논문 보충

17102070 Jaehun Hwang

데이터 설명

자세한 내용을 들어가기 전에 이 연구에 사용된 2개의 데이터셋(Dataset) 에 대하여 알아보도록 하겠습니다. 첫번째는 Kaggle 에서 배포되고 Data Science for COVID-14 (이하 DS4C) 에 의해 제작된 DS4C: Data Science for COVID-19 in South Korea 데이터셋 (Dataset) 에 속해있는 `PatientInfo.csv` 입니다. 본 데이터는 질병관리청 (Korea Centers for Disease Control & Prevention; 이하 KCDC) 에서 발표된 COVID-19 감염 정보를 바탕으로 각 지방정부의 확진자 리보트 내용을 함께 참고하여 작성되었으며, 확진자 각각의 역학조사 내용을 담고있습니다. 세부 컬럼(Column) 으로는 확진자 식별번호 (patient_id), 성별 (sex), 나이대 (age), 국적 (country), 시/도 (province), 시/군/구 (city), 감염 사례 (infection_case), 감염시킨 사람의 식별번호 (infected_by), 접촉 번호 (contact_number), 증상 발현 날짜 (symptom_onset_date), 확진 날짜 (confirmed_date), 퇴원/격리해제 날짜 (released_date), 사망 날짜 (deceased_date), 상태 (state) 가 있으며 총 14개의 열 그리고 5165개의 관측지로 이루어져 있습니다. 성별은 남 (male) / 여 (female) 로 구분되며, 1122개의 결측치를 보유하고 있습니다. 나이대는 0대, 10대 부터 90대, 100대까지 10살 단위로 나뉘어져 있으며 총 1380개의 결측치를 보유하고 있습니다. 국적은 확진자의 국적을 나타내며 대한민국 (Korea), 중국 (China), 미국 (United States), 프랑스 (France), 베트남 (Vietnam) 등 총 16개 나라로 이루어져 있습니다. 이 중 한국 환지가 5123 건으로 99%를 차지하였고 결측치는 한건도 없었습니다. 시/도는 확진자가 확진된 지역의 특별/광역시/도 를 나타내며 서울, 부산, 대구, 광주, 강원도, 제주도 등 국내 17개 지역을 전부 포함하고 있고, 결측치는 한건도 없었습니다. 시/군/구 는 시/도에 속한 행정 구역이며 서초구, 강서구 등 전국 162개 행정구역을 나타내며, 94건의 결측치가 있었습니다. 감염사례는 집단 감염 사건과 같은 확진자가 속한 감염 사건을 말하며, 해외 유입 (overseas inflow), 환자 접촉 (contact with patient), 신천지 (Shincheonji Church), 쿠팡 물류 센터 (Coupang Logistics Center) 등 총 51개의 값으로 이루어져 있으며 919개의 결측치를 가지고 있습니다. 감염시킨 사람의 식별번호는 같은 데이터 내에 있는 확진자에게 감염이 되었다고 판명되었을 경우에 감염시킨 환자의 ID 를 표시하며, 전체 5165명 중 1346명만 밝혀져 있고 나머니 3819명의 환자는 감염자가 알려지지 않았습니다. 접촉 번호의 경우는 환자가 접촉한 사람수를 나타내며 4374개의 결측치를 가지고 있었습니다. 날짜를 나타내는 열의 경우 전부 2020년 1월 23일 부터 당해년도 6월18일 까지 약 5개월간의 값을 가지고 있습니다. 증상 발현 날짜는 결측치가 4475개로, 정확한 발현 날짜조사가 거의 불가능했음을 보여주고 있습니다. 물론 무증상 확진자가 포함되었을 가능성도 있습니다. 확진날짜는 보건소 등에서 확자가 확진 판정을 받은 날짜로 결측치가 단 3개 밖에 존재하지 않았으며, 분석에 주된 정보로 활용되었습니다. 이외에 퇴원날짜는 3578개, 사망 날짜는 5099개의 결측치를 보여주었는데, 그렇다고 나머지 1587명의 환자가 퇴원/자가격리 해제를 하지 않았다거나 5099명 환자중에 추가 사망자가 없을 것이라고는 단언할수는 없습니다. 제작자가 데이터셋을 만들때 각 환자를 정확히 어느 시점까지 조사했는지 알수 없었기 때문입니다. 상태 컬럼(Column)는 격리 (isolated; 병원에서 격리되어 치료중인 경우), 해제 (released; 병원에서 퇴원한 경우), 사망 (decreased; 사망한 경우) 세가지 값을 가지며 결측치는 하나도 없었습니다. 다만, 해제된 상태를 가진 확진자가 2929명인 것에 비해서, 해제 날짜가 기록된 환자가 1346명 밖에 없었다는 점을 고려해볼때, 일부 데이터가 정확히 기록되지 않았음을 알수있습니다. 본 연구는 모든 열 중에서 확진자 식별번호, 성별, 나이대, 시/도, 감염시킨 사람의 식별번호, 확진날짜를 중심으로 코로나의 확산을 분석하였으며, 정보가 불확실한 증상 발현 날짜, 퇴원 날짜, 사망 날짜 등에 관련된 열은 사용하지 않았습니다.

두번째 데이터는 한국철도공사에서 제공하고 교통 데이터 거래소 에서 배포한 여객 일별/역별 승하차 데이터인 `KR_TB_TR_STN_DAY_CON_20190101_20190531_UTF8.csv` 입니다. 이름에서도 알수 있듯이 2019년 1월 1일 부터 2019년 5월 31일 6개월(152일)간 매일 각역의 총 승차인원수와 하차인원수를 기록한 것입니다. 컬럼(Column)은 날짜 (RUN_DT), 역코드 (STN_CD), 역명을 나타내는 역코드명 (STN_CD_NM), 해당역에서 승차한 총 인원수인 승차인원수 (ABRD_PRNB), 그리고 해당역에서 하차한 총 인원수인 하차인원수 (GOFF_PRNB) 로, 총 5개의 열 및 35040개의 관측치로 이루어져 있습니다. 역코드(명)의 경우 서울, 광명, 대전, 대구, 부산 등 코레일 소속 간선철도역사 243곳을 나타내며 모든 영역에서 결측치는 나오지 않았습니다.

재귀함수 추가설명

본 연구는 감염시킨 사람의 식별번호 컬럼(Column)을 바탕으로 환자간 감염 전파 계보를 구하였습니다. 정확한 방식은 감염시킨 사람의 식별번호가 기존 확진자 식별번호에서 왔다는 사실을 이용하여, 특정 환자 관측치의 감염시킨 사람 컬럼(Column) 이 결측치가 나올때까지 거슬러 올라가는것입니다. 이때 이를 자동화하기 위해서는 반복문으로는 한계가 있었고, 재귀함수를 사용할수밖에 없었습니다. 그 이유는, 첫번째로, 각 계보별 깊이가 동일하지 않습니다. 각 감염계보는 누군가를 감염시켰지만, 자신은 누군가로부터 감염되지 않은 확진자 (이하 최초확진자) 로부터 시작되고, 감염 사건이 똑같지 않는 한, 각 계보마다 n차 감염의 수치는 다르게 나타날수밖에 없습니다. 두번째로, 같은 차수에서도 감염자 숫자가 다르게 나타날수 있습니다. 예를 들어 같은 확진자로 부터 감염된 환자 갑과 을이 있을때, 갑은 추가로 3명을 감염시켰는데, 을은 1명만 감염시켰을수도 있습니다. 그래서 전체 반복해야하는 횟수가 불확실하기 때문에, 단순 반복문으로는 감염계보를 찾아내기 쉽지 않은것입니다. 반면, 재귀함수는 적절한 기저조건만 정해주면 특정 조건이 만족될때까지 같은 행동을 반복시킬수 있기때문에 깊이나 횟수에 구애받지 않습니다. 따라서 저희는 특정 확진자의 감염시킨 사람이 나오지 않는것을 기저조건으로 정하고, 그렇지 않을때에는 감염시킨 사람의 식별번호와 같은 확진자 식별번호를 가진 관측치를 찾는 방식의 재귀함수를 짜서 계보를 구하였습니다. 이때 감염시킨 사람의 식별번호를 Python dictionary 의 key 로 저장하고 그 확진자에 의해 감염된 환자들을 value 로 저장하여 상하관계를 표현하였습니다.

Appendix

1. `PatientInfo.csv` 예시

	patient_id	sex	age	country	province	city	infection_case	infected_by	contact_number	symptom_onset_date	confirmed_date	released_date	deceased_date	state
0	1000000001	male	50s	Korea	서울	강서구	overseas inflow	NaN	75	2020-01-22	2020-01-23	2020-02-05	NaN	released
1	1000000002	male	30s	Korea	서울	중랑구	overseas inflow	NaN	31	NaN	2020-01-30	2020-03-02	NaN	released
2	1000000003	male	50s	Korea	서울	종로구	contact with patient	2002000001	17	NaN	2020-01-30	2020-02-19	NaN	released
3	1000000004	male	20s	Korea	서울	마포구	overseas inflow	NaN	9	2020-01-26	2020-01-30	2020-02-15	NaN	released
4	1000000005	female	20s	Korea	서울	성북구	contact with patient	1000000002	2	NaN	2020-01-31	2020-02-24	NaN	released

2. `KR_TB_TR_STN_DAY_CON_20190101_20190531_UTF8.csv` 예시

	RUN_DT	STN_CD	STN_CD_NM	ABRD_PRNB	GOFF_PRNB
0	20200101	924	송도교	2	-
1	20200101	3900023	서울	38,306	50,440
2	20200101	3900025	용산	19,702	23,967
3	20200101	3900030	영등포	10,812	11,125
4	20200101	3900039	안양	387	269

3. 감염계보 예시 (`Python dictionary` → `.json` 변환)