I. Project 목표

인터넷 뉴스의 텍스트를 데이터로 사용하여 해당 텍스트가 어떤 주제에 속하는지 분류한다.

더불어 '워드 클라우드라드' 라는 시각화 기능을 추가한다. 그로 인해 단어의 빈도수를 대략적으로 확인하고 무엇이 중요한 단어인지 단번에 확인할 수 있게끔 한다.

II. Project 범위

카테고리 별 네이버 뉴스 (정치, 경제, 사회, 생활/문화, IT/과학)

III. 타깃 업무 설명

Column B: 기사 카테고리 – 옳고 그름을 판단하는 기준이 된다.

Column E: 기사 본문 – 본문의 단어 형태소 하나하나를 쪼개고 분석한다.

IV. 분석할 대상 (데이터), 분석의 방향 설정

1) Training data : 2019/12/01 ~ 2019/12/31 기간의 네이버 뉴스 (약 6만개)

   뉴스들을 주제별로 수집 -> 학습 진행

2) Test data : 2020/12/01 ~ 2020/12/31 기간의 네이버 뉴스 (약 1만개 - Training Data와 분리)

   특정 텍스트를 제공했을 때 해당 텍스트가 어떤 주제에 속하는지 분류 확인

3) 분석의 방향 : 웹 사이트 네이버에서 뉴스를 카테고리별로 웹 크롤링을 통해서 가져온다. 가져온 데이터에서는 카테고리 이름과 본문 내용으로 총 2개를 활용한다. 그 후 본문의 유사한 단어들을 비슷한 방향과 힘의 벡터를 갖도록 변환하여 사용한다. (Word2Vec 사용) 이렇게 한 뉴스마다 생성된 단어 리스트들로 미리 만들어 둔 모델과 연관 지어 기계학습을 진행한다. 기계학습이 모두 완료되었다면 평가를 한 후, 추가로 '워드클라우드'를 제작한다.

V. 진행절차

1) 데이터 가져오기

```
In [2]:
#https://github.com/lumyjuwon/KoreaNewsCrawler
#training data
from korea_news_crawler.articlecrawler import ArticleCrawler

Crawler = ArticleCrawler()
Crawler.set_category("정치", "경제", "사회","생활문화","세계","IT과학")
Crawler.set_date_range(2019, 6, 2019, 12)
Crawler.start()

{'start_year': 2019, 'start_month': 6, 'end_year': 2019, 'end_month': 12}
```

<그림 1 - 뉴스 크롤링을 위한 라이브러리 코드>

	A	B	C	D	E
3575	20191205	IT과학	연합뉴스	이희용의 ‘	바다에서 수거한 플라스틱 쓰레기로 만든 고래 조각상 이 헝가리 부다페스트의 국회의사당 앞
3576	20191205	IT과학	뉴시스	폴더블폰 ‘	갤럭시폴드 메이트X 출시로 관련 부품주 주목 내년 폴더블폰 시장 후발주자 참여로 확대 예상
3577	20191205	IT과학	아시아경제	물에 잠기는	일본의 간사이공항은 대표적인 인공섬입니다. 이제는 이런 바다를 매립해 만드는 인공섬이 아
3578	20191205	IT과학	아이뉴스24	청와대 행정	이스트시큐리티 악성파일 분석 통해 김수키 조직 확신 아이뉴스24 최은정 기자 청와대 관련
3579	20191205	IT과학	ZDNet Kor	산업에 인공	국가기술표준원 2019 AI 산업 표준화 워크숍 개최 지디넷코리아 주문정 기자 산·학·연 인공지
3580	20191205	IT과학	연합뉴스	칼 아이컨	HP 주주들에 공개서한…제록스와 한 배 타고 주주 공략 나선 듯 행동주의 투자자 칼 아이컨.
3581	20191205	IT과학	매일경제	亞 최다 지	국내 연구진이 주도한 아시아인 최다 유전체 분석 연구 성과가 국제 학술지 네이처 Nature 최
3582	20191205	IT과학	세계일보	나노 물결무	카이스트 이성빈 교수 연구팀 박문집 연구원 왼쪽 이성빈 교수 국내 연구진이 나노 물결무늬
3583	20191205	IT과학	세계일보	2020년부터	내년 1분기부터 점진적 확대 한국 대만 태국 일본 등 국경을 넘나드는 결제 경험 제공 ‘모바일
3584	20191205	IT과학	이데일리	현지 가정식	에어비앤비 제공 이데일리 한광범 기자 에어비앤비가 식탁에서 현지 문화를 경험하는 ‘쿠킹
3585	20191205	IT과학	전자신문	리인벤트19	한글과컴퓨터가 아마존웹서비스 AWS 를 통해 세계 소프트웨어 SW 시장을 공략한다. 클라우
3586	20191205	IT과학	세계일보	카카오엔터	카카오의 사내 독립기업 CIC AI LAB 분사 카카오엔터프라이즈가 3일 공식 출범했다. 카카오(
3587	20191205	IT과학	ZDNet Kor	동서발전 김	고 김용균 사고 발생 1주기…현장 설비보강 및 안전 점검 지디넷코리아 주문정 기자 한국동서
3588	20191205	IT과학	전자신문	강원도 정선	강원도는 오는 19일 춘천 더존비즈온 강촌캠퍼스에서 강원도와 함께 꿈꾸는 빅데이터 세상
3589	20191205	IT과학	서울신문	과학계는 기	서울신문 알츠하이머 이미지 픽사베이 제공 미국 웨스트버지니아대 의대 산하 록펠러 신경과
3590	20191205	IT과학	디지털데일	퀄컴서밋 개	공정위 ‘명분’ 퀄컴 ‘실리’ 챙겨…대법원 판단 ‘관심’ 디지털데일리 윤상호기자 우리나라 법원이
3591	20191205	IT과학	국민일보	삼성·LG 中	글로벌 스마트폰 업체들이 일본 시장 공략을 위한 채비에 분주하다. 내년 도쿄올림픽을 앞두
3592	20191205	IT과학	한국일보	퇴직 앞둔	지난달 26일 경기 용인시에 있는 KT SAT의 위성 관제실에서 이인호 가운데 위성관제팀 부장
3593	20191205	IT과학	머니S	출근길법원	사진 뉴시스 세계 최대 무선통신칩 제조사 퀄컴이 공정거래위원회 공정위 의 1조원대 과징금
3594	20191205	IT과학	머니투데이	애플도 참가	머니투데이 박효주 기자 애플 5G 아이폰 4종 출시 예상…삼성 갤럭시A 시리즈까지 5G 확대

<그림 2 - 크롤링된 데이터들 >

2) 데이터 전처리 (자연어 처리 - 사용된 단어의 목록 작성)

```
In [ ]:
# 코드 통합
import csv
import os

os.chdir("C:₩₩Users₩₩Yoo Jae Un₩₩데이터사이언스_실습₩₩Csv") // Csv가 있는 파일 위치 수정

category = ['IT과학', '경제','사회', '생활문화','세계','정치']

body_unity = open('body_unity.csv', 'w', encoding='UTF-8')
wcsv = csv.writer(body_unity)

count = 0

for category_element in category:
    file = open('Article_'+category_element+'.csv', 'r', encoding='UTF-8', newline="")
    line = csv.reader(file)
    try:
        for line_text in line:
            wcsv.writerow([line_text[1], line_text[4]])
    except:
        pass
```

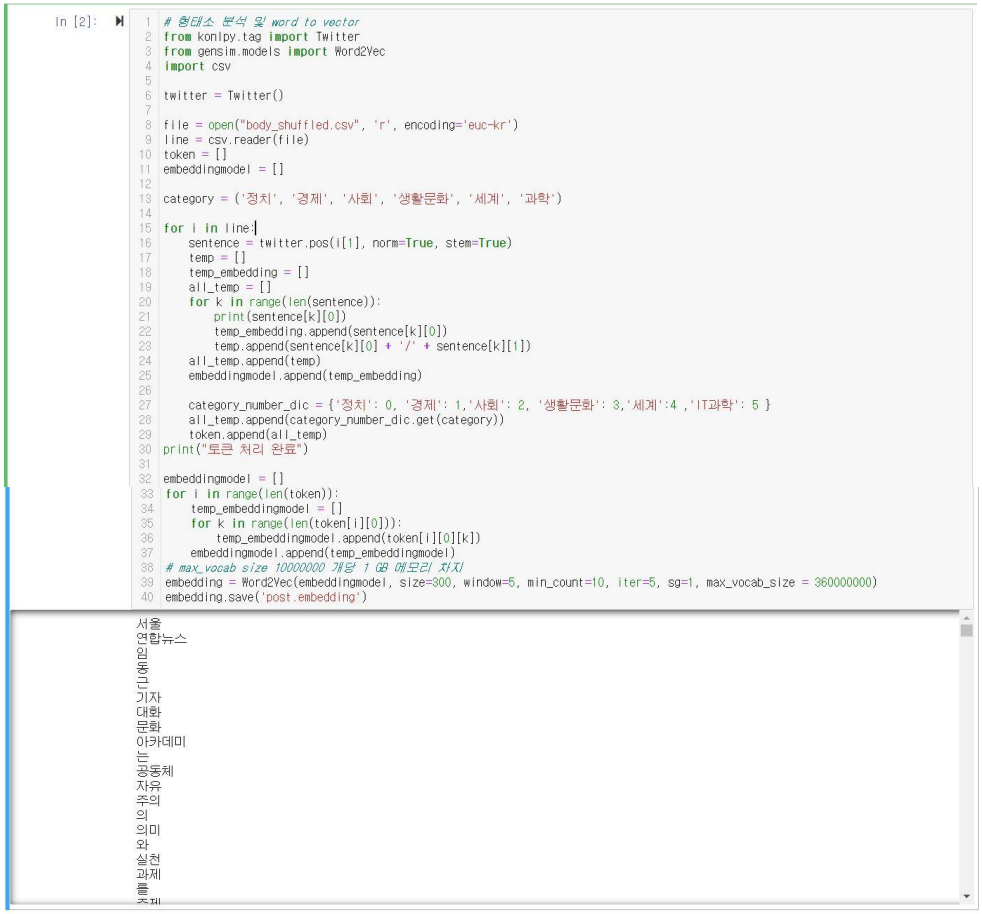
< 그림 3 - 카테고리별로 나눠진 csv파일을 통합 >

Csv 파일들을 통합 할 때, 프로젝트의 목표상 Column B(기사 카테고리), Column E(기사 본문) 만 필요하므로 두 column을 제외한 모든 것들은 삭제한다.

```
In [*]:
# 코드 셔플
import csv
import random
import os

os.chdir("C:₩₩Users₩₩Yoo Jae Un₩₩데이터사이언스_실습₩₩") # Csv가 있는 경로 설정

file = open('test_data.csv', 'r', encoding='euc-kr')
line = file.readlines()
random.shuffle(line)
rcsv = csv.reader(line)

body_write = open('test_data_shuffled.csv', 'w', encoding='euc-kr', newline="")
wcsv = csv.writer(body_write)

for i in rcsv:
    try:
        wcsv.writerow([i[1].strip(), i[4]])
    except:
        pass
```

<그림 4 - 데이터 셔플 >

카테고리별 연속으로 학습을 시킬 경우, 최적의 W를 특정 카테고리에서만 발견하게 되고 다음 카테고리 데이터에서는 별 효율이 없는 학습을 진행하게 된다. 따라서 셔플을 진행한다.



```
In [2]:
# 형태소 분석 및 word to vector
from konlpy.tag import Twitter
from gensim.models import Word2Vec
import csv

twitter = Twitter()

file = open("body_shuffled.csv", 'r', encoding='euc-kr')
line = csv.reader(file)
token = []
embeddingmodel = []

category = ('정치', '경제', '사회', '생활문화', '세계', '과학')

for i in line:
    sentence = twitter.pos(i[1], norm=True, stem=True)
    temp = []
    temp_embedding = []
    all_temp = []
    for k in range(len(sentence)):
        print(sentence[k][0])
        temp_embedding.append(sentence[k][0])
        temp.append(sentence[k][0] + '/' + sentence[k][1])
    all_temp.append(temp)
    embeddingmodel.append(temp_embedding)

    category_number_dic = {'정치': 0, '경제': 1,'사회': 2, '생활문화': 3,'세계':4 ,'IT과학': 5 }
    all_temp.append(category_number_dic.get(category))
    token.append(all_temp)
print("토큰 처리 완료")

embeddingmodel = []
for i in range(len(token)):
    temp_embeddingmodel = []
    for k in range(len(token[i][0])):
        temp_embeddingmodel.append(token[i][0][k])
    embeddingmodel.append(temp_embeddingmodel)
# max_vocab size 10000000 개당 1 GB 메모리 차지
embedding = Word2Vec(embeddingmodel, size=300, window=5, min_count=10, iter=5, sg=1, max_vocab_size = 360000000)
embedding.save('post.embedding')
```

```
서울
연합뉴스
임
동
근
기자
대화
문화
아카데미
는
공동체
자유
주의
의
의미
와
실천
과제
를
주제
```

<그림 5 - 형태소 분석 및 단어 vector로 전환>

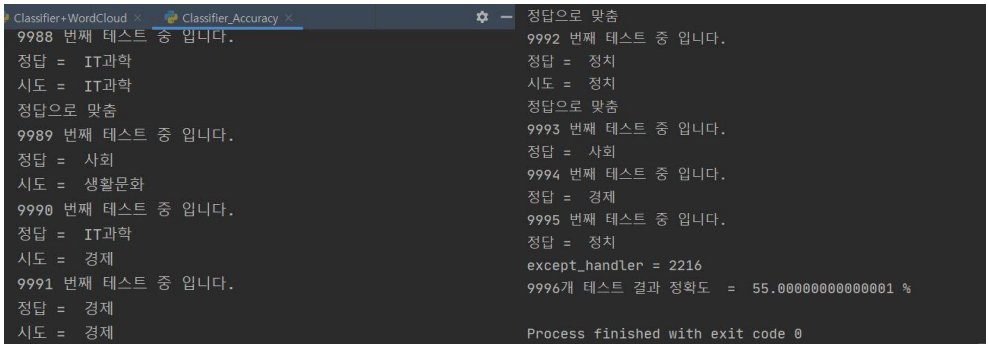
3) 머신러닝 학습 (Bi_LSTM 사용, tensorflow 버전 -> 1.15.3 사용)

Word2Vec에서 embedding된 자료를 기반으로 학습 (Bi_LSTM_csv_train.py)

4) 결과

4-1) 파일 이름 : Classifier_Accuracy.py

그림 6은 정확도 분석을 위해 testdata의 기사 본문을 input 데이터로, 태그는 결과비교로 하였다.



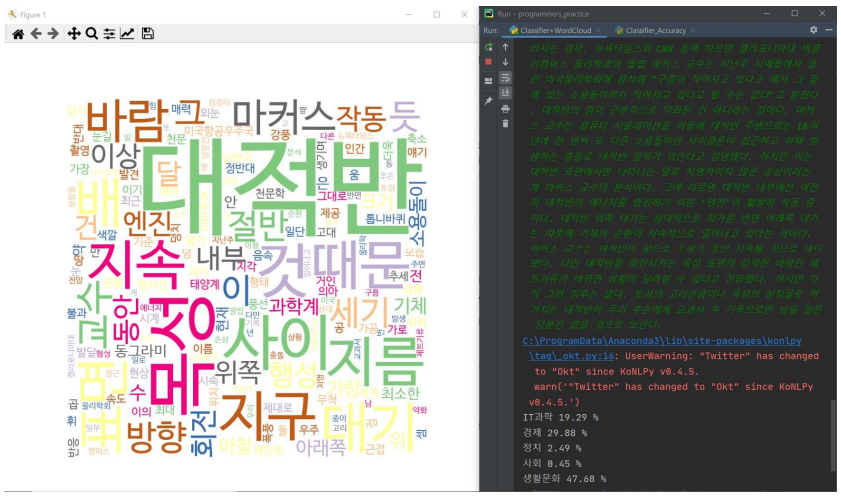
< 그림 6 >

그림 6에서 볼 수 있듯이 정확도가 55%를 기록했다.

(문장 분석 오류로 인해 실행이 되지 않는 2216개 제외)

4-1) 파일 이름 : Classifier+WordCloud.py

그림 7은 기존 태그를 유추하는 코드에 워드 클라우드를 띄워주게 하는 코드를 추가하였다.



< 그림 7 – 태그 유추와 워드 클라우드 >

VI. 평가

처음에 나만의 방식대로 코드를 작성하고 프로그램을 실행시켜 본 결과, 일정 수치만 들어가도 카테고리가 경제로 인식되는 경우가 많았다. 즉 기사의 내용만 보고 카테고리를 맞추는 정확도가 낮았다. 그 후 인터넷 검색을 통하여 내가 하는것과 비슷한 코드를 찾았고 실행을 하며 비교해보았다.

왜 정확도가 낮았는지 생각해본 결과는 아래와 같다.

1) 뉴스 원본을 봤을 때, 카테고리와 뉴스의 내용이 맞지 않는 것들이 있었다.

   - 내용은 정치에 대해 다루고 있는데, 카테고리가 IT로 되어 있음

2) 학습 데이터가 부족했다.

   - 6만개를 학습시켰을 때의 정확도 : 20% (나의 코드)

   (이후 50만개 학습데이터를 덮어 씌워서 캡처하지 못했습니다.)

   - 50만개를 학습시켰을 때의 정확도 : 55% (다른 사람의 코드)

   - 데이터를 학습시키는데 많은 시간이 소요 (6만개 -> 10시간 소요)

VII. 개선해야 할 점

1) 데이터 전처리 과정에서 특수문자들을 모두 지웠다고 생각했으나 '~' 등이 지워지지 않았음

   -> 웹 크롤링시에 추가적인 코드 작성 필요

2) 양질의 데이터 학습

   -> 양적으로 많고, 질적으로 옳은 데이터들을 기반으로 학습시켜야 함

[참고 사이트]

강의 07 네이버 뉴스 카테고리 예측 모델 전이 학습

https://wikidocs.net/75397

[Machine Learning] Word2Vec 소개 및 실습

https://m.blog.naver.com/PostView.nhn?blogId=wideeyed&logNo=221349385092&proxyReferer=https%2F%2Fwww.google.com%2F

머신러닝(TensorFlow)을 이용한 게시글 및 댓글 카테고리 분류

https://hyrama.com/?p=488

[Python] 텍스트 분류하기 (Text Classification)

http://arkainoh.blogspot.com/2017/09/python.text.classification.html

Python 한글워드클라우드 만들기

https://business-analytics.tistory.com/3