뉴스 기사 HEADLINE 추출
(SEQ2SEQ + ATTENTION)

4반 7조

김상아, 윤민희

정보경, 정진균

목차

1. 기사 크롤링
2. 전처리
3. 사용 모델
4. 결과
5.프로젝트 고찰

1. 기사 크롤링


```python
def crawling(soup):
    result = soup.find('div', class_='article_word').get_text().replace('\r', '').replace('\n', '').replace('                ', '').replace(
        ' (function(){var s="365863",w="760px",h="150px",a=document.createElement("script");a.src=\'https://native.mediacategory.com/servlet/adNativ
    return result


def get_href(soup):
    Href = []
    for i in soup.find('ul', class_='list_news').find_all('span', class_="tit"):
        i = i.find('a')["href"]
        Href.append(i)
    return Href


def main__contents():
    list_href = []
    result = []

    url = "https://sports.donga.com/ent"

    for i in range(0, 50):
        req = requests.get(url, params={'p': (i*20)+1})
        soup = BeautifulSoup(req.text, "html.parser")
        list_href += get_href(soup)

    for href in list_href:
        href_req = requests.get(href)
        href_soup = BeautifulSoup(href_req.text, "html.parser")
        result.append(crawling(href_soup))

    return result


if __name__ == "__main__":
    main__contents()
```

```python
def crawling(soup):
    result = []
    headline = soup.find('ul', class_="list_news").find_all(
        'span', class_='tit')
    for text in headline:
        text = text.find('a').get_text()
        result.append(text)
    return result


def main__headline():
    answer = []
    url = "https://sports.donga.com/ent"

    for i in range(0, 50):
        req = requests.get(url, params={'p': (i*20)+1})
        soup = BeautifulSoup(req.text, "html.parser")
        answer += crawling(soup)

    return answer


if __name__ == "__main__":
    main__headline()
```

기사 출처 : 스포츠 동아 및 네이버

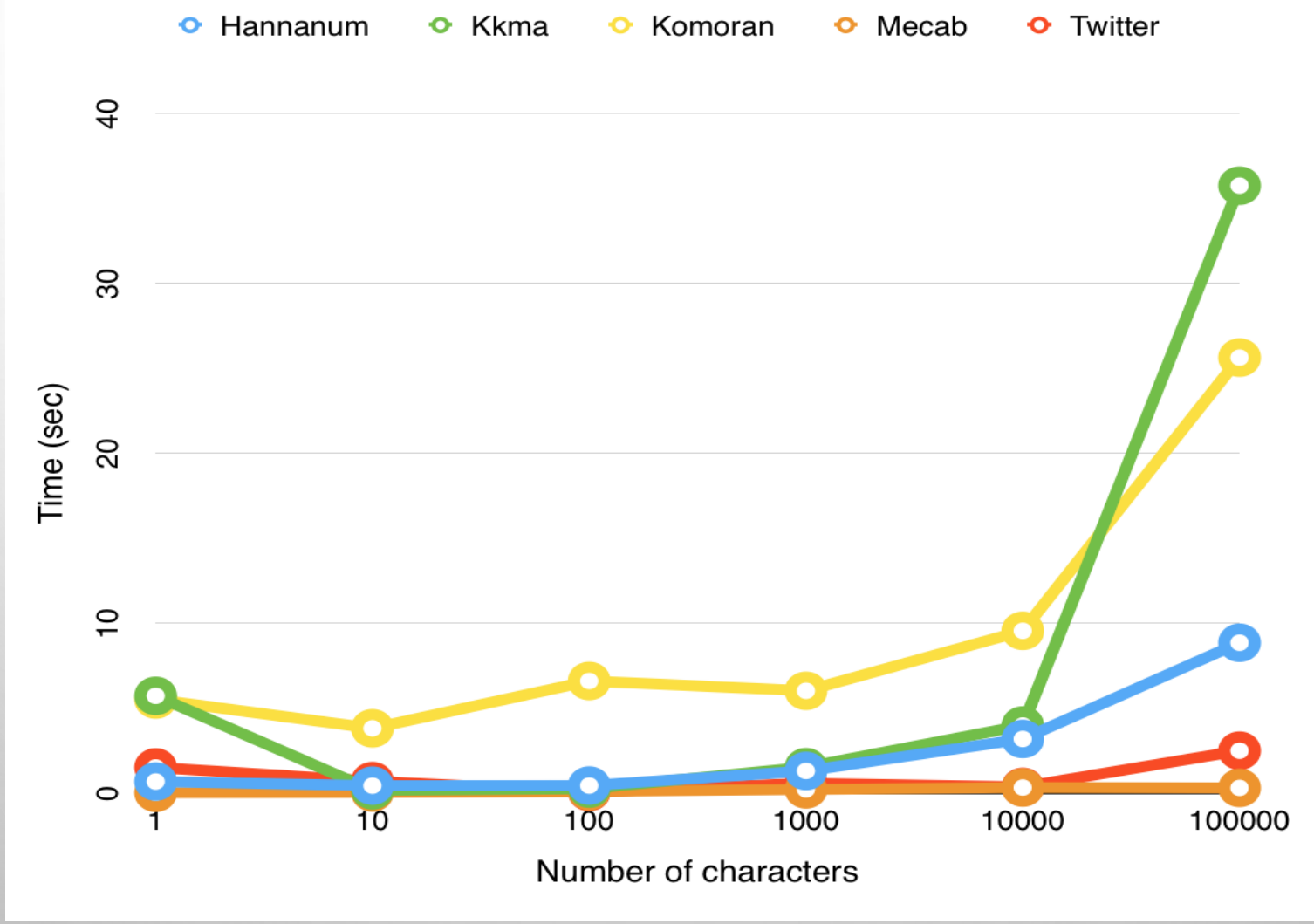
2. 전처리

```
# 데이터 불러오기
news = pd.read_csv("/content/drive/My Drive//23035news_content_headline.csv")
# news = news.append(dff, ignore_index=True)
news = news.loc[:, ['headline', 'content']]


# news.to_csv('23035news_content_headline.csv',encoding='utf-8-sig')

news = news.dropna(axis=0)
# 불용어 제거
# 영어 대 소문자 제거
sw = '[a-zA-Z]'
news['content'] = news['content'].str.replace(sw, ' ')
news['headline'] = news['headline'].str.replace(sw, ' ')

# 한글과 숫자를 제외한 특수문자 제거
news['content'] = news['content'].str.replace('[^ㅟw]', ' ' )
news['headline'] = news['headline'].str.replace('[^ㅟw]', ' ')

# 기사 스탑워드 제거
news['content'] = news['content'].str.replace('기자의 다른기사 더보기', ' ')



# 중복되는 헤드라인 행 제거
news = news.drop_duplicates(['headline'])
```

koNLpy 토크나이저 : Kkma 사용
중복을 제거하여 같은 헤드라인 약 2000개 제거

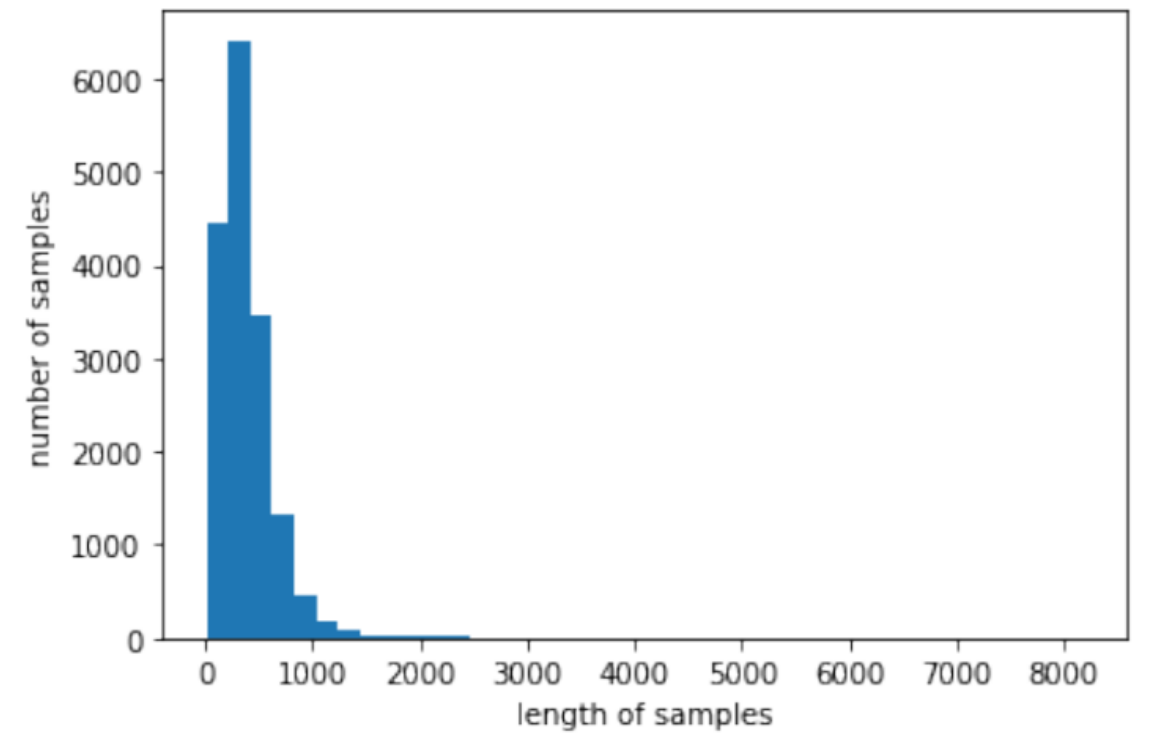
```
# 토큰화
# 형태소추출

encoder_input, decoder_input, decoder_output = [], [], []

i = 0

for stc in news['content']:
    morphs = kkma.morphs(stc)
    if i% 100 == 0:
        print(i)
    i+=1
    encoder_input.append(morphs)
i = 0
for stc in news['headline']:
    morphs = kkma.morphs(stc)
    if i% 100 == 0:
        print(i)
    i+=1
    decoder_input.append(["<start>"]+ morphs)
i = 0
for stc in news['headline']:
    morphs = kkma.morphs(stc)
    if i% 100 == 0:
        print(i)
    i+=1
    decoder_output.append(morphs + ["<end>"])
```

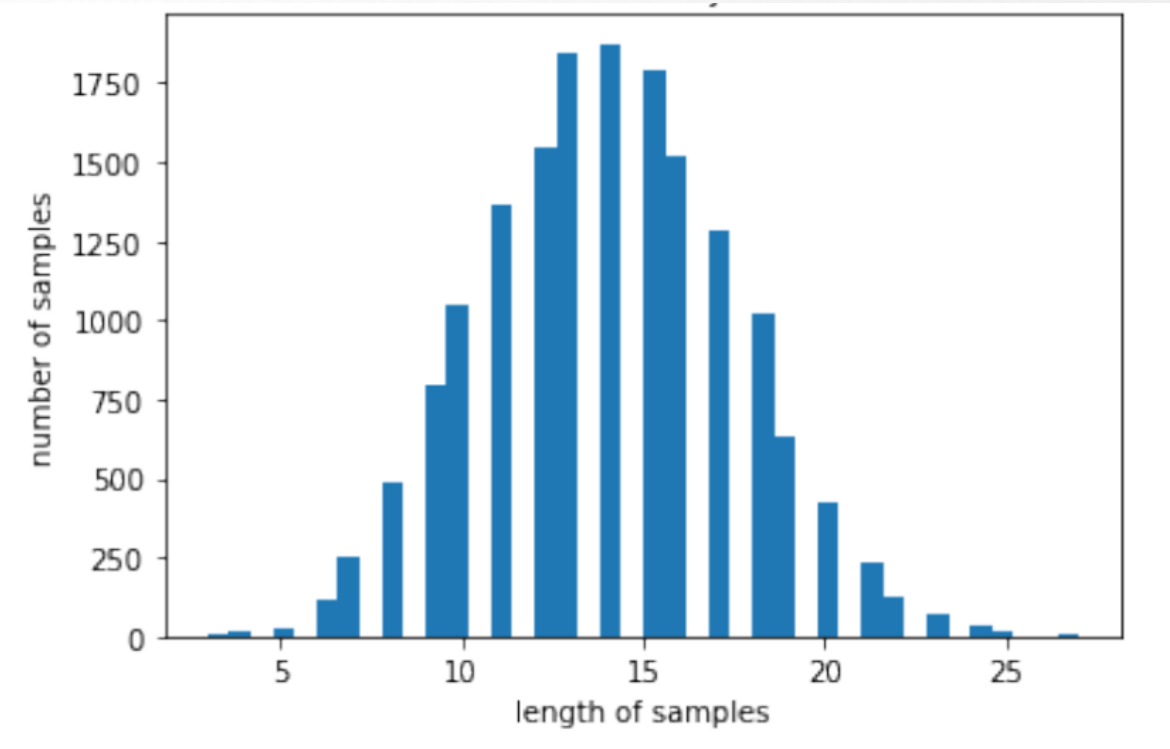
KONLPY의 토크나이저 비교



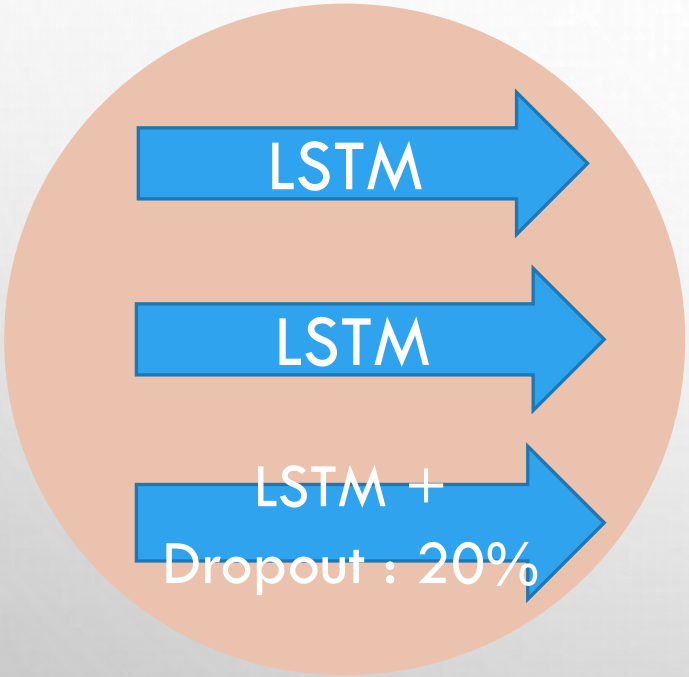
이미지 출처 : https://konlpy-ko.readthedocs.io/ko/v0.4.3/morph/

Kkma 사용 후 단어 분포 (padding : 1000/ 20)





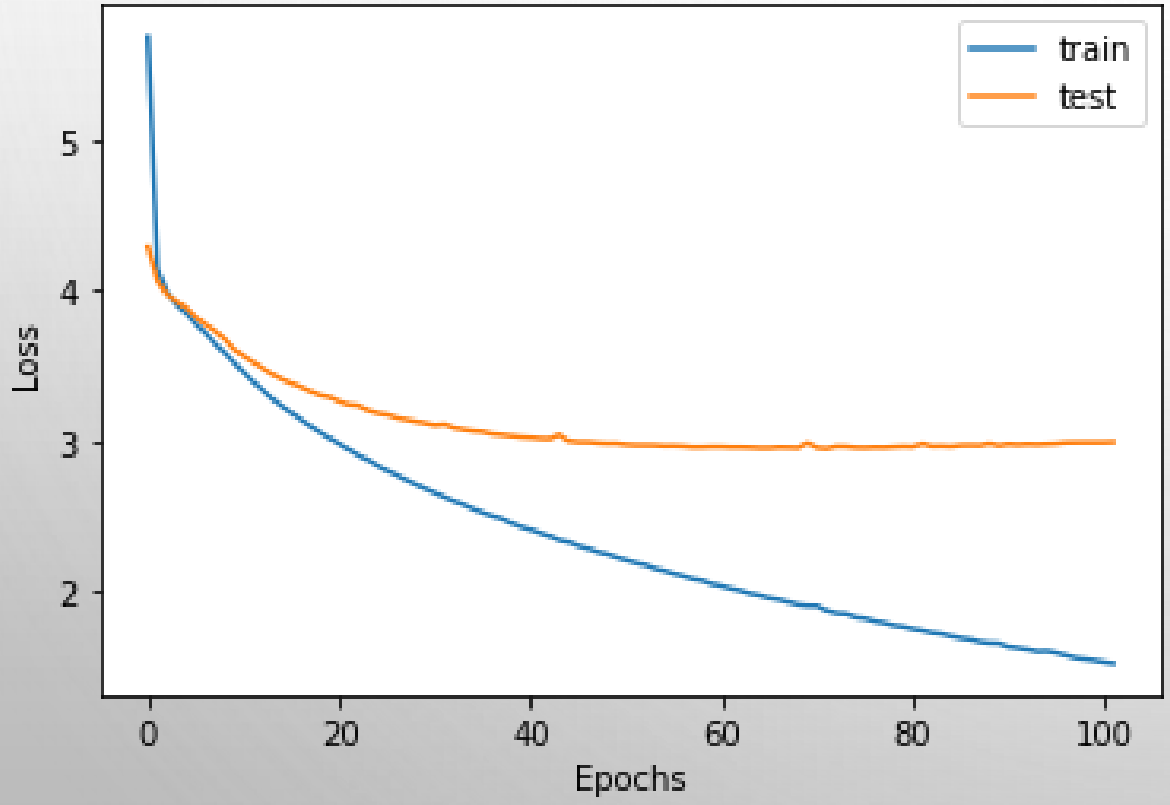
3. 사용모델

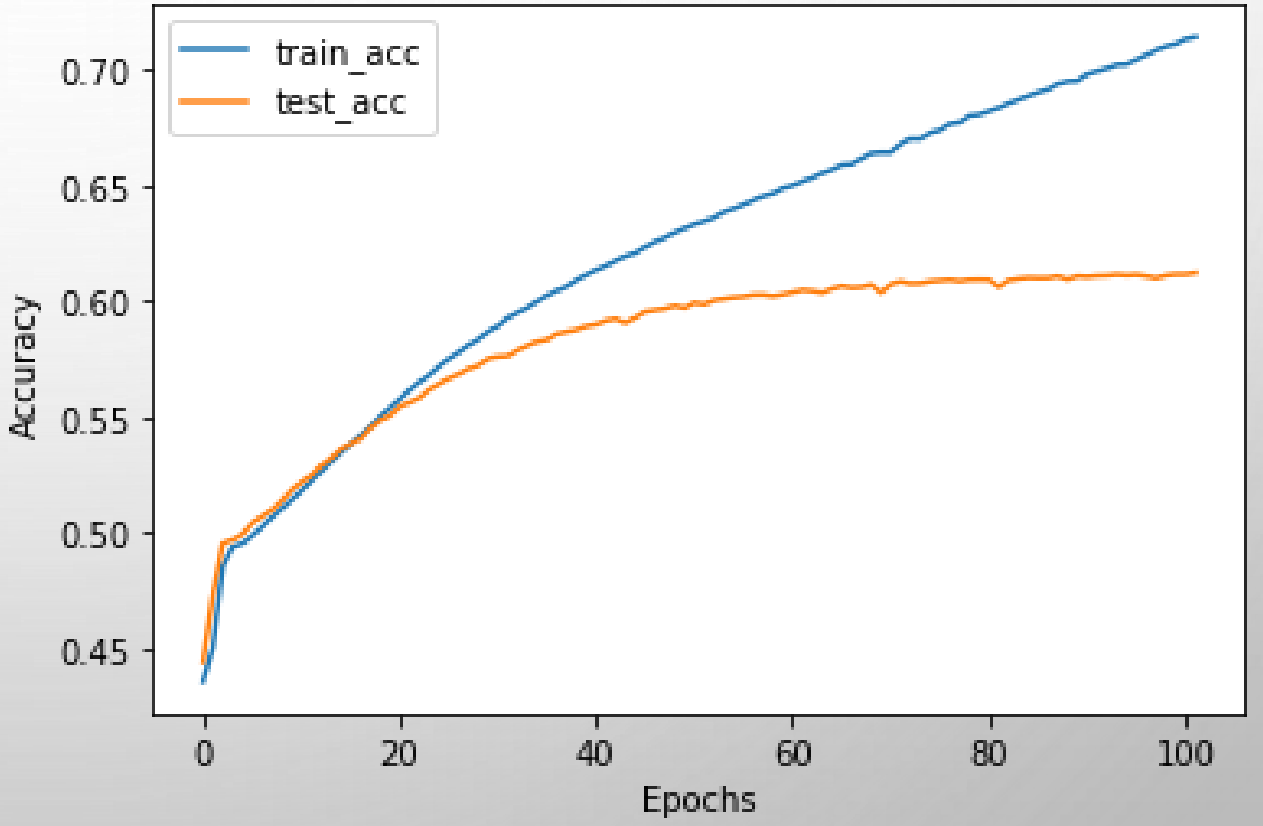


Seq2seq
3 - Layers

Attention

Train & Test 손실/ 정확도





4. 결과 – GOODCASE

Predicted title	label	contents
컴백 에이 핑크 정은지 7 월 컴백 확정 공식	에이핑크 정은지 7월 컴백 1년 9개월 만에 솔로 앨범 공식	에이핑크 정은지 7월 컴백 1년 9개월 만에 솔로 앨범 공식 다방면으로 활약하고 있는 에이핑크 정은지가 솔로 가수로 돌아온다 소속사 플레이엠엔터테인먼트 측은 23일 정은지가 오는 7월 중 솔로 앨범을 발표한다 며 자세한 일정과 음반의 형태는 추후 정리되는 대로 말씀드릴 예정.....
드림 캐쳐 컴백 첫 정규 앨범 콘셉트 포토 공개	드림캐쳐 새 콘셉트 티저 비비드 컬러 속 빛나는 비주얼	그룹 드림캐쳐의 콘셉트가 완성됐다 드림캐쳐컴퍼니는 지난 10일 오후 공식 어플리케이션과 채널에 드림캐쳐의 첫 정규앨범 의 네 번째 티저 이미지를 공개했다 드림캐쳐 멤버들은 버전으로 명명된 이번 티저...
광주 서 확 진자 접촉 감염 비상 광화문 집회 참석 종합	광주 성림침례교회 신도 등 32명 집단 감염 확진자 더 나올 듯 종합	광화문 집회 참석 후 확진된 신도 3차례 예배 참석야간 검사지난 25일 오후 광주 북구 각화동 성림침례교회 앞에서 교인 등을 검사하는 모습 연합뉴스 자료사진 광주 연합뉴스 손상원 기자 광화문 집회 관련 확진자가 다녀간 광주 한 교회에서 신도와 접촉자 등 30명을 웃도는 집단 감염이 발생했다 26일 광주시에 따르면 광주 북구 각화동 성림침례교회 신도.....
드림 캐쳐 컴백 첫 정규 앨범 티 저 공개	드림캐쳐 18일 컴백 확정 첫 정규 앨범 선보인다 공식	드림캐쳐 18일 컴백 확정 첫 정규 앨범 선보인다 공식 그룹 드림캐쳐가 새로운 이야기를 품고 돌아온다 드림캐쳐컴퍼니는 지난 3일 오후 6시 공식 어플리케이션과 채널에 첫 번째 정규 앨범 스케줄러를 공개하고 본격적인 컴백 카운트다운을 알렸다 스케줄러에 따르면 드림캐쳐는....
종합 김 원해 측 코로나 19 확 진 김 원해 측 자가 격리 중 공식 입장	서성종 코로나 확진 김원해 측 자가 격리 검사 결과 기다리는 중 공식입장	서성종 코로나 확진 김원해 측 자가격리 검사 결과 기다리는 중 공식입장 배우 서성종이 신종 코로나 바이러스 코로나19 확진 판정을 받았다 이가운데 오늘 19일 그와 함께 연극 짬뽕....

BADCASE

Predicted title	label	Contents (preprocessed)
종합 아이 돌 라디오 1 위 싹 쓰리 1 위 한 2 위 하 ㄴ 아이 돌 라디오 1 위	아는 형님 이진혁 젊은 강호동 같다는 말 들어 에너자이저 인증	이진혁이 젊은 강호동 같다는 말을 많이 들었다 라고 밝혔다 15일 방송되는 아는 형님 에 자타공인 연예계 절친 이준 정용화와 막 친해지고 있는 이진혁 정세운이 전학생으로 등장한다 연습생 시절부터 알고 지냈던 이준과 정용화의 오래된 추억부터 인연을 만들고 있는 이진혁과 정세운의 풋풋한 이야기까지 절친 두...
그것이 알고싶다 화성 초등생 실종사건 이춘재 심경고백	종합 김 주 ㄴ 김 성 현 측 김 민경 측 박지훈 측 성 과 성 추행	살인자의 자백 그리고 사라진 시신25일 방송되는 그것이 알고 싶다 에서는 이춘재를 직접 만난 화성 초등학생 실종 사건 피해자 가족을 통해 사건의 진실에 대해서 분석해본다 이춘재의 첫 심정 고백 제작진에 따르면 30년 전 실종된 막내딸을 살해한 것이 본인이라는 한 연쇄살인범의 고백 아버지가 지금껏 놓지 못했던 희망이 산산이 조각나는 순간이었다 막내딸을 죽인 살인범에게 꼭 들어야 할 말이 있다며....
공부가머니 신동엽 뛰어넘는 홍성흔子 입담 잔머리 폭소	놀 면 뭐하 니 싹 쓰리 결혼 결혼 발표 연예인 과 함께 하고프 ㄴ 인연	공부가머니 신동엽 뛰어넘는 홍성흔子 입담 잔머리 폭소 공부가 머니 에 신동엽을 능가하는 순발력 귀재가 등장한다 4일 방송될 공부가 머니 기획 박현석 프로듀서 선혜윤 에서는 홍성흔의 딸 화리와 달리 1분도 집중하지 못하는 아들 화철이의 해맑은 일상을 공개 모두를 폭소케....
악뮤 이수현 뮬란 가창 국내 공식 커버송 아티스트 공식	종합 원 드 어 나인 어 소속 사 아 전속 계약 만료 재 계약 공식 입장	악뮤 이수현 뮬란 가창 국내 공식 커버송 아티스트 공식 이수현이 영화 뮬란 을 불렀다 지난 14일 금 국내 공식 커버송 아티스트의 보이스 실루엣을 공개하며 폭발적인 궁금증을 불러일으켰던 영화 뮬란 이 드디어 오늘.....

5. 프로젝트 고찰

- 정확도 개선

초기 방향 : 정확도 개선

-- Split을 이용한 토큰화 – 정확도 42% (데이터 5000개) – 한글은 조사들이 붙어 있으므로 split으로는 구분하는 것이 어렵다.

-- Konlpy: 한나눔 모듈 명사추출 – 51%(데이터 5000개)

-- Konlpy: komoran 모듈 형태소 추출 – 54% (데이터 10000여 개)

-- Konlpy: kkma 모듈 형태소 추출 – 70%(데이터 10000여 개) – 정확도만 높고 결과가 좋지 않았다.

-- Konlpy: kkma 모듈 형태소 추출 – 62% (데이터 18000여 개) + 3 layers + dropout

- 데이터의 종류

- 동아스포츠의 연예, 스포츠, 포토, 아이돌 픽

- 중앙일보 실시간 뉴스

- 네이버 속보 등

데이터의 반 이상은 동아 스포츠였기 때문에 연예, 스포츠, 코로나관련 뉴스는 예측이 비슷하게 나온 것을 확인할 수 있었지만 훈련하지 않은 데이터에 대해서는 결과를 신뢰하기 어려웠다.

- 훈련 시간

데이터 양과 레이어 수가 많아질 수록 훈련시간이 증가

드롭아웃을 적용하면 훈련시간이 증가

데이터가 많을 수록 KKMA에서 토큰화 하는 시간이 증가

- 개선할 부분

1. 데이터 양을 늘리고 기사의 토픽을 더 다양하게 하여 다른 토픽들도 예측이 잘 수행되도록 한다.
2. SEQ2SEQ 레이어에서 드롭아웃 비율을 더 증가시키고 3개 레이어에서 모두 드롭아웃을 실행한다.

감사합니다.