

한국 3대 포털 기사 WordCloud 분석 자동화 시스템

고건호



Project Background & Goal

기획배경

정보의 홍수

너무 많이 배출되는 뉴스의 양

뉴스보다 SNS 선호하는 젊은 세대

뉴스 볼 여유가 없는 바쁜 직장인들

꾸준히 읽지 않으면 알기 힘든 트렌드

기획목표

매일 정해진 시간에 자동 실행

한국 3대 검색엔진의 기사 수집

한눈에 볼 수 있도록 WordCloud 분석

편의성 증대를 위해 Slack Messenger로 전송

Project Overview

개요

1. 한국 3대 포털엔진 기사 수집 - Naver, Daum, Nate
 2. Scrapy의 Pipeline 기능 활용 - MongoDB 저장
 3. MongoDB에 저장된 데이터 취합 (toWordCloud.py)
 4. WordCloud 분석 (toWordCloud.py)
 5. AWS CLI 명령어로 AWS S3에 WordCloud 이미지 저장
 6. Image URL 추출 후, Slack Messenger로 전송
- 위 모든 작업은 crontab 기능으로 시간 지정하여 실행

Skills

사용언어 : Python

시각화 : WordCloud

자료수집 : Scrapy, Xpath

자동화 : Crontab

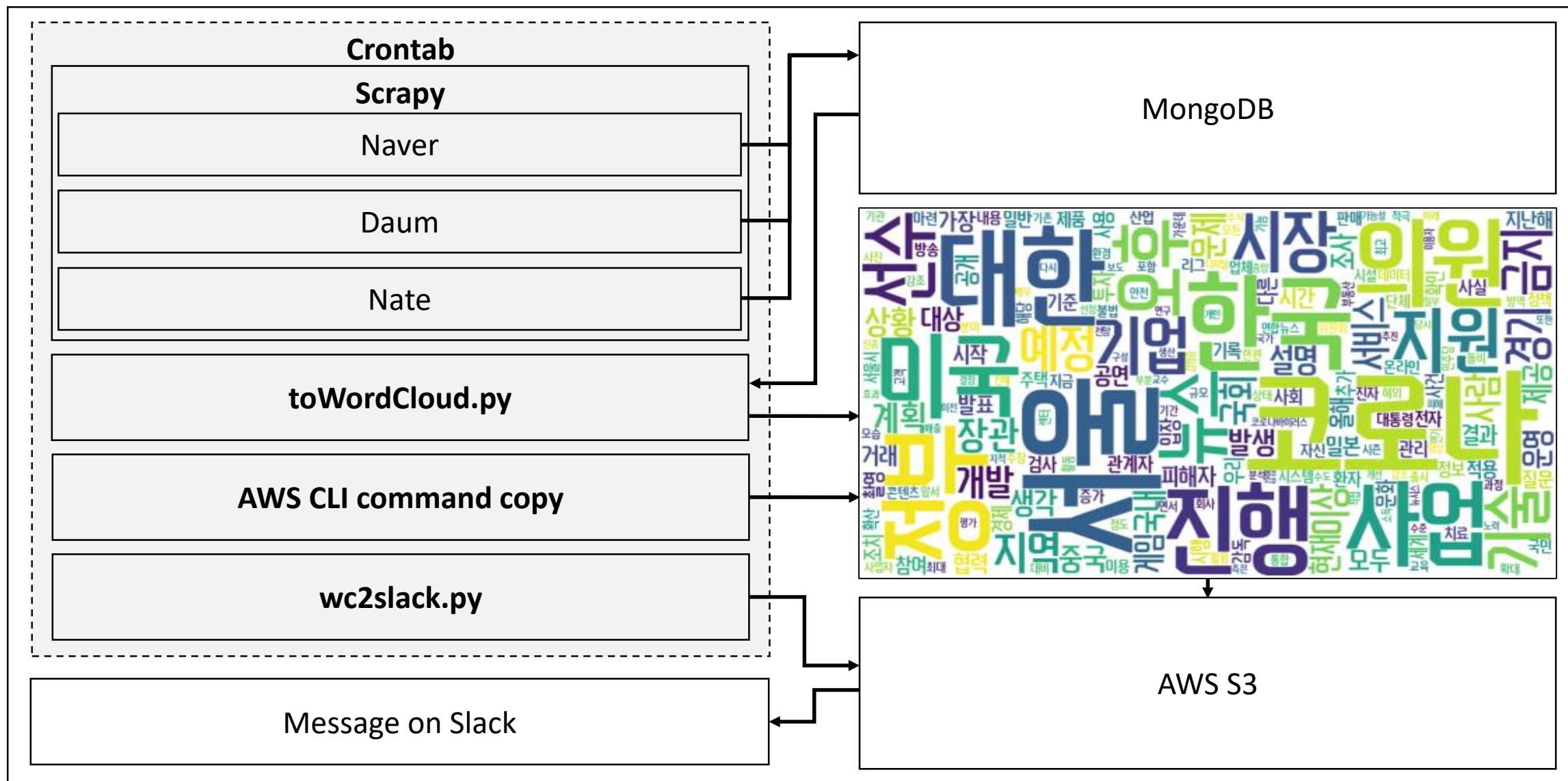
데이터베이스 : MongoDB, AWS S3

서버 : AWS EC2

언어분석 : konlpy, nltk

메신저 : Slack

Project Overview



Project Detail - 1. 기사 수집

Scrapy

1. 기사 수집 용 Scrapy 프로젝트 3개 생성
2. 파이프라인으로 수집된 데이터 MongoDB 저장

```
import scrapy
from newsDaum.items import NewsdaumItem
# from selenium import webdriver
from scrapy.http import TextResponse
import requests

class DaumSpider(scrapy.Spider):
    name = "NewsDaum"
    allow_domain = ["https://daum.net"]
    start_urls = ["https://news.daum.net/breakingnews/economic"]

    def parse(self, response):
        categories = ['society', 'politics', 'economic', 'foreign', 'culture', 'entertain', 'sports', 'digital']
        for name in categories:
            for page in range(1, 100):
                url = "https://news.daum.net/breakingnews/{}.format(name, page)
                req = requests.get(url)
                response = TextResponse(req.url, body=req.text, encoding="utf-8")
                links = response.xpath('//*[@id="mArticle"]/div[3]/ul/li/div/strong/a/@href').extract()
                for link in links:
                    yield scrapy.Request(link, callback=self.parse_content)

    def parse_content(self, response):
        item = NewsdaumItem()
        item['title'] = response.xpath('//*[@id="cSub"]/div/h3')[0].extract().split(">")[1].split("<")[0]
        item['category'] = response.xpath('//*[@id="kakaoBody"]')[0].extract().split(">")[1].split("<")[0]
        content = response.xpath('//*[@id="harmonyContainer"]/section/p/text()').extract()
        item['content'] = "".join(content)
        item['link'] = response.url
        yield item
```

```
from itemadapter import ItemAdapter
from .mongodb import collection

class NewsdaumPipeline:

    def process_item(self, item, spider):

        data = {
            "title": item["title"],
            "category": item["category"],
            "content": item["content"],
            "link": item["link"],
        }
        collection.insert(data)

        return item
```

Project Detail - 2. 데이터 불러오기

MongoDB

데이터베이스에 저장된 데이터 취합 (toWordCloud.py)

Robo 3T - 1.3

File View Options Window Help

db.getCollection('article').find({})

article 0.04 sec.

_id	title	category	content	link
1	미스터트롯 서울공연, 체조경기장 집합금지명령에 다시 취소	사회	(서울=뉴스1) 유승관 기자 = 22일 서울 송파구 올림픽공원 KSPO돔(체조경기장)에 ...	https://news.v.daum.net/v/20200722162657343
2	부산도시철도 청소노동자 고용전환 기대	사회	(부산=연합뉴스) 김재홍 기자 = 22일 오후 부산시의회 브리핑룸에서 열린 부산도시...	https://news.v.daum.net/v/20200722162656342
3	[경기] 경기도 내 산림휴양 시설 재개관..체험시설 등 보류	사회	경기도가 코로나19 확산 차단을 위해 휴관했던 산림휴양 시설들의 문을 열고 운영을...	https://news.v.daum.net/v/20200722162701350
4	부산도시철도 청소노동자 고용전환 기대	사회	(부산=연합뉴스) 김재홍 기자 = 22일 오후 부산시의회 브리핑룸에서 열린 부산도시...	https://news.v.daum.net/v/20200722162702352
5	박원순 피해자측 "비서관 20명에게 전보요청 목살"	사회	고 박원순 전 서울시장을 성추행 혐의로 고소한 A씨측이 22일 기자회견을 열고 "서...	https://news.v.daum.net/v/20200722162702351
6	새만금 그린인프라 구축 토대 마련	사회	전북대 산학협력단이 수행하는 이번 용역은 내년 4월까지 진행된다.새만금개발청은 ...	https://news.v.daum.net/v/20200722162703355
7	새만금 육상 태양광 3구역 발전 사업 본격화	사회	협약에 따라 중부발전건설사업은 이달 안에 군산에 사업시행법인을 설립해 인허가...	https://news.v.daum.net/v/20200722162702353
8	"폭설 속 터널서 과속 질주"...32대 연쇄추돌' 사매2터널 사고원인 보니	사회	터널 속 연쇄 추돌사고로 42명의 사상자를 낸 순천~완주고속도로 사매2터널 참사...	https://news.v.daum.net/v/20200722162720370
9	박경숙 Aquamarine 아트센터장, 전북대에 1000만 원 기탁	사회	전북대 김동원 총장은 22일 오전 박 센터장을 대학에 초청해 발전기금 기탁식을 갖...	https://news.v.daum.net/v/20200722162713365
10	민주당 전북도당 위원장 내달 9일 선출	사회	22일 전북도당 선거관리위원회는 회의를 열고 도당위원장 선출을 위한 시행 세칙 등...	https://news.v.daum.net/v/20200722162709362
11	남원시 홍보대사에 김리를 한복 정장 디자이너 위촉	사회	남원시 산동면에서 태어난 김리를 디자이너는 전주대학교를 졸업하고 현재 '리을'이...	https://news.v.daum.net/v/20200722162704357
12	민주당 전북도당 위원장 27일 후보등록, 8월9일 선출 확정	사회		https://news.v.daum.net/v/20200722162842422
13	농관원 전남지원, 10월까지 공익직불금 신청 농지 집중 점검	사회	[광주=뉴스1] 류형근 기자 = 국립농산물품질관리원 전남지원(농관원 전남지원)이 ...	https://news.v.daum.net/v/20200722162739376
14	괴산, 말산업 중심지로 뜬다	사회	충북 괴산군이 공공승마장 개장을 시작으로 말산업 육성 사업에 팔을 걷어붙였다.괴...	https://news.v.daum.net/v/20200722162757387
15	박원순 피해자측 "경찰 소장 접수 전 검찰에 문의..면담 거부당해"	사회	한국성폭력상담소-한국여성의전화와 피해자 측 변호인인 김재현 변호사는 22일 서...	https://news.v.daum.net/v/20200722162812398
16	송파구 사랑교회서 사흘 사이 4명 확진.. 모두 신도들	사회	정은경 질병관리본부 중앙방역대책본부(방대본) 본부장은 22일 오후 진행한 코로나...	https://news.v.daum.net/v/20200722162841421
17	창원시, 공원 물놀이장 오는 8월 1일부터 개장	사회	[창원=쿠기뉴스] 강종효 기자 = 경남 창원시(시장 허성무)는 여름철 인기 피서시설...	https://news.v.daum.net/v/20200722162802389
18	피해자 지원단체 2차 기자회견 관련 서울시 입장발표	사회	[서울=뉴스1] 전진환 기자 = 황인식 서울시 대변인이 22일 오후 서울시청 브리핑...	https://news.v.daum.net/v/20200722162820403
19	마약 투약 혐의 보람상조 회장 장남, 2심에서 집행유예 석방	사회	마약을 밀반입해 투약한 혐의로 기소된 상조업체 보람상조 최철홍 회장의 장남이 1...	https://news.v.daum.net/v/20200722162734373
20	故 박원순 전 서울시장 49재 참석하는 박주신 씨	사회	(서울=뉴스1) 민경석 기자 =故 박원순 전 서울시장의 아들 박주신 씨(왼쪽) 등 유...	https://news.v.daum.net/v/20200722162843424

Logs

Project Detail - 3. 언어 분석

Python

1. konlpy 패키지 활용하여 명사 추출
2. nltk 패키지 활용하여 추출된 명사 분석
3. WordCloud 이미지 추출 및 서버 저장



```
from konlpy.tag import Okt
from tqdm import tqdm
from nltk import FreqDist
from wordcloud import WordCloud
from datetime import datetime
import pandas as pd
import matplotlib.pyplot as plt
```

```
import pymongo

import requests, json
```

```
client = pymongo.MongoClient("mongodb://13.207.246.10:27020/")
result_nate = client['nate'].article
result_daum = client['daum'].article
result_naver = client['naver'].article
```

```
df_nate = pd.DataFrame(list(result_nate.find()))
df_daum = pd.DataFrame(list(result_daum.find()))
df_naver = pd.DataFrame(list(result_naver.find()))
```

```
df = df_nate.append(df_daum)
df = df.append(df_naver).reset_index()
```

```
df = df.drop(['_index', '_id'], axis=1)
```

```
def tokenize(doc):
    tagger = Okt()
    tokens = [t for t in tagger.nouns(doc)]
    return tokens
```

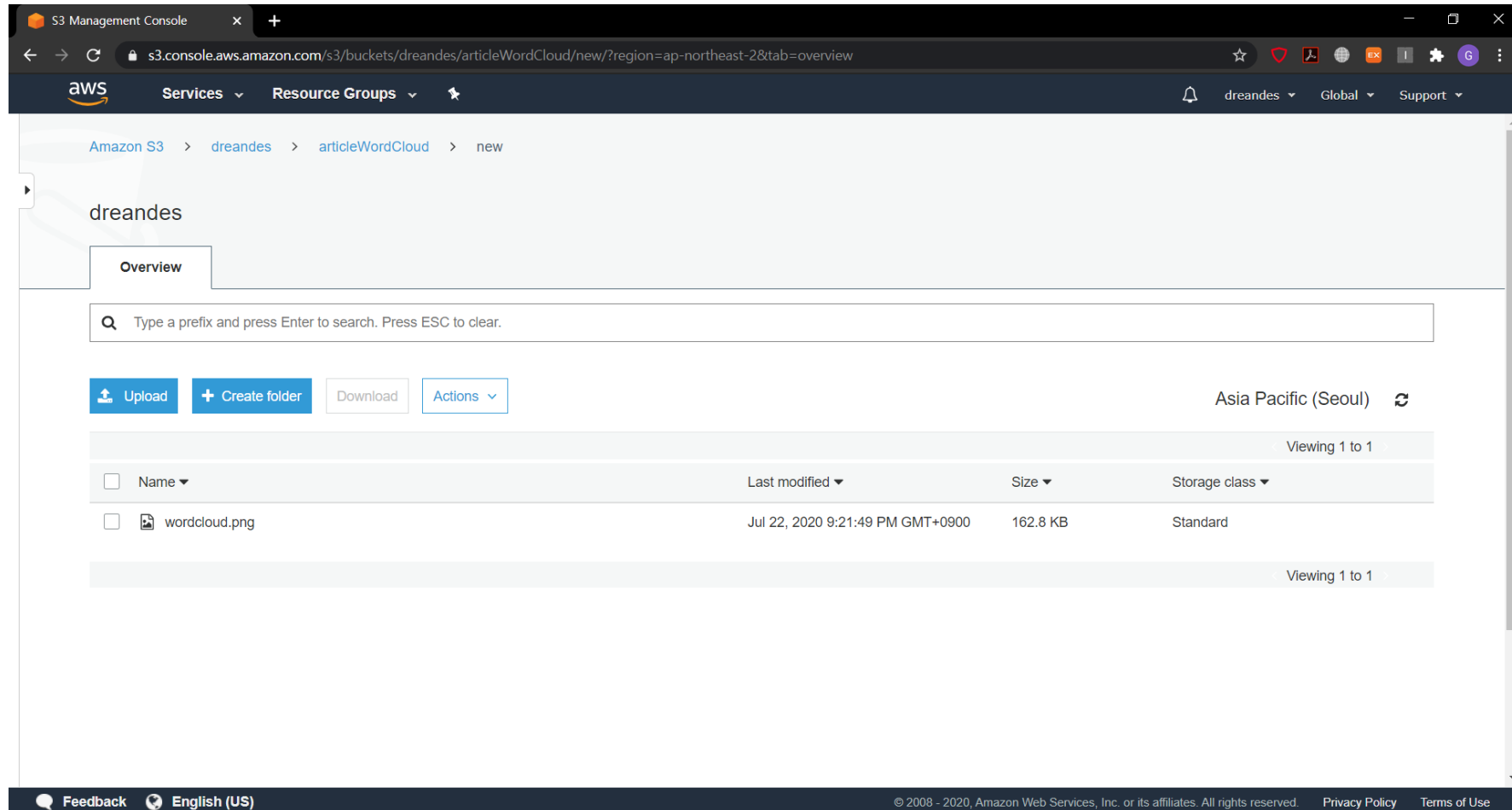
```
def towordcloud(df):
    df = df.dropna()
    docs = tuple([x for x in df.to_numpy()])
    sentences = []
    for d in tqdm(docs):
        tokens = [token for token in tokenize(d) if token.isalnum()]
        sentences.append(tokens)
    words = [word for sentence in sentences for word in sentence]
    words = [word for word in words if len(word) > 1]
    words_remove = ['으로', '에서', '에도', '했다', '있다', '이다', '무단', '배포', '위해', '대표', '매출', '그를', '틀해', '최근', '경우', '이번', '이후', '라며', '지난', '대해', '기자', '관련', '']
    words_r = [word for word in words if word not in words_remove]
    fd = FreqDist(words_r)
    # print(fd.most_common(20))
    font_path = '/home/ubuntu/python3/Crawling/koverwatch.ttf'
    wc = WordCloud(width=1000, height=600, background_color="white", random_state=0, font_path=font_path)
    plt.imshow(wc.generate_from_frequencies(fd))
    plt.axis("off")
    str = "/home/ubuntu/python3/Crawling/wordcloud_" + datetime.now().strftime("%Y.%m.%d_%H.%M.%S") + ".png"
    # plt.savefig(str)
    plt.savefig('/home/ubuntu/python3/Crawling/wordcloud.png')
    # plt.show()
```

```
!towardcloud(df['content'])
```

Project Detail - 4. Amazon S3 저장

AWS CLI

AWS CLI 명령어를 통해 wordcloud.png 서버 복사



Project Detail - 5. image_url 추출 후, Slack 전송

Slack

image_url 추출 후, Slack Messenger로 전송

```
import requests, json
import boto3

s3 = boto3.client('s3')
image_url = s3.generate_presigned_url('get_object',
                                     Params={'Bucket': 'dreandes',
                                             'Key': 'articleWordCloud/new/wordcloud.png'})

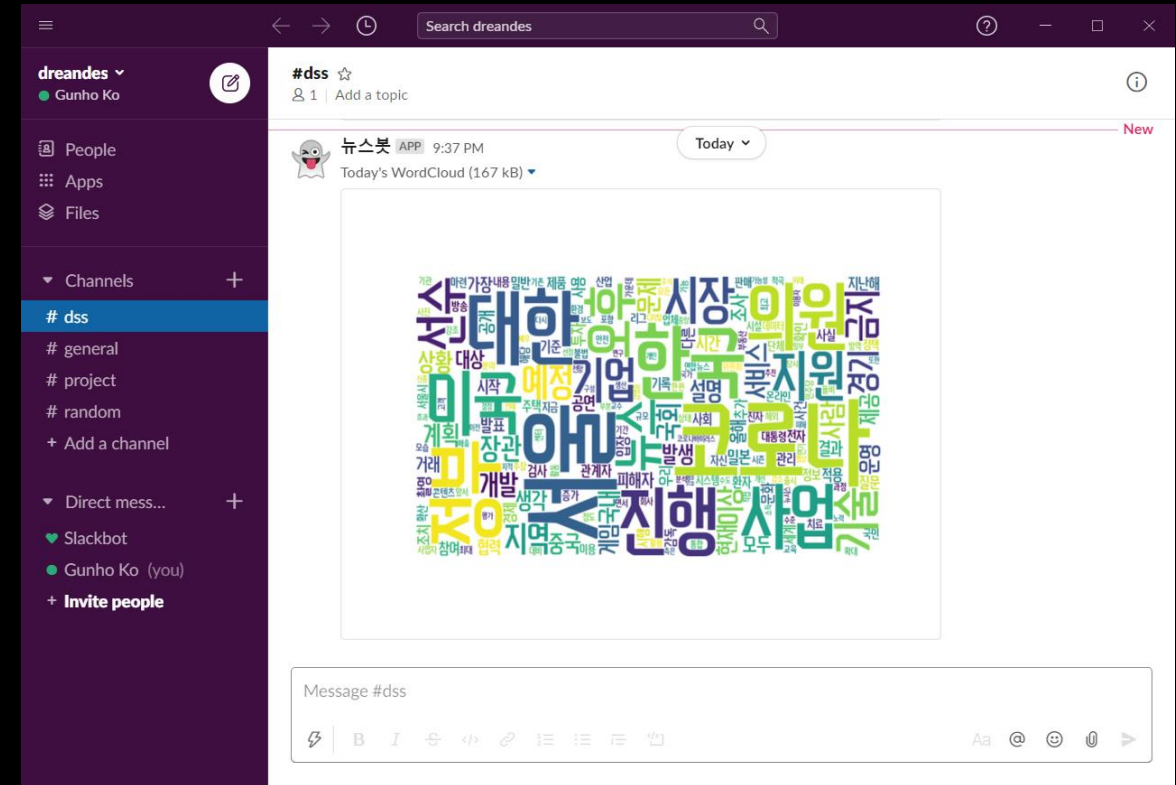
def send_msg(slack_webhook, block, channel="#dss", username="뉴 스 봇"):
    payload = {"channel": channel,
               "username": username,
               "text": msg,
               "icon_emoji": ":ghost:",
               "blocks": block
              }
    requests.post(slack_webhook, json.dumps(payload))

slack_webhook = "

block = [{"type": "image",
          "title": {
              "type": "plain_text", "text": "Today's wordCloud", "emoji": True
          },
          "image_url": image_url,
          "alt_text": "wordcloud"}]

msg = "Today's wordCloud"

send_msg(slack_webhook, block)
```



Project Detail - 6. 자동화

Crontab

Crontab 실행 시간 지정을 통한 자동화 시스템 구축

```
ubuntu@ip-~  
PATH=/home/ubuntu/.pyenv/versions/python3/bin/  
# Scrapy  
30 16 * * * cd /home/ubuntu/python3/Crawling/newsNate/newsNate && scrapy crawl NewsNate  
30 16 * * * cd /home/ubuntu/python3/Crawling/newsDaum/newsDaum && scrapy crawl NewsDaum  
30 16 * * * cd /home/ubuntu/python3/Crawling/newsNaver/newsNaver && scrapy crawl NewsNaver  
# Get wordcloud  
34 16 * * * cd /home/ubuntu/python3/Crawling && python toWordCloud.py  
# Copy to S3  
39 16 * * * /usr/bin/aws s3 cp /home/ubuntu/python3/Crawling/wordcloud.png s3://dreandes/articleWordCloud/new/  
# Send WordCloud to Slack  
40 16 * * * cd /home/ubuntu/python3/Crawling && python wc2slack.py  
# Move past WordClouds to old directory  
45 16 * * * /usr/bin/aws s3 mv s3://dreandes/articleWordCloud/new s3://dreandes/articleWordCloud/old --recursive  
45 16 * * * /bin/mv /home/ubuntu/python3/Crawling/wordcloud.png /home/ubuntu/python3/Crawling/wordcloud/  
# Edit this file to introduce tasks to be run by cron.  
#  
# Each task to run has to be defined through a single line  
# indicating with different fields when the task will be run  
# and what command to run for the task  
#  
# To define the time you can provide concrete values for  
# minute (m), hour (h), day of month (dom), month (mon),  
# and day of week (dow) or use '*' in these fields (for 'any').#  
# Notice that tasks will be started based on the cron's system  
# daemon's notion of time and timezones.  
#  
# Output of the crontab jobs (including errors) is sent through  
# email to the user the crontab file belongs to (unless redirected).  
#  
# For example, you can run a backup of all your user accounts  
# at 5 a.m every week with:  
# 0 5 * * 1 tar -zcf /var/backups/home.tgz /home/  
#  
# For more information see the manual pages of crontab(5) and cron(8)  
#  
# m h dom mon dow command
```

감사합니다.

