워드클라우드

이영석

lee@cnu.ac.kr

http://yslee.cs-cnu.org

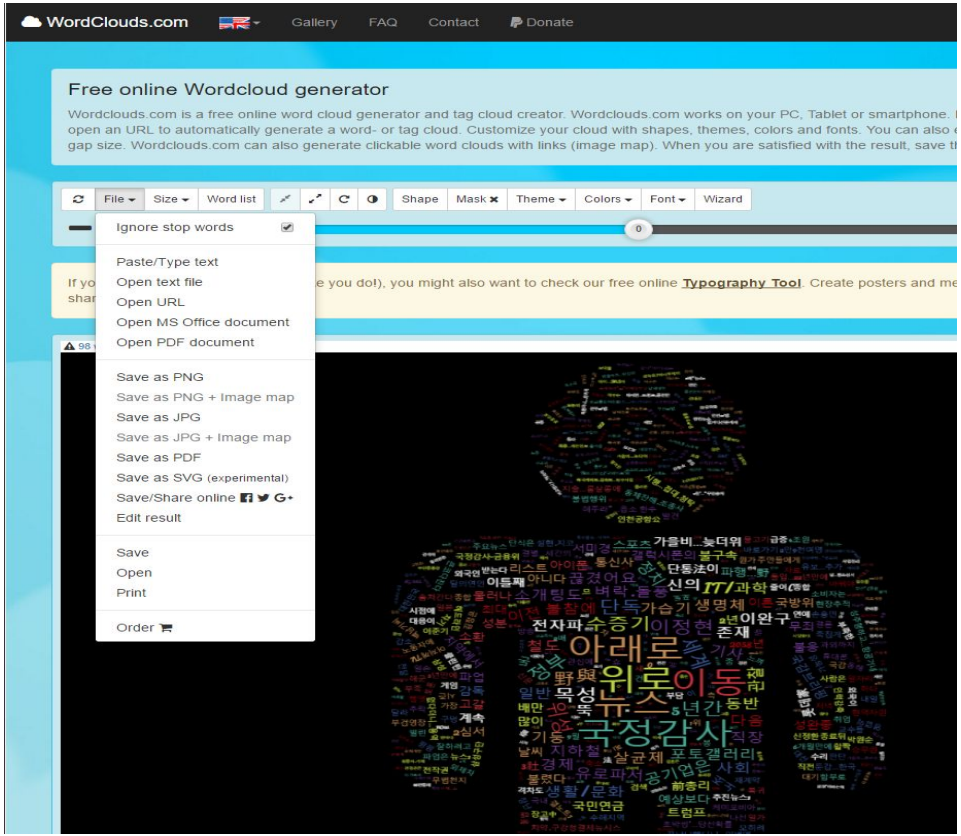
충남대학교 컴퓨터공학과

워드클라우드

•

http://www.wordclouds.com/

워드 클라우드를 체험해봅시다.



세상, 사람들의 이야기

- 매체
  - 종이: 신문, 책
  - 웹/앱: 트위터, 페이스북, 블로그, 웹 커뮤니티, ...

- 사람들의 생각 살펴보기
  - 유통되는 글을 보면 사람들의 생각을 볼 수 있다!
    - 여론 조사
  - 온라인의 글들 유통주기가 굉장히 빠르고, 양이 매우 큼
    - 모든 글들을 찾아서 읽어보기가 어려워요!

온라인 글들을 정리해주세요!

1. 온라인 웹/앱의 글들을 모아주세요 !

2. 이쁘게 정리해주면 좋겠어요!

온라인 정보 응용

- 여론조사
  - 선거, 신제품 반응
- 긴급메시지
  - 지진이 나면 지진희 갤에서 텔레그램 알림을 받는다

특정 텔레그램 채널에 가입해 메시지를 받는 방식인데, 메시지가 전송되는 조건은 '**디씨인 사이드 지진희 갤러리에서 1분 내에 글 20개 이상이 등록됐을 때**'다. 이 알림봇 개발자가 밝힌 이유를 원문에서 발췌하면 이렇다.

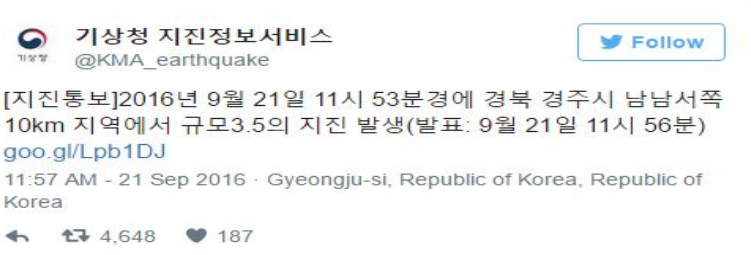
4. 시간 순으로 보자면 다음과 같습니다.

20:33 - 지진 발생

20:34 - 지진희갤 글 게시 시작

20:35 - 지진희갤 글 50개 돌파

20:37 - 기상청 발표



워드클라우드 작업 순서

1. 텍스트 데이터 수집
   - 웹 크롤링, 웹 문서 스크래핑
2. 텍스트 가공 및 자연언어처리
   - 한국어 품사 등의 형태소 분석
   - 사전에 의한 단어추출
3. 시각화
   - 원하는 형태의 워드 시각화

텍스트 데이터 수집하기

- 웹 브라우저에서 글을 캡쳐하기 : 컨텐츠에 대한 접근 API가 없는 경우
  - 웹 크롤링(Crawling) 소프트웨어 이용하기
    - 예) Scrapy, BeautifulSoup, Selenium

- 컨텐츠 제공업체가 제공하는 API 이용하기
  - Google, Naver, Twitter, Facebook 등에서는 API 형태로 온라인 컨텐츠에 대한 접근을 제공함
    - 예) Naver News API

Python 웹 크롤링 예제

디씨 “식물갤 ”

```
파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)
from bs4 import BeautifulSoup
import urllib2

page_url = 'http://gall.dcinside.com/board/lists/?id=tree'

url_open = urllib2.urlopen(page_url)

soup = BeautifulSoup(url_open, 'html.parser',

title_list = soup.findAll('td', attrs={'class'

for title in title_list:
    print(title.text)
~
```

```
yslee@yslee-ThinkPad-T410:~/python-wordcloud$ python webcrawl.py
[당첨자발표] 영단기 실전 1000제 이벤트 1차 당첨자 발표
식물 사진을 올려주세요.[265]
사라세니아종 잘아시는분있나요??
이름 모르는 꽃[2]
아주머니들이 이거 막 따가시던데 몰까요[1]
이 꽃 이름은 뭔가요[1]
안녕하세오 묘모긔에오
안녕하세요. 혹시 씨앗이름을 알 수 있을까요?
로즈마리가 원래 향이 짙은종인가요?[2]
포자류 겔러리는 없나요[1]
국화 소생시킬 방법좀요 ㅠㅠ[1]
여기에 제가기르는 커피 따서 말리고 우려먹는 논문쓰면 초념글 보내주나요?
이끼이름을 알고싶어요[1]
소국 키우는거좀 물어보려고 왔어요, 그리고 식물 키우면서 궁금한점[1]
이건 무슨 식물 열매인가요?[1]
혹시 외국 식물도 가능하세요?[2]
이 식물 이름이 뭔가요[1]
선인장에 붙은 점들 처리방법이 뭐에요?
이 꽃 뭔가요[2]
하나 더 질문드립니다[2]
이 식물 이름이 뭔가요?[2]
저희 집 화분에 이런 꽃이 폈는데 [4]
이거 이름 뭔가요?[1]
```

프로그램 설명

- 활용할 라이브러리 선언
  - BeautifulSoup
    - HTML과 XML 파일들에서 데이터를 추출하는 파이썬 라이브러리
  - Urllib
    - 웹 상의 문서나 파일을 가져올 수 있는 기본 파이썬 라이브러리

```
from bs4 import BeautifulSoup
import urllib2
```

```
page_url = 'http://gall.dcinside.com/board/lists/?id=tree'

url_open = urllib2.urlopen(page_url)
```

디씨 식물갤(tree) 웹 페이지를 열기

```
soup = BeautifulSoup(url_open, 'html.parser', from_encoding='utf-8')

title_list = soup.findAll('td', attrs={'class':'t_subject'})
```

BeautifulSoup으로 HTML 웹 페이지를 한글 디코딩하여 soup에 저장하기
't_subject' 라고 HTML 문서내에 존재하는 게시글 제목을 추출하기

```
for title in title_list:
    print(title.text)
```

게시글이 여러 개 이기때문에 title_list를 반복하여 글 제목을 출력하기!

네이버 API로 웹 크롤링

네이버에서 “충남대 ” 로 검색한 결과 모아보기

```
yslee@yslee-ThinkPad-T410:~/naver-api$ more naver-cnu.sh
curl -XGET "https://openapi.naver.com/v1/search/news.xml?query=%ec%b6%a9%eb%82%a
8%eb%8c%80&display=30&start=7&sort=sim"\
        -H "User-Agent: curl/7.43.0"\
        -H "Accept: */*"\
        -H "Content-Type: application/xml" \
        -H "X-Naver-Client-Id: cuoqfeqtc2zhNp7l6UyD" \
        -H "X-Naver-Client-Secret: PkTpUL_eeq" > cnu-news.xml
        Naver Open API - news ::'충남대'
xml_grep 충남농협, <b>충남대</b>병원과 함께 논산지역을 찾아 농업인 행복버스 진행
         <b>충남대</b>병원,국립대학교병원장 회의 개최
         <b>충남대</b>병원-<b>충남대</b>예술대학과 MOU 체결
         <b>충남대</b>병원-예술대학 인재양성 협약
         <b>충남대</b>학교병원 <b>충남대</b>예술대학과 MOU 체결
         <b>충남대</b>병원-<b>충남대</b>예술대학과 MOU 체결
         <b>충남대</b>병원, '어린이 대상 암예방' 캠페인 전개
         <b>충남대</b>병원 "문제가 된 선택진료비 전액 환불하겠다"
         충남농협 - <b>충남대</b>병원 논산지역 무료검진 실시
         <b>충남대</b>병원, 적 직원 대상 '김영란법' 교육
         <b>충남대</b>병원, 부정청탁 금지법 교육
         <b>충남대</b>병원, 국립대병원장 회의 개최
         대전 <b>충남대</b> 앞 신로데오거리 '유성 매드블럭' 스트리트몰 들어선다.
         <b>충남대</b>, 대전시·인니와 과학기술기반사업 협력 MOU
         병가 낸 의사 팔아 '특진료' 챙긴 <b>충남대</b>병원
         [충청브리핑] 환자 뒤통수 친 <b>충남대</b>병원
         <b>충남대</b> 수시 경쟁률 8.56대1 역대 최고… 국립대 강세 여전
         [단독] 유학 간 의사인데... <b>충남대</b>병원의 이상한 특진비 청구
         <b>충남대</b> 수시 8.56대 1 역대 최고… 을지대 의예과 44.5대 1
         <b>충남대</b>병원, 4년간 3만명에 진료비 수억원 부당징수 의혹
         <b>충남대</b>병원, 병원·의사 짜고 챙긴 '특진비'
```

네이버 개발자 등록


NAVER Developers
API 소개
개발가이드
Open Source
NAVER D2
Application
Support
API 상태
Search Here
Client ID
cuoqfeqtc2zhNp7l6UyD
Client Secret
••••••••••
로그인 오픈 API 통계
어제의 로그인 사용자수
0
최근 7일간 로그인 사용자수
0

```
yslee@yslee-ThinkPad-T410:~/naver-api$ more naver-cnu.sh
curl -XGET "https://openapi.naver.com/v1/search/news.xml?query=%ec%b6%a9%eb%82%a
8%eb%8c%80&display=30&start=7&sort=sim"\
        -H "User-Agent: curl/7.43.0"\
        -H "Accept: */*"\
        -H "Content-Type: application/xml" \
        -H "X-Naver-Client-Id: cuoqfeqtc2zhNp7l6UyD" \
        -H "X-Naver-Client-Secret: PkTpUL_eeq" > cnu-news.xml

xml_grep 'title' cnu-news.xml --text_only > cnu-news.txt
```

- cURL : URL을 입력으로 하여 웹 페이지를 다운받게 하는 명령어
- GET 은 웹 페이지를 다운받게 하는 HTTP 문법
- Naver 뉴스 검색 API 지정하는 곳 “https://openapi.naver.com/...
- 검색 키워드 “query=…”
  충남대학교 단어를 URL 인코딩 (http://www.convertstring.com)

- xml_grep ‘title’ 에서 뉴스의 제목을 추출하기

파이썬으로 워드클라우드 만들기 I

```
from collections import Counter
import urllib
import random
import webbrowser

from konlpy.tag import Hannanum
from lxml import html
import pytagcloud # requires Korean font support


r = lambda: random.randint(0,255)
color = lambda: (r(), r(), r())
```

웹 페이지 가져오기

```
def get_bill_text(billnum):
    url = 'http://pokr.kr/bill/%s/text' % billnum
    response = urllib.urlopen(url).read().decode('utf-8')
    page = html.fromstring(response)
    text = page.xpath(".//div[@id='bill-sections']/pre/text()")[0]
    return text
```

```
def get_tags(text, ntags=50, multiplier=10):
    h = Hannanum()
    nouns = h.nouns(text)
    count = Counter(nouns)
    return [{ 'color': color(), 'tag': n, 'size': c*multiplier }\
                for n, c in count.most_common(ntags)]
```

문서 단어 추출

```
def draw_cloud(tags, filename, fontname='Noto Sans CJK', size=(800, 600)):
    pytagcloud.create_tag_image(tags, filename, fontname=fontname, size=size)
    webbrowser.open(filename)
```

이미지파일에 한글폰트로 단어 그리기

```
bill_num = '1904882'
text = get_bill_text(bill_num)
tags = get_tags(text)
draw_cloud(tags, 'korean.png')
```

R 워드 클라우드

```
library(KoNLP)
library(wordcloud)
library(plyr)

# adr <- file("56 안철수.txt", blocking=F, encoding = "UTF-8") #글씨 깨어진다면 encoding 바꾸어야 한다.
# adr <- file("56 박근혜.txt", blocking=F, encoding = "UTF-8") #글씨 깨어진다면 encoding 바꾸어야 한다.
adr <- file("56 노무현.txt", blocking=F, encoding = "UTF-8") #글씨 깨어진다면 encoding 바꾸어야 한다.
adr2 <- readLines(adr); head(adr2)
# adr2  <- readLines("56 노무현.txt", , FileEncoding = "UTF-8")
close(adr)
# warnings()
# ?useSejongDic
useSejongDic() #세종 사전을 사용. 한글 단어 사전이다.

# ?mergeUserDic
#mergeUserDic(data.frame("안철수", "ncn")) # 세종 사전에 없는 단어를 추가할때.
#mergeUserDic(data.frame(c('안철수', '박근혜', '노무현', '정규직'), c('nqpc')))   # 세종 사전에 없는 단어
dics <- c('sejong', 'woorimalsam')
category <- c('news')
user_d <- data.frame(c('안철수', '박근혜', '노무현', '정규직'), c('nqpc'))
buildDictionary(ext_dic=dics, category_dic_nms = category, user_dic = user_d, replace_usr_dic=F)   # 세종
nouns <- sapply(adr2, extractNoun, USE.NAMES=F)   # 각 줄에서 명사만 추출해 낸다. KoNLP의 함수다
head(nouns)

# 불필요한 글자를 제거한다. 한 글자로 된 명사들은 분석에 불필요하다.
# nouns <- gsub("저ㅣ수ㅣ들","",nouns)
# 또는 아래와 같이 아예 한 글자로 된 모든 단어를 제거 한다.
c <- unlist(nouns) # 필터링을 위해 unlist 작업을 해서 저장합니다.
nouns <- Filter(function(x) {nchar(x) >= 2} ,c) # 두 글자 이상 되는 것만 필터링하기
head(nouns)

nouns <- gsub(" ㅣ\\d+","", nouns) # 숫자 없애기
c <- unlist(nouns) # 필터링을 위해 unlist 작업을 해서 저장합니다.
nouns <- Filter(function(x) {nchar(x) >= 2} ,c) # 두 글자 이상 되는 것만 필터링하기

wordcount <- table(unlist(nouns))
head(sort(wordcount, decreasing=T),30) # 그냥 확인해 보려고

pal <- brewer.pal(12,"Set3")
pal <- pal[-c(1:2)]
# 폰트 세팅
# windowsFonts(malgun=windowsFont("맑은 고딕"))
# 윈도우즈 폰트데이터베이스에서 찾을 수 없는 폰트패밀리입니다
# 라는 오류가 나타나면 설정한다.
wordcloud(names(wordcount), freq=wordcount, scale=c(7, 0.2), min.freq=4, random.order=F,  random.color=F
          colors=pal, family="AppleGothic")
# warnings()
```

자연언어처리

- 형태소 분석
  - “4월 벚꽃 축제가 열렸습니다.”

```
]: pprint(kkma.pos(u'4월 벚꽃 축제가 열렸습니다.'))

[(4, NR),
 (월, NNM),
 (벚꽃, NNG),
 (축제, NNG),
 (가, JKS),
 (열리, VV),
 (었, EPT),
 (습니다, EFN),
 (., SF)]
```

한글 형태소 분석기 SW

- 꼬꼬마
  - http://kkma.snu.ac.kr/
- 여러가지
  - http://konlpy.org/ko/v0.4.3/morph/
  - https://github.com/krikit/hanal/wiki/%ED%95%9C%EA%B5%AD%EC%96%B4-%ED%98%95%ED%83%9C%EC%86%8C-%EB%B6%84%EC%84%9D%EA%B8%B0-%EB%8F%99%ED%96%A5

Python, R 에서 한국어 처리

- KoNLP
  - R에서의 한국어처리 패키지
    - 품사처리, 형태소분석기, 세종사전
  - https://github.com/haven-jeon/KoNLP/blob/master/etcs/KoNLP-API.md
- KoNLPy
  - http://konlpy.org/ko/v0.4.3/examples/wordcloud/