T-SA:

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

과목	산학캡스톤디자인1(2019-1학기)
담당교수	정현숙 교수님
팀명	브이아이 (VI)
발표자	이석준
발표일자	2019.04.11.

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

T-SA: Contents

Twitter Keyword Search API based Tweet Analysis

1. T-SA: Team Introduction
2. T-SA: Purpose of Development
3. T-SA: Development Environment
4. T-SA: Program Flowchart
5. T-SA: Development Schedule
6. T-SA: Weekly Progress
7. T-SA: Github

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

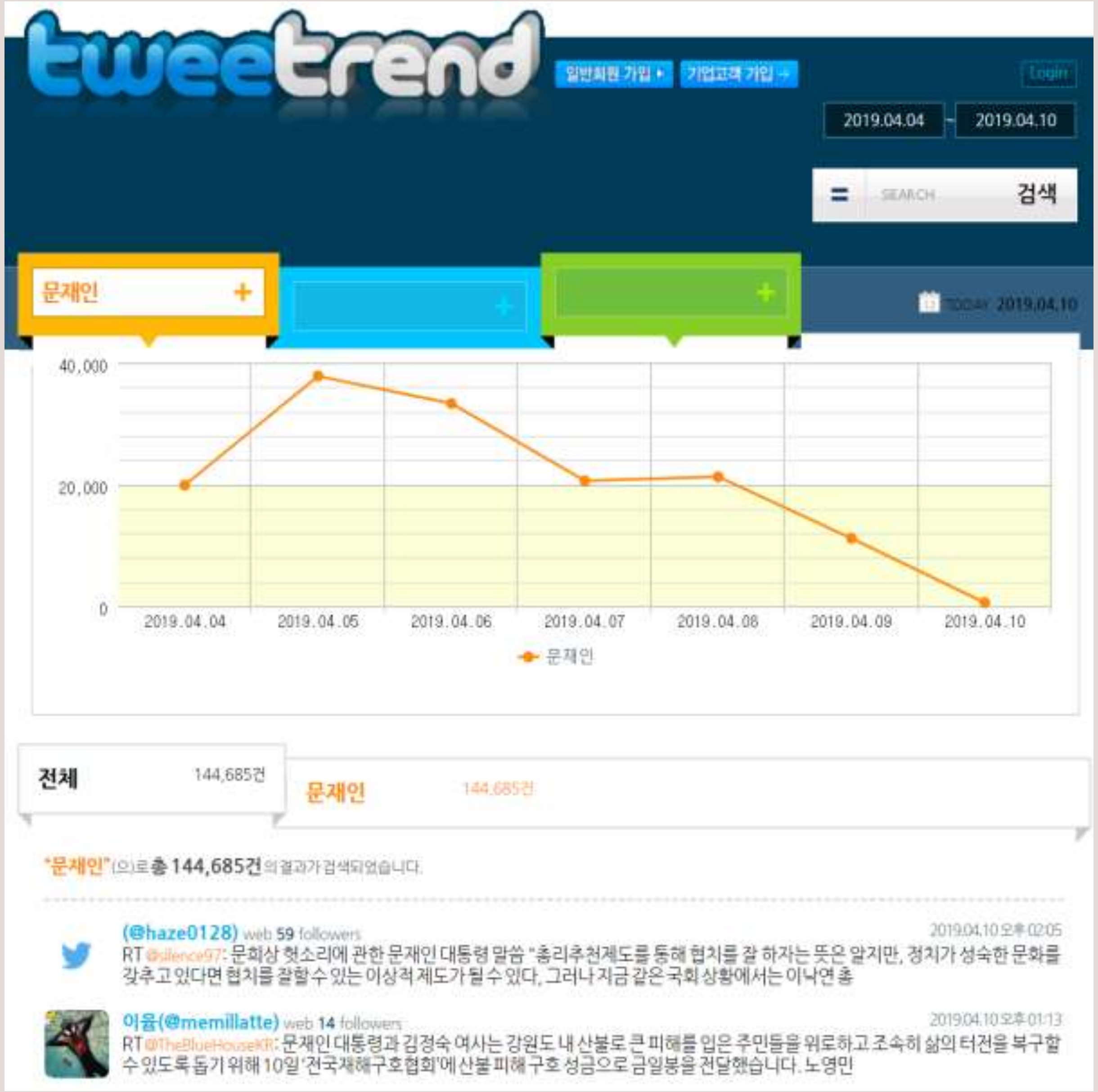
Twitter Keyword Search API based Tweet Analysis

대한민국 지역 및 특정 기간에 사용된 키워드 트렌드 분석

특정 인물의 트윗 스타일 분석

T-SA: Purpose of Development_W05

Twitter Keyword Search API based Tweet Analysis

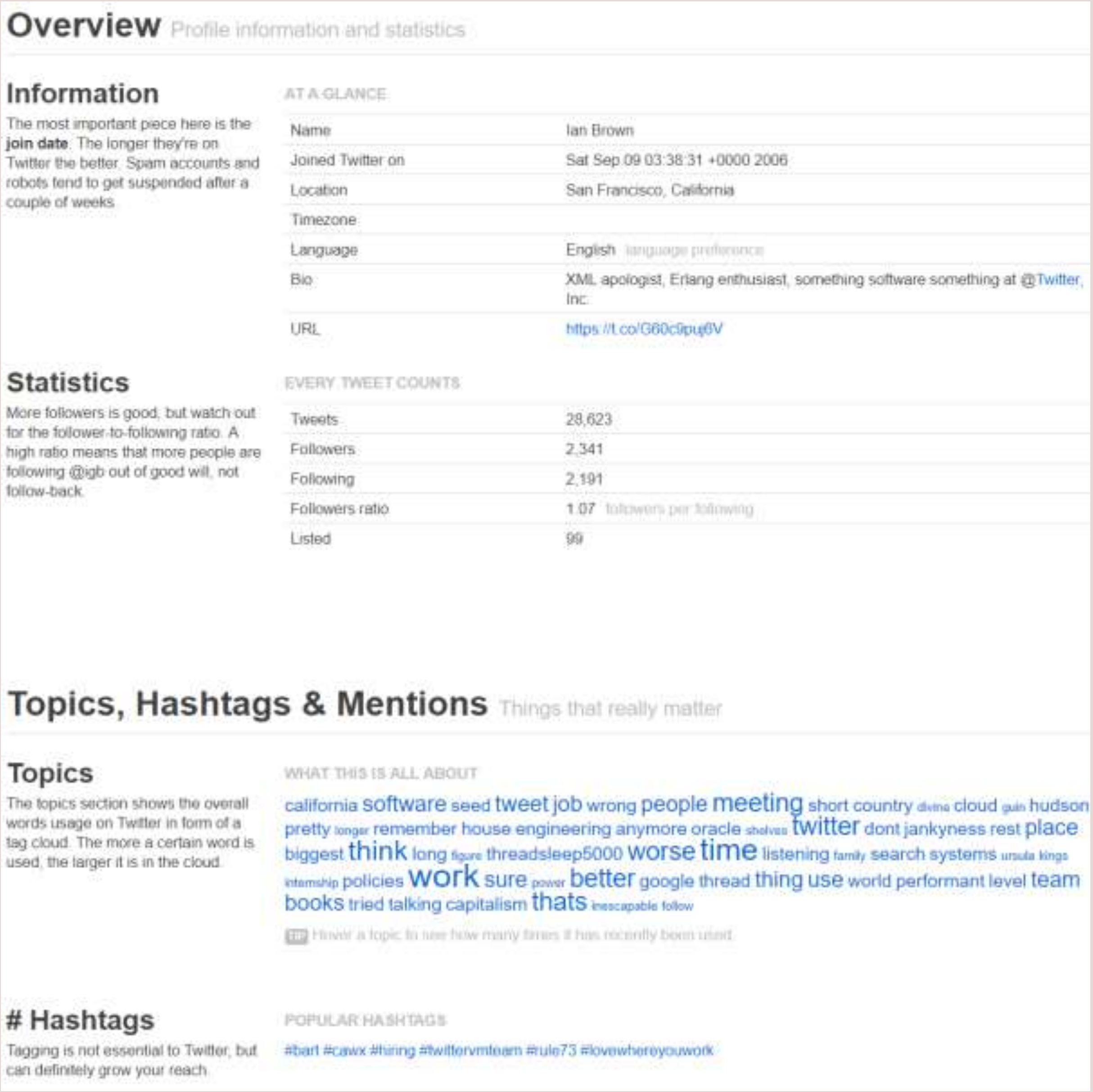


전체	144,685건
문재인	144,685건

"문재인"(으)로 총 144,685건의 결과가 검색되었습니다.

(@haze0128) web 59 followers 2019.04.10 오후 02:05
RT @silence97: 문희상 헛소리에 관한 문재인 대통령 말씀 "총리추천제도를 통해 협치를 잘 하자는 뜻은 알지만, 정치가 성숙한 문화를 갖추고 있다면 협치를 잘할 수 있는 이상적 제도가 될 수 있다, 그러나 지금 같은 국회 상황에서는 이낙연 총

이윰(@memillatte) web 14 followers 2019.04.10 오후 01:13
RT @TheBlueHouseKR: 문재인 대통령과 김정숙 여사는 강원도 내 산불로 큰 피해를 입은 주민들을 위로하고 조속히 삶의 터전을 복구할 수 있도록 돕기 위해 10일 '전국재해구호협회'에 산불 피해 구호 성금으로 금일봉을 전달했습니다. 노영민



Overview Profile information and statistics

Information

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.

AT A GLANCE

Name	Ian Brown
Joined Twitter on	Sat Sep 09 03:38:31 +0000 2006
Location	San Francisco, California
Timezone	
Language	English language preference
Bio	XML apologist, Erlang enthusiast, something software something at @Twitter, Inc.
URL	https://t.co/G60c9puj6V

Statistics

More followers is good, but watch out for the follower-to-following ratio. A high ratio means that more people are following @igb out of good will, not follow-back.

EVERY TWEET COUNTS

Tweets	28,623
Followers	2,341
Following	2,191
Followers ratio	1.07 followers per following
Listed	99

Topics, Hashtags & Mentions Things that really matter

Topics

The topics section shows the overall words usage on Twitter in form of a tag cloud. The more a certain word is used, the larger it is in the cloud.

WHAT THIS IS ALL ABOUT

california software seed tweet job wrong people meeting short country divina cloud guin hudson pretty longer remember house engineering anymore oracle shelves twitter dont jankyness rest place biggest think long figure threadsleep5000 worse time listening family search systems ursula kings internship policies work sure power better google thread thing use world performant level team books tried talking capitalism thats inescapable follow

TIP Hover a topic to see how many times it has recently been used

Hashtags

Tagging is not essential to Twitter, but can definitely grow your reach

POPULAR HASHTAGS

#bart #cawx #hiring #twittervmteam #rule73 #lovewhereyouwork

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

18.04.2 LTS

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

3.6.8

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

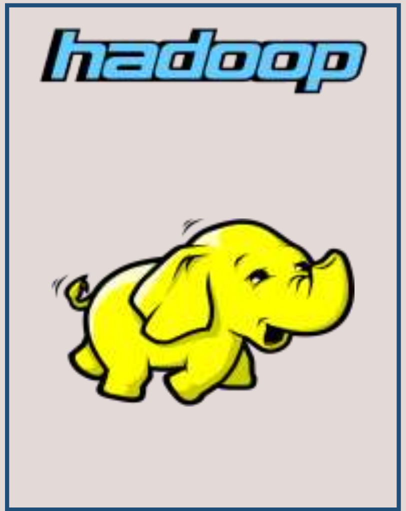
10.1.38

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



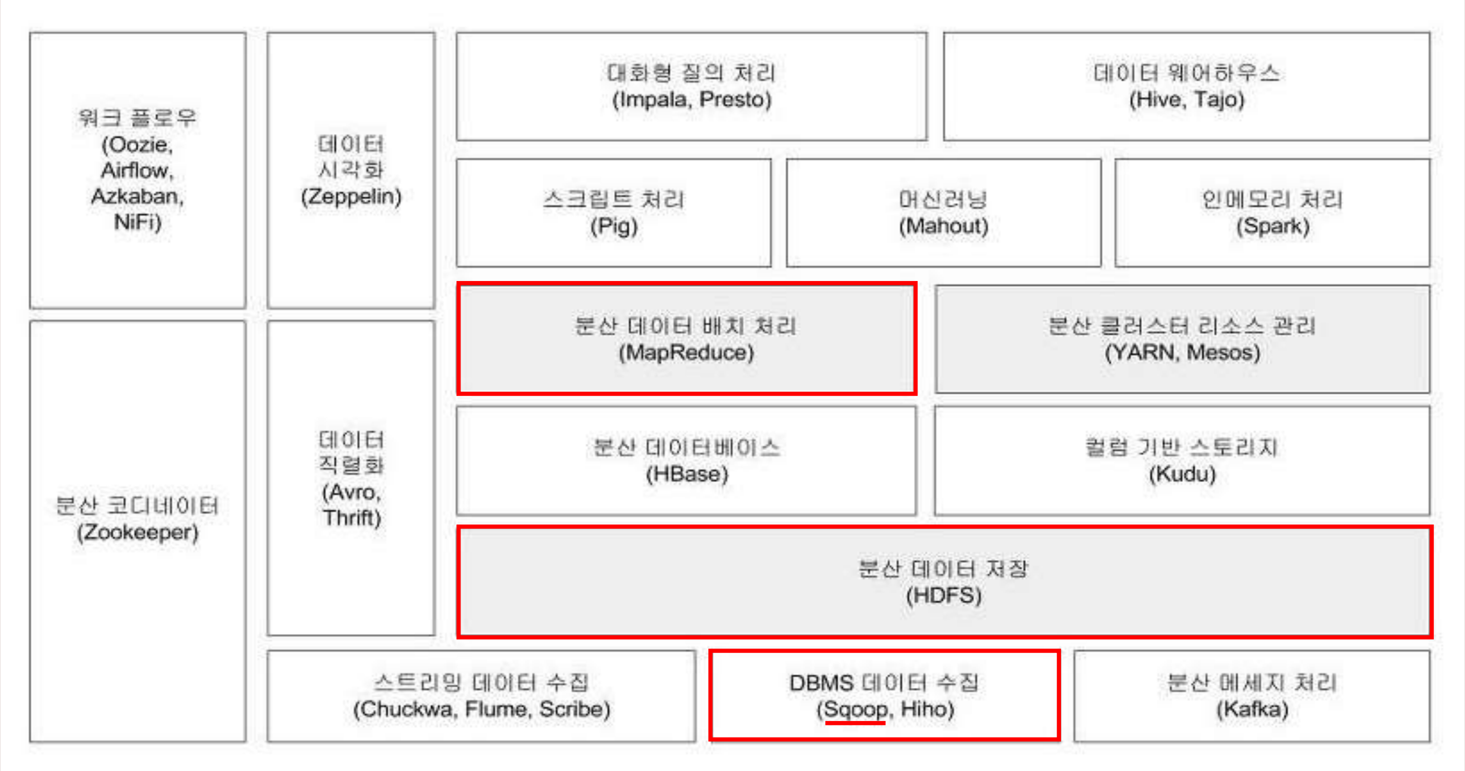
3.2.0

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

Using Hadoop Ecosysytem



워크 플로우 (Oozie, Airflow, Azkaban, NiFi)

데이터 시각화 (Zeppelin)

대화형 질의 처리 (Impala, Presto)

데이터 웨어하우스 (Hive, Tajo)

스크립트 처리 (Pig)

머신러닝 (Mahout)

인메모리 처리 (Spark)

분산 데이터 배치 처리 (MapReduce)

분산 클러스터 리소스 관리 (YARN, Mesos)

분산 코디네이터 (Zookeeper)

데이터 직렬화 (Avro, Thrift)

분산 데이터베이스 (HBase)

컬럼 기반 스토리지 (Kudu)

분산 데이터 저장 (HDFS)

스트리밍 데이터 수집 (Chuckwa, Flume, Scribe)

DBMS 데이터 수집 (Sqoop, Hiho)

분산 메세지 처리 (Kafka)

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

1.8.0_191

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

2019-03 (4.11)

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

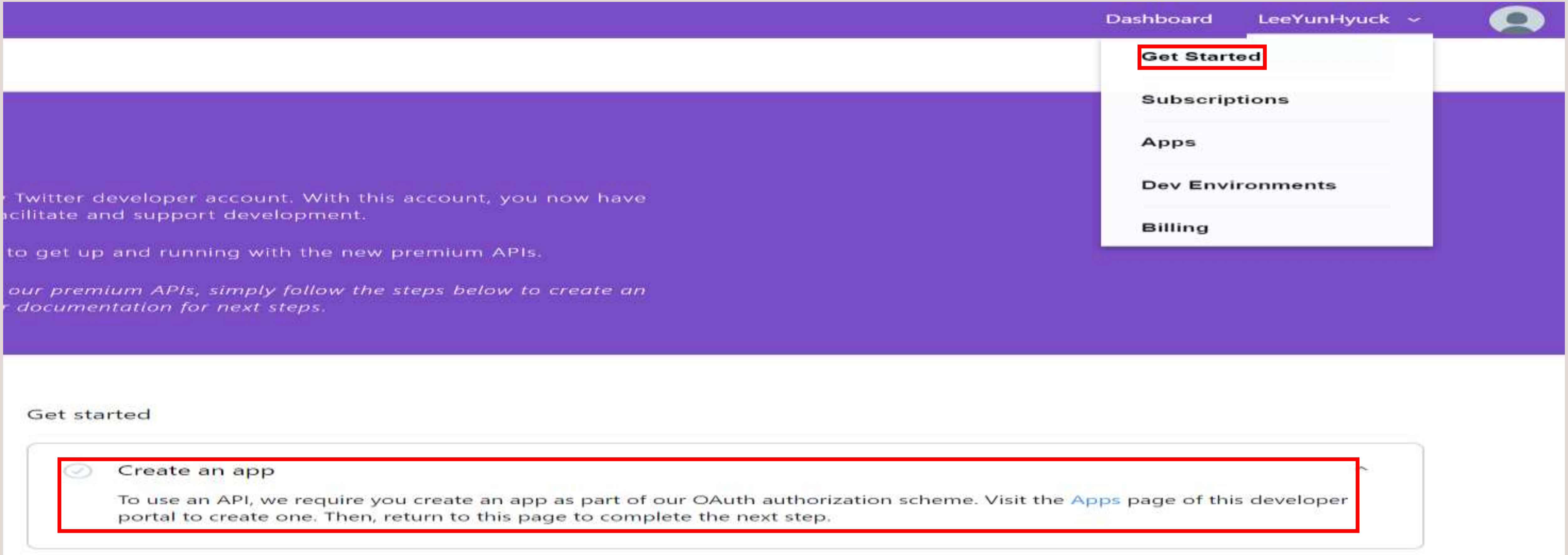
Twitter API furnish developer with ***publish and analyze of Tweets***, ***optimize ads***, and ***create unique customer experiences***.

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

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

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

Importing Twitter API Key



Twitter developer account. With this account, you now have
acilitate and support development.

to get up and running with the new premium APIs.

*our premium APIs, simply follow the steps below to create an
documentation for next steps.*

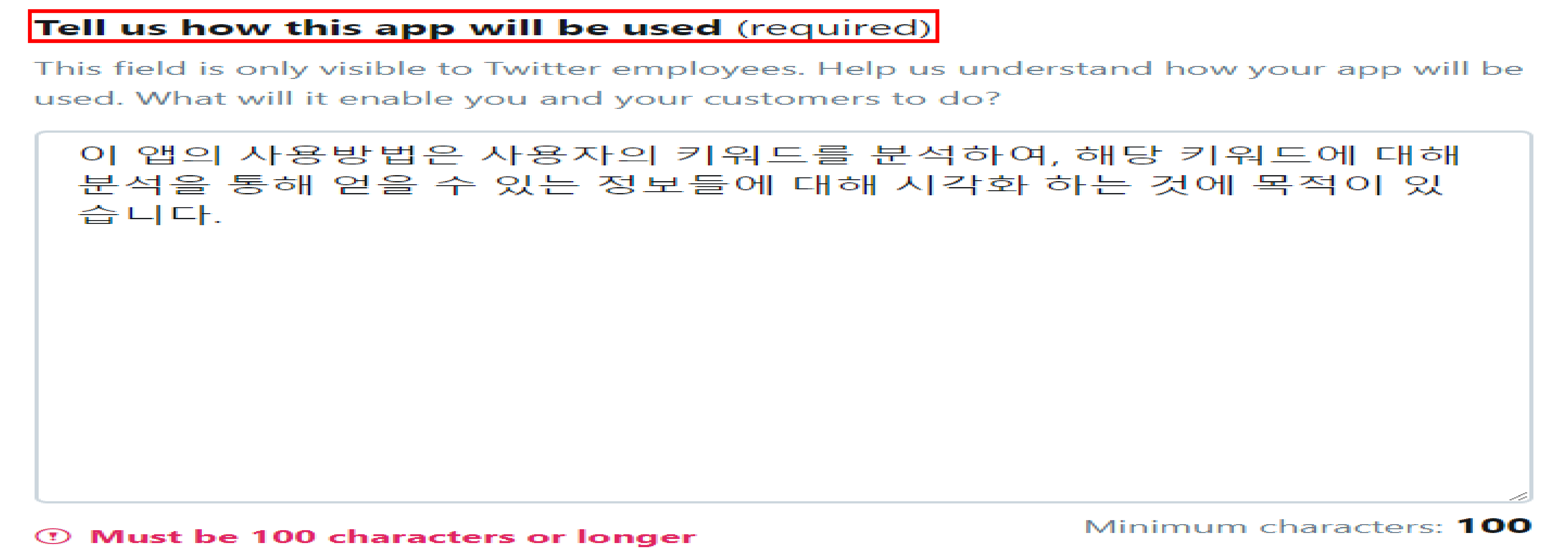
Get started

Create an app

To use an API, we require you create an app as part of our OAuth authorization scheme. Visit the Apps page of this developer portal to create one. Then, return to this page to complete the next step.

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.

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**

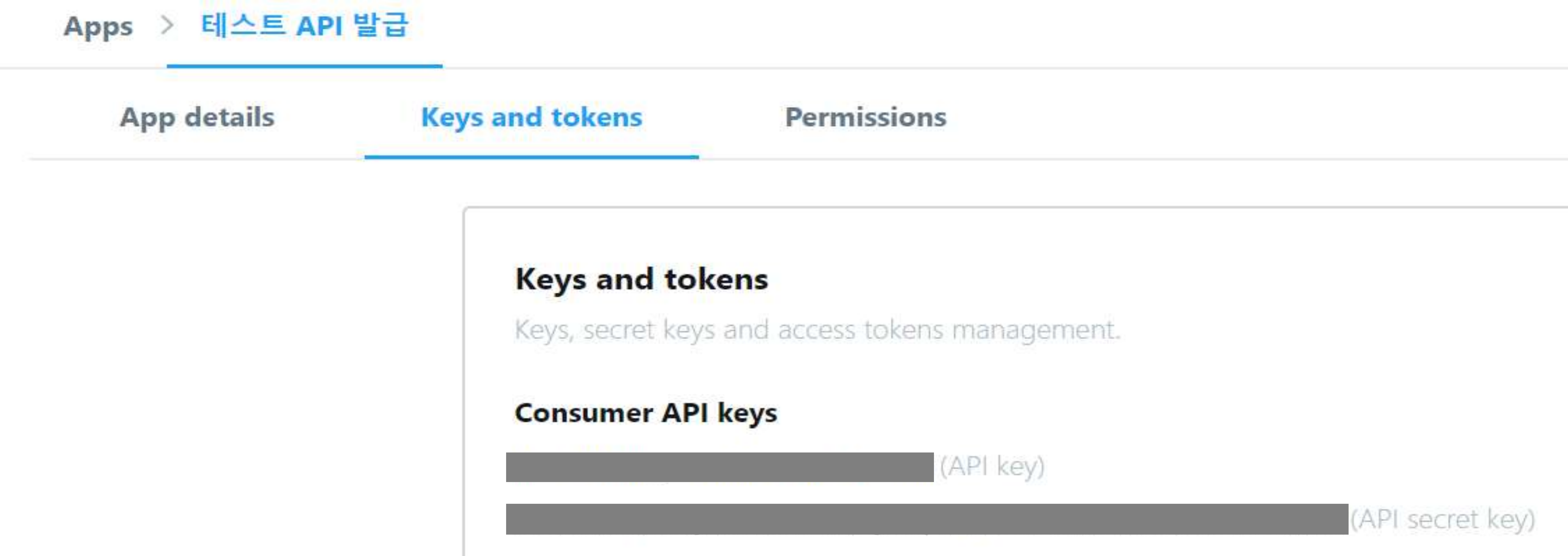
Minimum characters: **100**

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



Apps > 테스트 API 발급

App details | Keys and tokens | Permissions

Keys and tokens

Keys, secret keys and access tokens management.

Consumer API keys

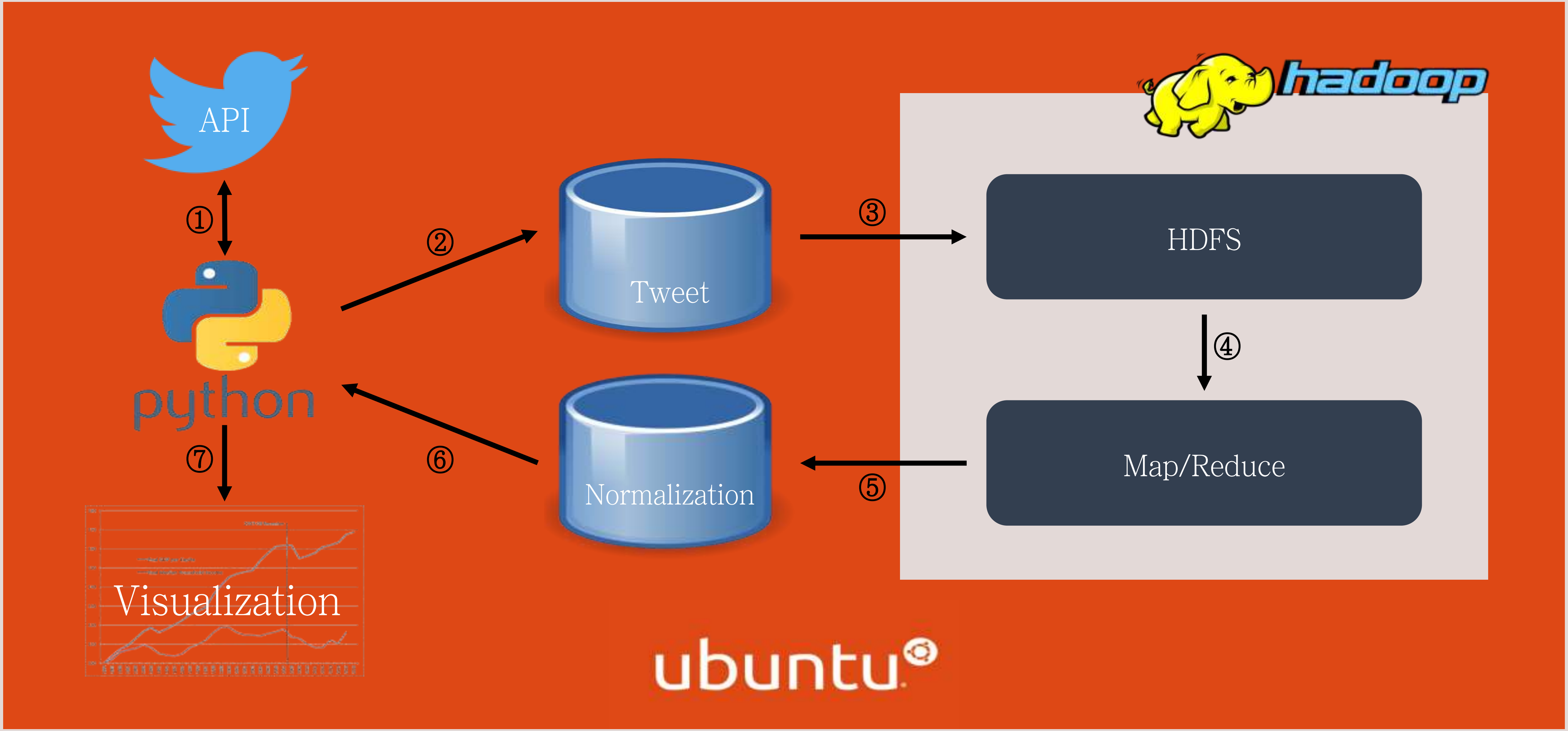
[redacted] (API key)

[redacted] (API secret key)

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

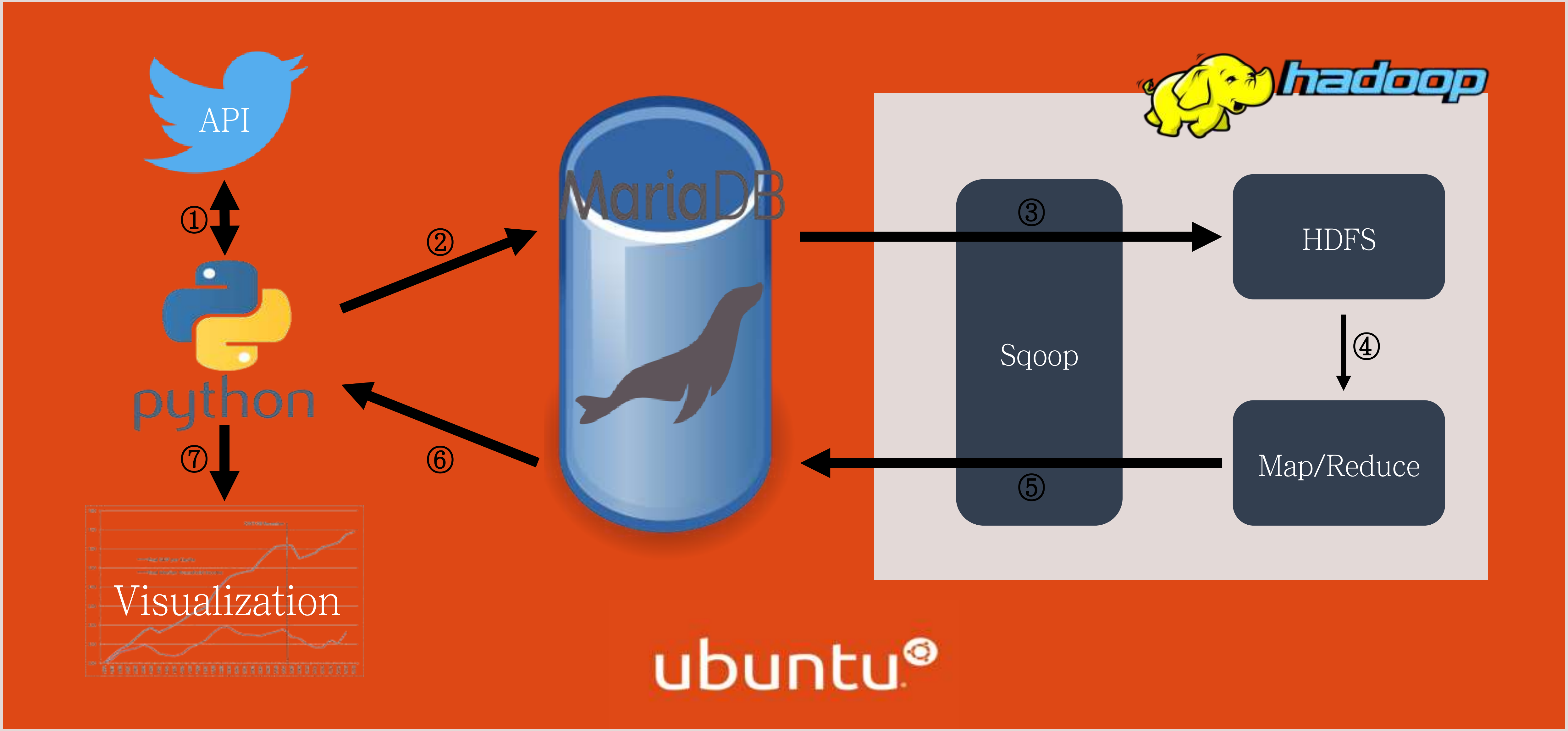
T-SA: Program Flowchart

Twitter Keyword Search API based Tweet Analysis



T-SA: Program Flowchart_W04

Twitter Keyword Search API based Tweet Analysis



API

①

python

②

MariaDB

hadoop

③

Sqoop

HDFS

④

Map/Reduce

⑤

⑥

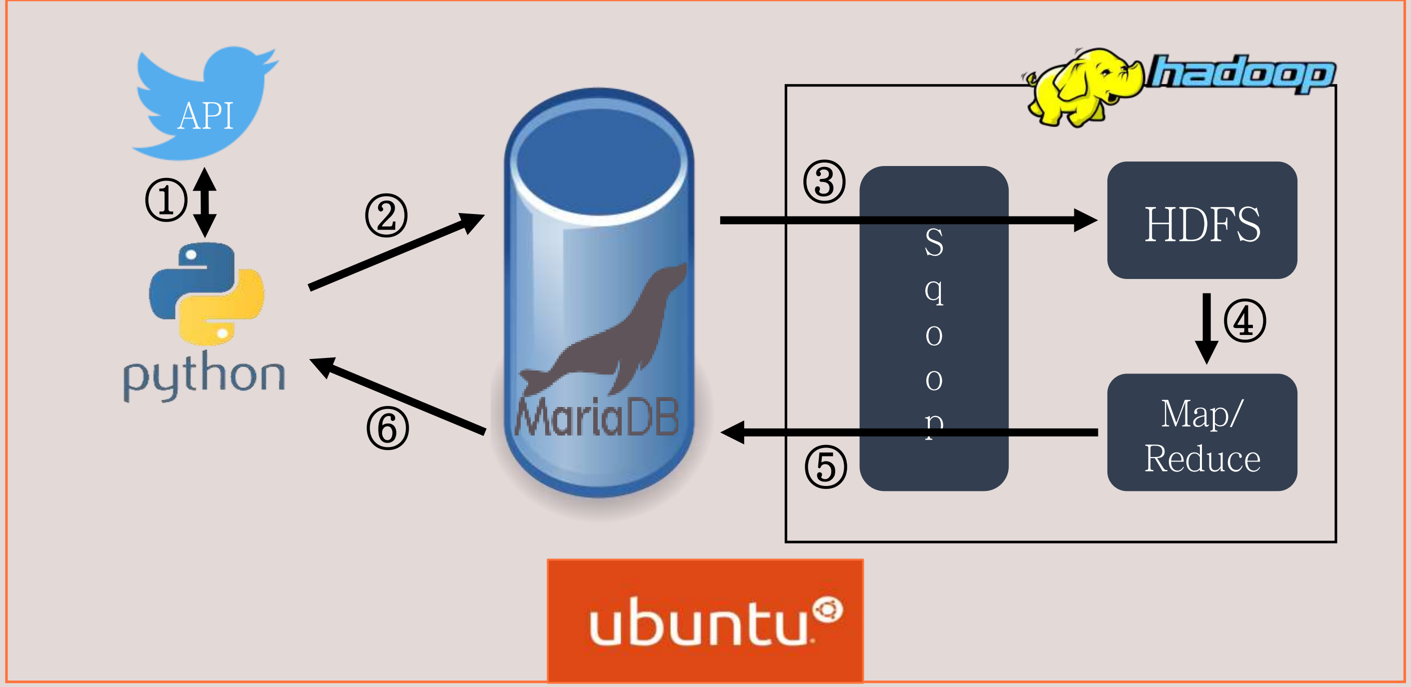
⑦

Visualization

ubuntu

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

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Conference	■	■	■	■	□	□	□	□	□	□	□	□	□	□	
Create document and PPT	■	■	■	■	□	□	□	□	□	□	□	□	□	□	
Create class of Twitter API in Python			■	■	□										
Create control class of MariaDB in Python					□	□	□								
Tasking of HDFS			■	■	□										
Tasking of Map/Reduce						□	□	□							
Normalized data save to MariaDB									□	□					
Visualization through by Python										□	□				
Test and Modification of T-SA											□	□	□	□	

T-SA: Weekly Progress_W05

Twitter Keyword Search API based Tweet Analysis

Lee SeokJune

문서 작성 및 수정

발표 준비

Lee YunHyuck

Map/Reduce 구현

DB, Hadoop 연동

Bae InGyu

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

Seo JaeIck

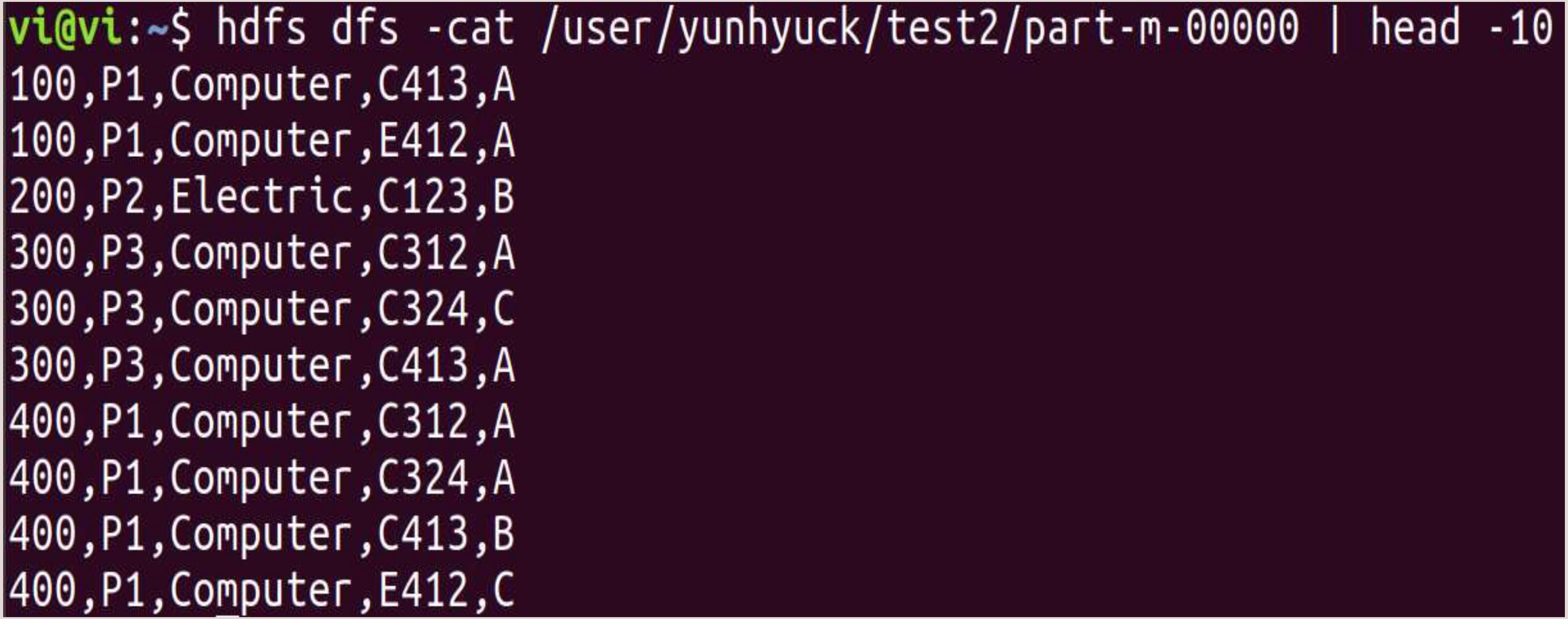
?

Hadoop
(+Sqoop)

T-SA: Weekly Progress_W05

Twitter Keyword Search API based Tweet Analysis

HDFS에 저장된 데이터 확인



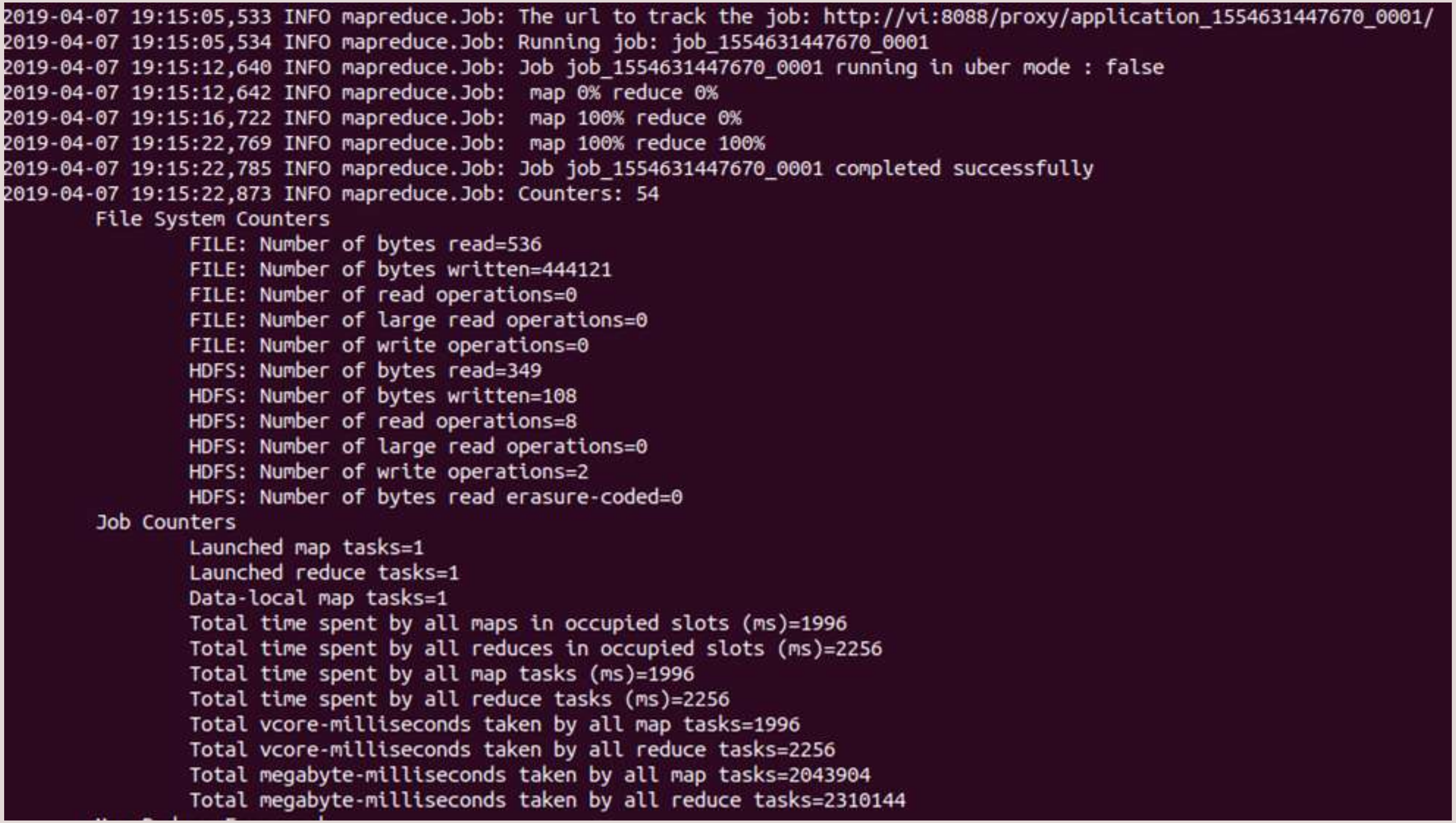
```
vi@vi:~$ hdfs dfs -cat /user/yunhyuck/test2/part-m-00000 | head -10
100,P1,Computer,C413,A
100,P1,Computer,E412,A
200,P2,Electric,C123,B
300,P3,Computer,C312,A
300,P3,Computer,C324,C
300,P3,Computer,C413,A
400,P1,Computer,C312,A
400,P1,Computer,C324,A
400,P1,Computer,C413,B
400,P1,Computer,E412,C
```

T-SA: Weekly Progress_W05

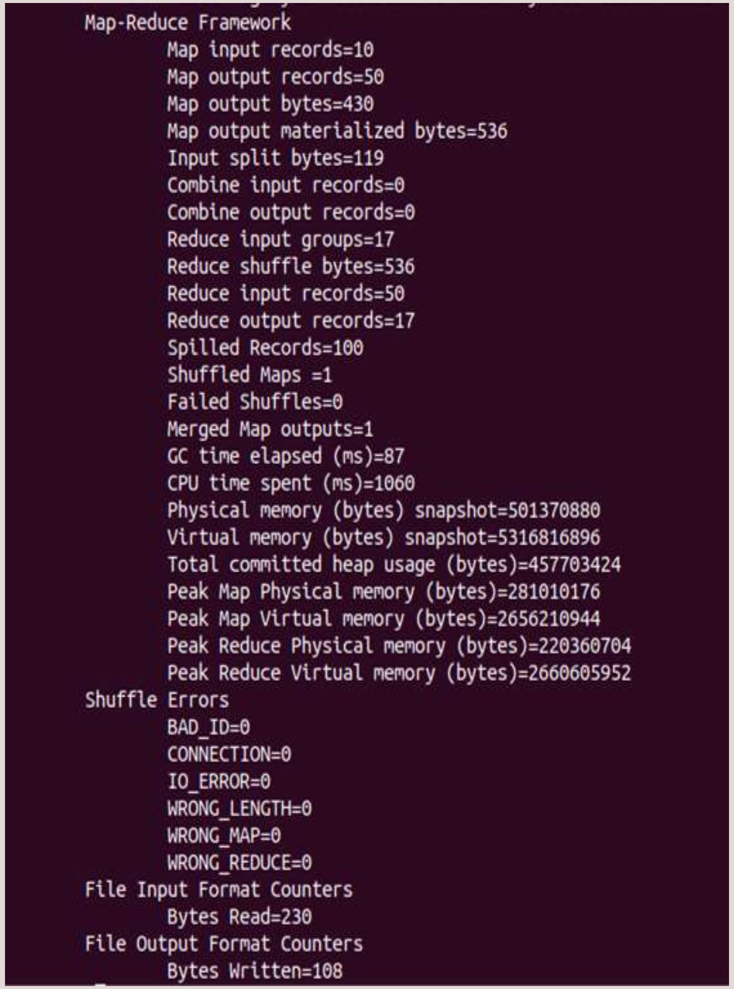
Twitter Keyword Search API based Tweet Analysis

Yarn 실행

yarn jar /home/vi/hadoop/jar/Wordcount.jar KeywordCount /user/yunhyuck/test2/part-m-00000 output



```
2019-04-07 19:15:05,533 INFO mapreduce.Job: The url to track the job: http://vi:8088/proxy/application_1554631447670_0001/
2019-04-07 19:15:05,534 INFO mapreduce.Job: Running job: job_1554631447670_0001
2019-04-07 19:15:12,640 INFO mapreduce.Job: Job job_1554631447670_0001 running in uber mode : false
2019-04-07 19:15:12,642 INFO mapreduce.Job:  map 0% reduce 0%
2019-04-07 19:15:16,722 INFO mapreduce.Job:  map 100% reduce 0%
2019-04-07 19:15:22,769 INFO mapreduce.Job:  map 100% reduce 100%
2019-04-07 19:15:22,785 INFO mapreduce.Job: Job job_1554631447670_0001 completed successfully
2019-04-07 19:15:22,873 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=536
                FILE: Number of bytes written=444121
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=349
                HDFS: Number of bytes written=108
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=1996
                Total time spent by all reduces in occupied slots (ms)=2256
                Total time spent by all map tasks (ms)=1996
                Total time spent by all reduce tasks (ms)=2256
                Total vcore-milliseconds taken by all map tasks=1996
                Total vcore-milliseconds taken by all reduce tasks=2256
                Total megabyte-milliseconds taken by all map tasks=2043904
                Total megabyte-milliseconds taken by all reduce tasks=2310144
```

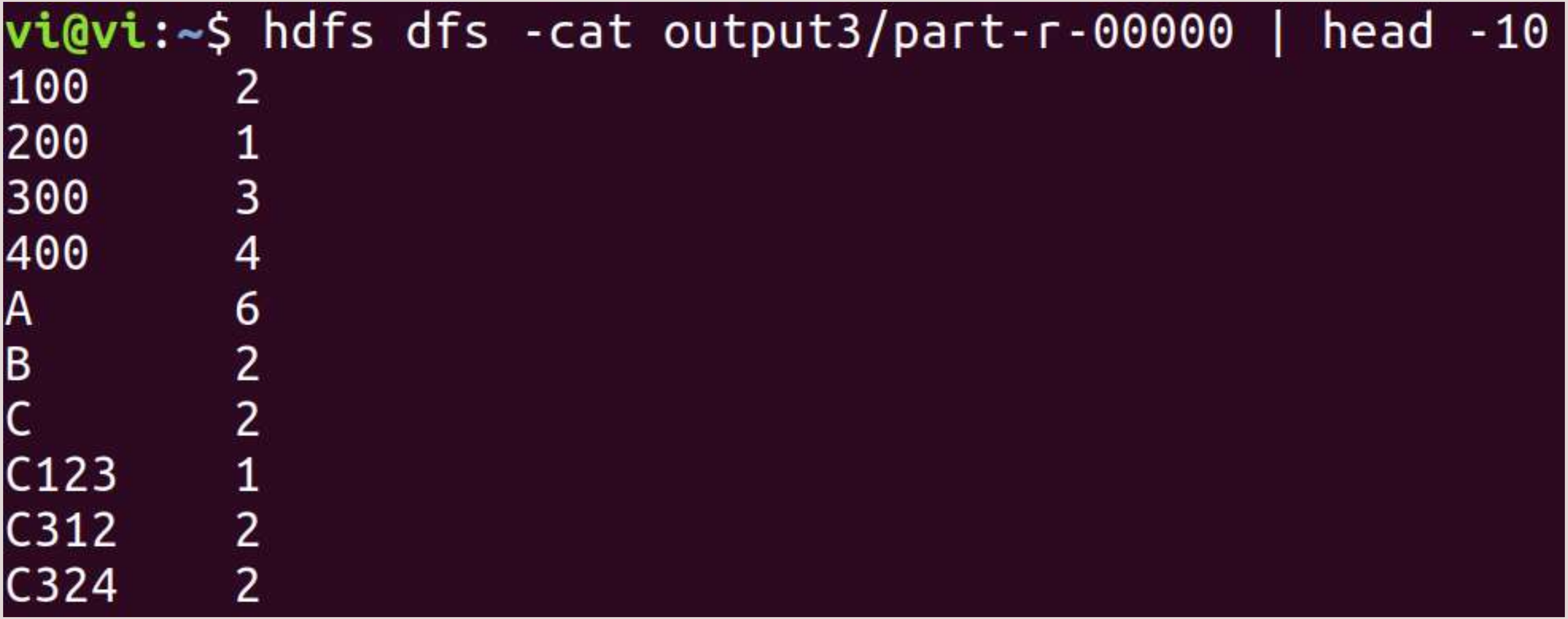


```
        Map-Reduce Framework
                Map input records=10
                Map output records=50
                Map output bytes=430
                Map output materialized bytes=536
                Input split bytes=119
                Combine input records=0
                Combine output records=0
                Reduce input groups=17
                Reduce shuffle bytes=536
                Reduce input records=50
                Reduce output records=17
                Spilled Records=100
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=87
                CPU time spent (ms)=1060
                Physical memory (bytes) snapshot=501370880
                Virtual memory (bytes) snapshot=5316816896
                Total committed heap usage (bytes)=457703424
                Peak Map Physical memory (bytes)=281010176
                Peak Map Virtual memory (bytes)=2656210944
                Peak Reduce Physical memory (bytes)=220360704
                Peak Reduce Virtual memory (bytes)=2660605952
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=230
        File Output Format Counters
                Bytes Written=108
```

T-SA: Weekly Progress_W05

Twitter Keyword Search API based Tweet Analysis

Yarn 상세 결과 확인



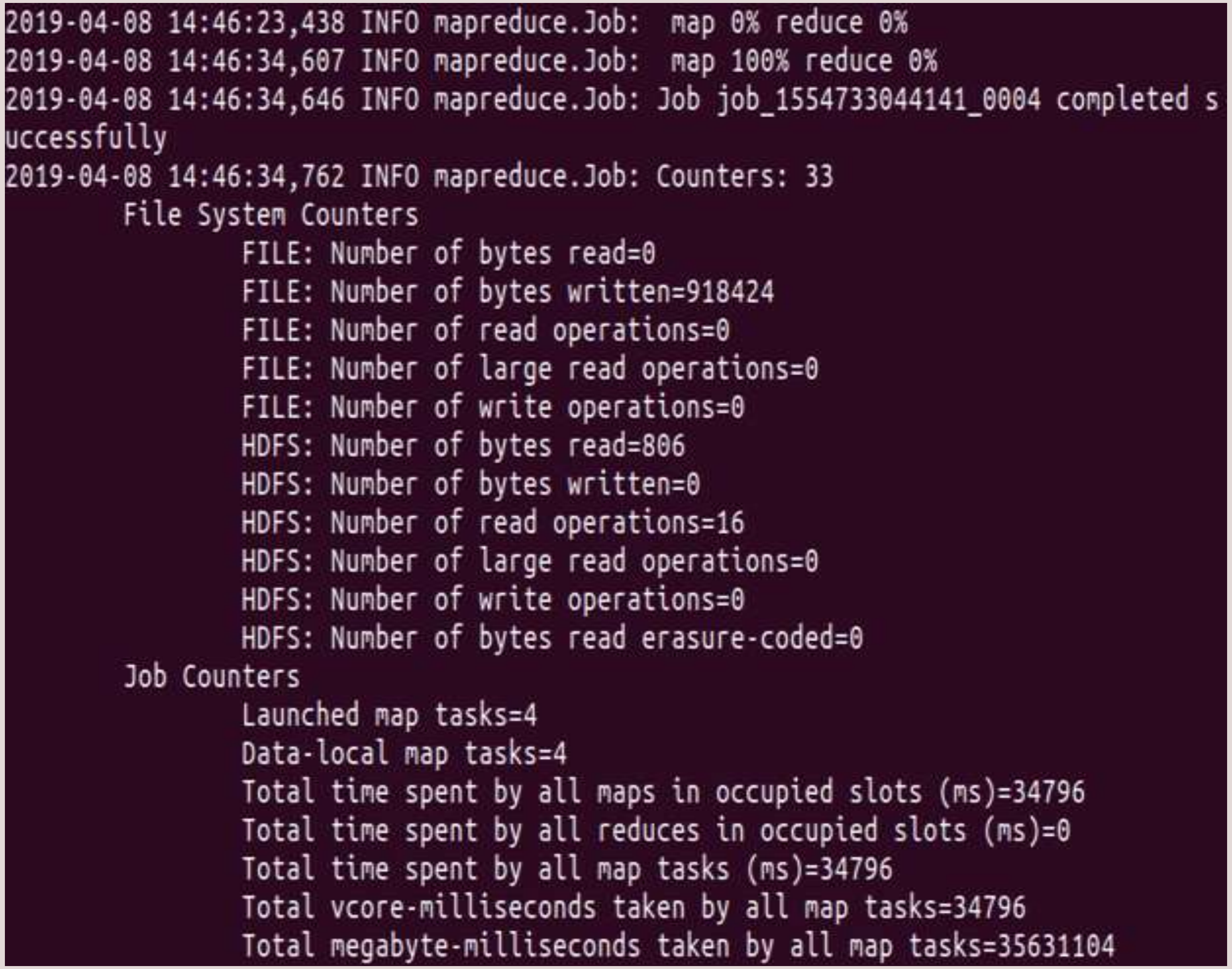
```
vi@vi:~$ hdfs dfs -cat output3/part-r-00000 | head -10
100     2
200     1
300     3
400     4
A       6
B       2
C       2
C123    1
C312    2
C324    2
```

T-SA: Weekly Progress_W05

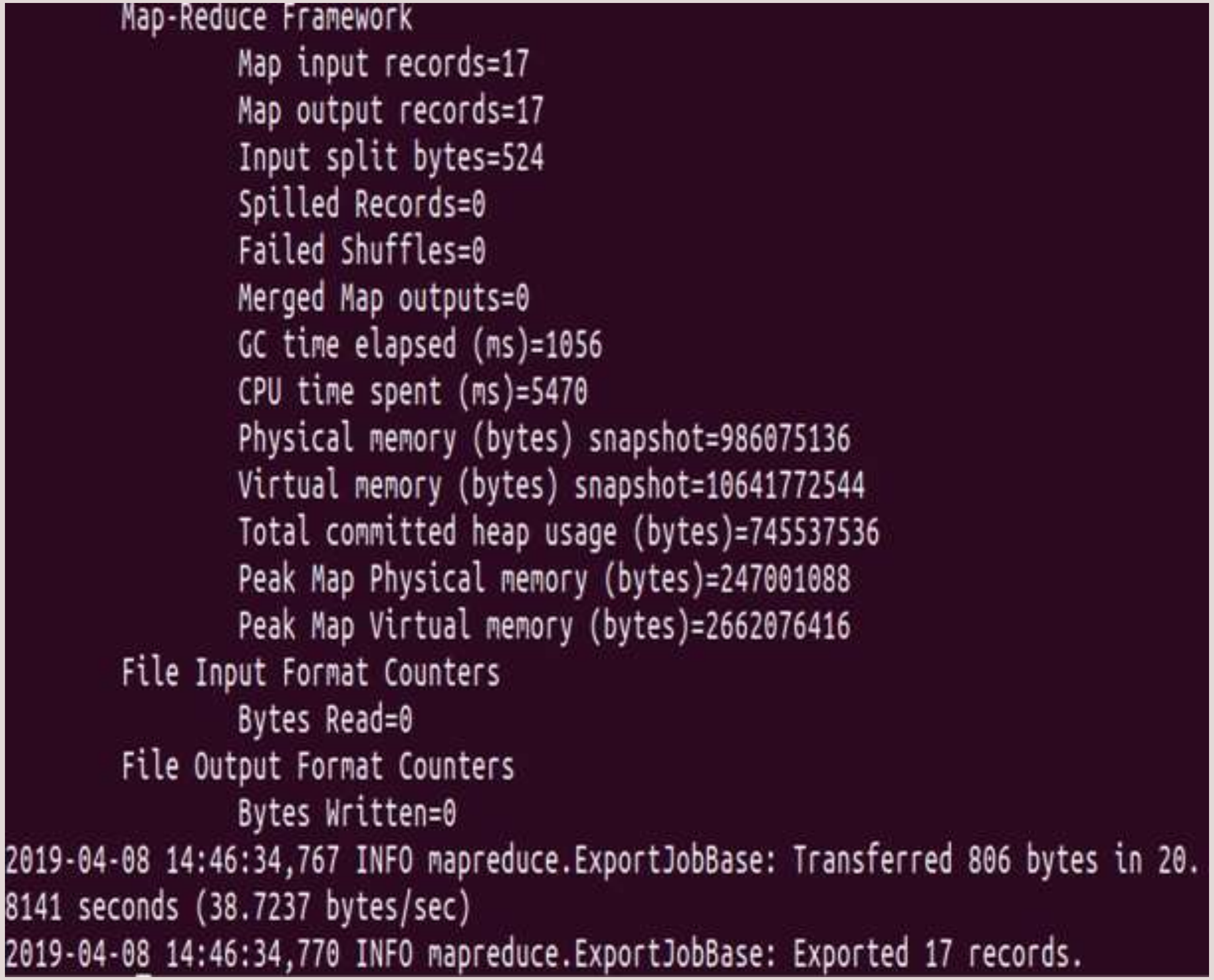
Twitter Keyword Search API based Tweet Analysis

Sqoop Job 실행

sqoop export --connect jdbc:mysql://localhost/mysql --username root -P --table test --export-dir hdfs://localhost:9000/user/vi/output3/part-r-00000 --columns a,b --input-fields-terminated-by "₩t"



```
2019-04-08 14:46:23,438 INFO mapreduce.Job:  map 0% reduce 0%
2019-04-08 14:46:34,607 INFO mapreduce.Job:  map 100% reduce 0%
2019-04-08 14:46:34,646 INFO mapreduce.Job: Job job_1554733044141_0004 completed s
uccessfully
2019-04-08 14:46:34,762 INFO mapreduce.Job: Counters: 33
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=918424
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=806
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=16
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=4
                Data-local map tasks=4
                Total time spent by all maps in occupied slots (ms)=34796
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=34796
                Total vcore-milliseconds taken by all map tasks=34796
                Total megabyte-milliseconds taken by all map tasks=35631104
```

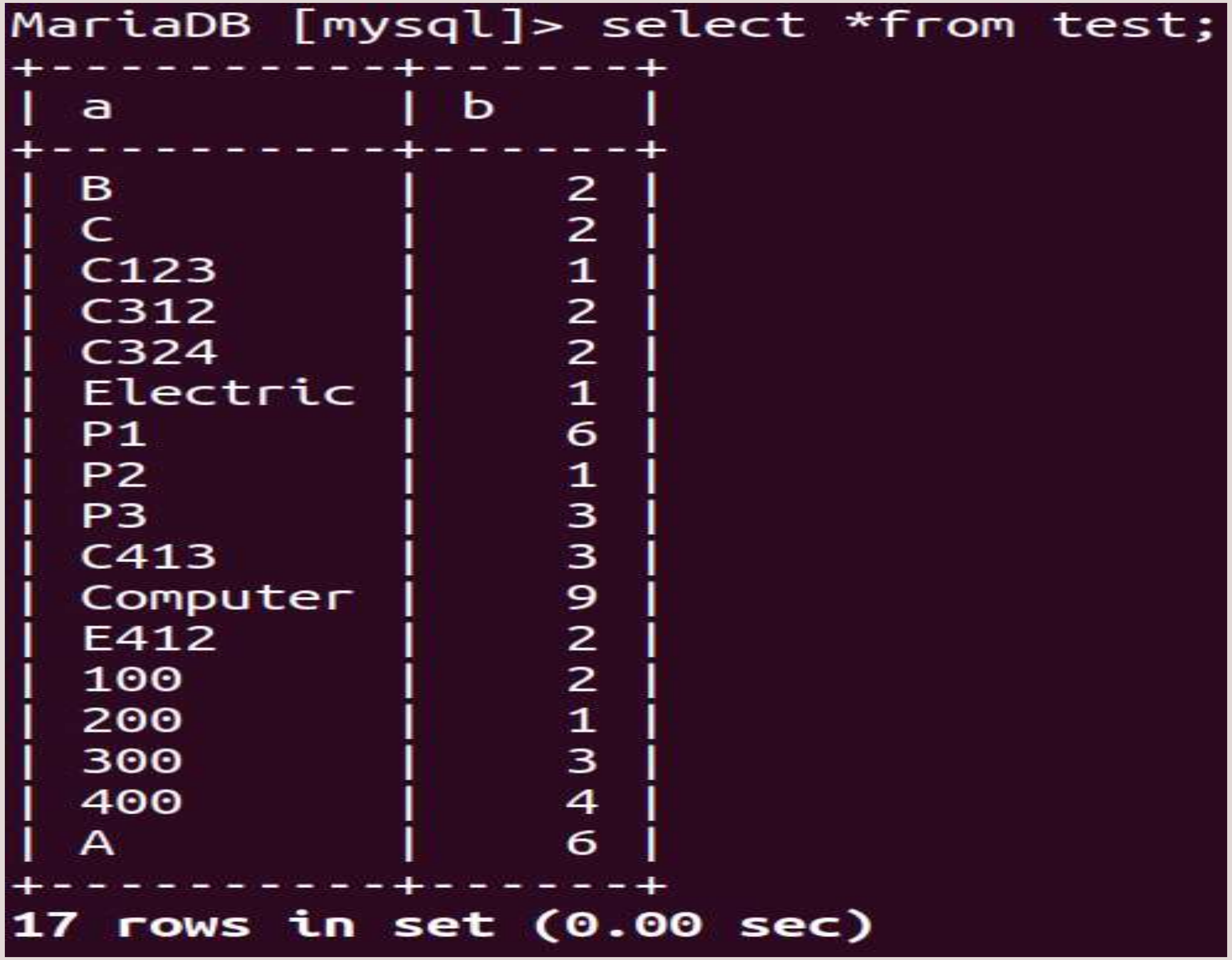


```
        Map-Reduce Framework
                Map input records=17
                Map output records=17
                Input split bytes=524
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=1056
                CPU time spent (ms)=5470
                Physical memory (bytes) snapshot=986075136
                Virtual memory (bytes) snapshot=10641772544
                Total committed heap usage (bytes)=745537536
                Peak Map Physical memory (bytes)=247001088
                Peak Map Virtual memory (bytes)=2662076416
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
2019-04-08 14:46:34,767 INFO mapreduce.ExportJobBase: Transferred 806 bytes in 20.
8141 seconds (38.7237 bytes/sec)
2019-04-08 14:46:34,770 INFO mapreduce.ExportJobBase: Exported 17 records.
```

T-SA: Weekly Progress_W05

Twitter Keyword Search API based Tweet Analysis

MariaDB 결과 확인



```
MariaDB [mysql]> select *from test;
+----------+-------+
| a        | b     |
+----------+-------+
| B        |     2 |
| C        |     2 |
| C123     |     1 |
| C312     |     2 |
| C324     |     2 |
| Electric |     1 |
| P1       |     6 |
| P2       |     1 |
| P3       |     3 |
| C413     |     3 |
| Computer |     9 |
| E412     |     2 |
| 100      |     2 |
| 200      |     1 |
| 300      |     3 |
| 400      |     4 |
| A        |     6 |
+----------+-------+
17 rows in set (0.00 sec)
```

T-SA: Weekly Progress_W05

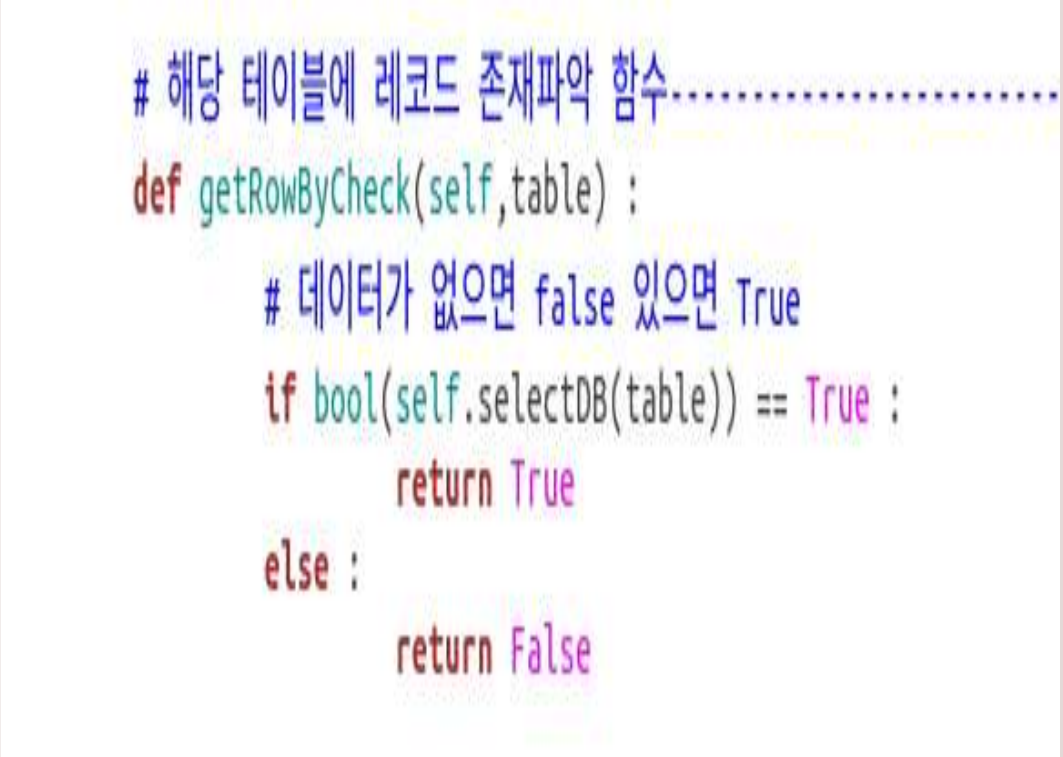
Twitter Keyword Search API based Tweet Analysis

Python(DB)

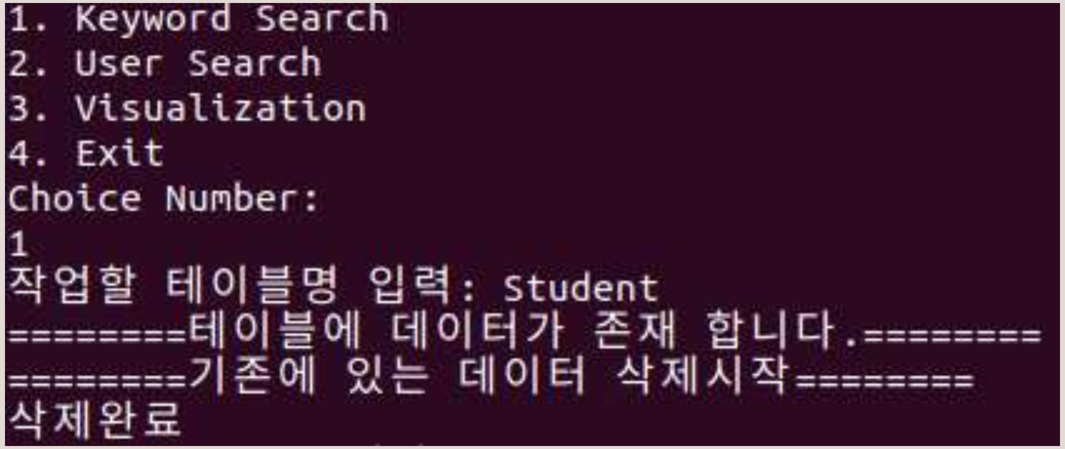
T-SA: Weekly Progress_W05

Twitter Keyword Search API based Tweet Analysis

DBModule.getRowByCheck



```python
# 해당 테이블에 레코드 존재파악 함수........................
def getRowByCheck(self,table) :
        # 데이터가 없으면 false 있으면 True
        if bool(self.selectDB(table)) == True :
                return True
        else :
                return False
```

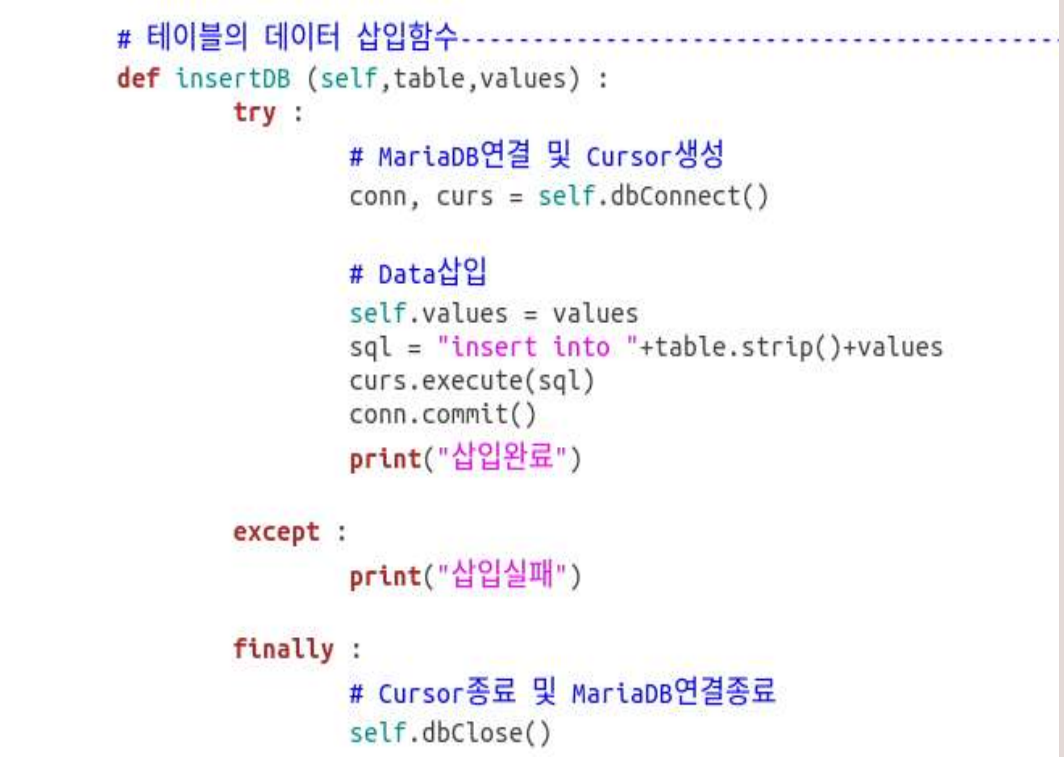


```
1. Keyword Search
2. User Search
3. Visualization
4. Exit
Choice Number:
1
작업할 테이블명 입력: student
========테이블에 데이터가 존재 합니다.========
========기존에 있는 데이터 삭제시작========
삭제완료
```

T-SA: Weekly Progress_W05

Twitter Keyword Search API based Tweet Analysis

DBModule.insertDB

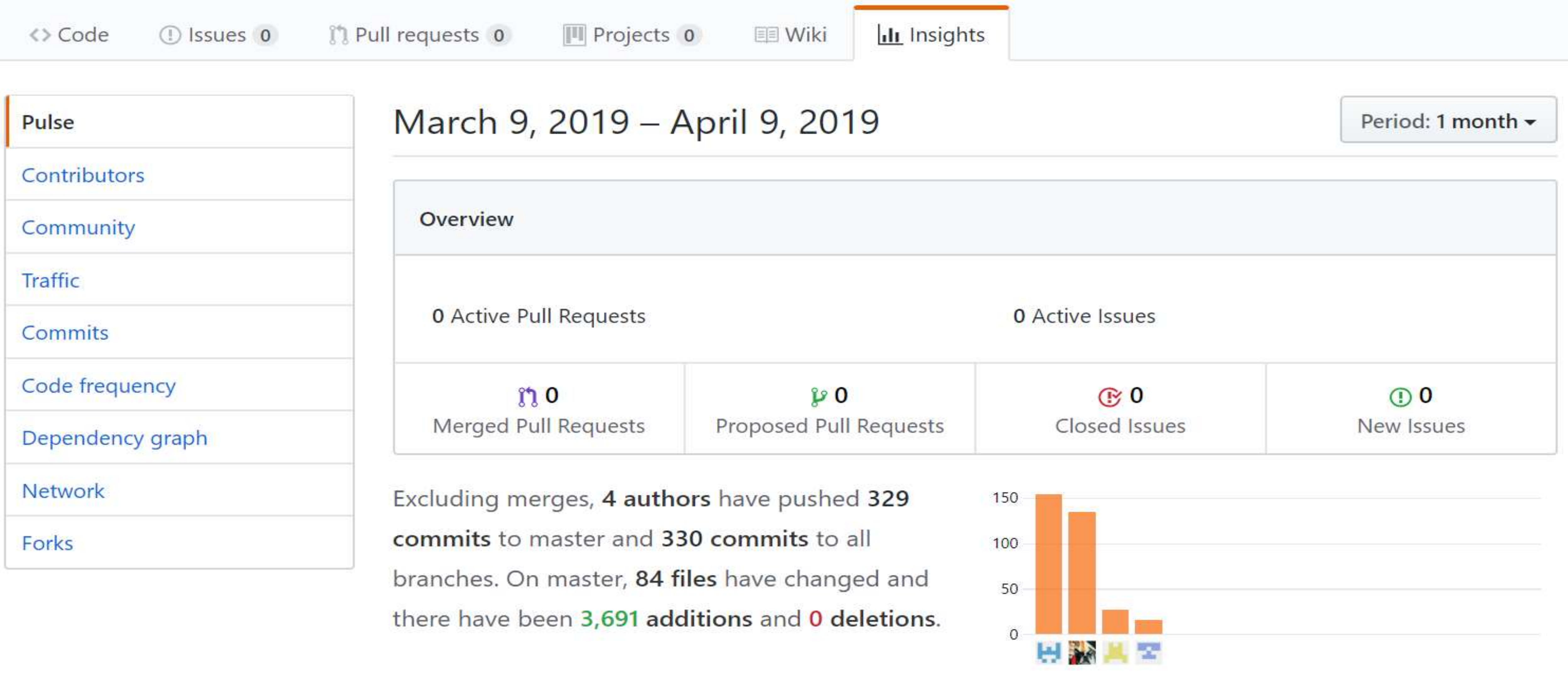


```
# 테이블의 데이터 삽입함수-----------------------------------
def insertDB (self,table,values) :
        try :
                # MariaDB연결 및 Cursor생성
                conn, curs = self.dbConnect()

                # Data삽입
                self.values = values
                sql = "insert into "+table.strip()+values
                curs.execute(sql)
                conn.commit()
                print("삽입완료")

        except :
                print("삽입실패")

        finally :
                # Cursor종료 및 MariaDB연결종료
                self.dbClose()
```

T-SA: Github_W05

Twitter Keyword Search API based Tweet Analysis

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



March 9, 2019 – April 9, 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 **329 commits** to master and **330 commits** to all branches. On master, **84 files** have changed and there have been **3,691 additions** and **0 deletions**.

Q & A

Thank you.