나 서버에서 데이터를 사용할 수 있으며 웹이 작동하는 방식의 기초입니다. URL 은 서버의 특정 리소스를 식별할 수 있게 합니다. 하이퍼텍스트 전송 프로토콜인 HTTP 는 GET, POST, DELETE, PUT 이라는 일련의 동사가 있습니다. 이러한 동사 중 하나는 URL 이 식별되는 리소스에서 수행하며 정상적인 브라우징을 위해서 GET 을 사용합니다. 클라이언트나 브라우저 소프트웨어에 GET 을 통해 요청하는 특정 형식의 데이터를 반환 받습니다.

사용자 또는 페이지의 다른 특정 리소스를 활용할 때는 주로 API 를 활용합니다. 그러나 API 를 활용할 때에는 주의해야 할 점이 몇 가지 있습니다.

모든 회사가 공개적으로 사용 가능한 API 에서 데이터를 사용할 수 있는 것은 아니며 가지고 있는 데이터가 원하지 않는 데이터이거나 필요하지만 사용할 수 없는 방대한 양의 데이터일 수 있습니다. 많은 경우, 이러한 API 는 인증 및 속도 제한을 보호됩니다. 따라서 분석을 실시할 때 많은 시간이 들게 됩니다. 또한 API 를 정기적으로 쿼리해야 하는 곳에서 사용하는 웹 어플인 경우 API 가 이용비용을 증가시키는 경우 비용이 계속해서 증가할 수 있습니다. 또한 API 의

라이선스로 인해 상업 용도로 API 사용이 불가능 할 수 있습니다. 따라서 사용약관을 잘 확인해야 합니다. 마지막으로 시간이 지나면서 API 가 변경될 수 있는데 변경 시, 비용 면에서 복잡해질 수 있거나 완전히 API 가 중단될 수 있습니다. 따라서 이에 대비할 계획을 잘 준비해야 합니다.

비디오: 데이터 수집, API 및 페이지 스크래핑

API 를 통해 수집하는 데이터는 주로 JSON 및 XML 형식입니다. JSON 은 자바스크립트의 일부이며 데이터를 표현하는 가벼운 수단으로 유용하게 사용되어 API 와 MongoDB 에서 주로 활용됩니다. JSON 은 일련의 정렬되지 않은 key 와 value 의 쌍으로 이루어집니다. XML 은 일련의 요소 및 속성인 데이터이며 JSON 과 같이 데이터를 표현하는 데 사용할 수 있으며 좀더 장황한 형식입니다. 데이터의 양이 많기에 공간을 많이 차지하므로 유용하게 사용되지는 않습니다. 웹 페이지의 데이터는 HTML 형식으로 저장되는데 이는 XML 과 비슷한 형태입니다. 웹 페이지를 방문하는 사용자의 프로세스를 자동화할 수 있으며, 이를 사용하여 웹 페이지에서 모든 데이터를 추출할 수 있습니다. 파이썬에서 이것을 사용하기 위해 많이 사용되는 라이브러리는 Python Request Library 입니다. Python 으로 웹 페이지를 스크래핑 할 때에는 Reqeust Library 와 Beautiful Soup 혹은 Scrapy 라는 라이브러리를 활용합니다. 페이지를 스크래핑 할 때에는 해당 페이지에 robot.txt 파일을 확인하여 스크래핑의 허용 유무를 확인하는 것이 중요합니다. 로봇 프로토콜은 자체적으로 법적 지위를 가지고 있지는 않으나 이를 무시하는 것은 비윤리적이며 법률 자문을 구해야 하며, 라이선스 또는 저작권 법을 위반할 가능성이 있기에 주의해야 합니다.

토론: 왜 일부 조직은 자사의 데이터가 수집되는 것을 적극 환영할까요?

가령 본인이 웹서비스를 제작한다고 가정을 하면 예를 들어 일기예보 정보를 자신이 만든 웹페이지에 띄우기, 지도를 이용한 길찾기, 맛집찾기, 자신의 웹서비스에서 사용자들로 하여금 결제가 가능하도록 만들기 등을 한다고 가정을 하겠습니다. 보통 일반인들에게는 위의 기능들을 제공할 만한 기반이 되는 데이터와 관련 프로그램이 없기 때문에 이러한 기반 데이터와 프로그램을 API 를 통해 공개하여 일반인들이 이러한 기능들을 구현할 수 있게끔 해줍니다. 따라서

자사의 API 를 공개함으로써 자사의 브랜드 이미지를 상승시킬 수 있으며 개발자들이 만든 프로그램의 기반이 되는 API 가 자사의 API 를 활용한 것이라면 홍보효과와 더불어 추후에 라이선스 같은 것을 체결하여 상생하는 관계로 발전해 나갈 수도 있습니다. 이러한 이유로 인해 기업 또는 정부에서 API 들을 공개하고 있습니다. 예시로는 SK 의 날씨정보, Daum 의 지도정보, 카카오페이 결제 등이 있습니다.

3 주차 실습 **'연습: HTML 및 페이지 스크래핑'**을 통해 HTML 의 Tag 와 구성요소를 알게 되었고, Requests 를 통해 페이지에 원하는 요청을 보내는 방법을 알게 되었습니다. 또한 Robots.txt 를 통해 스크래핑 허용여부를 확인하는 방법을 배웠으며, Beautiful Soup 라이브러리의 find 문을 활용하여 페이지에서 원하는 정보만 스크래핑 하는 방법을 알게 되었습니다.

비디오: 데이터 탐구 및 수정

데이터를 수집한 이후에는 데이터가 작업에 활용할 수 있는 형식으로 준비되어 있는지를 잘 확인해야 합니다. 이는 편향이나 허위 통계를 유발하지 않도록 주의해야 하며 데이터 조각들을 잘 통합하고 데이터를 요약하는 것이 중요합니다. 이를 위해 기술 통계를 사용합니다. 누락된 데이터를 수정하는 방법은 단순 무시, 알 수 없는 상수 넣기, 주위에 있는 다른 데이터 포인터를 기반으로 평균을 내거나 확률적인 모형을 사용하는 방법이 있습니다. 데이터에는 노이즈가 발생할 수 있습니다. 이는 정규분산 때문이거나 제한된 대역폭, 다운로드 과정에서의 유실 등 언제든지 발생할 수 있으므로 이동평균, 회귀분석, 이상치 제거, 수동 데이터 검사 등을 하여 노이즈를 없애는 것이 중요합니다.

비디오: 데이터 저장 및 관리

우리는 계산능력과 저장공간이 늘어나면서 데이터 사이언스를 진행하는 데에 좀 더 수월 해졌습니다. 분석을 수행할 수 있도록 적절한 인프라를 설정하는 것이 중요한데 클라우드 컴퓨팅이 인프라를 구축하는 데에 많이 활용됩니다. 클라우드 컴퓨팅은 월 단위로 비용을 지불하면 네트워크 기능과 하드웨어를 제공하는 서비스 혹은 저장장치로서 인프라 자체를 사용할 수 있습니다. 또한 데이터를 관리하는 데에 있어서 NoSQL DB 중의 한가지인 MongoDB 를 활용할 수

있습니다. MongoDB 는 BISON 이라는 JSON 스타일 형식으로 데이터를 저장하는 데이터베이스인데, 기존의 일관성과 가용성에 제약을 받지 않으면서 수평적으로 확장하는 기능이 있는 도구입니다. 파이썬에서도 활용할 수 있게 PyMongo 라이브러리를 사용하며, 쿼리를 하기 위해 find 함수를 사용합니다. 몽고디비는 스키마가 적용되지 않기 때문에 미리 설정해야 합니다. 때문에 다양한 컬렉션에서 데이터를 가져오려면 잠재적으로 여러 번 서버로 왕복 이동을 해야 할 수도 있습니다. 또한 데이터 중복문제가 발생할 수 있기 때문에 주의해야 합니다.

두 번째 실습 '연습: MongoDB 를 사용하여 정보 검색'에서는 PyMongo 라이브러리를 통해 파이썬에서 몽고디비를 활용하는 방법을 배웠으며 이를 활용해 영국의 식품위생 데이터를 처리하였습니다. db.collection_name() 함수를 이용해 데이터의 컬렉션들을 확인하는 방법을 배웠으며, 단일 조건 쿼리를 하는 find_one() 함수와 다중 조건 쿼리를 하여 Cursors 객체를 반환하는 find()함수를 활용하는 방법을 배웠습니다. 그리고 과제 '5. Data Management Assignment'를 통하여 배운 내용을 직접 실습해보았습니다.

4 주차에서는 데이터 사이언스에서의 통계와 필수적인 기술에 대한 수업과 Bokeh 라이브러리를 사용하는 방법을 배웠으며, 실습과 과제를 통해 숙달했습니다.

비디오: 데이터 세트의 통계적 속성

비디오에서는 검트리 광고 웹사이트의 로버가 BMW 로 만든 미니카를 예시로 보여주며 통계적 속성에 대한 내용을 설명했습니다. 데이터 집합의 일반적인 요소인 자동차의 대표 가격을 산출하기 위해 평균, 중앙값, 모드를 사용하는 방법에 대한 내용, 분산과 표준편차, 데이터 집합 내의 빈도 분포에 대한 내용을 다루었습니다. 평균은 값의 산술평균이며 스코어 합계를 총 개수로 나눈 값으로 계산됩니다. 그러나 실제로 데이터 집합의 중심이라고 예상하였던 평균은 데이터 집합내에 존재하지 않을 수 있다는 것이 평균산출의 문제점입니다. 대신에 데이터 집합에 있는 중심요소인 중앙값을 활용할 수 있습니다. 중앙값을 찾으려면 원하는 열에 따라 모든 요소를 정렬하고 요소의 수가 홀수인 경우에는 중앙요소의 값을 선택하고 짝수인 경우에는 중앙 쌍의 평균을 선택합니다. 또 다른 방법은 모드를

계산하는 것입니다. 모드는 데이터 집합에서 가장 빈번하게 나타나는 요소의 값입니다. 위와 같은 방법을 통해 중심 요소를 확인하면 범위를 구할 수 있습니다. 범위를 통해 집합 내의 최소 값과 최대 값을 구할 수 있으며, 모든 값에서 데이터 집합의 평균까지의 평균 거리를 설명할 수 있습니다. 데이터 집합 내에서 제곱된 거리의 평균을 구하여 분산을 구할 수 있으며, 이에 제곱근을 취하여 표준편차를 구할 수 있습니다. 또 관심 있게 다룰 특성은 데이터 집합 내의 빈도 분포입니다. 범위를 통해 이를 구하고 구한 결과를 그래프를 통해 시각화 합니다. 데이터 집합에 따라 분포의 형태가 다를 수 있습니다. 데이터 집합의 특성을 예측할 때에는 모든 값을 고려해야 하기 때문에 집합의 크기가 매우 크면 샘플링을 통해 특성을 예측합니다.

비디오: 샘플링 및 중심 극한 정리

샘플링은 더 큰 모집단에서 부분집합을 선택하는 과정으로서 모집단의 크기가 매우 클 때 데이터의 특성을 파악하기 위해 사용합니다. 샘플링을 할 때에는 무작위로 체계적으로 선택되어야 하며 충분한 크기를 갖고 있어야 합니다. 그런 다음 샘플 오류를 제한하는 데 도움이 되는 중심 극한 정리를 이용하여 오차를 측정하여 신뢰도를 산출할 수 있습니다.

토론활동: 통계를 사용하는 것으로 충분하다고 생각하십니까?

통계 수치가 같아도 데이터의 패턴이 다를 수 있음을 잘 보여주는 예시인 Anscombe’s quartet 을 통해 알 수 있듯이 데이터 조사 시에 통계적 수치에만 의존하면 잘못된 판단을 이끌어 낼 수 있기 때문에 데이터를 분석하기 전에 반드시 시각화를 통해 데이터의 패턴과 이상치를 파악하는 것이 중요합니다.

4 주차의 첫 실습인 ‘연습: Python 과 Bokeh 에서 통계 사용하기’를 통해 파이썬에서 통계를 사용할 수 있게 해주는 라이브러리인 Pandas 라이브러리와 시각화 라이브러리인 Bokeh 를 활용하여 데이터 프레임에서 mean, median, mode 등의 통계치를 이끌어 내고, Bokeh 를 통해 차트를 만들고 Random 함수를 통해 난수를 추출하는 방법과 사분위수, 이상치를 제거하는 방법을 배웠습니다.

비디오: 머신 러닝 및 선형 회귀 분석, 비디오: 데이터 분류

경험에 따라 자동으로 개선되고 주어진 데이터로 학습하며 데이터를 예측할 수 있는 알고리즘을 작성하는 것과 관련이 있는 머신 러닝 중에서 유입 변수를 기반으로 결과 변수를 예측하는 선형회귀와 알고리즘이 객체를 분류하려고 하는 분류 문제에 대한 내용을 배웠습니다. 그 중에서 자율 학습과 감독 학습에 대한 내용을 배웠으며 벡터 머신 및 나이브 베이스를 지원하는 알고리즘을 배웠습니다. 선형회귀를 진행하기 위해서는 데이터에서 유입변수와 예측변수를 정하고 산점도를 나타내고 선형방정식을 이끌어 내야 합니다. 일반적으로 모델과 함께 정확도와 신뢰구간을 같이 보고해주어야 합니다. 분류 문제는 객체가 특정 기능을 갖고 모델이 특정 카테고리를 갖도록 설정됩니다. 예를 들어 텍스트 분류에서, 대상의 기능은 텍스트에서 용어의 빈도가 될 수 있습니다. 좋은 모델은 보이지 않는 개체가 속한 범주를 예측할 수 있습니다. 예를 들어, 특정 문서의 언어는 무엇인지 정서 점수는 어떠한지 등을 예측 할 수 있습니다. 또 다른 인기 있는 분류 문제의 예시로는 스팸 필터링이 있습니다.

지원 벡터 기계로 분류 문제를 진행할 때에는 각 차원이 특정 용어인 벡터공간을 구성하고 데이터 집합의 모든 문서를 벡터공간에 배치합니다. 그리고 비슷한 범주의 객체들을 벡터 공간에 넣어 다른 범주의 객체들로부터 멀리 떨어뜨립니다. 그 후 알고리즘 작업을 통해 두 범주 사이의 분리 기호를 찾습니다. 분리 기호는 문제가 2 차원인 경우는 두개의 기호, 다차원인 경우엔 초평면이 됩니다. 훈련 프로세스를 통해 분리기호의 최적위치를 구합니다. 지원벡터기계는 고차원 문제에서 잘 수행되나 훈련 프로세스가 매우 복잡하다는 단점을 갖고 있습니다.

나이브 베이스는 확률에 기반하여 분류 알고리즘을 사용합니다. 나이브 베이스는 모든 객체에 대해 특정 집단에 속할 확률을 계산합니다. 특징 간의 독립성을 가정하기에 나이브 베이스라고 합니다. 예를 들어 New York 이라는 표현에서 New 와 York 이라는 용어는 별도로 처리되며 해당 용어 간의 종속 관계는 가정하지 않습니다. 나이브 베이스는 훈련 프로세스에서 먼저, 특정 집단에 속하는 인스턴스의 사전 확률을 계산하고 하나의 집단의 각 특징에 있어서 특징이 특정 집단에 속할 확률을 계산합니다. 확률은 특정 용어에 대한 특정 집단의 용어 빈도를 해당 특정 집단 모든 용어 빈도의 합으로 나눈 값으로

계산됩니다. 계산한 확률 중 가장 높은 확률을 가진 집단을 솔루션으로 선택합니다. 나이브 베이스는 훈련 세트가 다소 작거나 훈련 시간이 중요할 때에 주로 사용합니다. 따라서 데이터 집합의 속성에 따라 적절한 알고리즘을 선택하여 문제를 해결해야 합니다.

두 번째 실습 **'연습: 선형 회귀 분석 및 분류'**에서는 Bokeh 라이브러리와 sklean 라이브러리를 활용하여 선형회귀 분석과 잔차 분석, 모델 평가와 Bayesian 분류기를 사용하여 SMS 문자 메시지가 '스팸'인지 '햄'인지를 분류하는 실습을 진행했습니다. 실습을 통해 데이터프레임의 데이터를 통해 선형회귀 모델을 구성하고 산점도를 통해 나타내고 선형회귀식을 구했으며 sqrt()함수로 잔차분석을 실시했습니다. 그 후 train_test_split()함수를 통해 훈련세트와 테스트세트로 데이터를 나누었으며 score()함수로 머신 러닝의 점수를 측정하였습니다. 그리고 과제 **'4. Statistics and Machine Learning Assesment'**를 통해 배운 내용을 숙달하였습니다.

5 주차 강의에서는 다양한 데이터 시각화 기법을 사용하여 데이터 사이언스 작업의 결과물을 보고하는 방법을 배웠습니다. 또한 주요 결과물을 강조하고 보고서의 영향을 향상시키기 위해 특정 유형의 데이터를 표시할 수 있는 다양한 방법을 탐구했습니다.

비디오: 데이터 시각화 소개

보통 우리는 어떤 종류의 정보를 매우 자주 방대한 양의 정보를 취하고, 어느 집단의 사람들에게 이것이 의미하는 바를 전달할 수 있기를 바라며 정보를 시각적으로 표시합니다. 그 대상이 회사의 상사거나 대중일 수 있고 정부의 장관일 수도 있습니다. 이는 먼저, 데이터 사이언스 작업을 한 이유에 따라 시각화의 청중이 결정됩니다. 데이터 시각화는 통계 그래프, 플룻 및 정보 그래픽을 사용하여 명확하고 효율적으로 정보를 전달하는 것을 말합니다. 데이터 사이언스 작업의 결과물을 청중들이 잘 받아들이고 깨달아 각자의 역할이 무엇인지를 깨닫게 하기 위해 스토리를 구성해 데이터를 시각화하여 전달하는 것이 중요합니다. 그리고 시각화를 할 때에는 청중의 수준을 잘 고려하여 나타내는 것이 중요합니다. 또한 기술이 발전하며 오늘날에는 정적 그래픽에서

대화식 그래픽으로 발전하게 되어 사용자로 하여금 그래픽을 통해 스스로 정보를 얻을 수 있게끔 할 수 있게 되었습니다.

비디오: 데이터 시각화 유형

데이터 시각화의 형은 크게 두 가지 범주로 나눌 수 있습니다. 하나는 탐색적인 시각화입니다. 이는 사용자가 스스로 데이터를 탐색하게 합니다. 사용자는 어떤 각도로 데이터를 볼지 선택하고 데이터의 전반적인 스토리가 무엇인지, 데이터의 흥미로운 점은 무엇인지 알 수 있습니다. 다른 시각화의 유형은 설명적 시각화입니다. 이는 분석작업을 하여 얻은 결과가 무엇인지 청중에게 말하고자 할 때 사용하는 유형입니다. 더 넓은 층의 청중에게 핵심데이터를 전달하기 위해 주로 사용하며 전달하고자 하는 내용이 가능한 한 혼란을 줄이고 문맥에 맞게 시각화를 제공하는 것이 중요합니다.

비디오: 데이터 시각화를 위한 데이터

좋은 시각화를 만들기 위해서는 보유하고 있는 데이터 유형을 이해하는 것이 중요합니다. 데이터 시각화의 사용할 수 있는 데이터의 유형에는 연속, 이산, 명목 데이터가 있습니다. 연속 데이터는 실제 정확한 값을 갖고 있진 않지만 근사치를 통해 정확도를 높일 수 있는 데이터입니다. 이산 데이터는 항상 정수로 나타납니다. 명목 데이터는 특정 유형의 범주를 기반으로 분리할 수 있는 데이터입니다.

비디오: 데이터 인코딩

데이터의 유형에 따라 시각적 객체로 인코딩하는 방법을 결정하는 것이 중요합니다. 인코딩이란 특정 데이터 조각을 페이지의 시각적 객체에 매핑하는 것을 의미합니다. 주로 2D 평면에 인코딩 된 데이터를 가장 많이 볼 수 있습니다. 보통 x 축과 y 축의 두가지 차원을 사용합니다. 그래프에 더 많은 정보를 추가하기 위해선 차원을 늘려 z 축을 추가하여 3D 평면으로 인코딩 시킬 수도 있습니다. 차원을 더 늘려 4 차원에 구현하는 것도 가능하나 이는 매우 까다롭습니다. 따라서 레티날 변수를 활용합니다.

비디오: 레티날 변수

레티날 변수는 사람들이 느끼기 쉽고 어떤 차이를 감지할 수 있는 변수입니다. 예시로는 위치, 크기 및 색상 등이 있습니다. 보통 수치를 갖는 데이터인 이산 데이터나 연속데이터에 유용하게 사용됩니다. 레티날 변수를 사용하면 청중들로 하여금 시각 자료를 더욱 쉽게 이해하고 인식할 수 있도록 합니다.

비디오: 시각적 인코딩 순위

Cleave 와 McGill 은 1980 년대에 레티날 인코딩을 사용하여 사람들이 차이를 얼마나 잘 인지하는지 비교하는 다양한 방식을 살펴 성과를 얻었고 이를 통해 기본 지각 과제라고 부르는 10 가지 시각적 인코딩 순위를 측정했습니다. 이는 보고 있는 것의 차이점을 느끼고 이해할 수 있는 인간의 능력에 근거합니다. 순위에는 위치, 길이, 각도 및 경사 지역, 색상, 색조 등이 있습니다. 따라서 적절한 인코딩을 사용하여 전달하는 것이 중요합니다.

비디오: 차트 정크

사람들이 보고 인식할 수 있도록 차트를 보다 효과적이고 쉽게 만들기 위해서 차트 정크라는 것을 제거해야 합니다. 차트 정크는 차트의 의미 또는 차트에 영향을 미치거나 변경하지 않으면서 차트에서 제거할 수 있는 차트 내 혹은 이미지 내의 모든 것을 말합니다. 보통 차트에 차트 정크가 있으면 보는 사람의 주의를 흐리게 합니다. 따라서 뷰티 패러독스 현상에 빠져선 안되며 시각화를 할 때에 스토리에 집중할 필요성과 데이터가 보여주는 것의 균형을 잡아서 구성하는 것이 중요합니다.

비디오: 마음을 위한 그래픽 디자인

그래픽을 효과적으로 만들기 위해 할 수 있는 것은 우리가 데이터를 인식하도록 돕는 시각적 장치들입니다. 다양한 이론을 바탕으로 한 모범 사례들은 패턴인식과 페이지에 있는 것들을 구성할 수 있는 방법을 배우며 인간의 두뇌가 자연스럽게 사물을 그룹화하는 방식을 최대한 활용합니다. 이 개념은 독일 심리학자들에게 유래한 게슈탈트 이론에 기초합니다. 이의 예로는 그룹화의 원리, 연결성의 원리, 연속성의 원리, 폐쇄의 원리 등이 있습니다. 이러한 원리를 활용하면 두뇌의

자연적인 프로세스를 활용하고 보는 이가 이미지를 보고 쉽게 내용을 인식하도록 합니다.

비디오: 스토리 텔링

스토리 텔링이란 전하고 싶은 정보를 시각적인 것을 통해 전달하는 개념 중에 하나이며, 즉흥적으로 꾸며서 단어, 소리 또는 이미지로 이벤트를 전달하는 것을 말합니다. 스토리 텔링을 할 때에는 내가 전달하고자 하는 바와 독자의 수준을 고려하여 저자 중심 스토리 텔링을 할 지 독자 중심 스토리 텔링을 할지, 혹은 두가지를 합친 방식을 적용할지 등을 고려해야 합니다. 또한 분석 결과를 전달할 때에는 객관성을 유지하는 것이 매우 중요합니다.

토론 활동: 웹에서 접한 데이터 스토리텔링의 예는 무엇인가요?

https://www.newyorker.com/tech/annals-of-technology/mapping-new-york-noise-complaints

- 뉴욕에서 가장 시끄러운 동네 -

이 기사는 뉴요커 지에 벤 웰링턴이라는 사람이 기고한 글입니다.

그는 뉴욕시의 어떤 동네가 가장 시끄러운가를 지도에 나타냈고 그 동네가 왜 시끄러운지를 표현하기 위해 뉴욕시의 소음 민원 발생 데이터를 활용했습니다. 이러한 기사를 쓰게 된 계기는 뉴욕에는 예전부터 도시의 소음과 관련된 민원이 꾸준히 발생해왔고 이에 시의회가 주목하여 환경 보호국이 도시 전체에서 소음 샘플링을 시작하도록 하는 법안을 도입하였고 샘플링의 결과 도시 소음 공해의 주요 원인이 교통시스템이라는 결론을 내렸습니다. 그러나 교통시스템 뿐만 아니라 불도저, 공기 압축기, 로더, 덤프트럭, 소형 착암기, 포장 차단기, 확성기, 에어컨, 선풍기, 진공 청소기 등 교통 시스템뿐만 아니라 공사로 인한 소음과 생활 소음 역시 소음 관련 민원에서 큰 비중을 차지하고 있었고, 글쓴이는 도시의 OpenData 포털을 통해 소음이 어디에서 오는지 더 정량적으로 파헤치기 위해 이 기사를 작성하였습니다. 이러한 분석을 통해 소음 규제를 담당하는 두 기관인 뉴욕 경찰국과 환경보호국에게 도시의 여러 지역에 알맞는 맞춤형 솔루션을 개발할 수 있도록 합니다.

5 주차 실습 '연습: Bokeh 를 사용한 데이터 시각화'에서는 Bokeh 와 ipywidgets 라이브러리를 활용하여 시각화를 할 때 여러가지 widget 과 tool 등을 다루는 방법과 지도에 매핑하는 방법을 배웠습니다. 또한 과제 **'3. Visualisation Assignment** 를 통해 데이터 프레임으로부터 데이터를 추출하고 소스 데이터를 변환시켜 지도에 매핑하는 방법과 여러가지 도구들을 다룰 수 있도록 숙달하였습니다.

6 주차에서는 데이터 사이언스의 미래와 악용 가능성에 대해 알아보았습니다.

비디오: 데이터 사이언스가 '악용'될 수 있을 까요?

데이터가 수집된 방법, 데이터가 수집된 이유, 데이터의 한계가 무엇인지, 방법의 한계가 무엇인지에 대해 신경 쓰지 않는다면 데이터 사이언스를 악용할 수 있습니다. 그렇기 때문에 이를 방지하기 위해 데이터를 악의적으로 사용하는 사람들을 찾기 위해 데이터를 분석할 수 있어야 하며 정부 차원에서 보안 서비스를 강화해야 합니다. 따라서 올바른 윤리 의식을 갖고 데이터 사이언스를 진행하는 것이 중요합니다.

비디오: 데이터 사이언스의 미래는 어떨까요?

미래에는 우리의 생활방식이 변화할 것이며 모든 종류의 센서 및 데이터 제작자와 데이터에 대한 풍부한 지원을 통해 데이터 및 새로운 산업과 비즈니스, 새로운 직업 등이 많이 창출될 것입니다. 또한 기존의 데이터 사이언티스트들이 더 성숙해질 것입니다. 스스로가 하고 있는 일이 무엇인지 잘 이해하게 될 것이며, 데이터 사이언스 창출물의 투명성에 대한 규칙이나 규정이 생겨나게 될 것입니다.

토론: 데이터 과학자들에게 미래가 무엇이라고 생각합니까?

미래에는 데이터가 보다 더 세분화 되고 다양해지고 접할 수 있는 기회도 많아질 것입니다. 또한 국가 차원에서도 이러한 데이터 과학자와 데이터 수집 및 활용을 위한 지원정책이 많이 펼쳐져 관련 직종이 다수 창출될 것입니다. 현재는 데이터 사이언스 창출물에 대한 보호가 잘 이루어지고 있지 않지만 이와 관련된 정책도 등장하여 데이터 사이언티스트들로 하여금 서로의 창출물에 대한 존중이 높아질

것입니다. 물론 이런 과정에서 데이터를 악용하는 사례들도 많이 나타날 것입니다. 이를 막기 위해 데이터 사이언티스트들이 데이터 윤리의식을 갖고 프로젝트를 진행하여야 하며 양성 과정에서도 데이터 윤리 관련 교육이 강조될 것이라고 예상됩니다.

2) 데이터 사이언스 적용 연구계획

2-1. 데이터

https://www.data.go.kr/data/15043378/openapi.do

보건복지부_코로나 19 시도발생_현황 조회 서비스 오픈 API 를 활용하여 각 시,도별로 코로나 19 로 인한 사망자 수, 전일대비 증감 수, 격리 해제 수, 10 만명당 발생률 등을 파악할 수 있습니다.

요청변수는 서비스키, 페이지 번호, 한 페이지 결과 수, 데이터 생성일 시작범위, 데이터 생성일 종료범위로 구성되어 있으며 Python 샘플 코드는 다음과 같습니다.

```
# Python 샘플 코드 #
from urllib2 import Request, urlopen
from urllib import urlencode, quote_plus

url = 'http://openapi.data.go.kr/openapi/service/rest/Covid19/getCovid19SidoInfStateJson'
queryParams = '?' + urlencode({ quote_plus('ServiceKey') : '서비스키',
quote_plus('pageNo') : '1', quote_plus('numOfRows') : '10', quote_plus('startCreateDt') :
'20200410', quote_plus('endCreateDt') : '20200410' })

request = Request(url + queryParams)
request.get_method = lambda: 'GET'
```

```
response_body = urlopen(request).read()
print response_body
```

한가지 더 구하고 싶은 데이터는 코로나 19 로 인한 사회적 거리두기 단계 별 데이터를 수집하려 해보았으나 공공 데이터 포털에서는 해당 데이터를 찾을 수 없었으며 필요 시 담당기관인 질병관리청에 문의하거나 공식 절차인 공공데이터 제공신청을 이용하라는 답변을 받았습니다.

2-2. 문제정의

2019 년 12 월 중국 우한에서부터 시작된 호흡기 감염 질환인 코로나 19 로 인해 전세계가 약 14 개월간 고통을 받고 있습니다. 각 나라 별로 방역 대책을 구축하여 방역을 실시하였음에도 종식을 바라는 요즈음 특히 K 방역이 크게 화제가 되었습니다. 우리나라는 다중이용시설 방문 시 QR 코드 체크인 및 방문기록을 남기게 하고 대중 교통 및 시설 이용시 마스크 필수 착용과 특히 사회적 거리두기 캠페인을 통하여 단계 별로 국민 행동요령 및 수칙을 적용하고 있습니다.

저는 이러한 사회적 거리두기 조치가 단계 별로 실질적인 코로나 19 확진자 감소에 얼마나 효과를 주었는지를 확인하고 분석하기 위해 주제를 선정하였습니다.

2-3. 적용방법

전국의 모든 효과를 알아보기에는 무리가 있을 것이므로 확진자가 가장 많이 발생한 서울지역을 대상으로 거리두기 단계 별 확진자 증감을 파악합니다.

Python 을 활용하여 사회적 거리두기 단계 별 데이터 csv 을 pandas 라이브러리의 read 문을 통해 DataFrame 형식으로 읽어들입니다.

또한, 오픈 API 를 통해 시, 도 별로 확진자 증감 데이터를 얻은 후 서울특별시의 데이터 만을 필터링합니다.

초기 데이터는 거리두기 단계 별 실질적 효과를 나타내기 힘들 것으로 판단하여 단계 조치 적용 후 3 일 간의 데이터는 제거한 후 활용합니다. 그 후, 시각화 라이브러리인 Bokeh 를 활용하여 읽어들인 데이터를 응답변수는 거리두기 단계 별 데이터로, 예측변수는 서울시 코로나 19 확진자 증감데이터로 할당하여 차트를 만듭니다. 다음으로는 numpy 라이브러리의 train_test_split()와 random_state()를 활용하여 seed 값을 기준으로 응답변수와 예측변수를 train set 와 test set 로 나누어 줍니다. 다음으로 train set 데이터를 LinearRegression()과 fit() 함수를 통해 선형회귀 시킨 후 머신 러닝을 시킵니다. 그리고 올바르게 학습이 되었는지 test set 데이터로 score()를 측정합니다.

마지막으로, 얻은 결과를 시각화하기 위해 STAMEN_TERRAIN 과 같은 Bokeh.tile_providers 을 활용하여 사회적 거리두기 단계 별로 서울시의 코로나 확진자 증감 추이를 지도에 매핑합니다.

2-4. 기대효과

사회적 거리두기 단계 별 코로나 19 증감 효과분석을 통하여 실질적으로 사회적 거리두기라는 방역조치가 어느 정도의 효과가 있었는지 알 수 있으며 단계 별 조치 간의 차이를 확인하여 코로나 확산에 보다 직접적인 영향을 미치는 요인이 무엇인지 파악할 수 있으며 방역 조치를 보완시키는 등의 효과가 있을 것이라고 예상됩니다. 또한 이를 시각화하여 결과를 누구나 열람할 수 있도록 공개하여 다수가 접할 수 있도록 합니다.

분석 결과를 통해 국민들로 하여금 코로나 종식을 위해 스스로가 할 수 있는 일이 무엇인지를 상기시키고 경각심을 일깨우는 효과가 있을 것이라 예상되며 최종적으로는 국민 다수의 방역 협조를 높여 코로나 19 의 종식을 앞당기는 발판이 될 수 있다고 생각합니다.