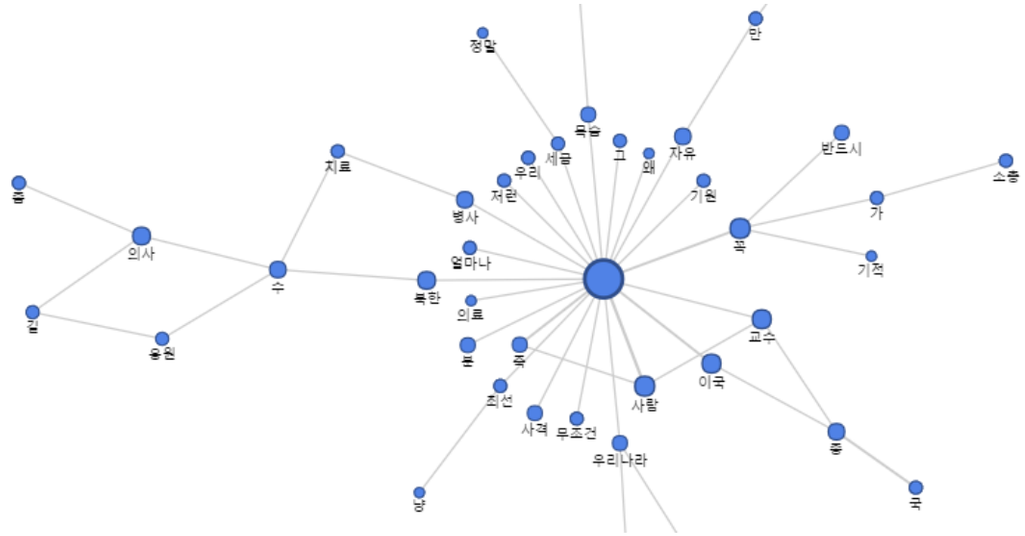
5) Python 인터넷 기사 댓글 크롤링 및 분석 프로젝트

Programming Project Portfolio

Python 인터넷 기사 댓글 크롤링 및 분석 프로젝트



프로젝트 설명

수행 기간 : 2017.10 ~ 2017.11

내용 :

Python을 이용해 **인터넷 기사 URL을 받아와 댓글 내용들을 크롤링**하여 **대중의 주요 의견, 트렌드들을 손쉽게 파악**이 가능한 툴

프로그래밍 단계에서는 인터넷 기사의 html 댓글의 태그 속성으로 **검색해 댓글 데이터를 얻어와 형태소로 분리, 명사 빈도수 체크**

구글의 spread-sheet, fusion table 툴을 통해 데이터 시각화

Git Repository :

https://github.com/JisooJang/python_lab

Programming Project Portfolio

Python 인터넷 기사 댓글 크롤링 및 분석 프로젝트

1) 웹 기사 페이지에서 댓글 데이터를 읽어와 리턴하는 함수 정의

def getCommentFromURL(*url*, *tagName*, *attrType*, *attrName*, *moreButton*) :

매개변수 : 기사 URL, 댓글의 html태그 종류, 댓글 태그 속성, 댓글 태그 속성 명, 댓글 더보기 버튼 태그 속성명

1. 웹에서 **비동기식으로 로드되는 데이터**를 가져오기위해 **webdriver의 PhantomJS()**를 통해 브라우저를 킴
2. 매개변수 URL로 브라우저를 이동하고, 비동기 데이터가 모두 로딩되기 위해 time.sleep메소드를 통해 기다림

3. 댓글에 더보기 버튼이 있을시, 버튼 태그 속성명으로 버튼 요소를 가져와 버튼이 존재하지 않을 때까지 계속 더보기 버튼을 클릭

4. **webdriver**객체를 **BeautifulSoup**와 연계사용하여 편리하게 html 데이터를 읽어올 수 있음
   page.find_all(*tagName*, attrs={*attrType*: *attrName*}) 메소드를 통해 태그이름, 태그 속성값으로 태그 내 데이터를 가져옴

```
# test case =〉    temp = page.find_all('span', attrs={'span': 'u_cbox_contents'})
```

Python 인터넷 기사 댓글 크롤링 및 분석 프로젝트

2) 문자열을 받아 단어를 추출하여, 각 단어와 빈도수를 체크하여 리턴하는 함수

```
def getKeyword(text, ntags=50):
```

구현 방법

1. konlpy의 Twitter객체를 생성하여 nouns(text) 함수를 통해 text에서 **명사만 추출**하여 저장
2. Counter **객체**에 위 명사만 추출한 데이터를 매개변수로 생성하여 **빈도수를 체크 가능하게 저장** Counter(nouns)
3. 반복문 내에서 **Counter 패키지의 most_commons(ntags)** 함수를 통해 **명사중 빈도수가 큰 명사부터** 순서대로 입력받은 정수 갯수만큼 저장되어있는 객체를 반환시켜 리턴

```
for n, c in count.most_common(ntags):
    temp = {'tag' : n, 'count' : c }    # tag : 단어, count : 빈도 수
    return_list.append(temp)
```

ex) [{'tag': '꼭', 'count': 70}, {'tag': '교수', 'count': 58}, {'tag': '자유', 'count': 36}, {'tag': '이국', 'count': 35}….]

Programming Project Portfolio

Python 인터넷 기사 댓글 크롤링 및 분석 프로젝트 - 실행

```
C:₩Users₩user₩AppData₩Local₩Programs₩Python₩Python36₩python.exe C:/Users/user/PycharmProjects/untitled/test3.py
2,084
<selenium.webdriver.remote.webelement.WebElement (session="47f14540-d01e-11e7-9d36-e588ce6fcff8", element=":wdc:1511421422867")>
젊은 사람이 죽을 각오하고 사선을 넘어 왔는데 꼭 살았으면 좋겠습니다
이건 아직 살아있는것도 기적이고 살려내는것도 기적이다..
교수님도 부담스럽겠습니다..그냥 최선만 다해주시면됩니다귀순 군인 힘내세요!
관통상만 3군데 장기조직 오염도 심하고 살려내는게 기적중에 기적이다
꼭 살려주세요
이런 것까지 이념 논쟁화 하는게 안타깝네.... 죽기를 각오하고 넘어온 사람을 무조건 살려야 하는거 아닌감? 인간으로서 가져야 하는 기본적인 마인드까지 이념 논쟁으로 더럽히지 맙시다.....
어렵게 넘어왔는데 꼭살아서 대한민국에 같은국민으로써 자유를 누리길.....
응원합니다. 누구는 목숨을 걸어가며 자유를 위해 오는데 우리는 자유의 의미와 기쁨을 망각한채 불평만 하지 않은지 돌아봐야할듯 합니다.
이분이 아덴만에서 납치되어 총상입은 선장도 살려냈었지. 어렵지만 이국종교수는 해낼것같다. 그가 대한민국국민인게 다행이고 자랑이다.
잘견뎌서 꼭 살았음좋겠어요. .
북한사람 응호하는 건 처음인데 저 북한군인은 꼭 살았으면 좋겠다. 이대로 그냥 가 버리면 마음 아플거 같다. 왜 북한에서 태어났니?. 꼭 기적적으로 살아서 아무도 없지만 한국에서 새로운 삶을
이국종 교수 카리스미 넘 멋지다 어찌됐든 한국의 외과1인자에게 수술받는 행운아네
탈북한 의지면 살수있다.어떻게 해서 탈출했는데 죽으면 억울하지 않겠냐...꼭 살아남아라.
내가 살고 있는 이 나라는 저 청년이 빗발치는 총탄을 맞고 죽을 각오로 찾아온 자유의 나라이었구나! 열심히 살아야겠다! 그리고 그 자유와 평등이 훼손되지 않도록 잘 보존해야겠다
귀순병사 운 좋네. 한국 최고로 꼽히는 의사한테 맡겨져서
외국논문 공부하며 환자 수술중 아마 밤에 잠도 못주무시고 공부하시는듯 대단하시다 그 위치에서도 끊임없이 연구하는 정신이 ㄷㄷㄷ
이런분들께 국가적지원좀 해줘라 환자살리려는 사명감하나로 여기까지오신분
목숨 걸고 왔는데.. 살려야지.도망가는 뒷사람 에게 총질이나 해대니..
꼭 건강한 모습으로 볼수있길 바랍니다~이국종 박사님 항상 감사드립니다.
이젠 민간병원이  아닌  군 병원서   수술이  가능한  집도의도  있기를
그대로 짚차로 질주해서 넘어오려다 배수로에 바퀴가 빠져 못움직인거지. 계획과 달리 돌발상황에서 내려서 달리기 시작.
이국종교수 신변보호해줘라!!
우리가 친구들과 카페에서 하하호호 수다도 떨고, 취업은 언제 할 수 있을까 걱정하는 동안 또 다른 우리의 또래는 우리에게는 너무나도 흔한, 그래서 소중한 것을 잘 알지 못하는 '자유'를 위해
이국종교수 수술팀은 항상 적자에 시달린다... 이런건 정부에서 지원해줘야한다...안그러면 중증외상센터를 누가 운영하려고 하겠는가??
꼭 살아나면 좋겠음 ㅜㅜ 근데 냉정하게 개인으로만 생각하면 살아나도 문제긴 하고...저렇게 큰 수술을 여러번 받고 했는데 휴유증이 없는건 욕심이고 분명 휴유증과 계속되는 병원치료가 필요할
경계선 하나 그어놓고 이게 무슨 짓인가.. 사람목숨이 파리 목숨보다 못하네..에구..
 "특히 한국인에게서 발견하기 어려운 이상 소견이 있어 미국 논문을 연구하며 치료하고 있다"  <<이국종교수님 고생많으십니다. 감사합니다
40발 씀????/////////실화냐?북한 진짜 미친거 아냐?김정은 이 새끼 용서안되 우리나라 건드리면 넌 뒤져
나이도 어린데 죽을 각오로 넘어왔는데 꼭 살아라! 살아서 넘어 온 이유를 말해줘! 나라면 너처럼 죽을 각오도 못했을것 같다.. 진짜 대단하다.
```

Programming Project Portfolio

Python 인터넷 기사 댓글 크롤링 및 분석 프로젝트 - 실행

```
지금까지 살아있는것도 기적이다..권총 과 AK소총을 다섯발 시상 맞고 복부를 관통 내장도 다 터졌다 ㅠ 제발 살아나라  너무 불쌍하다ㅠ
남한군인들도 제대로된치료와 보상이있기를
빨리 회복해서,  나중에 얘기 좀 듣고싶다
자유 대한민국으로 오신걸 환영합니다. 꼭 살아나셔서 북한의 악한 실체에 대해 알려주시고 우리나라에서 북한 추종하는 세력들에게 따끔한 한마디 부탁드립니다. 우리나라 군대도 너무하네요. 자
우리 군은 뭐했냐? 그지경 될때까지 ..... 엄호사격 했으면 저정도는 안되었을듯...맘대로 총질하니 배가 벌집이 되지...국방장관 해당 부대 사단장 모두 물러나라
개복상태로 ㅠ 사경을 헤메고 있다고 생각하니 마음이 아프네요
꼭 회복되기를 간절히 기원합니다...
이거 치료비 또 이국종 교수님이 내시는거 아니냐??
반드시 살려내서 저 병사에게 총질한 추격조넘들 싹다 아오지로 갔으면 좋겠다
목숨걸고 자유를찾아 남한으로 월남한 병사입니다반드시살려야합니다  그리고 파이팅!!하십시오!!
목숨걸고왔는데 살아나야지요 .
오느라 고생했는데 여기서 행복하게 살아야지. 힘내라
꼭살아라!!
정말 꼭 살아서  영원히 자유의  품에 살길 바랍니다.
참 의사 이신분
교수님 감사드립니다. 귀순한 병사 분도 힘내주세요.
잘 회복해서 갖고 있던 한 풀고 잘 살기 바란다.
공동구역에서 총질당해도 직접적 피해가 없어 대등사격 안했다고? 그 좁은 구역도 경계를 살피지 않고 40발이 넘는 총탄이 날라와도 몰랐다는 그딴 군정신으로 38선을 어떻게 지키고 있는지 안봐도
역시 우리와 북한은 좀 다르긴하네... 우리나라에서 저렇게 입북했다면 .. 그냥 불구경하듯하고 있었을듯... 그나 저나... 청기와집에 주사파놈들좀 똘똘 말아서 보내버리고 싶다... ㅡㅡ^
의료진이 브리핑보다는 좀 쉬어야할듯
대응사격 엄호사격 못한 부대장들 얼마나 답답했을까.......문재인과 주체사상파 임종석 한테 징계 먹을까봐 사격 명령을 못한게 분명하다.....
이국종 교수님 진짜 좋아하는데멋지세요^^
살려고 남으로 왔는데 꼭 살아나길
이국종 교수 돈 안되는 수술만 한다고 병원 내에서 아웃사이더 취급 받는다는데 꼭 수술 성공하길 빕니다
의료진 환자 힘내십시오
재인이가  말같지도 않은 평화를 내세워  돌려보낼까  걱정된다

[{'tag': '꼭', 'count': 70}, {'tag': '교수', 'count': 58}, {'tag': '자유', 'count': 36}, {'tag': '이국', 'count': 35}, {'tag': '종', 'count': 34}, {'tag': '목숨', 'count': 29}, {'tag

Process finished with exit code 0
```

Programming Project Portfolio

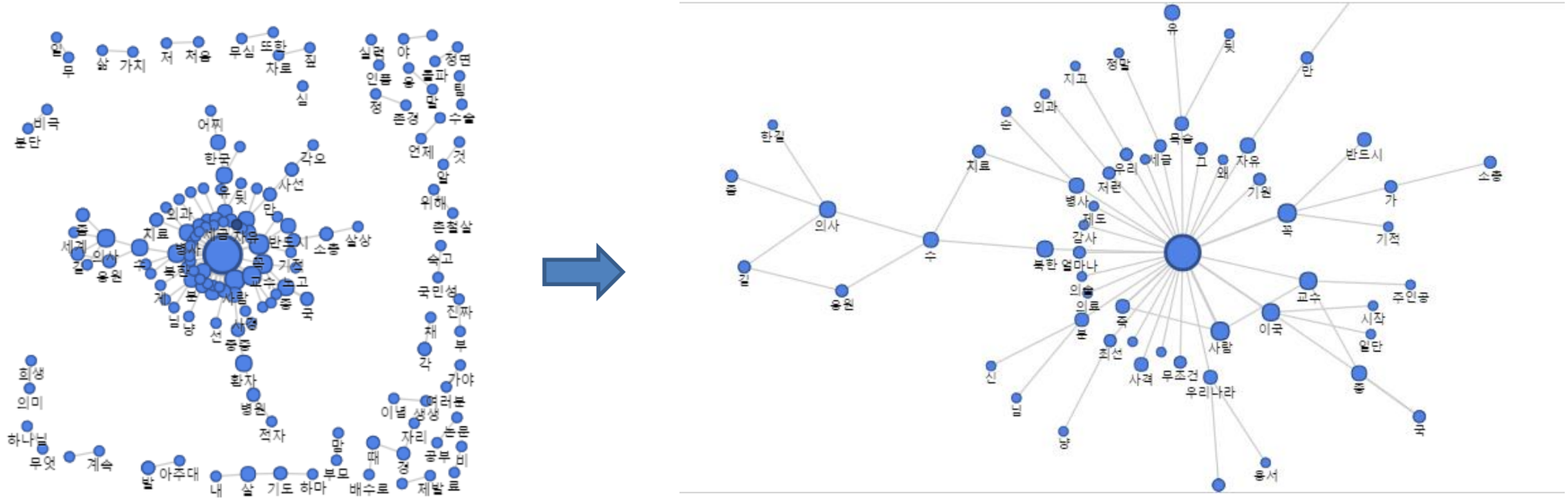
Python 인터넷 기사 댓글 크롤링 및 분석 프로젝트 - 데이터 시각화

내	살	이	나라	저	청년	빗발	총탄	죽	각오	자유	그
사람	죽	각오	사선	꼭	이건	아직	기적	기적	교수	그	
냥	최선	귀순	군인	관통상	군데	장기	조직	오염	기적		
기적	꼭	것	이념	논쟁	각오	사람	무조건	감	인간	기본	
	마인드	이념	논쟁	맙시	꼭	대한민국	국민	자유	누리		
길	응원	누구	목숨	자유	위해	우리	자유	의미	기쁨	망	
각	채	불평	이분	아덴만	납치	총상	선장	이국	종	교수	
	그	대한민국	국민	다행	자랑	꼭	북한	사람	응호	건	
처음	저	북한	군인	꼭	그냥	마음	왜	북한	꼭	기적	
무도	한국	삶	살	아직	사람	이국	종	교수	카	리스	
어찌	한국	외과	인자	수술	행운	탈북	의지	살수	꼭		
내	살	이	나라	저	청년	빗발	총탄	죽	각오	자유	그
	자유	평등	훼손	귀순	병사	운	한국	최고	의사	외국	
논문	공부	환자	수술	아마	밤	잠도	공부	그	위치	정	
신	분	국가	지원	좀	환자	사명	감	하나로	여기	오신	
목숨	뒷	사람	총질	해대	꼭	모습	이국	종	박사	항상	

1. 결과로 얻은 명사들을 google spread-sheet에 저장

Programming Project Portfolio

Python 인터넷 기사 댓글 크롤링 및 분석 프로젝트 - 데이터 시각화



2. 1번의 data-sheet를 이용하여 구글의 fusion-table chart를 제작

전체 데이터 차트에서 단어 노드 수를 조절하고 확대하여
특정 기사에 대해 대략적인 대중의 의견들을 한눈에 파악할 수 있음