공공 데이터 분석팀
결과 발표

허찬, 이지현, 조민제,
이나영, 이정진

KU-BIG

Contents

KU-BIG

NLP For Sentiment Analysis

Data description

- 2020년 1월부터 네이버 뉴스에 등재된 코로나 관련 기사 분석
- 수집 키워드 지정 → 뉴스 제목, 내용, 댓글 등을 크롤링

	Unnamed: 0	URL	댓글작성자	댓글	공감수	비공감수	댓글작성시간
0	0	https://news.naver.com/main/read.nhn?mode=LSD&...	jue7****	다행이네요. 근데 희한하게 하필 많고많은 직업중에 어린이집 교사들이 ㅜㅜ	3	1	2020-02-01 01:23:22
1	1	https://news.naver.com/main/read.nhn?mode=LSD&...	blue****	한동안 미세먼지는 없겠구만	6	0	2020-02-01 08:04:04
2	2	https://news.naver.com/main/read.nhn?mode=LSD&...	love****	1차 전세기에 못탄 한국인 1명 있었다는데 그심정은 말로 표현 못할거같아요 한국으...	23	1	2020-02-01 02:10:29
3	4	https://news.naver.com/main/read.nhn?mode=LSD&...	hoh3****	도대체 이분들 데리러 가고 오고가 왜 속보며 왜이리 호들갑이냐?? 우한서 들어온 6...	226	22	2020-02-01 00:07:09
4	5	https://news.naver.com/main/read.nhn?mode=LSD&...	dldl****	7번째확진자도 음성이다해서 격리해제로퇴원시켰지만 그뒤로 안좋아져서다시 확진자로 되었...	102	2	2020-02-01 00:58:57

KU-BIG

EDA(Word Cloud)

✓ 뉴스 기사 제목에서의 이슈 관측

Keyword: 코로나

Keyword: 사회적 거리두기

EDA(Word2Vec)

```
model.wv.most_similar("코로나/Noun", topn=10)
```

```
[('코로나바이러스/Noun', 0.6338039636611938),
 ('폐렴/Noun', 0.4644114673137665),
 ('감염병/Noun', 0.3794344663619995),
 ('전염병/Noun', 0.37329983711242676),
 ('COVID/Alpha', 0.3475281596183777),
 ('전세계/Noun', 0.3244360387325287),
 ('CV/Alpha', 0.3215518891811371),
 ('감염증/Noun', 0.3162800669670105),
 ('ASF/Alpha', 0.29650524258613586),
 ('팬데믹/Noun', 0.29388681054115295)]
```

```
model.wv.most_similar("자가/Noun", topn=10)
```

```
[('무단이탈/Noun', 0.5117320418357849),
 ('자택/Noun', 0.4610103368759155),
 ('증상/Noun', 0.4322056174278259),
 ('자각/Noun', 0.42662400007247925),
 ('확진/Noun', 0.4179067015647888),
 ('완치/Noun', 0.4144691526889801),
 ('접촉/Noun', 0.41152948141098021),
 ('귀가/Noun', 0.40454149246215821),
 ('입원/Noun', 0.39343252778053284),
 ('해제/Noun', 0.39183616638183594)]
```

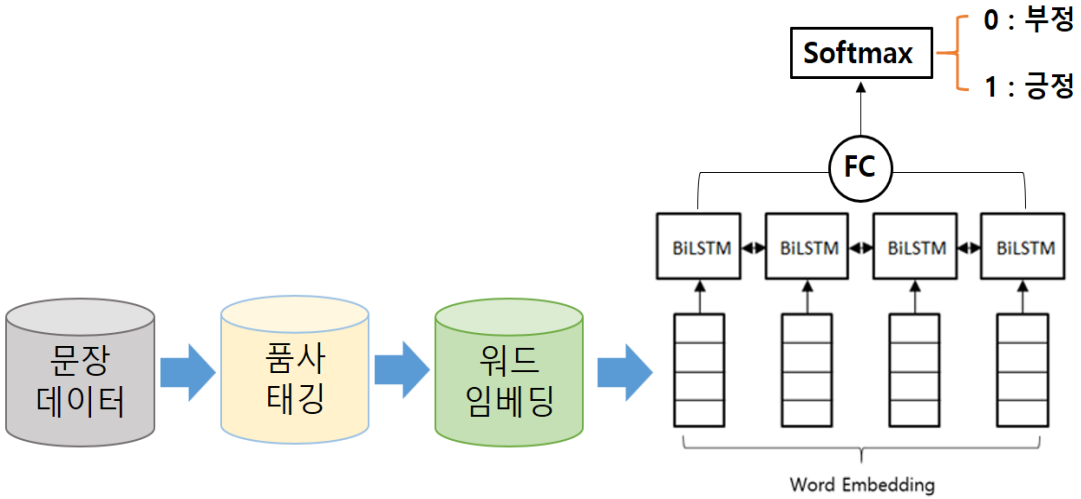
```
model.wv.most_similar("격리/Noun", topn=10)
```

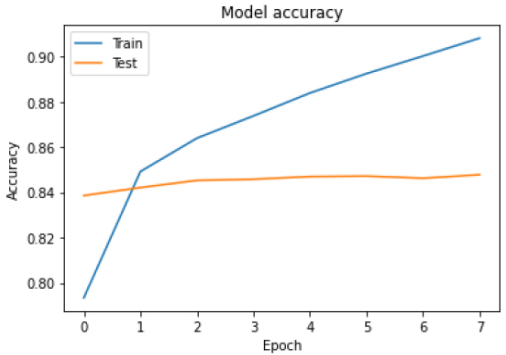
```
[('당착/Noun', 0.6667430400848389),
 ('입소/Noun', 0.5311622619628906),
 ('입원/Noun', 0.5010474920272827),
 ('귀가/Noun', 0.48056691884994507),
 ('대기/Noun', 0.4654673337936401),
 ('완치/Noun', 0.45034873485565186),
 ('증상/Noun', 0.4500478506088257),
 ('퇴소/Noun', 0.4497900605201721),
 ('하선/Noun', 0.4488641619682312),
 ('귀국/Noun', 0.4456971287727356)]
```

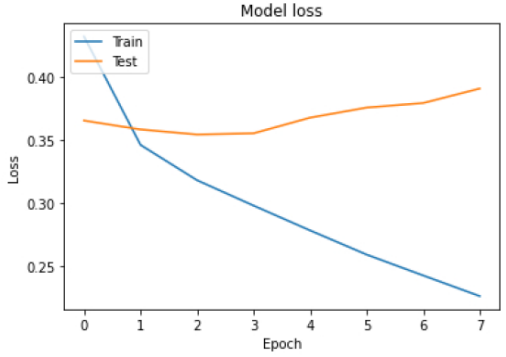
KU-BIG

Sentiment Analysis(BiLSTM)

- 네이버 영화 리뷰 파일: train set 15만 건, test set 5만 건 학습
- 학습한 모델을 뉴스 댓글 긍정 부정 분류에 적용



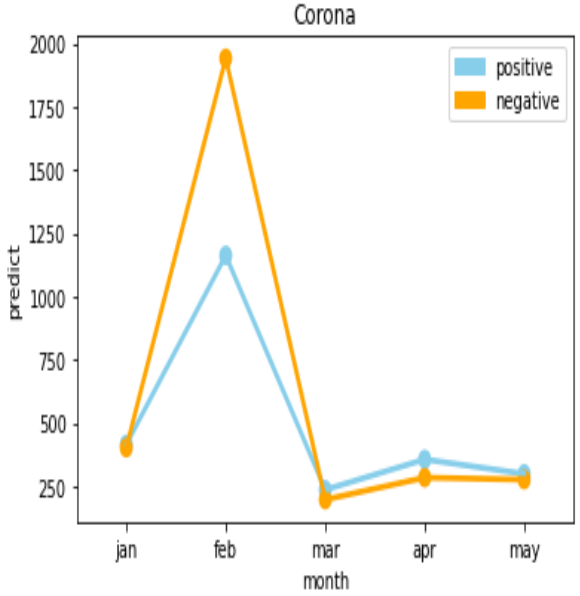


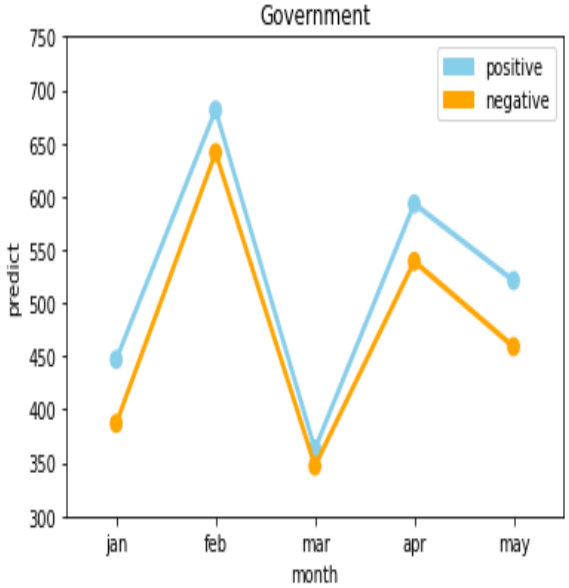


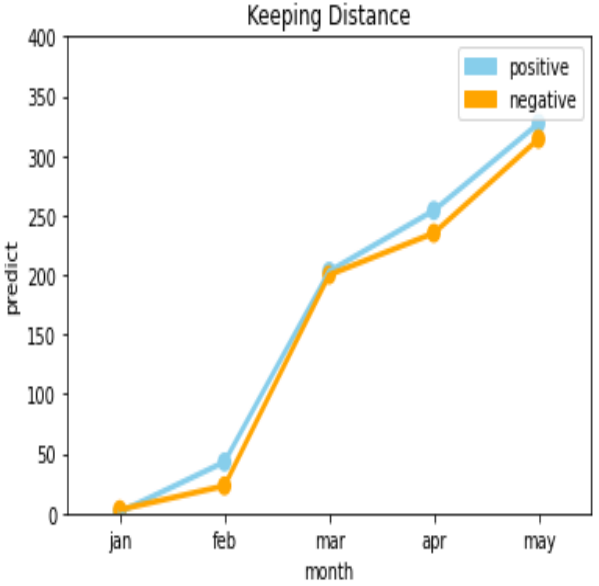
KU-BIG

Sentiment Analysis(BiLSTM)

댓글 수집 키워드: 코로나, 정부, 사회적 거리두기







Sentiment Analysis(Text-CNN)

불용어 제거, 토큰화 후 감성사전을 통한 댓글 분석

```
# 정규 표현식을 통한 한글 외 문자 제거

data['comments'] = data['comments'].str.replace("[^ㄱ-ㅎㅏ-ㅣ가-힣 ]","")
```

```
# 불용어 정의

stop_words = []

with open('stopwords.txt', 'r', encoding = 'utf-8') as f:
    lines = f.readlines()
    for line in lines:
        stop_words.append(line.rstrip('\n'))
```

```
# 형태소 분석기 OKT를 사용한 토큰화 작업

okt = Okt()
tokenized_data = []
for sentence in data['comments']:
    temp_X = okt.morphs(sentence, stem=True) # 토큰화
    temp_X = [word for word in temp_X if not word in stop_words] # 불용어 제거
    tokenized_data.append(temp_X)
```

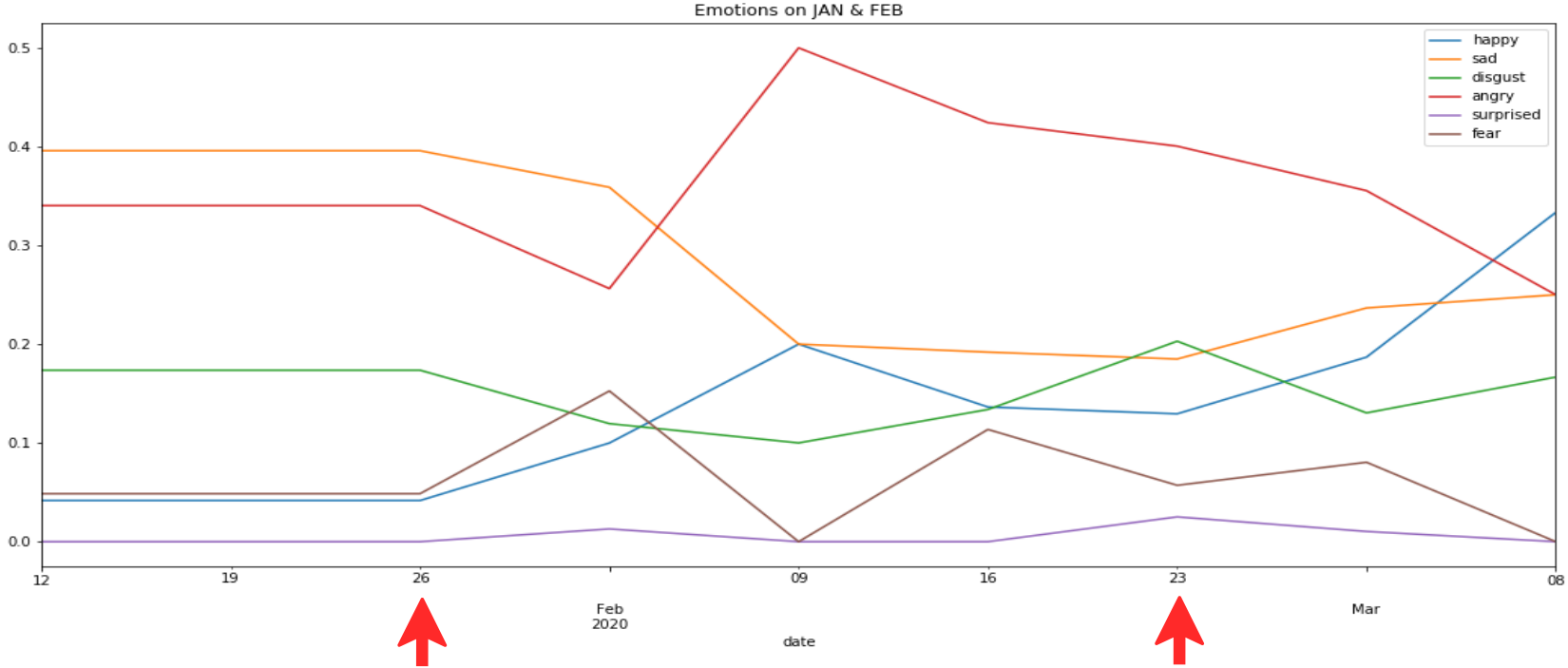
```
series_token_data = pd.Series(tokenized_data)
series_token_data[:5]
```

```
0    [황금, 연휴, 보내다, 개인, 방역, 수, 칙, 지키다, 일상, 돌아가다, 분수령...
1    [행동, 몇몇, 사람, 때문, 피로, 감, 느끼다, 사람, 계속, 느끼다, 야하다,...
2    [지금, 최소한, 마트, 가지, 않다, 생활, 설연휴, 시작, 계속, 지치다, 그렇...
3              [거리, 두기, 피로, 감, 느끼다, 아니다, 코로나, 폐, 질환, 느끼다]
4                    [꼭, 필요하다, 지키다, 야하다, 지치다, 사실, 이다]
dtype: object
```

	happy	sad	disgust	angry	surprised	fear	comments
1	0.0	1.0	0.0	0.0	0.0	0.0	그러나 마음대로 행동하는 몇몇 사람들 때문에 피로감을 느끼는 사람은 계속 느껴야한다...
7	0.0	0.0	0.6	0.4	0.0	0.0	버스에서전화질 수다 마스크미작용 민폐개진상 들 엄청많아요 해이해진분위기 확실하게...
10	1.0	0.0	0.0	0.0	0.0	0.0	대한국민이 최고예요지금 다른나라에선 한달도 못버티고 해수욕장에서 바글바글 모여다니고...
12	0.0	1.0	0.0	0.0	0.0	0.0	거리두기는 괜찮은데 바이러스가 사라지지않을까 봐 걱정스럽고 우울해지네요
16	0.0	0.0	0.0	1.0	0.0	0.0	대한민국 확진자 명 나왔을때 중국의 명절로 인하여 세계 각국으로 우한폐렴 전파가 될...

KU-BIG

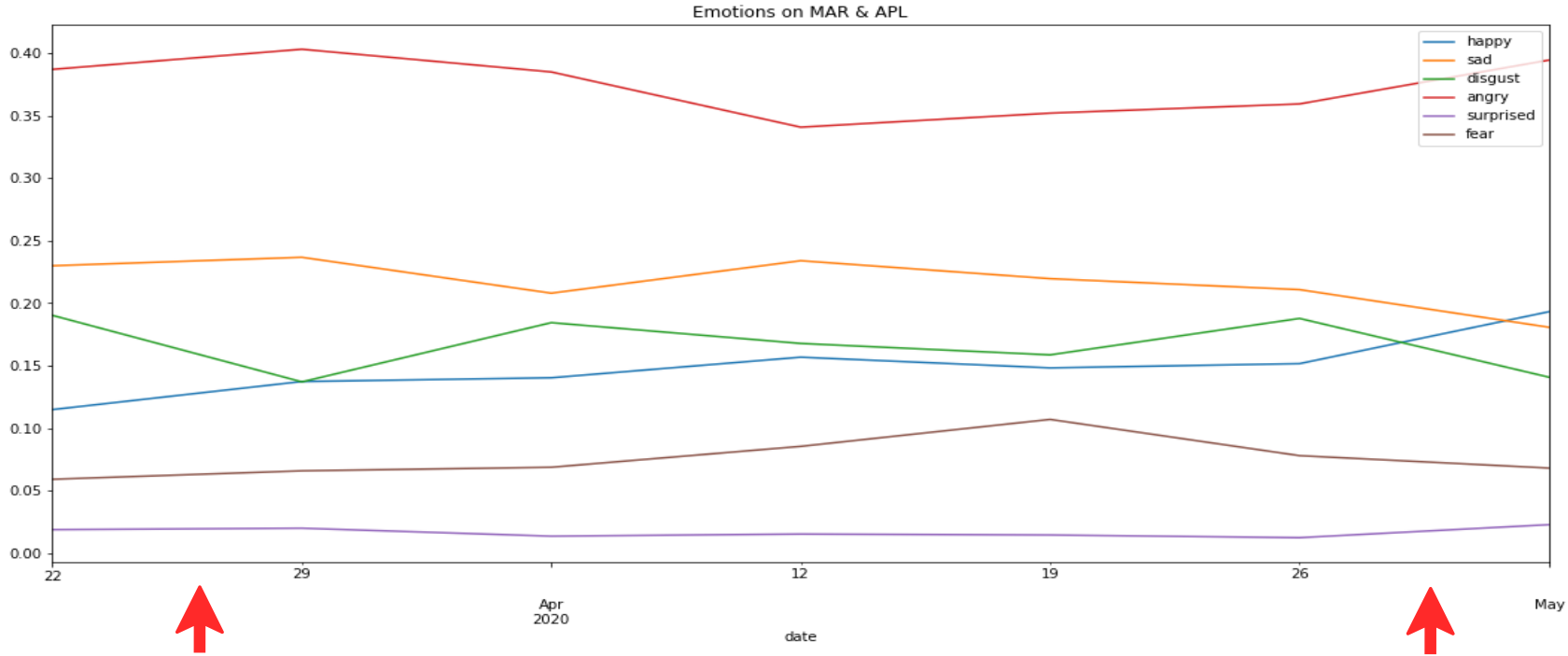
Sentiment Analysis(Text-CNN)



위기 경보 수준 `주의 ' → ' 경계 '

위기 경보 수준 `경계 ' → ' 심각 '

Sentiment Analysis(Text-CNN)



사회적 거리두기 시작

사회적 거리두기 연장

Limitations

Dataset

- 데이터의 개수 부족
- 뉴스 댓글의 특수성

	happy	sad	disgust	angry	surprised	fear
0	가뿐하다	가슴앓이	가소롭다	갈기갈기	갑작스럽다	가혹하다
1	감개무량하다	가엽다	거북하다	개새끼	경악하다	강압적
2	감격스럽다	가엾다	경박하다	개자식	경이	겁쟁이
3	감격하다	가혹하다	괴상하다	격노하다	급작스럽다	공포감
4	감동스럽다	각박하다	괴팍하다	격분하다	기겁하다	공포스럽다

BiLSTM

- 모델 데이터와 댓글 데이터의 차이

```
print("{}개의 댓글 중 {} 개의 댓글이 감정 단어를 보유."
```

856개의 댓글 중 198 개의 댓글이 감정 단어를 보유.

Text-CNN

- 감정 사전 단어의 개수 부족

KU-BIG

Visualization & Variable Selection

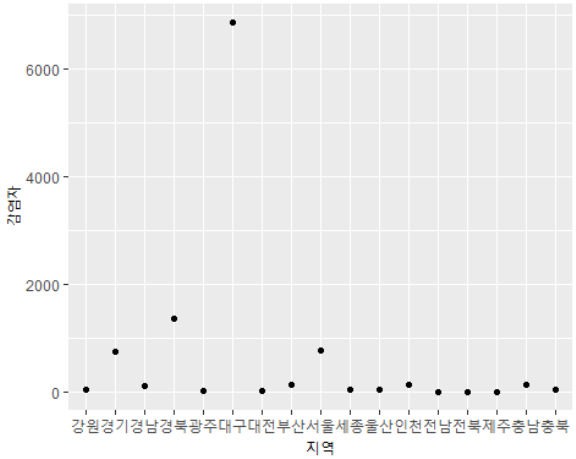
KU-BIG

Visualization

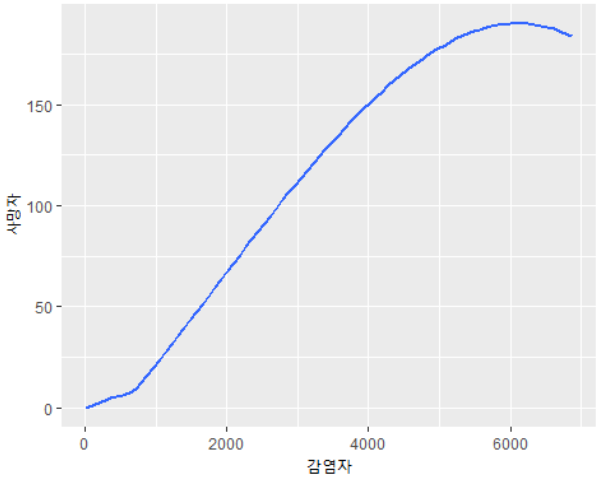
✓ 변수의 설정

Visualization

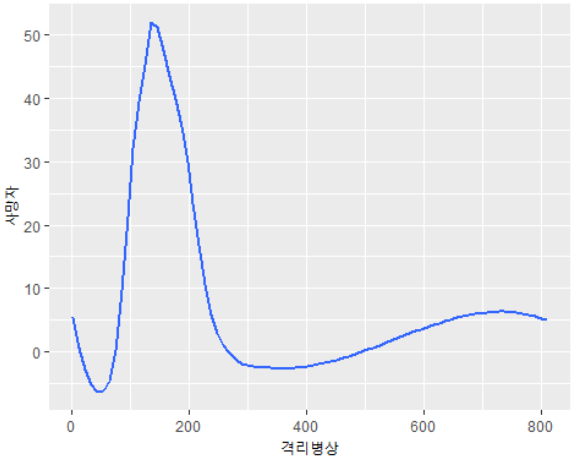
변수들 간의 상관관계



지역 & 감염자



감염자 & 사망자

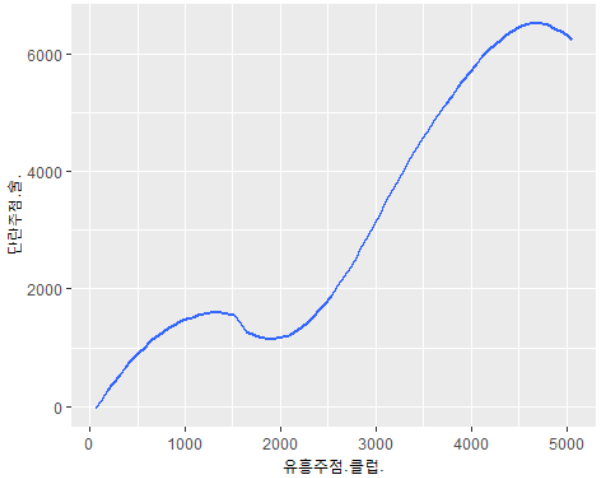


격리병상 & 사망자

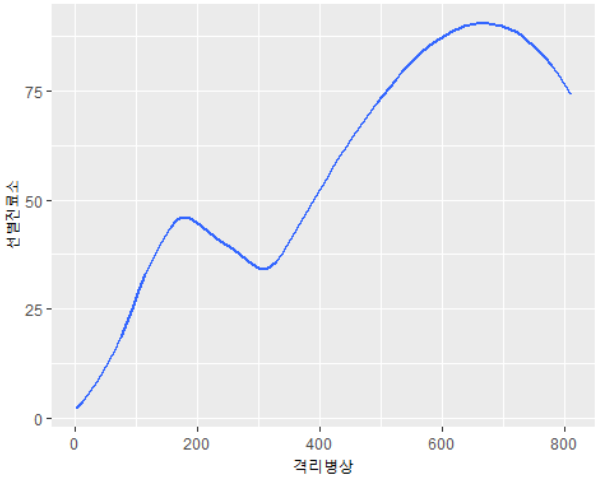
KU-BIG

Visualization

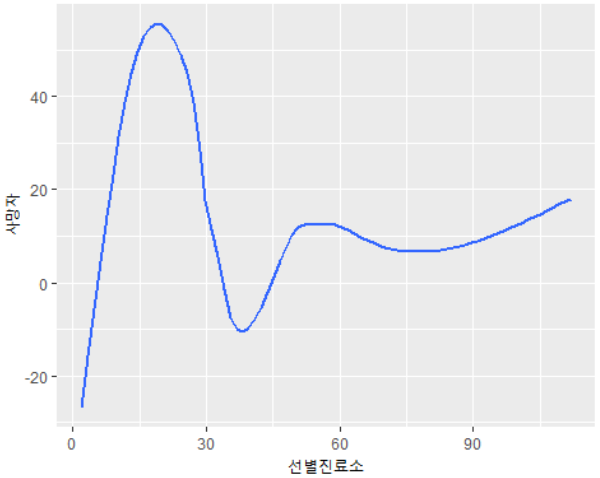
변수들 간의 상관관계



술집 & 클럽



선별진료소 & 격리병상



선별진료소 & 사망자

KU-BIG

Variable Selection

✓ Lasso Regression

감염자 = -8.34(격리병상) -31.19(선별진료소) -0.69(술집)

-0.05(신천지) +0.91(인구) +0.66(인구밀도) -0.91(1인당 소득수준)

격리병상과 선별진료소가 감염자와 가장 연관성↑

Variable Selection

✓ Random Forest

감염자 ← 술집 > 선별진료소 > 격리병상 > 인구밀도

> 인구 > 클럽 >1인당 소득 수준 > 신천지 수

술집 & 격리병상 & 선별진료소가
감염자와 가장 연관성↑

KU-BIG

Insights

변수 선택을 통한 인사이트 도출

- **술집**: 지역사회에서 만연한 집단감염 및 무증상 감염자의 확산을 의미
- **격리병상**: 음압병상과 집중치료시설이 방역당국의 감염률 관리에 중요한 역할
- **선별진료소**: 자발적으로 검사를 받고 지역감염의 연결고리를 끊는 게 가장 중요함

“의료시스템이 감당 가능한 선 아래로 확진자 수 유지”

선별진료소 및 의료 인프라 확보

의심환자가 스스로 검사를 받아 지역사회 감염 차단

03 Q&A

KU-BIG