[ 주제분석 정보 – 1. 흥행영화 예측, 2. 감염병예측 모델 ]

1. 흥행영화 예측 : OTT 시장 관련 분석.
2. OTT 플랫폼: 넷플릭스, 왓챠, 웨이브, 티빙, 네이버TV
3. 무엇을 크롤링할 수 있는지:
4. 넷플릭스 : 제목, 장르, 개봉연도, 시놉시스

: <https://www.4flix.co.kr/board/netflix/4414>

* 이 홈페이지에 랭킹도 있긴한데,, 어떤 사람이 크롤링 해놓은건 위 4가지뿐이고, 랭킹은 크롤링을 안하셨음.. 그래서 랭킹도 크롤링 되는진 모르겠음.

<https://codlingual.tistory.com/10> <- 이사람이 함.

* 그외: 제가 열심히 찾아봤는데,, 정말로 못찾았어요… [푹, 웨이브, 왓챠, 티빙, 디즈니플러스, 옥수수] 모두 전멸…. 죄송,, 대신.. 주제와 안맞는건 알지만,,ㅎㅎ,, OTT가 아닌 그냥 ‘영화’라면 네이버 영화에서 많은 것들을 크롤링할 수 있더라는걸 그냥 증정품마냥 넣어놨어요..ㅎㅎ 증정품이 어째 진짜 필요한 물건보다 크기가 많이 큰 것 같지만

1. 네이버 영화: 제목, 시놉시스, 영화 순위, 장르, 감독&배우, 댓글 리뷰
2. 제목, 시놉시스: <https://codlingual.tistory.com/11>
3. (조회별, 평점별) 영화 순위& 평점: <https://velog.io/@kks1342320/ARTCinema-Project-200313TIL>, <https://hmkim312.github.io/posts/Beautiful_Soup%EC%9D%84_%EC%9D%B4%EC%9A%A9%ED%95%9C_%EB%84%A4%EC%9D%B4%EB%B2%84_%EC%98%81%ED%99%94_%ED%8F%89%EC%A0%90_%ED%81%AC%EB%A1%A4%EB%A7%81/>
4. (판매량) 영화 순위 & 순위별 장르, 시놉시스 : <https://somjang.tistory.com/entry/Python%EB%84%A4%EC%9D%B4%EB%B2%84-%EC%98%81%ED%99%94-%EB%8D%B0%EC%9D%B4%ED%84%B0-%ED%81%AC%EB%A1%A4%EB%A7%81%ED%95%98%EA%B8%B0>
5. 댓글 리뷰: <https://ysyblog.tistory.com/59>
6. 감독, 배우: <https://benfatto.tistory.com/18>
7. 흥행 지표인 영화 관객수를 ott에선 어떤 데이터로 대체할 수 있을지& 그 데이터 수집 가능? :
8. 영화의 스트리밍 수 : 흥행지표를 OTT에선 “영화의 스트리밍 수”로 대응하는게 가장 적합할 것 같음.
9. 넷플릭스: FlixPatrol 사이트에서 찾을 수 있음.
10. 왓챠: 데이터 수집 방법 찾지 못했음.
11. 그 외 여러 플랫폼: FlixPatrol 사이트

* FlixPatrol 사이트를 크롤링 할 수 있다면 이 지표 사용 가능할 듯 (but 크롤링한 예시를 못 찾음,,)

1. OTT 플랫폼 안에서의 평점과 평점을 매긴 사람 수: 평점이 높다는 점과 흥행은 상관관계가 있다고 볼 수 있지만 평점만으로 판단하기는 부족하다고 판단. (누적 관객수는 적지만 평점은 매우 높은 작품 수 많기 때문.) 그래서 평점을 흥행 지표로 삼는다면 평점을 매긴 사람 수를 함께 고려해서 “누적 관객수” 지표를 “평점을 매긴 사람 수”로 대체하는 방법을 생각해볼 수 있을 것 같음.
2. 넷플릭스: 넷플릭스 내에선 별점이 없고 “좋아요”가 있지만 “좋아요” 수 데이터 찾지 못했음.
3. 왓챠: 왓챠의 평점과 평점 매긴 사람수는 왓챠피디아에 존재하므로 크롤링만 가능하면 왓챠피디아에서 긁어올 수 있을 듯. (근데 크롤링 가능 여부 모름.)

* 넷플릭스 데이터는 못 찾아서 불가능해보임, 왓챠 데이터는 왓챠피디아 크롤링만 가능하다면 가능.

1. 관객수 예측의 분석기법:
2. 데이터 분석을 활용한 흥행 영화 예측 논문 참고. :

:<https://scienceon.kisti.re.kr/srch/selectPORSrchArticle.do?cn=DIKO0014861016&dbt=DIKO>

* 단계선택법을 통해 유의한 변수들을 추출하고 나이브 베이즈 분류 or 로지스틱 회귀 or 인공신경망 or 랜덤포레스트 네가지 모델 중 몇 개 선택해서 예측하고 비교해볼 수 있을 듯. 그리고 랜덤포레스트의 MeanDecreaseGini 값을 보고 흥행 예측에 중요한 변수가 무엇일지 선별할 수 있을 것.

1. 감염병모델 예측:
2. 모델: 대표적인 전염확산 모형은
3. SIR모형: 각 개체가 건강한 개체(S) -> 감염된 개체(I) -> 회복된 개체(R) 의 순서로 진화한다고 가정. I는 S를 감염시키거나 회복되어 R이 되고, R 상태가 된 경우 영원히 면역을 가진다고 가정.
4. SIS모형: 평형상태를 이루는 전염확산 과정을 기술할 때 적합한 모형. S -> I -> S 과정을 순환하는 모형. 감염된 개체는 S를 감염시키거나 특정 회복률로 감염에서 회복되는데, SIR 모형과 다른 점은 감염된 개체가 회복되었을 때 다시 감염이 가능한 S상태가 된다는 것을 가정한다는 점.
5. SEIR모형: 노출된 그룹을 SIR모델에 추가한 모형. 노출된 그룹: 감염에 노출되었지만 아직 감염되지 않은 사람들의 그룹 🡪 SIR 모형보다 더 현실적.
6. SIRS모형: S -> I -> R -> S 과정을 순환하는 모형. R 상태의 개체들이 다시 S상태로 갈 수 있는 모형.
7. SEIRS모형: S -> E -> I -> R -> S 과정을 순환하는 모형.
8. Contact process: ‘모든 개체가 건강한 상태를 유지하는 상’과 ‘전염상태가 유지되는 상’의 두 상 사이의 상전이를 설명하는 모형. SIS모형의 일종.
9. 참고할만한 글들
10. 감염병 모델에 대해 개괄적으로 소개하는 글 : <http://fracton.khu.ac.kr/~syook/reprints/complex_special/epidemic.pdf>
11. 감염병 모델 사용시에 필요한 변수들과 실제 코로나 데이터를 활용한 전염병 모델 시각화:

<http://www.asaninst.org/contents/%EC%A0%84%EC%97%BC%EB%B3%91-%EB%AA%A8%EB%8D%B8%EA%B3%BC-covid-19/> - 아산정책연구원

1. SIR, SEIR 모델의 수식과 그 수식의 각각의 항들이 의미하는 바, 나오게 된 논리 등을 자세히 설명 : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7270519/#CR5>
2. 기존의 SEIR 모형의 S를 “감염전파율 감소를 위해 노력하는 행동변화 그룹”과 “그렇지 않은 그룹” 두 그룹으로 나눠서 분석한 보고서도 존재.

* <https://www.e-epih.org/upload/media/epih-42-e2020026-suppl.pdf>

1. 변수& 데이터 확보 가능 여부: 감염병 모델 적용을 위해선 다음의 변수가 필요하다.

(1) 감염자가 건강한 사람에게 전염시킬 가능성을 나타내는 β : β = γ R0로 구할 수 있다.

(2) 감염자가 회복되는 변화율을 나타내는 γ = : 코로나는 확진부터 완치까지 걸리는 시간이 평균 20일

(출처: <https://m.health.chosun.com/svc/news_view.html?contid=2020062902672>)

🡪 따라서 γ = 0.05

(3) 기초감염재생산수 R0 : 질본 보고서에 주차별로 존재.

🡪 질본 “감염재생산지수 개념 및 방역정책에 따른 변화” 보고서에 시간별로 존재. (뒷 페이지 참고)

\*\*\* 위의 세 변수를 구한 후 SIR의 세 미분방정식의 식에 대입해서 적분하면 시간에 따른 S, I, R 그룹의 비율이 나오므로 그래프를 그릴 수 있다. 그런데 SIR은 한계가 존재하는 모형이므로 SEIR, SEIQR 모형 등을 사용해야 할 수도 있다. 그럴 경우, 아래의 변수가 필요.

1. 발현율 σ = 을 구하면 SEIR 모형 사용 가능. 평균 잠복기는 5.2일 이므로 코로나 발현율은 약 0.2

(출처: <https://www.medicaltimes.com/Users/News/NewsView.html?ID=1131695>)

1. R에서 다시 S로 돌아가는 비율인 ε 을 구하면 SEIRS 모형 사용 가능.

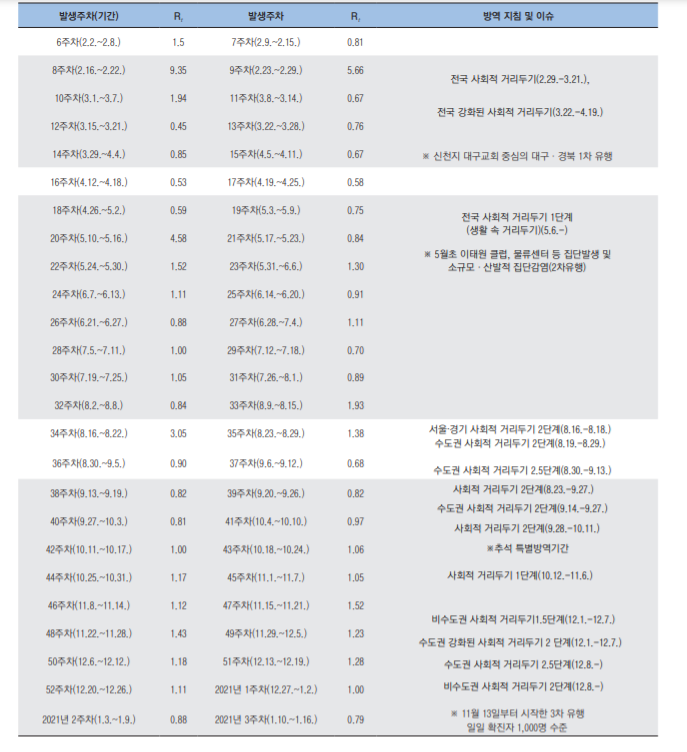
* 재감염율을 찾지는 못했다.

다른 정보를 찾았는데, 가볍게 앓고 회복한 환자의 6%에선 중화항체가 생기지 않았다. ( = R 집단에서 가볍게 병을 앓았던 사람이 S로 돌아가는 비율이 0.06이다. 병을 심하게 앓았던 R 집단의 사람이 S로 돌아가는 비율은 구하지 못했다. )

출처: <https://www.snu.ac.kr/coronavirus/research?md=v&bbsidx=128818>

1. 초기값 : 가 주어지면 도 구해진다, () 초기값은 우리가 설정하는 값이므로 모델 사용에 문제X.

이 모든 모형들의 미분방정식 수식& 시각화 정리, 감염병 모형 코드 in python

* <https://commonengineerr.tistory.com/19>
* 주차별 기초감염재생산수