****

**CSR:** 코로나 감염 현황에 따른 언론의 감정변동추이

Team **막강하니**

📊Changes in media Sentiment according to coRona infection status  
#DeepLearning#NLP#emotional analysis#Web Crawling

현재 가장 큰 이슈인 팬대믹(pandamic) 코로나 바이러스 기사에서 **정치편향적이고 불필요**한 정보를 명시하여 사회분란을 조장하는 사례가 발생하고 있습니다.  
같은 내용의 사건을 기사로 내는데 언어표현은 민중심리를 좌지우지하는데 많은 책임감을 지닙니다. 한국기자협회의 윤리강령 제 2항 ‘공정보도’: ‘우리는 뉴스를 보도함에 있어서 진실을 존중하여 정확한 정보만을 취사선택하며, 엄정한 객관성을 유지한다.’ 라는 내용에 부합하게 **정당한 보도가 이루어지고 있는 것인지** AI기술을 사용하여 그 실태를 조사합니다.

# Dataset

공공API의 두 XML파일을 사용합니다.  
`보건복지부\_코로나19 감염\_현황`: 누적 확진자수, 사망자수, 검사자수, 누적 확진률 등의

**코로나 감염 현황 정보**

`한국언론진흥재단\_뉴스빅데이터\_메타데이터\_코로나`: 일자, 언론사, 제목, 키워드, 본문 등의

**코로나 관련 뉴스 분석 자료**

위 두 자료를 이용하여 3개월간의 코로나 키워드 뉴스 데이터를 csv파일 형태로 받았습니다. 원하는 칼럼(일자/본문/키워드)만 슬라이싱하여 새로운 데이터셋을 구성했습니다.

# Realated Work

* NLP(Natural Language Processing)  
  인간의 언어를 기계가 모사할 수 있도록 하는 인공지능의 한 분야입니다. 언어의 형태소를 분석하고 구절 단위를 분석합니다. 기사의 본문을 자연어 분석(토큰화, 감정사전 대조)을 통해 핵심 단어를 추출하고 감정변화의 추이를 그려봅니다. 우리는 **KoNLPy**라는 한국어 형태소분석기를 사용했습니다.

#KoNLPy가 지원하는 모듈 여러 개(Hannanum, Okt 등)를 테스트하며 어떤게 가장 감성탐지에 효과적일지 판단했습니다.

* Emotional analysis : 감정분석  
  한국어 감정 사전을 이용하여 기사를 세 가지 기준으로 감정을 분석합니다. 14,182개의 한글감성어를 지원하는 **EmoLex**감성낱말사전을 사용하여 긍부정 감정, 8 감정을 분석했습니다. 객관성주관성분석은 textblob모듈을 사용했습니다  
  +긍부정분석  
  +8감정(분노, 기쁨 등)분석  
  +객관성주관성분석
* Web Crawling  
  인증키를 받아 공공API 데이터를 가져와 분석합니다. 코로나 감염 현황과 코로나 관련 기사 데이터셋을 받아와 감정분석, 그래프 시각화에 필요한 정보만 슬라이싱하여 사용했습니다.

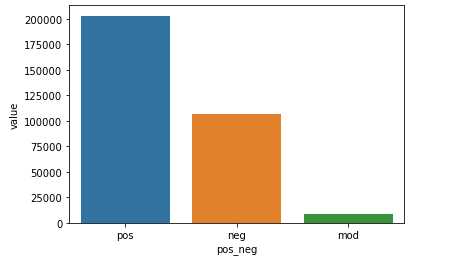
# Methodology

1. **긍부정 감정분석**

EmoLex감정사전을 긍부정 분석에 사용할 수 있도록 단어, positive, negative 총 세 칼럼으로 잘라 사용하였습니다. 중복데이터 및 value가 없는 값은 삭제 시켜 크기를 대폭 줄였습니다.( 14,182 words > 4,001 words)

뉴스 기사 본문 text 또한 특수기호를 제거하는 전처리 과정을 거쳤습니다. 정제된 본문을 형태소 분석기 KoNLPy의 Okt모듈을 사용하여 형태소별로 잘라 저장했습니다.

\*key idea를 사용하여 런타임을 12hours에서 6mins로 단축시켰습니다.



시각화는 matplotlib를 사용하였습니다. x축은 긍정,부정,중립 3가지의 감정이고, y축은 뉴스기사의 개수입니다.

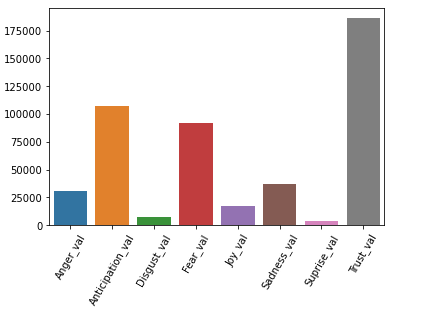
긍부정감정을 분석해본 결과, **긍정의 감정을 가지는 뉴스기사가 부정기사의 약 2배 가까이 쓰여지고 있다**는 것을 확인할 수 있었습니다.

뉴스데이터의 긍부정을 판단하는데 어떤 단어가 많이 쓰였는지 알아보기 위해 감정판단에 사용된 단어들의 빈도수를 조사하여 WordCloud로 시각화 하였습니다. 대표적인 긍정단어는 **‘대통령’, ’지원’, ‘참여’**가 있고, 대표적인 부정단어는 **‘정부’, ‘감염’, ‘위기’**가 있었습니다.

[](https://github.com/seawavve/CSR/blob/main/pos_keyWordCloud.png)[](https://github.com/seawavve/CSR/blob/main/neg_keyWordCloud.png)

1. 8감정 분석

EmoLex감정사전의 8가지 감정. 긍부정 분석과 같이 본문 텍스트 전처리, Okt 형태소 분석을 거쳐 3개월 간의 코로나 뉴스 기사를 8가지 감정으로 분석해 보았습니다.

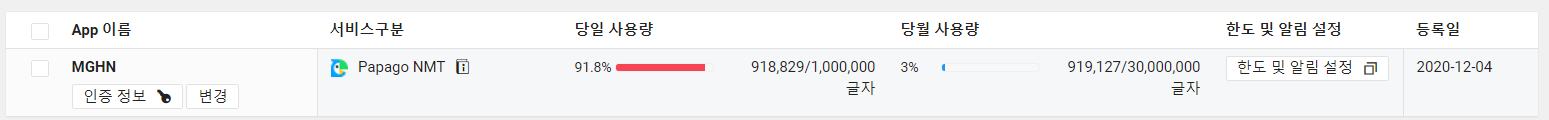


시각화는 matplotlib를 사용하였습니다. x축은 Fear, Anticipation, Joy 등의 8가지 감정이고, y축은 뉴스기사의 개수입니다.

8감정을 분석해본 결과, 코로나 뉴스기사에서는 **믿음(Trust)감정이 8감정 중에 가장 빈도수가 높게 나타난다**는 것을 확인할 수 있습니다.

1. 객관성 분석

객관성주관성 분석을 지원하는 모듈 TextBlob을 사용했습니다. 영문만 지원하기에 데이터셋을 적용시키기위해 문장 번역기를 사용했습니다. 본문의 size가 크기에 Naver Cloud PlatForm의 Papago API-Translation을 사용했습니다. 인증 아이디와 키를 받아와 약 920,000개의 글자를 번역했습니다.

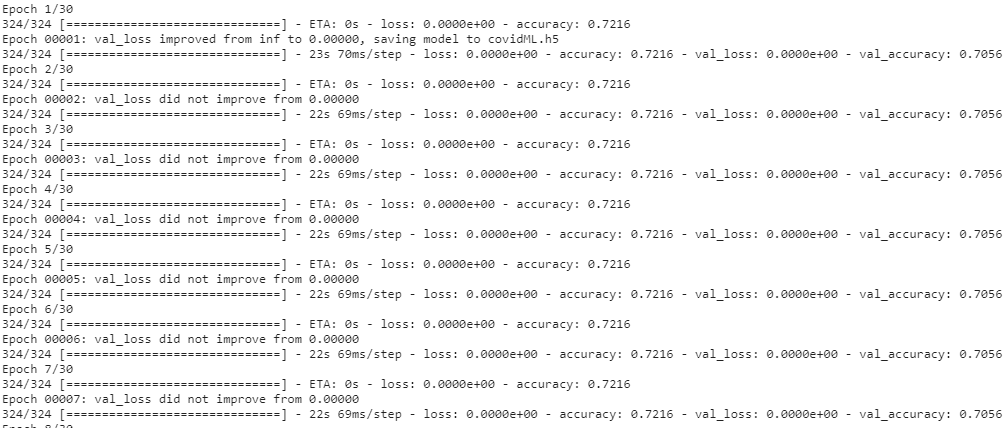


0~1사이의 값으로 객관성을 표현하는 textblob을 이용하여 감정분석을 했습니다. 값이 0에 가까울수록 객관적이고 1에 가까울수록 주관적입니다. 뉴스기사들의 객관성 전체 평균을 뽑은 결과 0.3824664050591109로 **어느정도 주관적인 의견이 들어간 기사를 작성했음**을 알 수 있었습니다.

1. 긍부정 감정학습

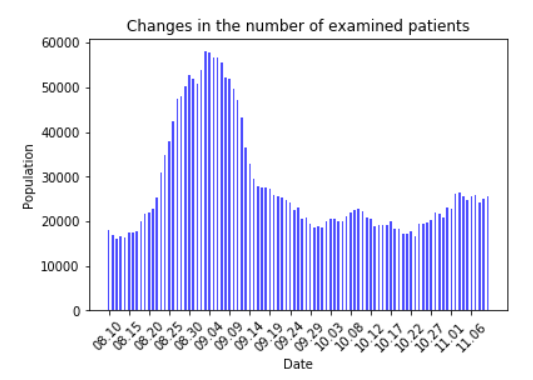
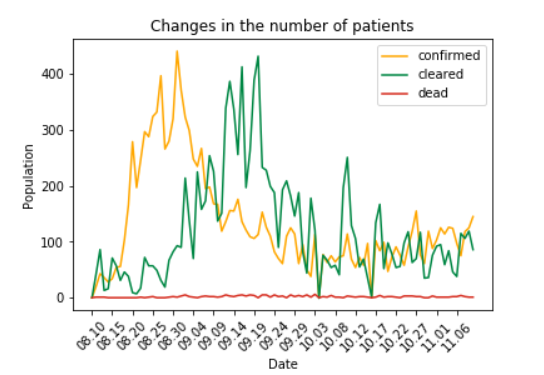
뉴스 본문의 키워드를 가져와 긍부정을 부착하고 학습을 실시했습니다. text들을 토큰화하여 train set은 1800개, test set은 200개로 설정했습니다.

학습 능력향상을 위해 하이퍼파라미터 조정을 하고, optimizer를 바꿔서 실험을 해보았습니다. 또, earlystopping 기법을 사용하여 과적합이 일어나기 전에 학습을 멈춰서 modelcheckpoint로 최고성능의 모델을 반환하도록하였습니다.



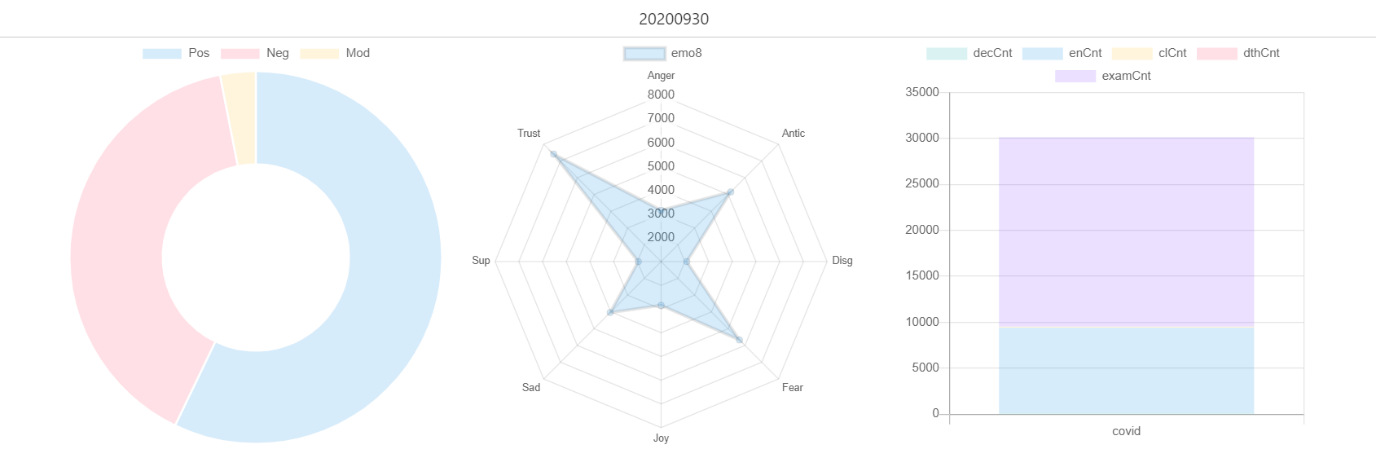
학습 결과 71%로 높은 정확도를 보임을 알 수 있습니다.

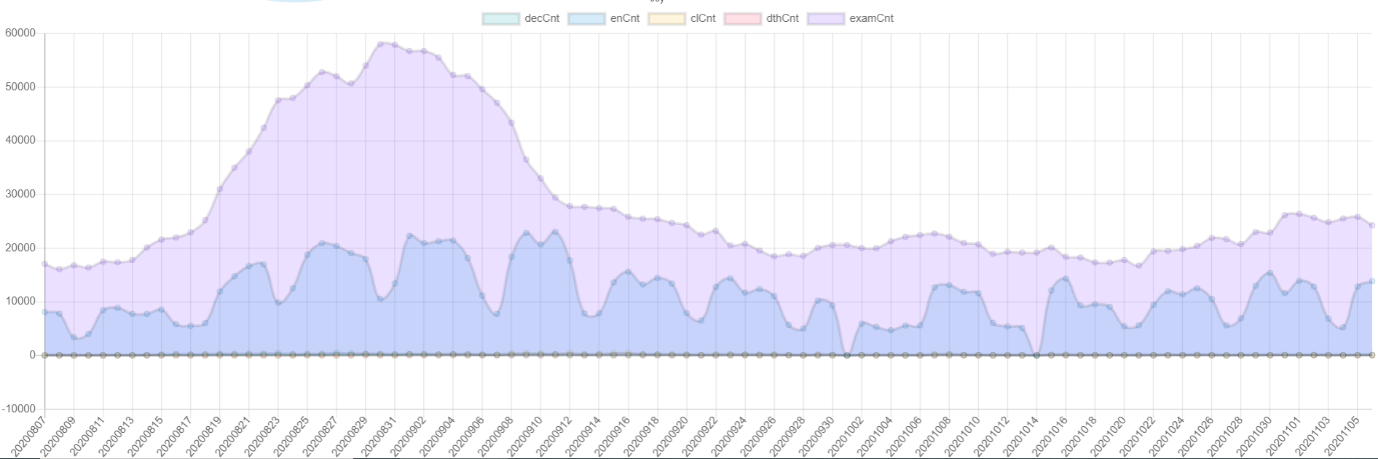
1. COVID-19현황 크롤링



공공API `보건복지부\_코로나19 감염\_현황` 데이터를 사용하여2020.08.07~2020.11.06 총 3개월 간의 환자 추이를 그렸습니다. Matplot 라이브러리를 활용하여 확진자, 완치자, 치료중환자, 사망자 분포도를 그려 전체적인 현황을 파악해보았습니다.

1. COVID-19 감염 현황에 따른 언론 감정변동 UI





url: <https://csr-demo.herokuapp.com/>

긍부정 감정분석, 8감정분석, COVID-19현황 크롤링을 사용해 수집한 모든 데이터를 날짜별로 잘라 UI로 구축했습니다. 긍부정은 파이그래프, 8감정은 방사형그래프, COVID-19현황 그래프는 라인그래프로 각 수치를 표현했습니다. 각 그래프는 2020.08.07 ~ 2020.11.06 날짜를 선택함에 따라 당일의 감정 데이터를 보입니다.

긍부정그래프에서 pos는 긍정, neg는 부정, mod는 중립을 의미합니다. 대부분의 긍정지표가 부정지표보다 높은 것을 알 수 있습니다.

8감정그래프에서 anger는 분노, Antic은 기대, Fear는 두려움, Joy는 즐거음, Sad는 슬픔, Disg는 혐오 sup은 놀라움, trust는 신뢰를 의미합니다. 대부분의 일자에서 가장 높은 감정으로는 신뢰, 두번째 높은 감정으로는 기대였습니다.

COVID-19현황 그래프에서 decCnt는 확진자수, enCnt는 음성결과수, clCnt는 완치자수, dthCnt는 사망자수, examCnt는 검사자수입니다. 가장 검사자수가 높은 일자(2020.08.30)의 뉴스 감정표를 보면 부정의 기사가 600>1000개로 평소의 약 2배 가까이 증가했고, 8감정 중에서는 혐오의 감정이 평소보다 30%정도 더 증가함을 확인할 수 있습니다.

# Conclusion

전체적으로 **긍정**의 기사가 부정의 기사보다 많았습니다. 8감정 중에서는 **Trust**와 **Anticipation** 감정이 다른 감정보다 눈에 띄게 빈도가 높았습니다. 주관성 수치에 따라 어느정도 **객관성을 따르지 않는 뉴스기사들이 많이 존재한다**는 것을 알 수 있습니다. COVID-19 **확진자수, 검사자수가 증가함에 따라 부정적기사, 혐오 두려움 등의 부정적인 감정을 가지는 뉴스기사의 수가 증가함**을 확인했습니다.

# Ending

# -한국어 감성사전의 부재가 프로젝트를 진행하는데 큰 어려움이었습니다. 그로인해 영어로 된 EmoLex라는 영문 사전을 번역하여 감정분석에 사용했습니다.

# 그로인해 발견한 문제점 첫번째는, 영어로 된 서로다른 단어가 한글로 번역하자 같은 단어로 번역이 된다는 것입니다. (ex. Apple, Apology -> 사과 ) 이로 인해 긍부정 value의 정확도가 떨어졌고, 중복되는 단어가 많았습니다.

# 두번째는, 영어단어가 번역되면서 아예 다른 뜻을 갖는 한글단어로 오역되는 경우입니다..

# 

# 긍부정 감정분석에서 긍정 단어 WordCloud자료입니다. ‘환자’라는 단어가 긍정으로 분류되는 점이 의문스러웠지만 ‘patient(참을성있는)’라는 단어가 한국어로 번역이 되면서 긍정 감성수치를 가지는 ‘patient(환자)’로 변환된 것을 확인할 수 있었습니다. 이러한 여러 개의 번역오류들이 감정을 분석하는데 정확도를 낮췄습니다.

# 세번째로는, 번역이 깨져서 ‘NO TRANSLATION’이라는 값이 많았습니다.

# 위와 같은 쓰레기 값들이 정확한 감정분석을 하는 데 걸림돌이 되었습니다. ‘한국어 감성사전’을 직접 구축하여 재실험을 진행하거나 open으로 배포한다면 한국의 ML기술을 발전시키는데 크게 이바지할 것으로 예상합니다.

# -한글문장의 객관성을 형태소를 잘라 한글 그대로 분석할 수 있는 방법을 찾지 못해 영어로 번역하여 분석한 점이 아쉽습니다. Textblob은 영문문장을 받아 객관성과주관성을 수치로 분석하여 반환하여 주는 모듈입니다. 긍부정, 8감정, 객관성 판단까지 한 번에 지원하는 한글 모듈을 만든다면 한글NLP발전을 이끌 것입니다.

-윤리강령에 부합하여 이루어지는지 데이터 분석을 통해 살펴보았습니다. 편향적인 감정을 이끌어내는 '미끼단어'를 파악하여 많은 사람들이 인지할 수 있게 시각화했습니다. 객관성을 지키고 정당한 사실을 알리는 의무를 가진 언론의 공정보도 실천 사명을 강조하며 중립성을 지키도록 독려할 수 있겠습니다. 또, 시민에게 자료를 공개하여 자극적인 기사에 휘둘리지 않고 객관적인 시선을 가질 수 있도록 일깨우겠습니다.

# Built with

Team 막강하니  
하니 @seawavve

* 아이디어 및 구상
* lexicon dataset 구성
* 2감정, 8감정 분석
* 공공API 코로나 확진자 추이 크롤링 및 시각화
* Github repo 관리
* 학습 아웃라인 구상
* 학습 능률 개선(EarlyStopping, ModelCheckpoint)
* 학습 시각화(epoch에 따른 Accuracy동향 그래프, loss동향 그래프, 네트워크 구조 그림)
* 긍부정 WordCloud
* 날짜별 데이터 전처리
* Papago NMT API 사용
* CSR 해설 문서 작성
* 발표 대본 작성 및 발표

막강 @mintai09

* 2감정 키워드추출
* key 아이디어를 이용한 감정 분석 효율 개선(12 hours -> 6 mins)
* 2감정 학습
* Papago NMT API를 사용한 번역
* 0806-1106 뉴스데이터 구축
* 0806-1106 뉴스본문 키워드 전처리
* 2감정 8감정 수치화 후 기준 설정
* 발표 ppt 제작

url: https://github.com/seawavve/CSR