

# T-SA:

Twitter keyword Search API based Tweet Analysis  
(트위터 키워드 검색 API기반 트윗 분석)

과 목 산학캡스톤디자인1(2019-1학기)

담당교수 정 현 숙 교수님

팀 명 브 이 아 이 (VI)

팀 장 이 석 준 (20165072)

팀 원 이 윤 혁 (20165062)

서 재 익 (20144773)

발 표 자 배 인 규 (20165073)

발표일자 2019.05.30.



# T-SA: Contents

Twitter Keyword Search API based Tweet Analysis

1. T-SA: Team Introduction
2. T-SA: Purpose of Development
3. T-SA: Related Works & Control Group
4. T-SA: Development Environment
5. T-SA: Program Flowchart
6. T-SA: Implementation
7. T-SA: Result
8. T-SA: Reference
9. T-SA: Impression

# T-SA: Team Introduction

Twitter Keyword Search API based Tweet Analysis

## 이석준 (Lee SeokJune)

조선대학교 컴퓨터공학과(20165072)

MariaDB 환경 구축 및 관리

Twitter API 구현

문서 작성 및 수정

op2se1@gmail.com



## 이윤혁 (Lee Yunhyuck)

조선대학교 컴퓨터공학과(20165062)

Hadoop3, Sqoop 환경 구축

Hadoop(Map/Reduce)구현

leeyh5134@naver.com



## 서재익 (Seo JaeIck)

조선대학교 컴퓨터공학과(20144773)

Twitter API 구현

Visualization 구현

nero8879 @naver.com



## 배인규 (Bae InGyu)

조선대학교 컴퓨터공학과(20165073)

Python, DB 연동 구현

Visualization 구현

happykkk789@naver.com



# T-SA: Purpose of Development

Twitter Keyword Search API based Tweet Analysis

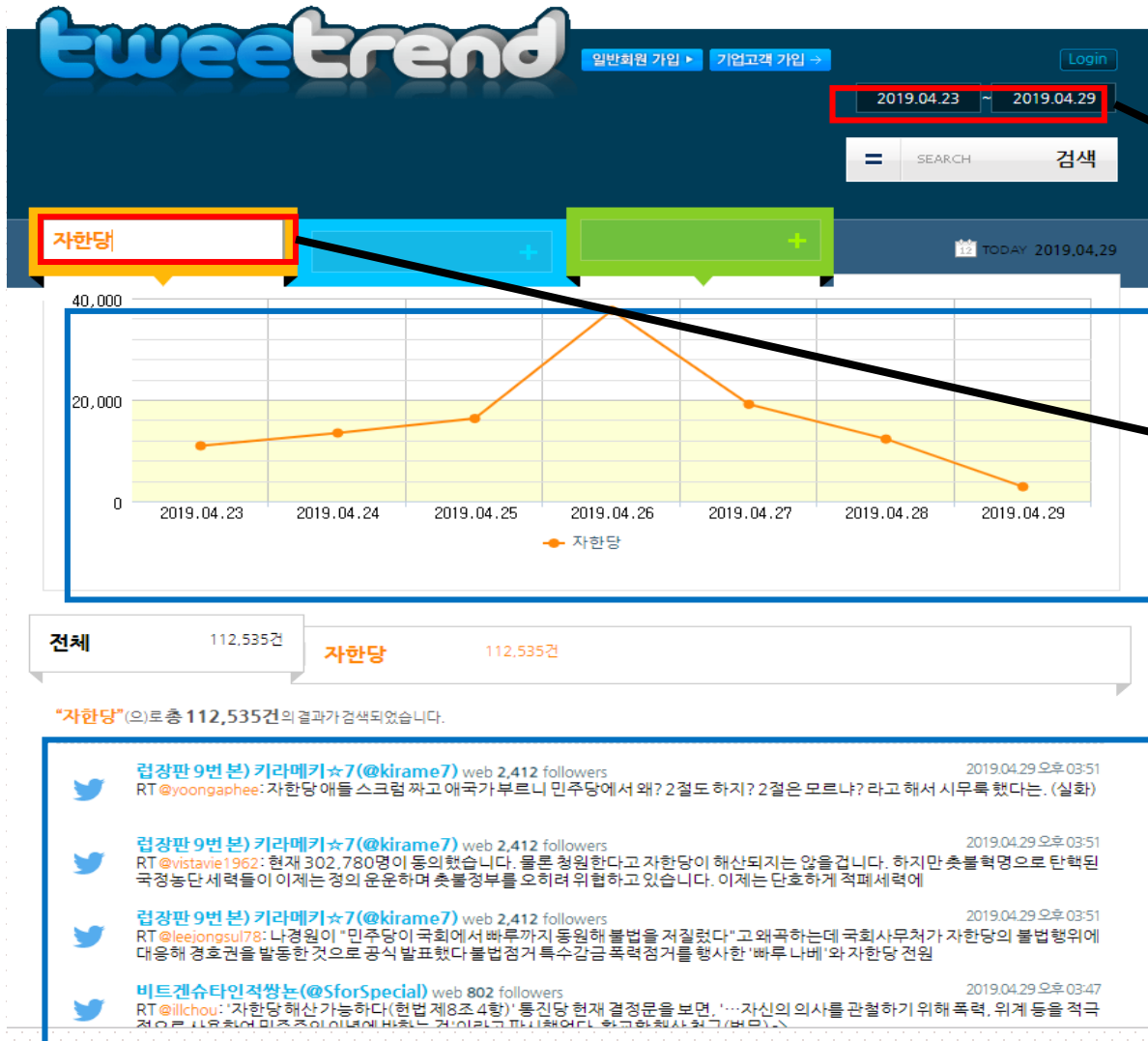
19대 대통령 선거 기간(2017.4.15 ~ 2017.05.09 ; 24일)에 후보 관련 키워드가 포함된 트윗 수집

19대 대통령 선거 후보 언급 횟수와 후보자들의 득표율 비교, 분석

19대 대통령 선거 후보들이 작성한 트윗 분석(Hashtag, 상세 정보 등)

# T-SA: Related Works

Twitter Keyword Search API based Tweet Analysis



[Tweetrend : http://tweetrend.com/](http://tweetrend.com/)

## 1. 검색할 기간 선택

비로그인 : 최대 7일 간의 검색 가능  
일반회원 (무료) : 최대 30일 간의 검색 가능  
일반회원 (유료) : 최대 6개월 간의 검색 가능

## 2. 검색할 키워드 입력

비로그인 : 1개의 키워드 입력 가능  
로그인 : 최대 3개의 키워드 입력 가능

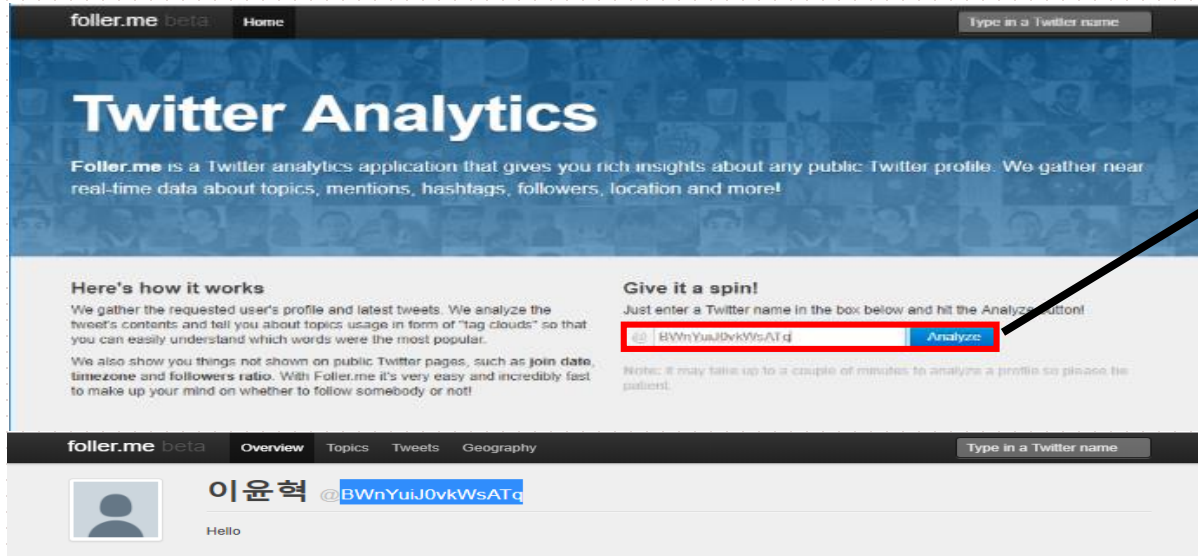
## 3. 그래프를 통한 날짜 별 트윗 수

## 4. 전체 트윗 개수 및 최근 트윗부터 리스트 출력

키워드를 검색하여 연관되는 키워드를 색출하고자 함.

# T-SA: Related Works

Twitter Keyword Search API based Tweet Analysis



[Foller.me : https://foller.me/](https://foller.me/)

## 1. 검색할 아이디 입력



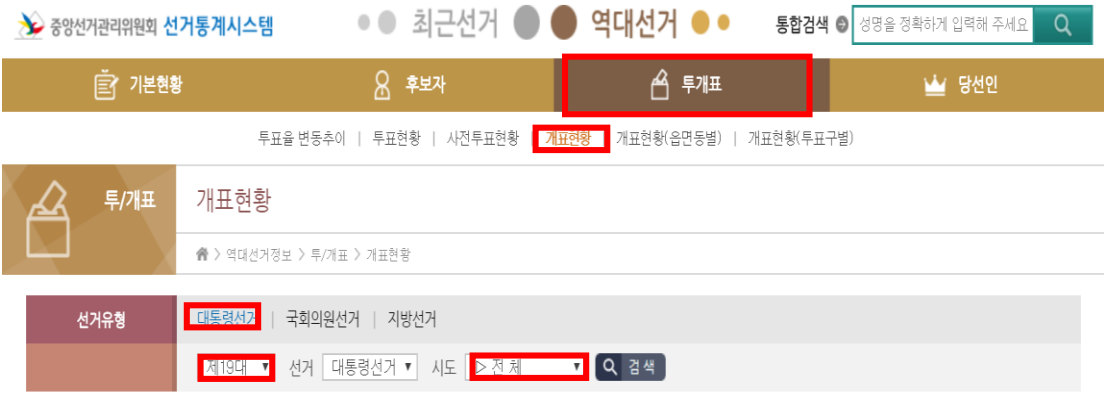
\*트위터 로그인 시 자신의  
프로필에 있는 아이디

## 2. 자신의 기본적인 정보 출력

정보 : 사용자 이름, 지역, 언어, 가입날짜  
상태 : 트윗 개수, 팔로잉 수  
시간 : 사용자가 활동하는 시간대  
해시태그를 포함한 언급한 사람들의 정보

Overview Profile information and statistics	
<strong>Information</strong> The most important piece here is the join date. The longer they're on Twitter the better. Spam accounts and robots tend to get suspended after a couple of weeks.	
<strong>AT A GLANCE</strong>	
Name	이윤혁
Joined Twitter on	Tue Mar 19 09:19:27 +0000 2019
Location	
Timezone	
Language	English language preference
Bio	Hello
URL	
<strong>Statistics</strong> More followers is good, but watch out for the follower-to-following ratio. A high ratio means that more people are following @BWnYuiJ0vkWsATq out of good will, not follow-back.	
<strong>EVERY TWEET COUNTS</strong>	
Tweets	9
Followers	0
Following	3
Followers ratio	0.00 followers per following
Listed	0

사용자의 정보를 수집하여, 사용자의 트윗 스타일을 분석하고자 함.



중앙선거관리 위원회 선거통계시스템

<http://info.nec.go.kr/>

19대 대통령 선거 후보자별 득표 현황 데이터 가져오기.

15명의 후보자별 전체 득표율과 지역별 득표율을 확인 할 수 있다.

시도명	선거인수	더불어민주당 문재인	자유한국당 홍준표	국민의당 안철수	바른정당 유승민	정의당 심상정
합 계	42,479,710	13,423,800	7,852,849	6,998,342	2,208,771	2,017,458
		(41.08)	(24.03)	(24.03)	(6.76)	(6.17)
서울특별시	8,382,999	2,781,345	1,365,285	1,492,767	476,973	425,459
		(42.34)	(20.78)	(22.72)	(7.26)	(6.47)

득표율을 기반으로 상위 5명 후보에 대한 정보를 시각화를 하여, 이를 트윗과 비교를 하고자 함.

# T-SA: Development Environment

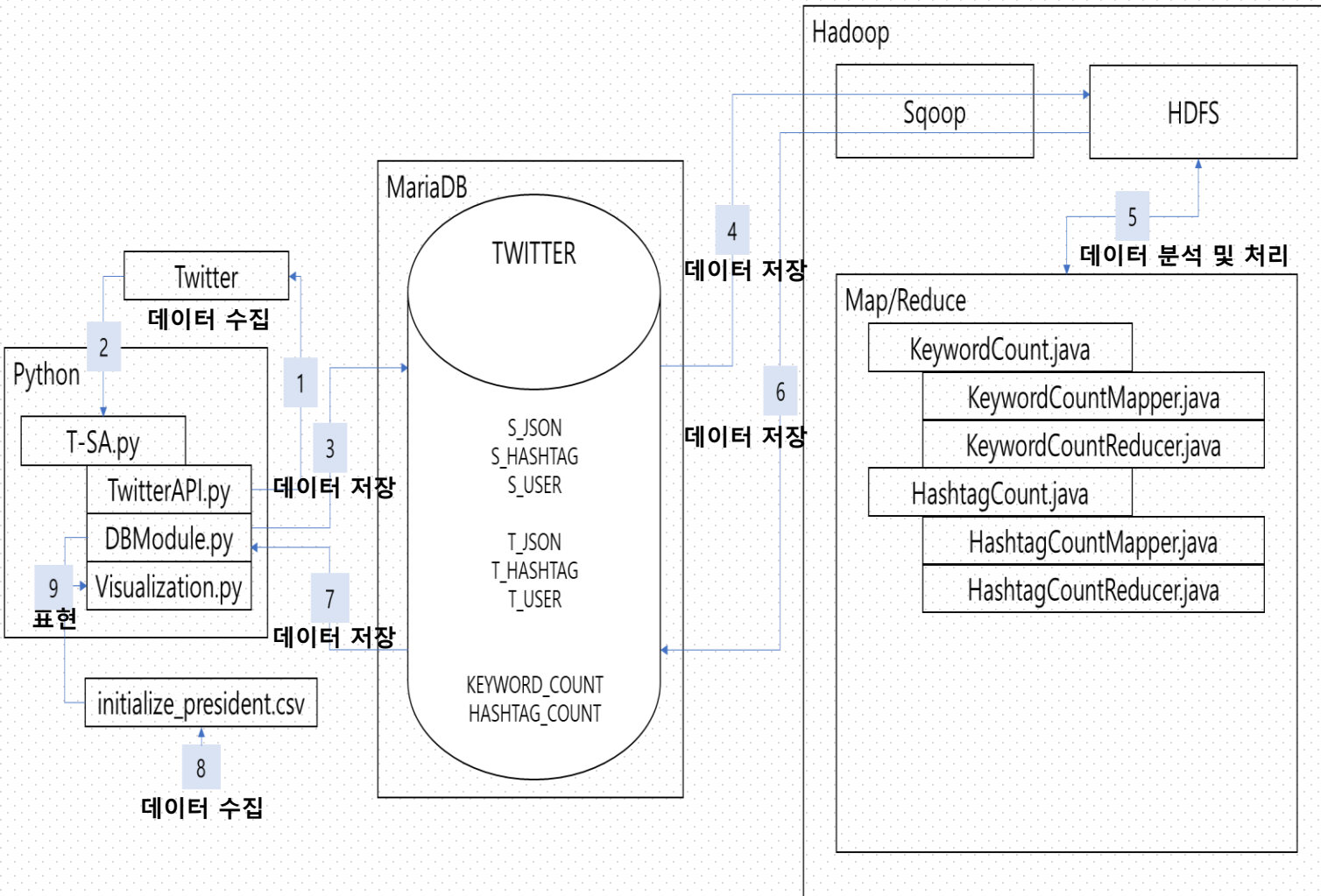
Twitter Keyword Search API based Tweet Analysis

설치순서	이름	버전	사 용 이 유
1	Python	3.6	수집한 데이터를 데이터 베이스에 저장 하며, 시각화에 필요한 모듈들을 조합하여 사용. <b>CLI(Command Line Interface)</b> 기반이기때문에 테스트와 구현이 용이.
<a href="https://www.python.org/">https://www.python.org/</a>			
2	MariaDB	10.1.38	수집된 데이터가 <b>정형데이터</b> 여서 대표적인 DBMS로 <b>RDBMS</b> 를 선택. RDBMS의 대표적인 소프트웨어에는 Oracle, MySQL, MSSQL 등이 있다. 오라클 소유의 MySQL의 불확실한 라이선스로 인해 MariaDB 선택.
<a href="https://mariadb.com/kb/ko/mariadb">https://mariadb.com/kb/ko/mariadb</a>			
3	OpenJDK	1.8.0_191	Java 애플리케이션을 실행하기 위한 <b>JVM</b> , 컴파일을 위한 <b>JDK</b> 필요. 하둡 3.2.0버전 과 스쿱 1.4.7 구동을 위해 JDK 8 선택.
<a href="https://openjdk.java.net/">https://openjdk.java.net/</a>			
4	Eclipse	2019-03(4.11)	IDE(Integrated Development Environment)로서, 무료로 사용할 수 있는 개발 툴 하둡(자바기반 오픈소스 프레임 워크) 구현.
<a href="https://www.eclipse.org/">https://www.eclipse.org/</a>			
5	Hadoop	3.2.0	대량의 자료를 처리하는 분산 응용 프로그램을 지원하는 자바 소프트웨어 프레임 워크. 기존 2.0 버전 보다 맵, 리듀스의 콜렉터를 기본 구현해주며, <b>성능이 30%이상 향상.</b>
<a href="https://hadoop.apache.org/">https://hadoop.apache.org/</a>			
6	Sqoop	1.4.7	Sqoop(SQL-to-Hadoop)은 <b>하둡과 RDBMS간 대량의 데이터를 전송</b> 하기 위해 만들어진 툴.
<a href="https://sqoop.apache.org/">https://sqoop.apache.org/</a>			



# T-SA: Program Flowchart

Twitter Keyword Search API based Tweet Analysis



[1, 2] (데이터 수집) TwitterAPI를 이용해서 정보(트윗 내용 (작성 시간, 트윗, 해시태그 등), 사용자 정보(아이디, 닉네임, 위치정보, 팔로우 수, 팔로잉 수, 언어 등)) 수집

[3] (데이터 저장) 크롤링 된 데이터를 MariaDB에 저장

[4] (데이터 저장) Sqoop을 이용하여 MariaDB에 저장된 데이터를 HDFS에 저장

[5] (데이터 분석 및 처리) HDFS에 업로드 된 데이터를 Map/Reduce과정을 통해 정규화하고 결과를 HDFS에 저장

[6] (데이터 저장) Sqoop을 이용하여 HDFS에 저장된 정규화된 데이터를 MariaDB에 저장

[7] (데이터 저장) MariaDB에 저장된 데이터를 Python으로 불러온다.

[8] (데이터 수집) 비교할 데이터(선관위 데이터 (initialize\_president.csv)) 수집  
\*\*중앙선거관리 위원회 선거통계시스템 (<http://info.nec.go.kr/>)

[9] (분석 및 표현) 불러온 데이터(선관위 데이터 (initialize\_president.csv), 정규화 데이터) 시각화

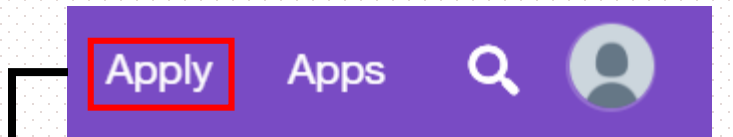
# T-SA: Implementation(Twitter API Issue)

Twitter Keyword Search API based Tweet Analysis

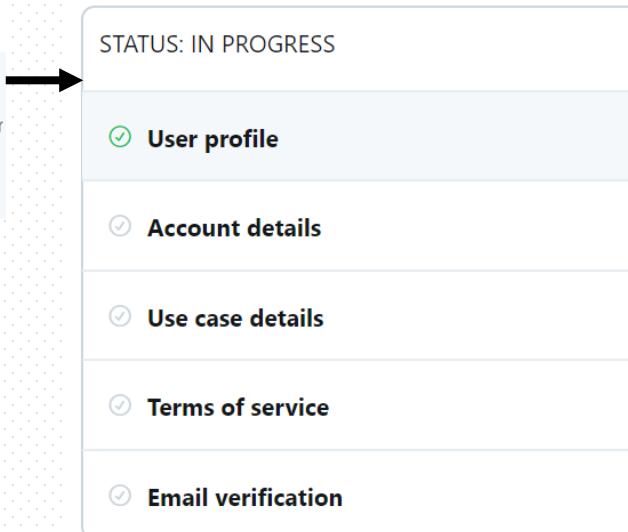
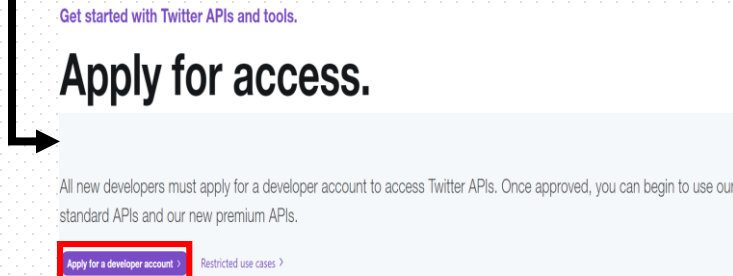
## Twitter API 발급을 위한 개발자 등록

트위터 계정에 가입된 상태에서 진행, **미 진행 시 API 발급 불가능.**

트위터 개발자 사이트 : <https://developer.twitter.com/>



개발자 등록 전



User profile – 핸드폰 번호와 이메일 주소 업데이트  
Account details – 계정 정보 선택 입력 : 단체 or 개인  
Use case details – 개발자 계정 목적과 용도 입력  
Terms of service – 트위터 개발자 정책 동의  
Email verification – 이메일 확인



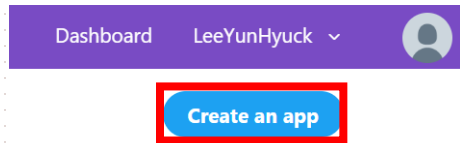
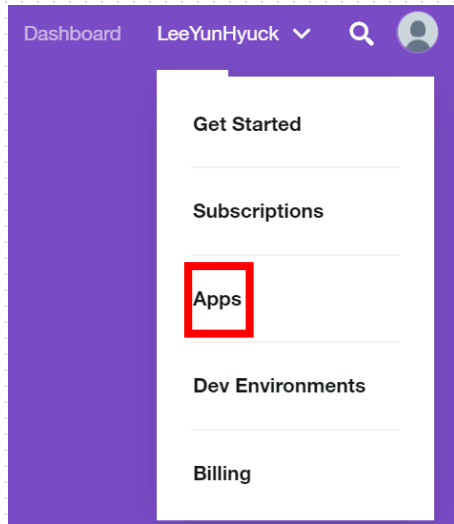
개발자 등록 후

# T-SA: Implementation(Twitter API Issue)

Twitter Keyword Search API based Tweet Analysis

## Twitter API 발급

개발자 등록이 완료된 상태에서 진행



### 필수 작성

App name : 앱 이름(~32자)

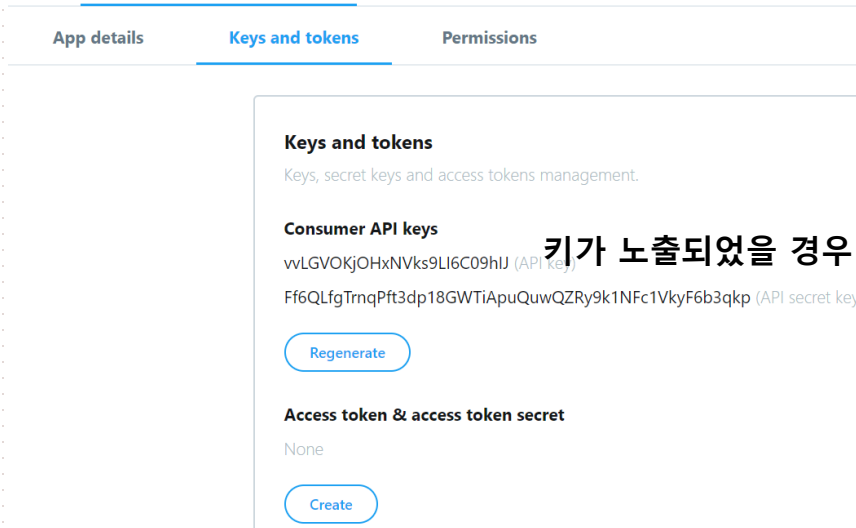
Application description : 앱에 대한 설명(10~200자)

Website URL : 웹 사이트 (작성한 트윗의 출처 표시 기능)

Tell us how app will be used : 앱의 사용 방법 작성 (100자 이상)



Apps > API for CapstonePresentation



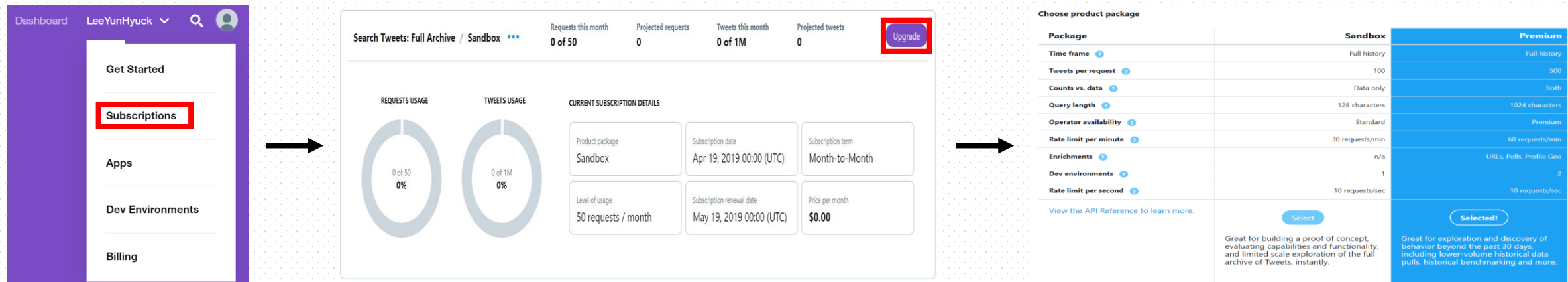
키가 노출되었을 경우 재발급을 받을 수 있다.

# T-SA: Implementation(Twitter API Issue)

Twitter Keyword Search API based Tweet Analysis

## Twitter API의 접근권한-1

API의 발급이 완료된 상태에서 진행



# T-SA: Implementation(Twitter API Issue)

Twitter Keyword Search API based Tweet Analysis

## Twitter API의 접근권한-2

API의 발급이 완료된 상태에서 진행

### Standard(기본 Free)

지난 7일간 게시된 최근 트윗을 제공

### Premium(50\$, 99\$)

30-days: 지난 30일간 게시된 트윗을 제공  
Full-archive: 2006년부터 게시된 트윗 제공

### Enterprise

Premium과 같은 두 가지를 제공하며 기업에서 주로 이용

### Standard 와 Premium의 차이점

1. 1번 요청해서 가져올 수 있는 트윗의 개수가 100개에서 500개로 확대
2. 조회할 수 있는 트윗의 글자수가 128자에서 1024자로 증가
3. 1분당 요청할 수 있는 횟수가 30회에서 60회로 증가

**하지만,** 트위터 결제 카드 등록 시, **대한민국을 차단**하여, 다른 방법으로 진행

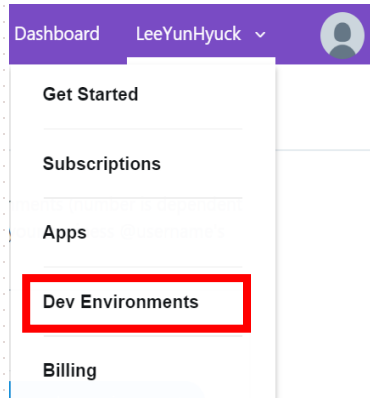
```
><select id="country" name="country" data-value="shortcode" class="non-hosted-field crs-country" data-  
default-option="Country" data-region-id="state" data-recurly="country" data-blacklist=  
"DJ,BA,BO,BV,BW,AX,FM,PR,RU,KR,TW,TN,SK,PF,LS,CN,GQ,IR,SY,VN,SX,KG,CU,GH,OM,MF,MA,MM,MU,SS,SI,KP,KW,SD"  
data-crs-loaded="true">...</select> == $0
```

크롬에서 제공하는 개발자 환경에서 웹 페이지 소스를 통해 확인

# T-SA: Implementation(Twitter API Issue)

Twitter Keyword Search API based Tweet Analysis

## Twitter 개발 환경 설정 : 트위터가 설립된 2006년 부터 데이터를 가져오기 위한 과정 트위터 API 키를 발급 받은 상태에서 진행



Search Tweets: 30-Days / Sandbox

최근 30일 트윗을 가져오기 위한 환경

NOT SET UP

This dev environment has not been set up.

Set up dev environment

Search Tweets: Full Archive / Sandbox

2006년 이후의 트윗을 가져오기 위한 환경  
(2017년 대선 기간의 데이터를 가져오기 위해 선택)

NOT SET UP

This dev environment has not been set up.

Set up dev environment

개발 환경의 이름 : TSA  
발급 받은 키의 이름 : 테스트 API발급

### Set up Search Tweets: Full Archive dev environment

Dev environment label

TSA

App

테스트 API 발급

테스트 API 발급

API for CapstonePresentation

Create a new app...

Search Tweets: Full Archive / Sandbox

Dev environment label

TSA

App Name

테스트 API 발급

App owner

@BWnYuiJ0vkWsATq

Delete environment

Dashboard

LeeYunHyuck ▾

🔍

👤

대쉬보드를 통해 키의 사용량 확인

Search Tweets: Full Archive / Sandbox

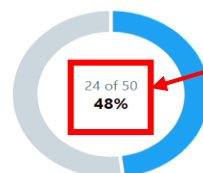
Requests this month  
24 of 50

Projected requests  
32

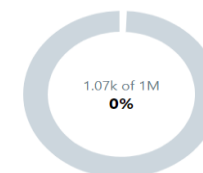
Tweets this month  
1.07k of 1M

Projected tweets  
1.4k

REQUESTS USAGE



TWEETS USAGE



May

May

# T-SA: Implementation

Twitter Keyword Search API based Tweet Analysis

## TwitterAPI : 데이터 수집 단계

<https://developer.twitter.com/en/docs/tweets/search/guides/standard-operators>

앱을 통해 발급받은 키를 통해 파이썬을 이용하여 데이터를 가지고 오는 방법

### <기본적인 트윗의 정보를 JSON으로 가져오기>

curl --request POST

```
--url https://api.twitter.com/1.1/tweets/search/fullarchive/<ENV>.json //fullarchive 모드 적용, 개발환경 이름
--header 'authorization: Bearer <BEARER_TOKEN>' //Access Token 입력
--header 'content-type: application/json' //Auth 인증을 통해 받는 데이터 타입
--data '{ "query": "from:TwitterDev lang:en", //TwitterDev 계정에서 언어는 영어로
        "maxResults " : " 100 " , //가져오는 최대 트윗을 100으로 설정
        " fromDate " : " <YYYYMMDDHHmm> " , //가져오는 날짜의 시작점(201708081111)
        " toDate " : " <YYYYMMDDHHmm> " }' //가져오는 날짜의 마지막점(201808080000)
```

# T-SA: Implementation

Twitter Keyword Search API based Tweet Analysis

## 19대 대통령 선거 득표율 : 데이터 수집 단계 <http://info.nec.go.kr/>

중앙선거관리 위원회 선거통계시스템을 통한 상위 5명의 후보자별 득표율 데이터 획득

시도명	선거인수	투표수	후보자별 득표수(득표율)														무효 투표수	기권수
			더불어민주당 문재인	자유한국당 홍준표	국민의당 안철수	바른정당 유승민	정의당 심상정	새누리당 조원진	경제자유 당 오영구	국민대통 합당 장성미	들쭉한 국당 이재오	민중연합 당 김서동	한국국민 당 이경희	홍익당 윤홍식	무소속 김민찬	계		
합계	42,479,710	32,807,908	13,423,800 (41.08)	7,852,849 (24.03)	6,998,342 (21.41)	2,208,771 (6.76)	2,017,458 (6.17)	42,949 (0.13)	6,040 (0.01)	21,709 (0.06)	9,140 (0.02)	27,229 (0.08)	11,355 (0.03)	18,543 (0.05)	33,990 (0.10)	32,672,175	135,733	9,671,802
서울특별시	8,382,999	6,590,646	2,781,345 (42.34)	1,365,285 (20.78)	1,492,767 (22.72)	476,973 (7.26)	425,459 (6.47)	9,987 (0.15)	789 (0.01)	3,554 (0.05)	1,938 (0.02)	3,416 (0.05)	1,277 (0.01)	2,177 (0.03)	3,950 (0.06)	6,568,917	21,729	1,792,353
부산광역시	2,950,224	2,261,633	872,127 (38.71)	720,484 (31.98)	378,907 (16.82)	162,480 (7.21)	109,329 (4.85)	2,651 (0.11)	276 (0.01)	1,316 (0.05)	465 (0.02)	981 (0.04)	496 (0.02)	1,041 (0.04)	2,156 (0.09)	2,252,709	8,924	688,591
대구광역시	2,043,276	1,581,347	342,620 (21.76)	714,205 (45.36)	235,757 (14.97)	198,459 (12.60)	74,440 (4.72)	4,057 (0.25)	259 (0.01)	563 (0.03)	324 (0.02)	804 (0.05)	401 (0.02)	986 (0.06)	1,501 (0.09)	1,574,376	6,971	461,929
인천광역시	2,409,031	1,820,091	747,090 (41.20)	379,191 (20.91)	428,888 (23.65)	118,691 (6.54)	129,925 (7.16)	2,646 (0.14)	374 (0.02)	1,618 (0.08)	410 (0.02)	1,230 (0.06)	594 (0.03)	625 (0.03)	1,681 (0.09)	1,812,963	7,128	588,940
광주광역시	1,166,901	957,321	583,847 (61.14)	14,882 (1.55)	287,222 (30.08)	20,862 (2.18)	43,719 (4.57)	152 (0.01)	111 (0.01)	655 (0.06)	103 (0.01)	2,265 (0.23)	136 (0.01)	264 (0.02)	614 (0.06)	954,832	2,489	209,580

### 상위 5명(문재인, 홍준표, 안철수, 유승민, 심상정)

1. 투표수와 득표율이 묶여 있기 때문에 따로 분류.
2. 5명을 기준으로 한, 전체 득표율은  $99.45(41.08+24.03+21.41+6.71+6.17)$ 이기 때문에 득표수를 이용.
3. 득표수는 막대 그래프와 꺾은선 그래프로 분류하고, 득표율은 파이 그래프로 확인.



# T-SA: Implementation

Twitter Keyword Search API based Tweet Analysis

## Map/Reduce : KeywordCount의 문제점

키워드의 연관성을 분석하기 위한 작업

가동	1	
가짜뉴스 <u>에</u>	1	
강원	1	
강조 <u>하는</u>	1	
것들..	2	
것을	1	
게	2	
경수짱	1	
고마 <u>하고</u>	1	
고민정	1	
고발 <u>을</u>	1	
공정하고	1	
공지	1	

← 공백을 기준으로 문장을 나눴을 때, 나오는 결과

데이터에서 **의미 있는 단어의 빈도를 분석**하여 시각화를 제공하기 위한 목적에 위반

가짜뉴스에, 강조하는, 고발을, 공정하고 → 가짜뉴스, 강조, 고발, 공정, 공지

명사 + 조사 형태에서 명사만을 추출하여 나타내고자 함.

자바 언어로 짜여진 오픈소스 들 중 KOMORAN을 선택

꼬꼬마 형태소 분석기, **KOMORAN**, OPEN-KOREAN-TEXT

→ <https://www.shineware.co.kr/products/komoran/>

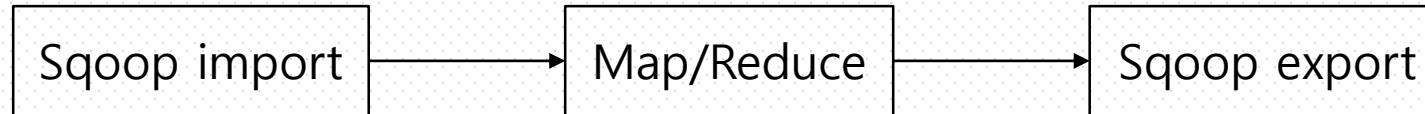
다른 형태소 분석기들 보다 속도, 정확성이 높으며,  
사용자 사전을 지원하여, 사용자가 원하는 단어를 명사로 추가 할 수 있다.  
Ex) 아이언맨 , 슈퍼맨 등

# T-SA: Implementation

Twitter Keyword Search API based Tweet Analysis

## Sqoop, Map/Reduce : 순차적 명령 실행 <https://docs.python.org/ko/3/tutorial/stdlib.html>

순서대로 명령어를 실행 시켜 주기위한 작업



기능에 맞는 명령어를 터미널에 직접 입력해야 하는 **사용의 불편함**

### Python의 표준 라이브러리 **OS** 사용

```
>>> import os
>>> os.system('start-all.sh') # 하둡실행 명령어
>>> os.system('jps') # 프로세스 상태 확인 명령어
```

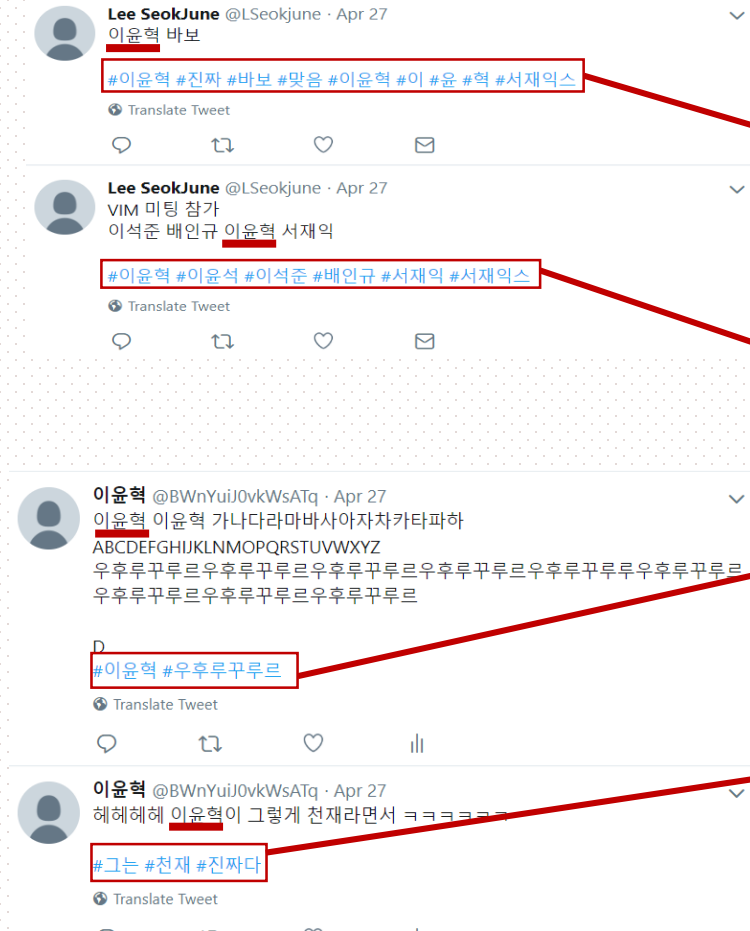
```
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [vi]
Starting resourcemanager
Starting nodemanagers
3668 Jps
2552 DataNode
3272 NodeManager
3082 ResourceManager
2364 NameNode
2845 SecondaryNameNode
```

Python os 모듈이란? 운영 체제와 상호 작용하기 위한 함수를 제공하는 라이브러리이다.

# T-SA: Result

Twitter Keyword Search API based Tweet Analysis

트위터에 '이윤혁'이라는 키워드가 들어간 트윗 작성



```
vi@vi:~$ jps
2994 DataNode
3539 ResourceManager
5927 Jps
3752 NodeManager
3289 SecondaryNameNode
2811 NameNode
vi@vi:~$
```

```
MariaDB [TWITTER]>
MariaDB [TWITTER]>
MariaDB [TWITTER]> SELECT * FROM KEYWORD_HASHTAG;
```

HCODE	START	END	TEXT
0004	0	4	이윤혁
0005	7	9	이
0007	108	112	이윤혁
0007	113	120	우후루꾸루르
0008	8	12	이윤혁
0008	13	16	진짜
0008	17	20	바보
0008	21	24	맞음
0008	25	29	이윤혁
0008	30	32	이
0008	33	35	윤
0008	36	38	혁
0008	39	44	서재익스
0009	28	31	그
0009	32	35	천재
0009	36	40	진짜다
0010	27	31	이윤혁
0010	32	36	이윤석
0010	37	41	이석준
0010	42	46	배인규
0010	47	51	서재익스
0010	52	57	서재익스
0011	18	21	천재
0011	22	26	이윤혁
0011	27	31	크지림

25 rows in set (0.00 sec)

트위터 내에 '이윤혁'이라는 키워드를 통해 해시태그를 maria 디비에 저장

# T-SA: Result

Twitter Keyword Search API based Tweet Analysis

```
MariaDB [VISUAL]> SELECT * FROM HASHTAG;
```

HASHTAG	COUNT
그놈	1
맞음	1
바보	1
배인규	1
서재익	1
서재익스	2
윤이	1
이이	2
이석준	1
이윤석	1
이윤희	6
진짜	1
진짜다	1
천재	2
크지림	1
혁	1

16 rows in set (0.01 sec)

진짜다 혁 진짜 천재 진짜 윤이 바보  
이윤희 서재익 그는  
서재익스 배인규 이윤석

가장 많이 사용된 키워드 순으로 글자 크기가 결정되어 시각화.

# T-SA: Result (Real DATA)

Twitter Keyword Search API based Tweet Analysis

## Sqoop(Import) 처리 결과

Map input records=68408

Map output records=68408

## Map/Reduce 처리 결과

Map input records=68408

Map output records=317389

Reduce input records=10924

Reduce output records=10924

## Sqoop(Export) 처리 결과

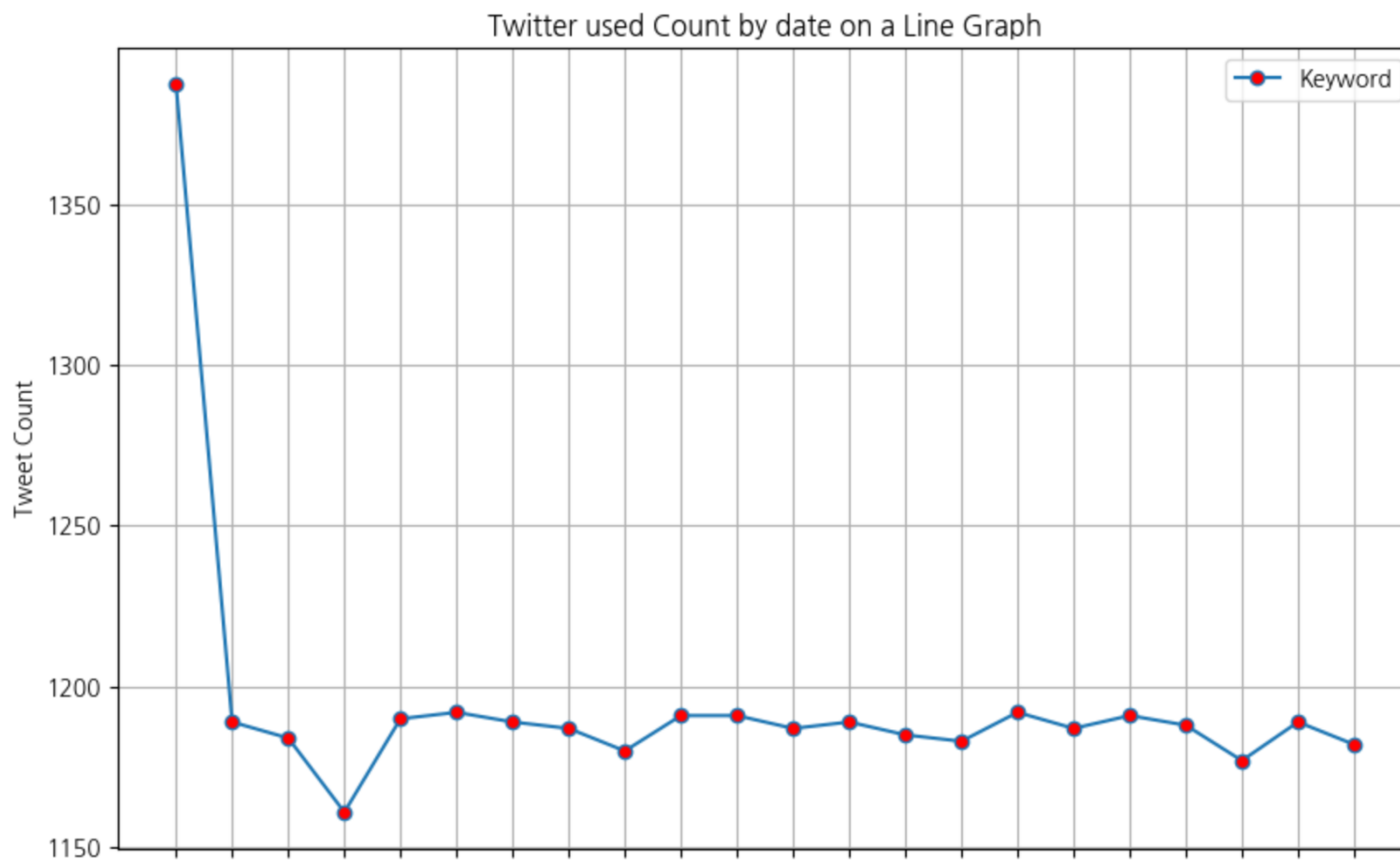
Map input records=10924

Map output records=10924

# T-SA: Result (Real DATA-Line Graph)

Twitter Keyword Search API based Tweet Analysis

19대 대통령 선거기간(2017.4.15 ~ 2017.05.09; 24일) 동안  
특정 키워드(문재인 or 홍준표 or 안철수 or 유승민 or 심상정)가 언급된 일별 트윗의 수



# Twitter Keyword Search API based Tweet Analysis

\*\*단어 사용빈도 수가 많을수록 글씨크기가 크다

(기간: 2017-04-18 ~ 2017-05-10)  
(쿼리: "문재인" OR "홍준표" OR "안철수" OR "유승민" OR "심상정")

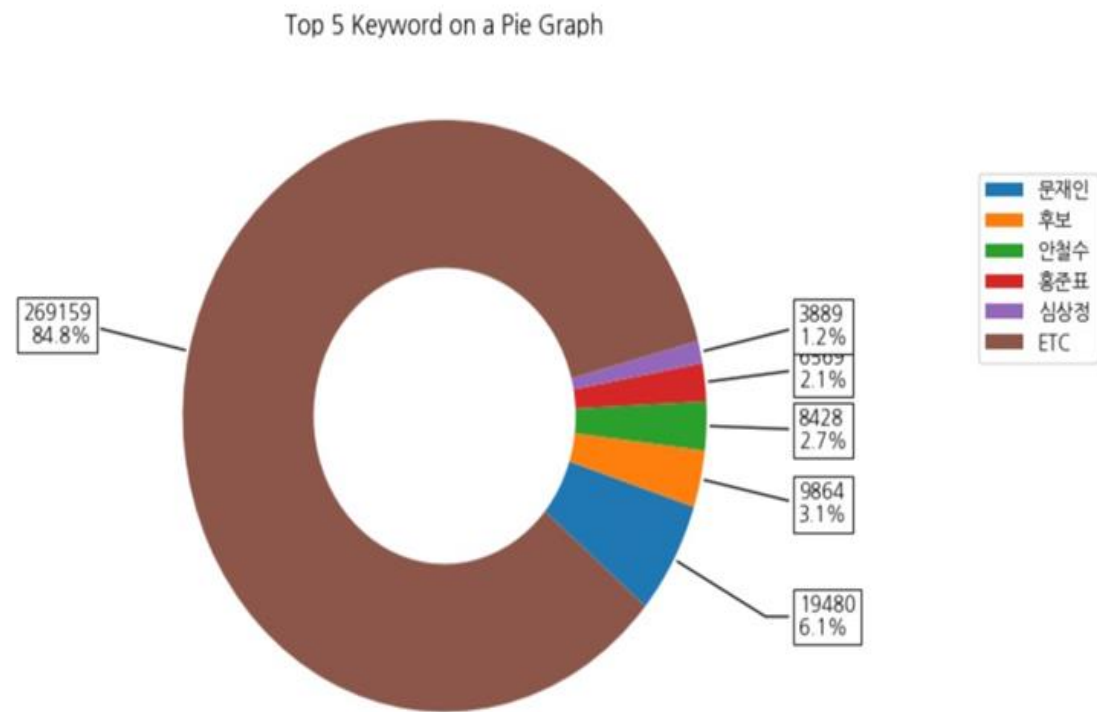
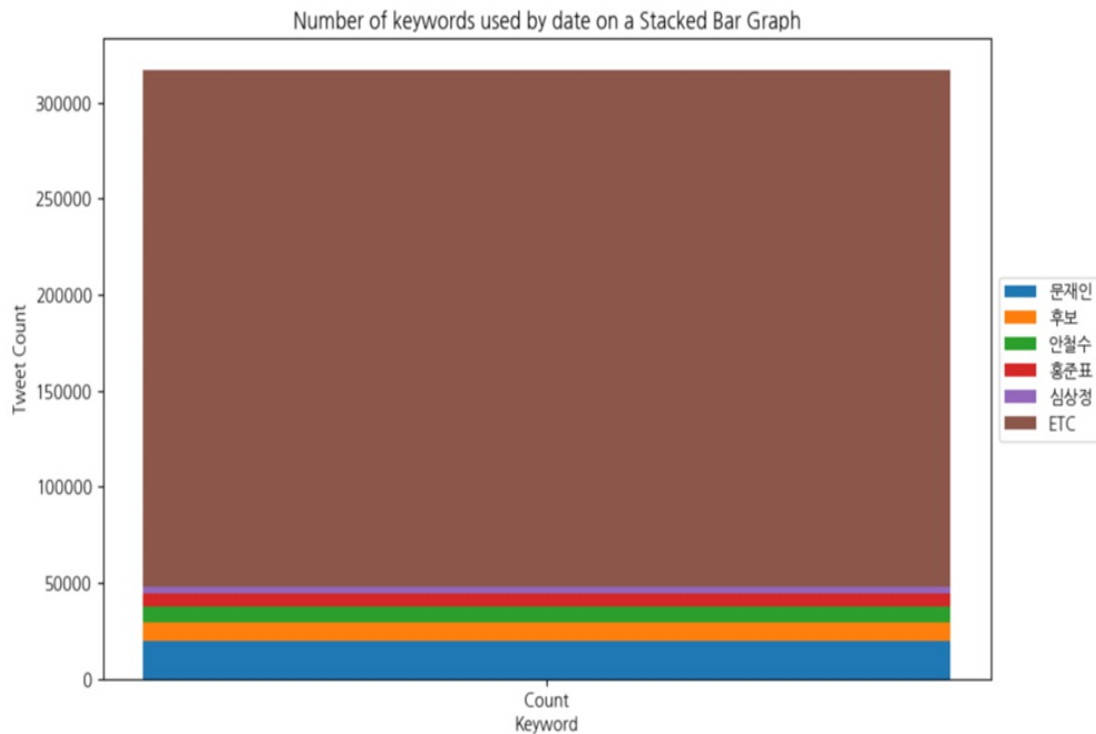




# T-SA: Result (Real DATA-Stacked Bar Graph)

Twitter Keyword Search API based Tweet Analysis

19대 대통령 선거기간(2017.4.15 ~ 2017.05.09; 24일) 동안  
특정 키워드(문재인 or 홍준표 or 안철수 or 유승민 or 심상정)가 언급된 트윗들의 단어 사용빈도  
(상위 5개, 나머지(etc))



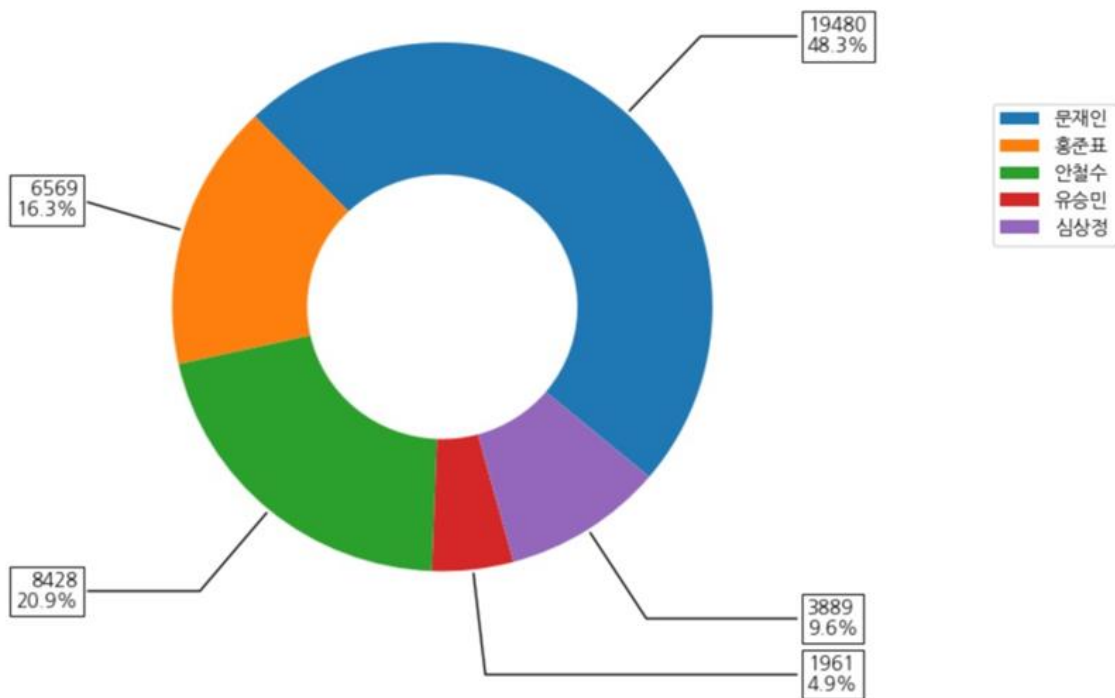


# T-SA: Result (Real DATA-Pie Graph)

Twitter Keyword Search API based Tweet Analysis

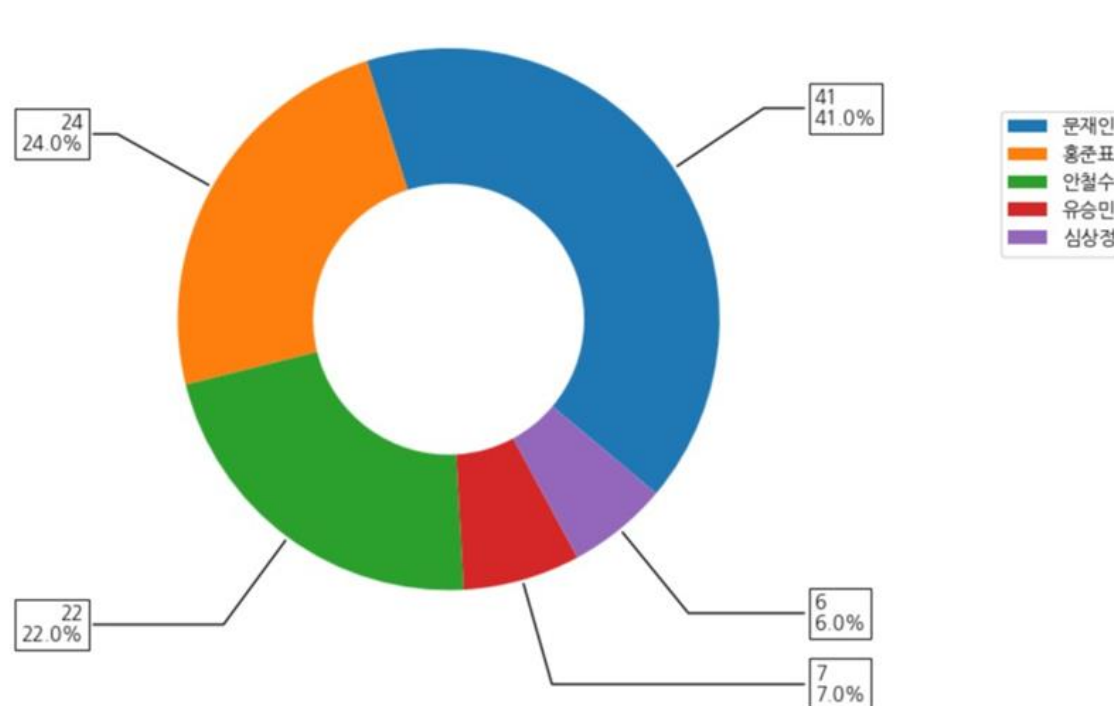
19대 대통령 선거기간(2017.4.15 ~ 2017.05.09; 24일) 동안  
특정 키워드(문재인 or 홍준표 or 안철수 or 유승민 or 심상정)가 언급된 트윗들의 백분율과 득표율

Tweet Count on a Pie Graph



<트윗들의 백분율>

Voting Rate on a Pie Graph



<실제득표율>

# T-SA: Reference

Twitter Keyword Search API based Tweet Analysis

Tweetrend, <http://tweetrend.com/>

Foller.me beta, <https://foller.me/>

중앙선거관리위원회 선거통계시스템, <http://info.nec.go.kr/>

Twitter Developer, <https://developer.twitter.com/>

Search Tweets, <https://developer.twitter.com/en/docs/tweets/search/guides/standard-operators>

정재화, 시작하세요! 하둡 프로그래밍 빅데이터 분석을 위한 하둡 기초부터 YARN까지[개정2판], 2016.05.13, 위키북스

**Thank you**

# T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis



3.6

<https://www.python.org/>



10.1.38

<https://www.python.org/>



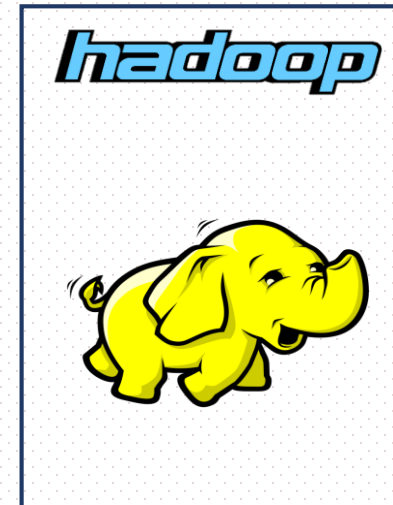
1.8.0\_191

<https://www.python.org/>



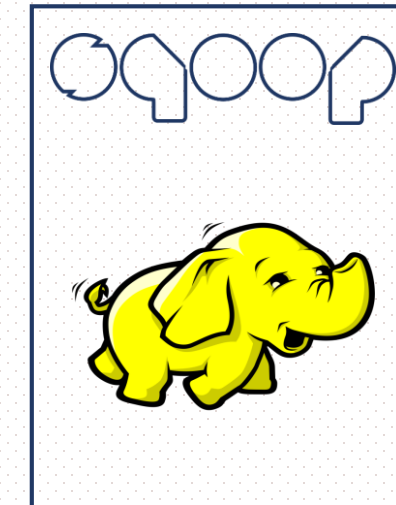
2019-03(4.11)

<https://www.python.org/>



3.2.0

<https://www.python.org/>



1.4.7

<https://www.python.org/>

# T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis



- 파이썬(Python)은 1991년 프로그래머인 귀도 반 로섬(Guido van Rossum)이 발표한 고급 프로그래밍 언어로, 플랫폼 독립적이며 인터프리터식, 객체 지향적, 동적 타이핑 대화형 언어이다.
- 파이썬은 비영리의 파이썬 소프트웨어 재단이 관리하는 개방형, 공동체 기반 개발 모델을 가지고 있다.

## Python site

<https://www.python.org/>

## Python 설치

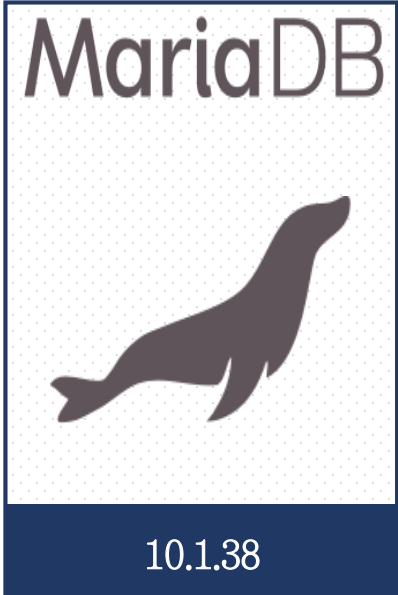
```
~$ sudo apt-get install python3
```

## Version Check

```
~$ python3 --version  
Python 3.6
```

# T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis



- MariaDB는 MySQL의 발전된 형태의 대체제로써, <https://downloads.mariadb.org/>에서 다운로드 받을 수 있으며, GPL v2 라이선스로 유지되고 있고, MariaDB 커뮤니티와 MariaDB 재단이 주축이 되어 개발되고 있다.
- MariaDB는 현재까지 최신의 MySQL과 같은 브랜치로부터 릴리즈되며, 대개의 경우 MySQL과 마찬가지로 동작한다. MySQL의 모든 명령어, 인터페이스, 라이브러리와 API가 MariaDB에도 존재한다. 또한 MariaDB로 데이터베이스를 변환할 필요도 없다.

**MariaDB site**

<https://www.python.org/>

# T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis

## MariaDB 설치

~\$ sudo apt-get install mariadb-server

## MariaDB 권한 테이블 설정

~\$ sudo mysql\_secure\_installation

## Version Check

~\$ mariadb --version

mariadb Ver 15.1 Distrib 10.1.38-MariaDB, for  
debian-linux-gnu (x86\_64) using readline 5.2

## Enter current password for root (enter for none)

→ MariaDb의 root계정은 쉘 인증이 기본적으로 설정되므로 root계정으로 실행했다면 비밀번호 없이 (Enter) 아니면 비밀번호 입력

**Set root password? [Y/n]** → 따로 패스워드를 설정하고 싶으면 Y, root 그대로 사용 할려면 n

**Remove anonymous users? [Y/n]** → 익명 사용자를 삭제 할지 여부

**Disallow root login remotely? [Y/n]** → 원격 접속으로 루트 로그인 허용 여부

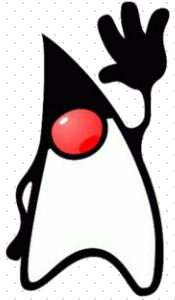
**Remove test database and access to it? [Y/n]** → 테스트 데이터베이스 삭제 여부

**Reload privilege tables now? [Y/n]** → 지금까지 작성한 권한 테이블을 적용 여부

# T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis

OpenJDK



1.8.0\_191

eclipse



2019-03(4.11)

- 이클립스(Eclipse)는 다양한 플랫폼에서 쓸 수 있으며, 자바를 비롯한 다양한 언어를 지원하는 프로그래밍 통합 개발 환경을 목적으로 시작하였으나, 현재는 OSGi(Open Service Gateway initiative)를 도입하여, 범용 응용 소프트웨어 플랫폼으로 진화하였다.

- OpenJDK는 Java SE (Standard Edition) 기반의 오픈 소스 JDK다. 2006년 Sun Micro System 은 Java를 오픈 소스화한다고 발표하였다. 그리고 그해 11월 HotSpot VM과 컴파일러를 GNU General Public License(이하 GPL)로 풀었다.

## OpenJDK 설치

```
~$ sudo apt-get install openjdk-8-jdk
```

## Version Check

```
~$ java -version
```

```
openjdk version "1.8.0_191"
```

```
OpenJDK Runtime Environment (build 1.8.0_191-8u191-b12-2ubuntu0.18.04.1-b12)
```

```
OpenJDK 64-Bit Server VM (build 25.191-b12, mixed mode)
```



# T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis



- 아파치 하둡(Apache, High-Availability Distributed Object-Oriented Platform)은 대량의 자료를 처리할 수 있는 큰 컴퓨터 클러스터에서 동작하는 분산 응용 프로그램을 지원하는 프리웨어 자바 소프트웨어 프레임 워크이다.
- 스쿱(Sqoop)은 구조화된 관계형 데이터베이스와 아파치 하둡 간의 대용량 데이터들을 효율적으로 변환하여 주는 CLI(Command-Line Interface) 애플리케이션이다.

# T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis

## Hadoop

### 1. SSH(Secure Shell) 설정

```
~$ sudo apt-get install ssh
~$ ssh-keygen -t rsa -P "" -f ~/.ssh/id_rsa
~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
~$ chmod 0600 ~/.ssh/authorized_keys
```

### 2. Protobuf 2.5.0 설치

```
~$ sudo apt-get install g++ pentium-builder
~$ cd /usr/local/
~$ sudo wget
https://github.com/google/protobuf/releases/download/v2.5.0/protobuf-2.5.0.tar.gz
~$ sudo tar xvzf protobuf-2.5.0.tar.gz
~$ sudo tar xvzf protobuf-2.5.0.tar.gz
~$ cd protobuf-2.5.0
~$ ./configure
~$ make
~$ make install
```

## 3. Hadoop3.2.0 설치

```
~$ tar xvzf hadoop-3.2.0.tar.gz
~$ ln -s hadoop3.2.0 hadoop
~$ sudo gedit ~/.bashrc
```

```
export HADOOP_HOME=/home/yunhyuck/hadoop
export PATH=$PATH:$HADOOP_HOME/bin
export PATH=$PATH:$HADOOP_HOME/sbin
export HADOOP_MAPRED_HOME=${HADOOP_HOME}
export HADOOP_COMMON_HOME=${HADOOP_HOME}
export HADOOP_HDFS_HOME=${HADOOP_HOME}
export YARN_HOME=${HADOOP_HOME}
```

```
~$ vi hadoop-env.sh
```

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
```

```
~$ vi core-site.xml
```

```
<configuration>
<property>
<name>fs.defaultFS</name>
<value>hdfs://localhost:9000</value>
</property>
</configuration>
```

## 하둡 호환성 확인

<https://hadoop.apache.org/docs/r3.2.0/hadoop-mapreduce-client/hadoop-mapreduce-client-core/dependency-analysis.html>

## 하둡 다운로드 경로

<https://www-eu.apache.org/dist/hadoop/common/hadoop-3.2.0/>

# T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis

## 3. Hadoop3.2.0 설치

~\$ vi hdfs-site.xml

```
<configuration>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.namenode.name.dir</name>
<value>/home/yunhyuck/Hadoop3_data/NameNode</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>/home/yunhyuck/Hadoop3_data/DataNode</value>
</property>
</configuration>
```

~\$ vi yarn-site.xml

```
<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.nodemanager.aux-services.mapreduce_shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
</configuration>
```

~\$ vi mapred-site.xml

```
<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
<property>
<name>mapreduce.admin.user.env</name>
<value>HADOOP_MAPRED_HOME=$HADOOP_COMMON_HOME</value>
</property>
<property>
<name>yarn.app.mapreduce.am.env</name>
<value>HADOOP_MAPRED_HOME=$HADOOP_COMMON_HOME</value>
</property>
</configuration>
```

~\$ bin/hdfs namenode -format

```
2019-03-29 20:07:15,364 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = yunhyuc/127.0.1.1
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 3.2.0
STARTUP_MSG: classpath =
```

~\$ sbin/start-all.sh

# T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis

## Sqoop

```
~$ wget http://apache.tt.co.kr/sqoop/1.4.7/sqoop-1.4.7.bin\_hadoop-2.6.0.tar.gz
```

```
~$ tar xvzf sqoop-1.4.7.bin_hadoop-2.6.0.tar.gz
```

```
~$ ln -s sqoop-1.4.7.bin_hadoop-2.6.0 sqoop
```

```
~$ sudo gedit ~/.bashrc
```

```
export SQOOP_HOME=/home/vi/sqoop
export PATH=$PATH:$SQOOP_HOME/bin
```

```
~$ source ~/.bashrc
```

```
~/sqoop/conf$ cp sqoop-env-template.sh sqoop-env.sh
```

```
#Set path to where bin/hadoop is available
export HADOOP_COMMON_HOME=/home/vi(계정이름)/hadoop
#Set path to where hadoop-*-core.jar is available
export HADOOP_MAPRED_HOME=/home/vi(계정이름)/hadoop
```

## ~/sqoop/conf\$ sqoop

Warning: /home/vi/sqoop/./hbase does not exist! HBase imports will fail.  
Please set \$HBASE\_HOME to the root of your HBase installation.

Warning: /home/vi/sqoop/./hcatalog does not exist! HCatalog jobs will fail.  
Please set \$HCAT\_HOME to the root of your HCatalog installation.

Warning: /home/vi/sqoop/./accumulo does not exist! Accumulo imports will fail.

Please set \$ACCUMULO\_HOME to the root of your Accumulo installation.

Warning: /home/vi/sqoop/./zookeeper does not exist! Accumulo imports will fail.

Please set \$ZOOKEEPER\_HOME to the root of your Zookeeper installation.

/home/vi/hadoop/libexec/hadoop-functions.sh: 줄 2364:

HADOOP\_ORG.APACHE.SQOOP.SQOOP\_USER: bad substitution

/home/vi/hadoop/libexec/hadoop-functions.sh: 줄 2459:

HADOOP\_ORG.APACHE.SQOOP.SQOOP\_OPTS: bad substitution

**Try 'sqoop help' for usage. # 이렇게 뜬다면 설치 완료.**

## 스쿱 사이트

<https://sqoop.apache.org/>

## 스쿱 다운로드 경로

[http://apache.tt.co.kr/sqoop/1.4.7/sqoop-1.4.7.bin\\_hadoop-2.6.0.tar.gz](http://apache.tt.co.kr/sqoop/1.4.7/sqoop-1.4.7.bin_hadoop-2.6.0.tar.gz)

# T-SA: Implementation

Twitter Keyword Search API based Tweet Analysis

## MariaDB에서 Hadoop으로 데이터 저장

```
sqoop import --connect jdbc:mysql://localhost/TWITTER --username T-SA --password 1234 --table  
KEYWORD_HASHTAG --columns TEXT --target-dir hdfs://localhost:9000/user/vi/HASHTAG_INPUT -m 1
```

connect	jdbc:DB종류://IP주소/DB이름
username	DB 계정
password	DB 암호
table	데이터를 가져올 테이블
columns	테이블에서 가져올 컬럼 리스트
target-dir	저장될 HDFS 디렉토리 경로

# T-SA: Implementation

Twitter Keyword Search API based Tweet Analysis

## Map/Reduce : 자연어 처리를 통한 키워드 개수 저장

```
yarn jar /home/vi/hadoop/jar/HashtagCount.jar HashtagCount /user/vi/HASHTAG_INPUT/part-m-00000  
HASHTAG_OUTPUT
```

```
// -----자연어 처리 부분-----
```

```
// 코모란 객체 생성 DEFAULT_MODEL 기본 사전 사용
```

```
Komoran komoran = new Komoran(DEFAULT_MODEL.FULL);
```

```
// 사용자 사전 경로 추가.(사용자 명사 정의 가능)
```

```
komoran.setUserDic("/home/vi/eclipse-workspace/KeywordCount/src/dic.user");
```

```
// 읽어온 단어 분석
```

```
KomoranResult analyzeResultList = komoran.analyze(token);
```

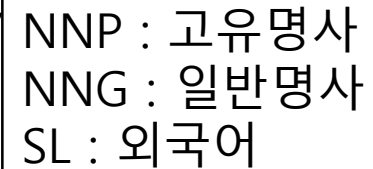
```
// tokens 리스트 정의 후, 명사에 대해 분류하여 적재.
```

```
List<String> tokens = analyzeResultList.getMorphesByTags("NNP","NNG","SL");
```

```
// 요소들을 읽어오기 위한 Iterator 생성 후, tokens 내용 적재.
```

```
Iterator<String> itr = tokens.iterator();
```

```
// -----
```



NNP : 고유명사  
NNG : 일반명사  
SL : 외국어

# T-SA: Implementation

Twitter Keyword Search API based Tweet Analysis

## Hadoop에서 MariaDB로 데이터 저장

```
sqoop export --connect jdbc:mysql://localhost/VISUAL --username T-SA --password 1234 --table  
HASHTAG --export-dir hdfs://localhost:9000/user/vi/HASHTAG_OUTPUT/part-r-00000 --columns  
HASHTAG,COUNT --input-fields-terminated-by "\t"
```

connect	jdbc:DB종류://IP주소/DB이름
username	DB 계정
password	DB 암호
table	데이터를 가져올 테이블
export-dir	데이터를 가져올 HDFS 디렉토리 경로
columns	테이블에서 매핑될 컬럼 리스트
input-fields-terminated-by	구분자

# T-SA: Implementation

Twitter Keyword Search API based Tweet Analysis

## Visualization.py : 워드 클라우드를 통한 데이터 시각화

# 한글폰트 적용

```
path = '/home/vi/.local/lib/python3.6/site-packages/matplotlib/mpl_data/fonts/ttf/NanumBarunGothicUltraLight.ttf'
```

```
fontprop = fm.FontProperties(fname=path, size=18)
```

```
# 워드 클라우드 설정
```

```
wc=WordCloud(font_path=path,background_color='white',max_words=2000)
```

```
wc=wc.generate_from_frequencies(b)
```

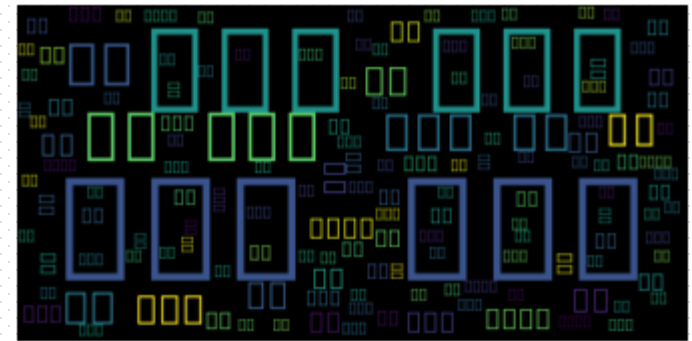
```
# 시각화 이미지 설정
```

```
plt.figure(figsize=(12,12))
```

```
plt.imshow(wc, interpolation='bilinear')
```

```
plt.axis('off')
```

```
plt.show()
```



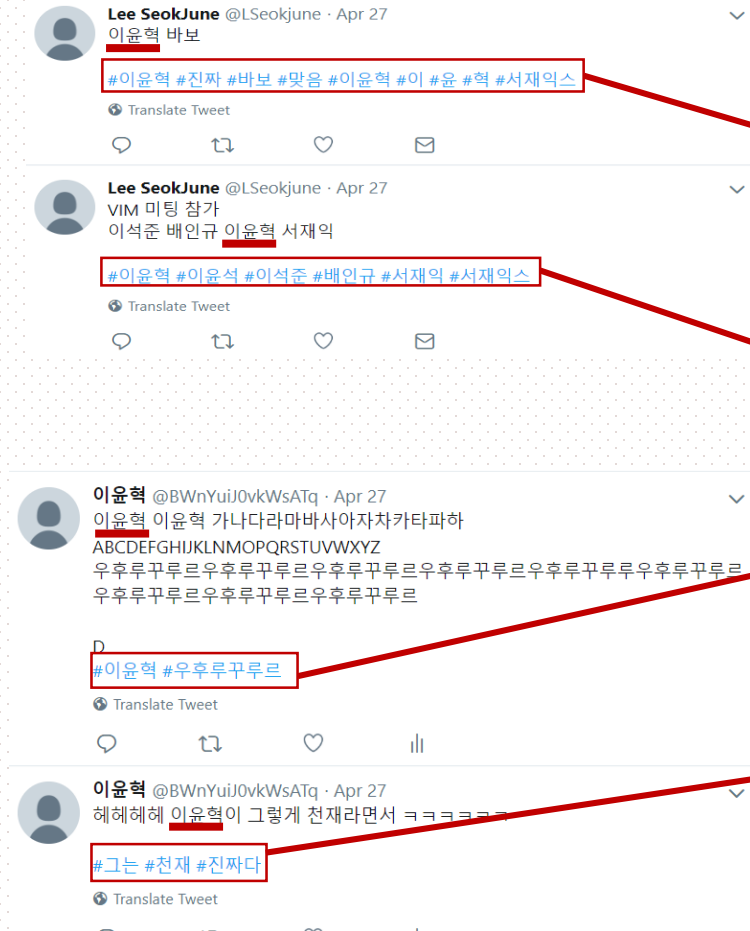
한글 폰트를 적용하지 않았을 경우



# T-SA: Result

Twitter Keyword Search API based Tweet Analysis

트위터에 '이윤혁'이라는 키워드가 들어간 트윗 작성



```
vi@vi:~$ jps
2994 DataNode
3539 ResourceManager
5927 Jps
3752 NodeManager
3289 SecondaryNameNode
2811 NameNode
vi@vi:~$
```

```
MariaDB [TWITTER]>
MariaDB [TWITTER]>
MariaDB [TWITTER]> SELECT * FROM KEYWORD_HASHTAG;
+-----+-----+-----+-----+
| HCODE | START | END | TEXT |
+-----+-----+-----+-----+
| 0004 | 0 | 4 | 이윤혁 |
| 0005 | 7 | 9 | 이 |
| 0007 | 108 | 112 | 이윤혁 |
| 0007 | 113 | 120 | 우후루꾸루르 |
| 0008 | 8 | 12 | 이윤혁 |
| 0008 | 13 | 16 | 진짜 |
| 0008 | 17 | 20 | 바보 |
| 0008 | 21 | 24 | 맞음 |
| 0008 | 25 | 29 | 이윤혁 |
| 0008 | 30 | 32 | 이 |
| 0008 | 33 | 35 | 윤 |
| 0008 | 36 | 38 | 혁 |
| 0008 | 39 | 44 | 서재익스 |
| 0009 | 28 | 31 | 그는 |
| 0009 | 32 | 35 | 천재 |
| 0009 | 36 | 40 | 진짜다 |
| 0010 | 27 | 31 | 이윤혁 |
| 0010 | 32 | 36 | 이윤석 |
| 0010 | 37 | 41 | 이석준 |
| 0010 | 42 | 46 | 배인규 |
| 0010 | 47 | 51 | 서재익스 |
| 0010 | 52 | 57 | 서재익스 |
| 0011 | 18 | 21 | 천재 |
| 0011 | 22 | 26 | 이윤혁 |
| 0011 | 27 | 31 | 크지림 |
+-----+-----+-----+-----+
25 rows in set (0.00 sec)
```

트위터 내에 '이윤혁'이라는 키워드를 통해 해시태그를 maria 디비에 저장

# T-SA: Result

Twitter Keyword Search API based Tweet Analysis

```
MariaDB [VISUAL]> SELECT * FROM HASHTAG;
```

HASHTAG	COUNT
그놈	1
맞음	1
바보	1
배인규	1
서재익	1
서재익스	2
윤이	1
이이	2
이석준	1
이윤석	1
이윤혁	6
진짜	1
진짜다	1
천재	2
크지림	1
혁	1

16 rows in set (0.01 sec)

진짜다 천재 진짜 윤이 바보  
혁  
이윤혁 서재익 그는  
크지림 이석준 맞음  
서재익스 배인규 이윤석

가장 많이 사용된 키워드 순으로 글자 크기가 결정되어 시각화.

