T-SA:

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

과 목	산학캡스톤디자인1(2019-1학기)
담 당 교 수	정현숙 교수님
팀 명	브이아이(VI)
발 표 자	서재익
발 표 일 자	2019.04.18.

Project Github URL: https://github.com/SeokJune/BigData_VI_T-SA/

T-SA: Contents

Twitter Keyword Search API based Tweet Analysis

T-SA: Team Introduction

Twitter Keyword Search API based Tweet Analysis

Name	Lee SeokJune
Student ID	20165072
Cell Phone	010-4020-5717
E-mail	op2se1@gmail.com
Major Lang	Java
GitHub	https://github.com/SeokJune
Part	- MariaDB 환경 구축 및 관리 - Hadoop(Map)구현 - 문서 작성 및 수정

Name	Lee YunHyuck
Student ID	20165062
Cell Phone	010-4220-5134
E-mail	leeyh5134@naver.com
Major Lang	Python
GitHub	https://github.com/yunhyuck
Part	- Hadoop3 환경 구축 - Hadoop, DB 연동 구현 - Sqoop 환경 구축 - Hadoop(Reduce)구현

Name	Bae InGyu
Student ID	20165073
Cell Phone	010-4679-4968
E-mail	happykkk789@naver.com
Major Lang	Python
GitHub	https://github.com/BaeInGyu
Part	- Python, DB 연동 구현 - Visualization 구현

Name	Seo JaeIck
Student ID	20144773
Cell Phone	010-2460-7617
E-mail	nero8879@naver.com
Major Lang	Python
GitHub	https://github.com/nero8879
Part	- Twitter API 구현 - Visualization 구현

T-SA: Purpose of Development

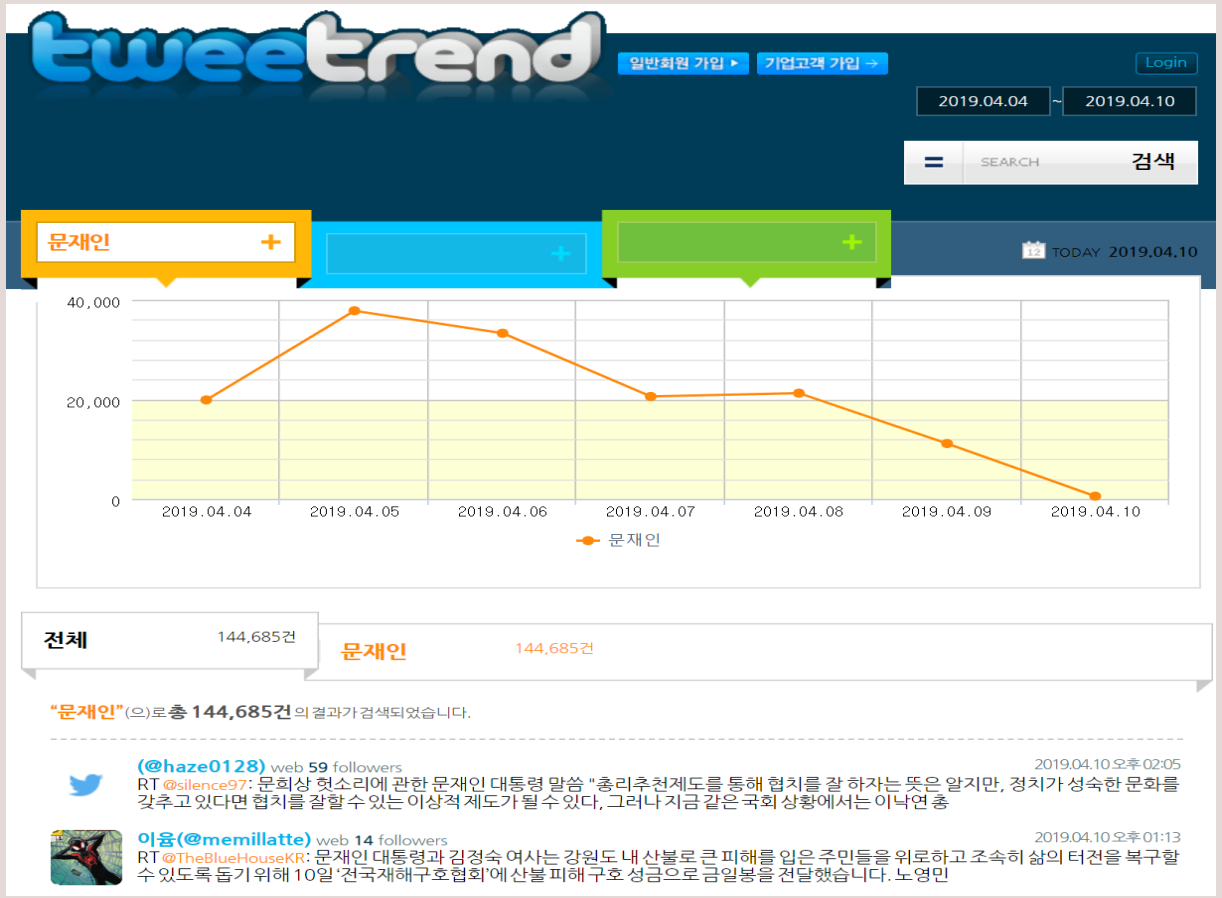
Twitter Keyword Search API based Tweet Analysis

대한민국 지역 및 특정 기간에 사용된 키워드 트렌드 분석

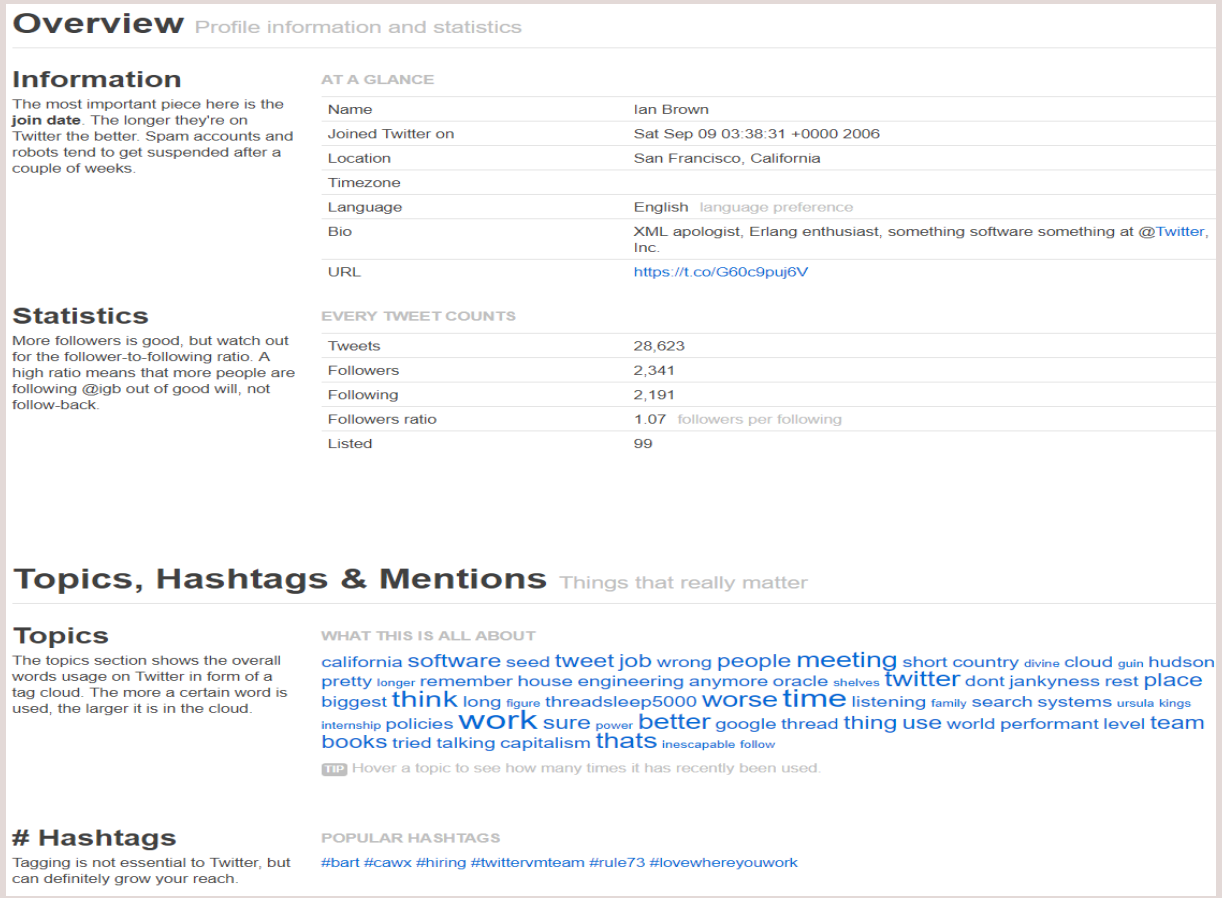
특정 인물의 트윗 스타일 분석

T-SA: Related Works_W06

Twitter Keyword Search API based Tweet Analysis



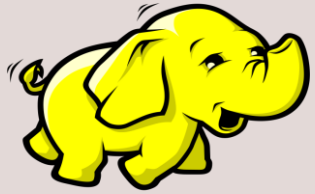
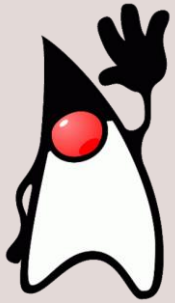
- http://tweetrend.com/
- 기간 별 특정 키워드에 해당하는 트윗 검색 기능



- https://foller.me/
- 사용자에 따른 정보, 주 토픽과 해시태그 내용 표시

T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis

ubuntu	python	MariaDB	hadoop	OpenJDK	eclipse
					
18.04.2 LTS	3.6.8	10.1.38	3.2.0	1.8.0_191	2019-03(4.11)

T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis

Ubuntu is an *open source software operating system* that runs from the desktop, to the cloud, to all your internet connected things.

Ubuntu Site:
- https://www.ubuntu.com/

T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis

Python features a *dynamic type system* and *automatic memory management*.
It supports multiple programming paradigms, including *object-oriented*, *functional* and *procedural*.

Python Stie:
- https://www.python.org/

민형기, 파이썬으로 데이터 주무르기, 2017.12.29, 비제이퍼블릭

파이썬으로 데이터 주무르기 저자의 블로그 중 파이썬 목록
- https://pinkwink.kr/category/Software/Python

T-SA: Development Environment

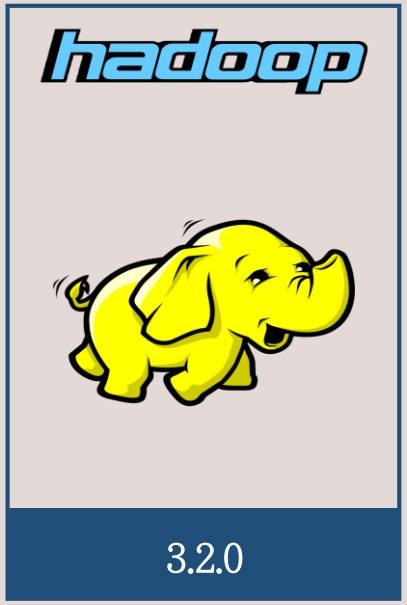
Twitter Keyword Search API based Tweet Analysis

MariaDB is an open source *relational database management system (RDBMS)*. Based on the same source code as MySQL, follow the *GPL v2 license*.

MariaDB Stie:
- https://mariadb.com/kb/ko/mariadb

T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis



Hadoop software library is a framework that allows for the *distributed processing of large data sets* across clusters of computers using simple programming models.

Hadoop Site:
- https://hadoop.apache.org/

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

T-SA: Development Environment_W04

Twitter Keyword Search API based Tweet Analysis

Using Hadoop Ecosysytem

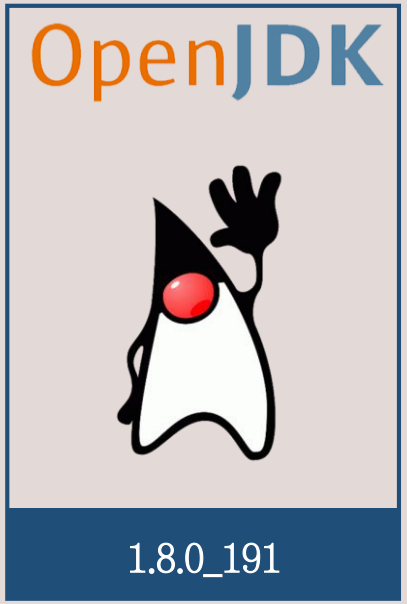
Sqoop
- RDBMS에서 데이터 수집
- 배치 처리 후 RDBMS에 데이터 저장

HDFS
- 분산 데이터 저장

Map/Reduce
- 분산 데이터 배치 처리

T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis

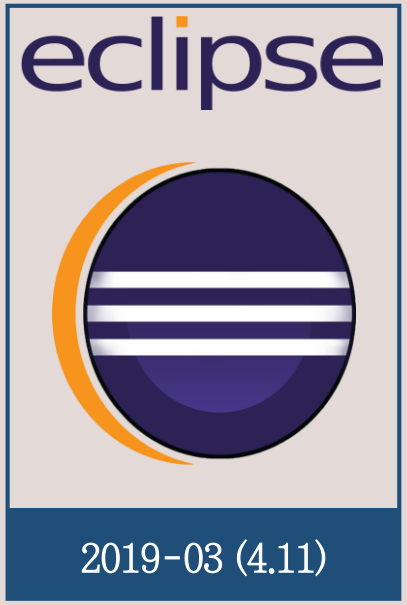


OpenJDK is a ***free and open-source*** implementation of the Java Platform, Standard Edition. also produces the ***virtual machine***, the ***Java Class Library***, the ***Java compiler*** and etc.

OpenJDK Site:
- https://openjdk.java.net/

T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis



Eclipse is an *integrated development environment(IDE)* used in computer programming, is the *most widely used Java IDE*, may also be used to develop applications in other programming languages via various plug-ins

Eclipse Site:
- https://www.eclipse.org/

T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis

Twitter API

5.0

Twitter API furnish developer with *publish and analyze of Tweets*, *optimize ads*, and *create unique customer experiences*.

Twitter Developer Site:
- https://developer.twitter.com

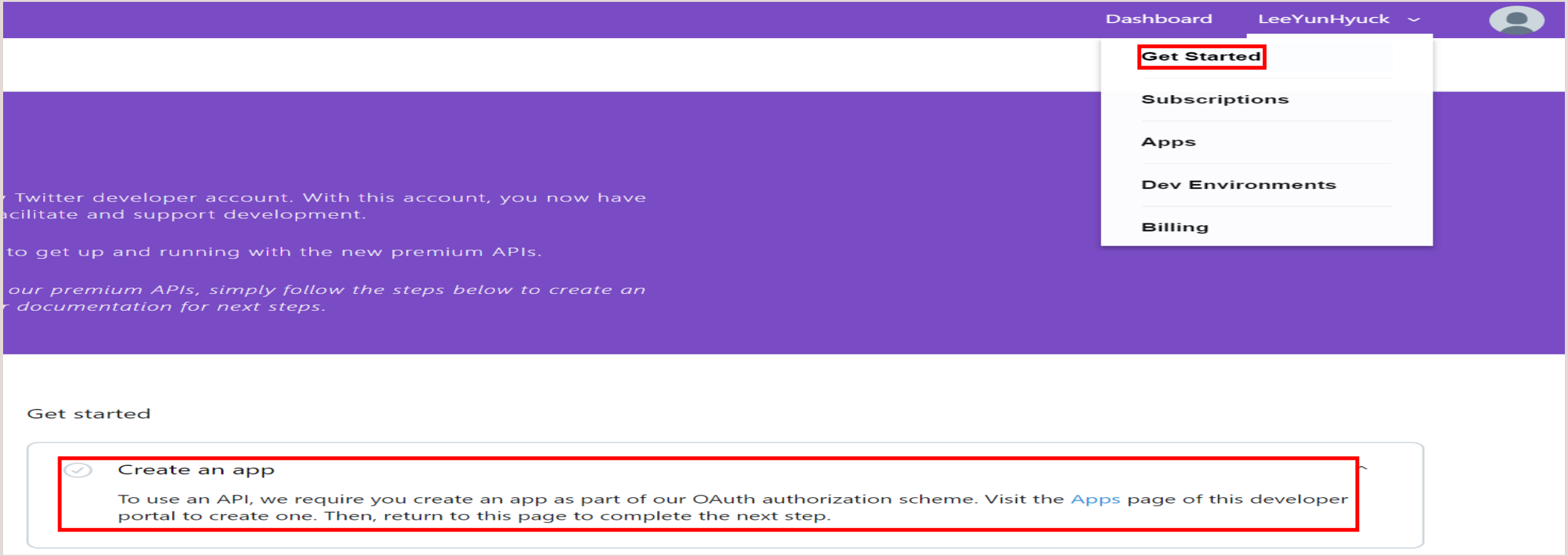
Tweepy Site:
- http://www.tweepy.org

Twitter Analysis Site:
- http://tweetrend.com/
- https://foller.me/

T-SA: Development Environment_W04

Twitter Keyword Search API based Tweet Analysis

Importing Twitter API Key



Get Started 를 누르면 아래와 같이 Twitter API를 사용하기 위한 인증키 발급을 받을 수 있는 목록을 받을 수 있다.
Create an app 을 제외한 나머지는 유료 이므로 무료로 사용하기 위한 인증키를 발급 받는다.

T-SA: Development Environment_W04

Twitter Keyword Search API based Tweet Analysis

Importing Twitter API Key

Tell us how this app will be used (required)

This field is only visible to Twitter employees. Help us understand how your app will be used. What will it enable you and your customers to do?

이 앱의 사용방법은 사용자의 키워드를 분석하여, 해당 키워드에 대해 분석을 통해 얻을 수 있는 정보들에 대해 시각화 하는 것에 목적이 있습니다.

Must be 100 characters or longer

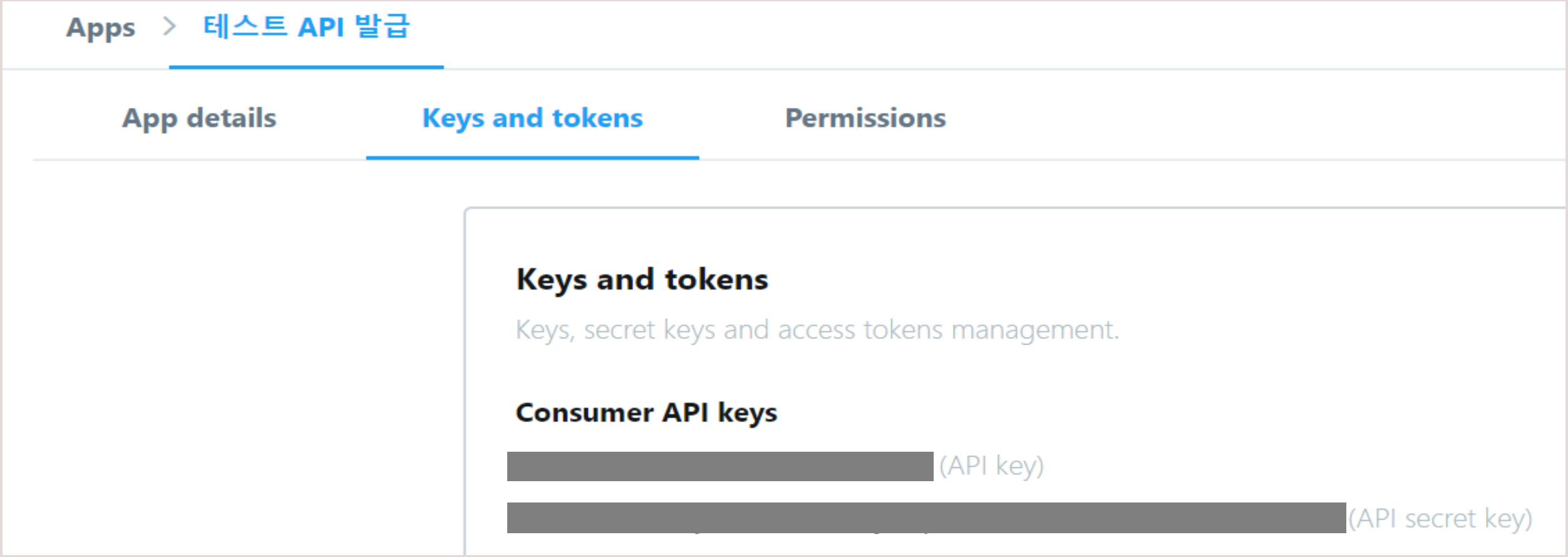
Minimum characters: **100**

앱의 이름, 앱의 설명, 사용하는 주소, 앱의 사용 방법에 대한 필수적인 요소를 작성합니다.

T-SA: Development Environment_W04

Twitter Keyword Search API based Tweet Analysis

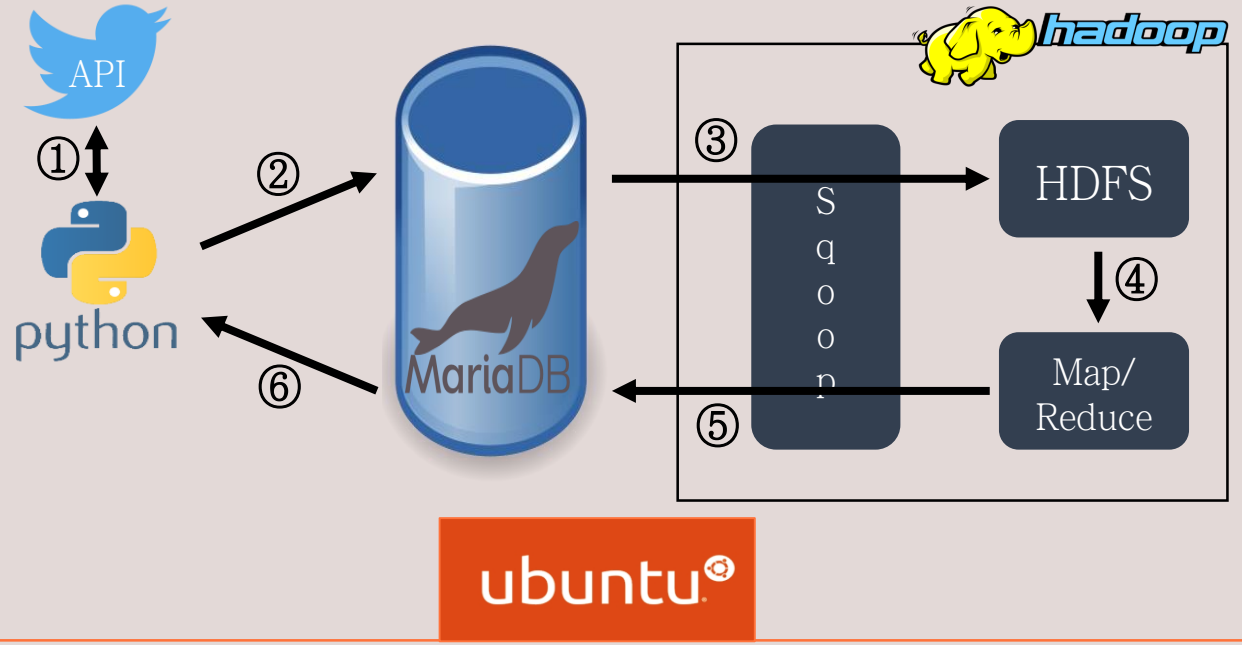
Importing Twitter API Key



개인 정보의 Apps를 통해 본인이 사전에 작성한 제목을 통해API 키가 발급이 된 것을 확인 할 수 있다.

T-SA: Program Flowchart_W05

Twitter Keyword Search API based Tweet Analysis



TwitterAPI: To import data from Twitter

Python: Provides tweepy which is twitterAPI, Visualization of Data

MariaDB: Open source R-DBMS, Based on the same source as MySQL

Hadoop: Distributed storge and Processing of big data, Pseudo-distributed

Sqoop: For BigData Transfers between Hadoop and MariaDB

① Twitter API를 이용한 데이터 크롤링

② 크롤링 된 데이터를 MariaDB에 저장

③ Sqoop을 이용해 HDFS에 분산 저장 처리

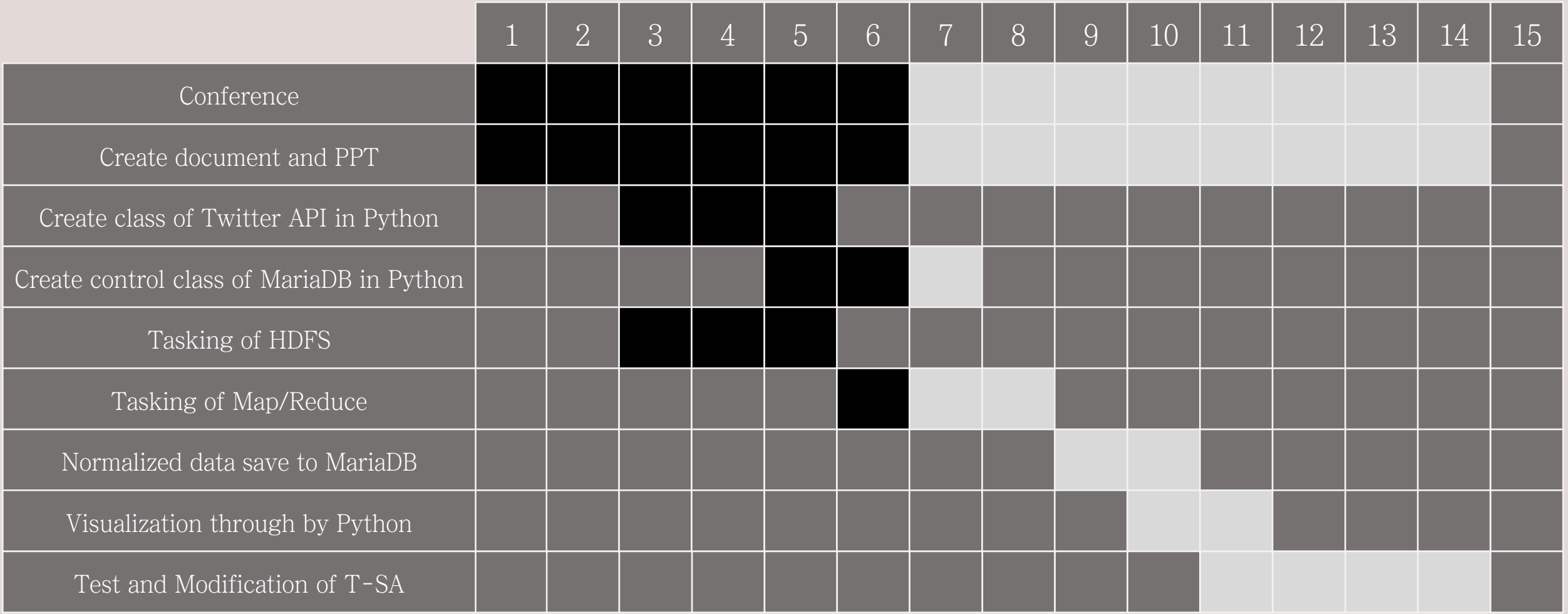
④ Map/Reduce를 통한 분산 데이터 배치 처리

⑤ Sqoop을 이용해 MariaDB에 저장

⑥ 저장된 데이터를 Python에 로드 및
시각화 라이브러리를 이용한 데이터 시각화

T-SA: Development Schedule

Twitter Keyword Search API based Tweet Analysis



T-SA: Weekly Progress_W06

Twitter Keyword Search API based Tweet Analysis

Lee SeokJune

문서 작성 및 수정

TwitterAPI Python 구현

Lee YunHyuck

Map/Reduce 자연어 처리
구현 및 오류 수정 작업

문서 작성 및 수정

Bae InGyu

Python, MariaDB 의
DML(Insert, RowCheck)
구현 및 전체 오류 수정 작업

Seo JaeIck

발표 준비

T-SA: Weekly Progress_W06

Twitter Keyword Search API based Tweet Analysis

Hadoop
(+Sqoop)

T-SA: Weekly Progress_W06

Twitter Keyword Search API based Tweet Analysis

KeywordCount 를 하기 위해 형태소의 품사를 구분 해야 할 필요성 발견

정규화, 토큰화, 어근화 과정을 통해 맵을 수행 예정.

가동	1
가짜뉴스에	1
강원	1
강조하는	1
것들..	2
것을	1
게	2
경수찡	1
고마하고	1
고민정	1
고발을	1
공정하고	1
공지	1

정규화

한국어를 처리하는 예시입니닼ㅋㅋㅋㅋㅋ ->
한국어를 처리하는 예시입니다 ㅋㅋ

토큰화

한국어를 처리하는 예시입니다 ㅋㅋ ->
한국어Noun, 를Josa, 처리Noun, 하는Verb, 예시Noun, 입니다Adjective, ㅋㅋ KoreanParticle

어근화 (입니다. -> 이다)

한국어를 처리하는 예시입니다 ㅋㅋ ->
한국어Noun, 를Josa, 처리Noun, 하다Verb, 예시Noun, 이다Adjective, ㅋㅋ KoreanParticle

Python(DB)

T-SA: Weekly Progress_W06

Twitter Keyword Search API based Tweet Analysis

TwitterAPI 를 통해 받아온 JSON을 테이블 형식으로 정리

_json	hashtags	metadata	user
created_at	Key	Key	Key
id	text	iso_language_code	id
id_str	indices	result_type	id_str
text			name
truncated			screen_name
hashtags			location
metadata			description
source			url
in_reply_to_status_id			
in_reply_to_status_id_str			protected
in_reply_to_user_id			followers_count
in_reply_to_user_id_str			friends_count
in_reply_to_screen_name			listed_count
user			created_at
geo			favourites_count
coordinates			utc_offset
place			time_zone
contributors			geo_enabled
is_quote_status			verified
retweet_count			statuses_count
favorite_count			lang
favorited			contributors_enabled
retweeted			is_translator
~~possibly_sensitive~~			is_translation_enabled
lang			following
			follow_request_sent
			notifications
			translator_type

```
def result_Keyword(self, tweets):
keywordJson = [] # Table(KEYWORD_JSON)
keywordHashtags = [] # Table(KEYWORD_HASHTAGS)
keywordMetadata = [] # Table(KEYWORD_METADATA)
keywordUser = [] # Table(KEYWORD_J_USER)
keyNum = 0
#tweet 하나씩 가져오기 ----------------------------------------------
for t in tweets:
# _json 선택 ----------------------------------------------------
json = t._json
# Table(KEYWORD_JSON) ---------------------------------------
keywordJson.append([json['created_at'],
json['id'],
json['id_str'],
json['text'],
…
json['lang']])
```

테이블 형식으로 정리된 데이터들을 가지고 MariaDB에 저장할 예정.

T-SA: Github_W06

Twitter Keyword Search API based Tweet Analysis

Project Github URL: https://github.com/SeokJune/BigData_VI_T-SA/

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Insights

- Pulse
- Contributors
- Community
- Traffic
- Commits
- Code frequency
- Dependency graph
- Network
- Forks

March 16, 2019 – April 16, 2019

Period: 1 month

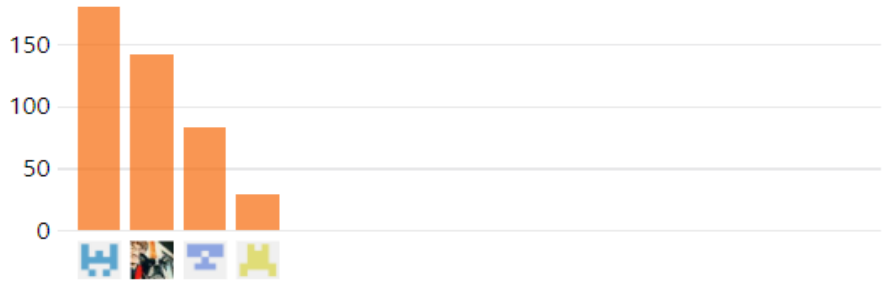
Overview

0 Active Pull Requests

0 Active Issues

0 Merged Pull Requests	0 Proposed Pull Requests	0 Closed Issues	0 New Issues

Excluding merges, **4 authors** have pushed **432 commits** to master and **433 commits** to all branches. On master, **102 files** have changed and there have been 3,328 **additions** and 0 **deletions**.



Q & A

Thank you.