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	이윤혁
발 표 일 자	2019.05.02.

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

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

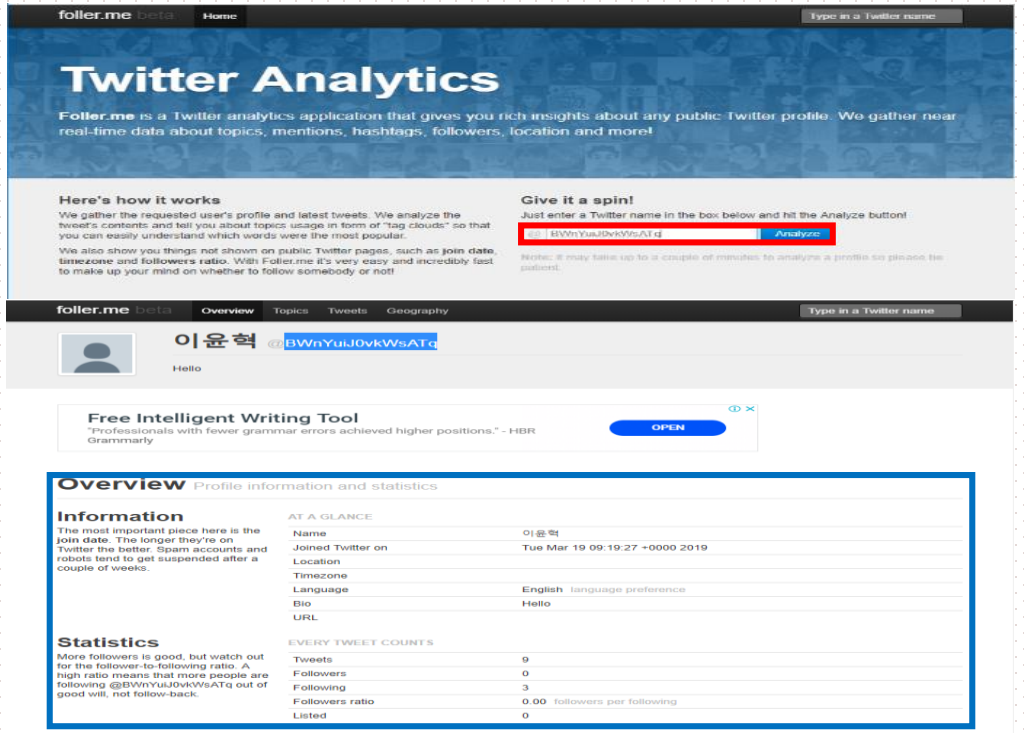
대한민국 지역 및 특정 기간에 사용된 키워드 트렌드 분석

특정 인물의 트윗 스타일 분석

데이터 시각화

T-SA: Related Works

Twitter Keyword Search API based Tweet Analysis

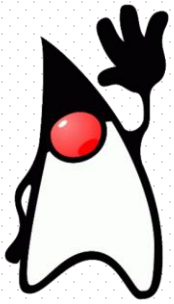
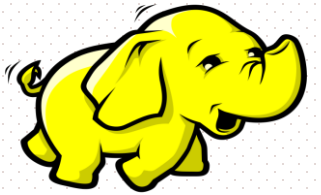
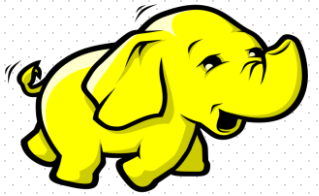


- http://tweetrend.com/
- 기간을 지정하여 해당 기간의 키워드를 통해 트윗 검색
- 그래프를 통해 해당 기간에 트윗이 올라온 결과 확인 가능

- https://foller.me/
- 메인 페이지에서 사용자의 아이디를 검색을 통해 사용자의 정보(이름, 지역, 언어, 트윗 개수, 팔로잉, 팔로워)를 파악

T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis

python	MariaDB	OpenJDK	eclipse	hadoop	sqoop
					
3.6	10.1.38	1.8.0_191	2019-03(4.11)	3.2.0	1.4.7

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

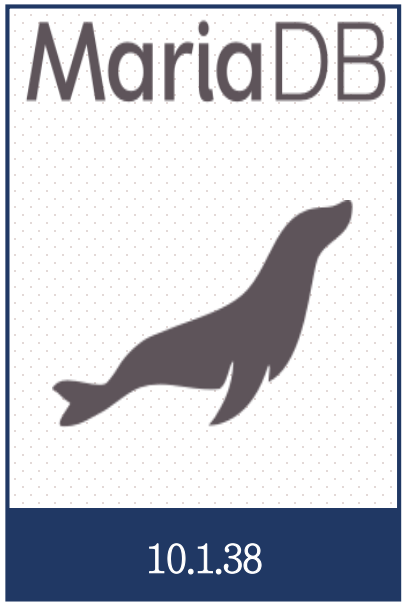
Version Check

~$ 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 MaySQL과 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

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 실행됐다면 비밀번호 없이 (Enter) 아니면 비밀번호 입력

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

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

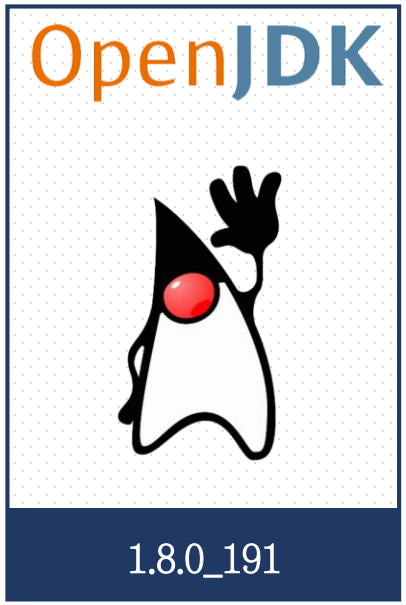
Disallow root login remotely? [Y/n] → 원격 접속으로 루트 로그인 허용 여부

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

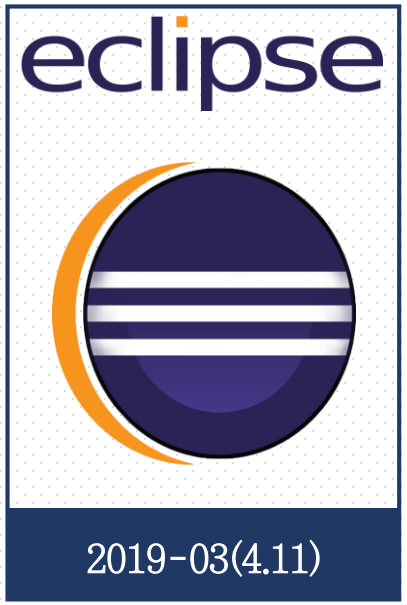
Reload privilege tables now? [Y/n] → 지금까지 작성한 권한 테이블을 적용 여부

T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis



1.8.0_191



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

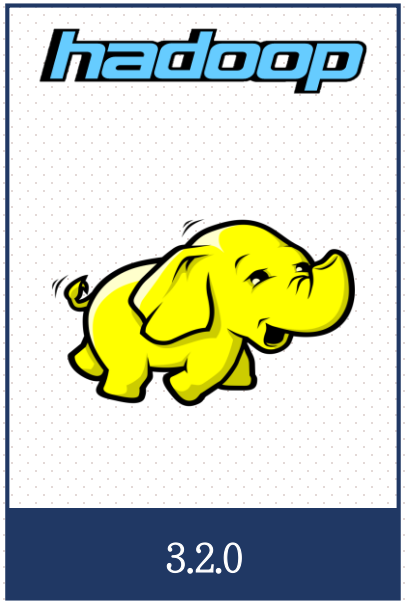
Version Check

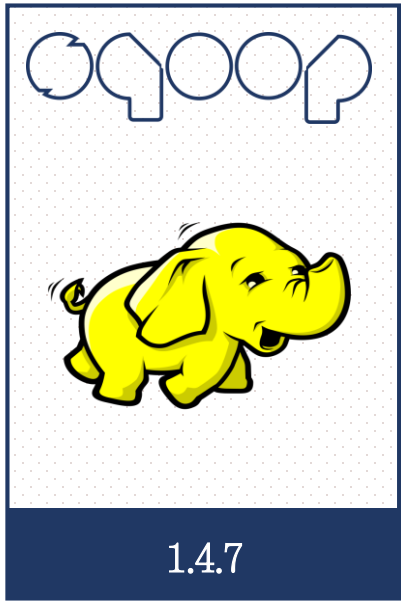
~$ java –version

```
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 프레임워크이다.

- 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 available일
export HADOOP_MAPRED_HOME=/home/vi(계정이름)/hadoop
```

~/sqoop/conf$ sqoop

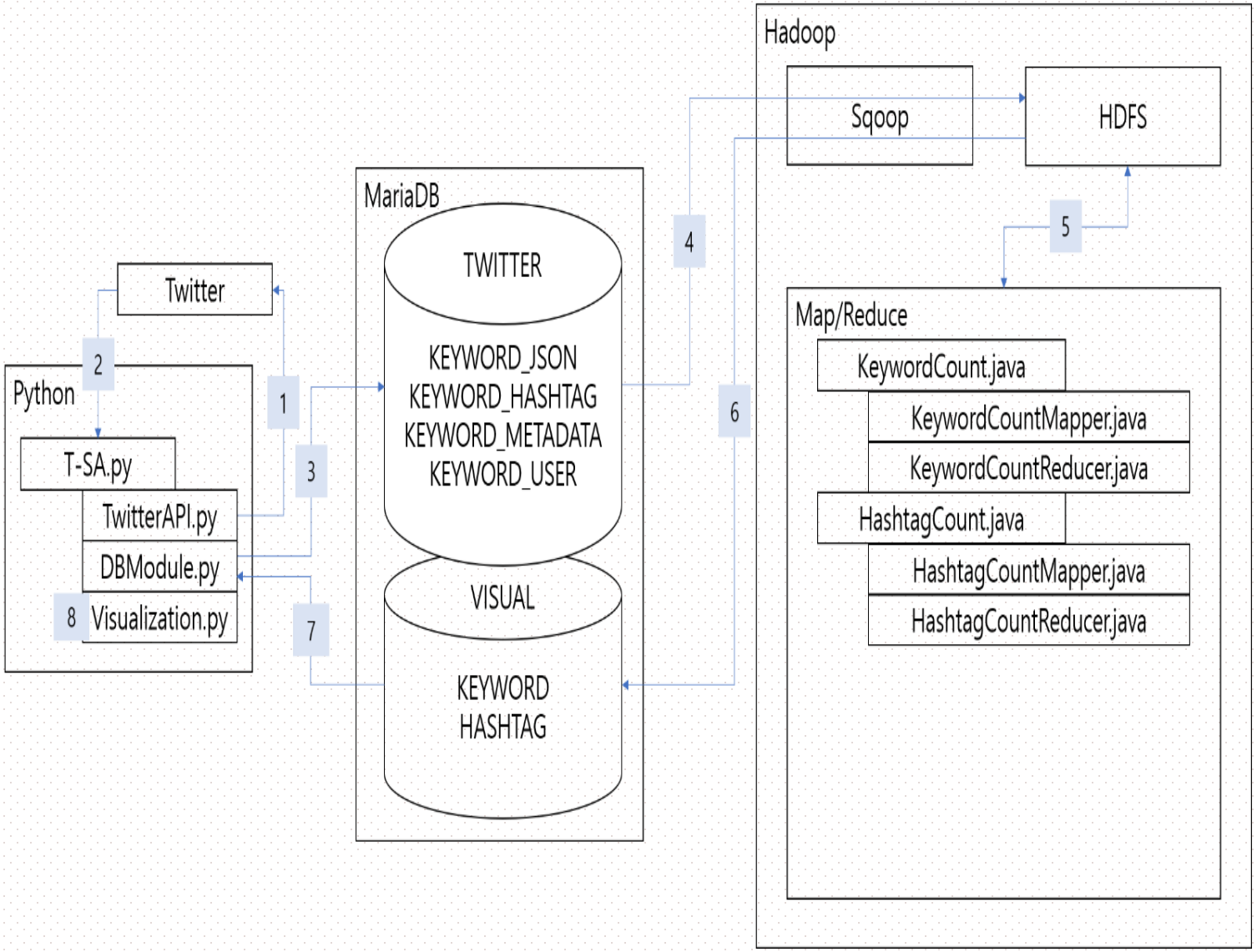
Warning: /home/vi/sqoop/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /home/vi/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/vi/sqoop/../accumulo 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

T-SA: Program Flowchart

Twitter Keyword Search API based Tweet Analysis



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

[3] 크롤링된 데이터를 MariaDB에 저장

[4] Sqoop을 이용하여 MariaDB에 저장된 데이터를 HDFS에 저장

[5] HDFS에 업로드된 데이터를 Map/Reduce과정을 통해 정규화하고 결과를 HDFS에 저장

[6] Sqoop을 이용하여 HDFS에 저장된 정규화된 데이터를 MariaDB에 저장

[7] MariaDB에 저장된 데이터를 Python으로 불러온다.

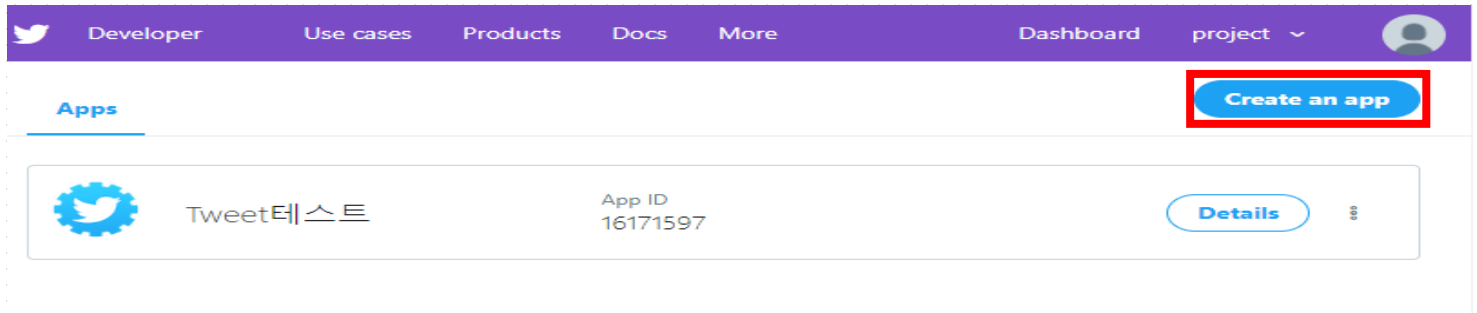
[8] 불러온 데이터 시각화

T-SA: Demonstrate

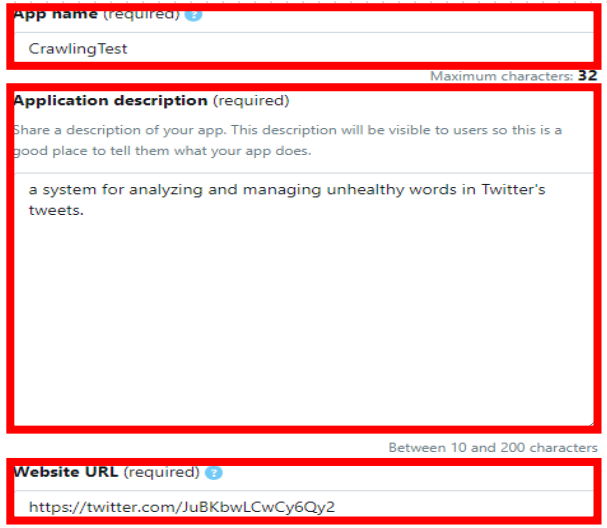
Twitter Keyword Search API based Tweet Analysis

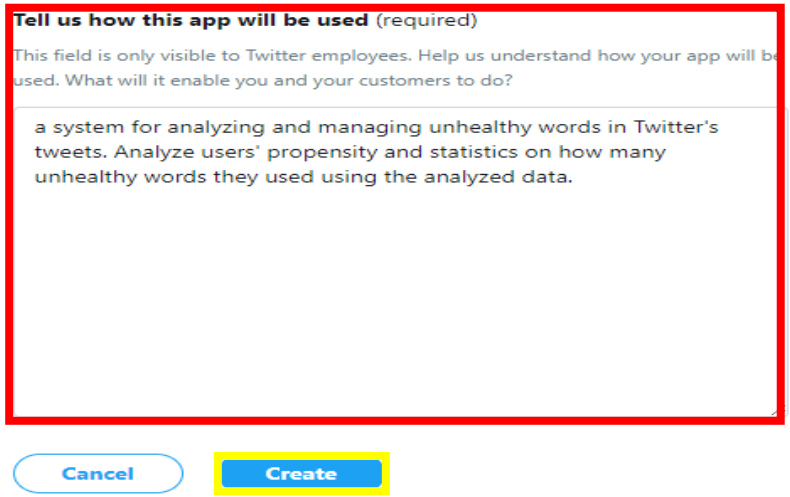
Twitter API 발급

1) https://developer.twitter.com/ 접속 » Project » Apps » Create an app



2) App name, Application description, Website URL, Tell us how this app will be used 작성

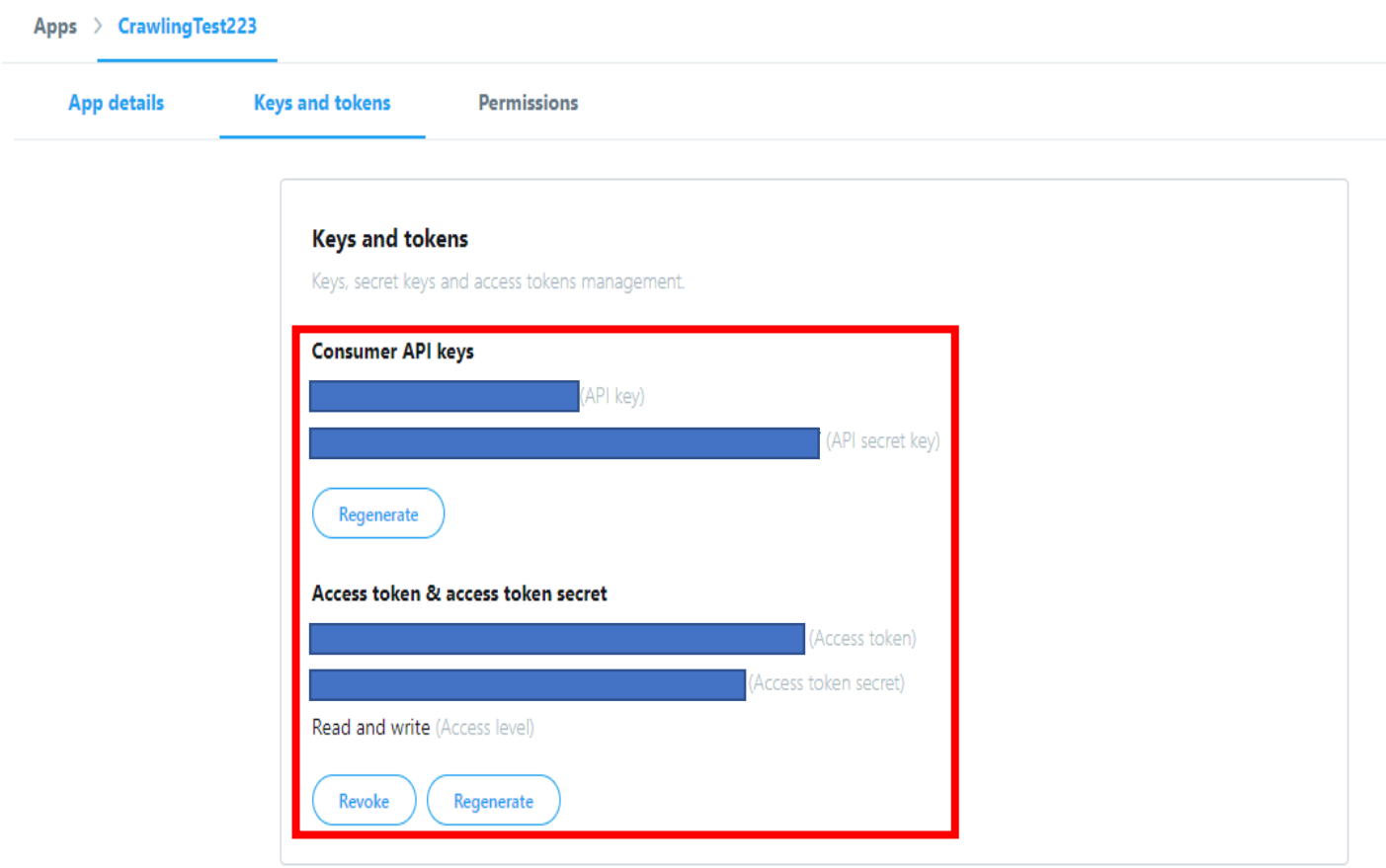




T-SA: Demonstrate

Twitter Keyword Search API based Tweet Analysis

3) 앱이 등록되었다면, Keys and tokens를 눌러서 API 키와 Access token을 발급받는다.
   - 다시 받고 싶다면 ‘Regenerate’를 눌러서 다시 새롭게 받을 수 있다.



T-SA: Demonstrate

Twitter Keyword Search API based Tweet Analysis

TwitterAPI.py : 키워드를 통한 데이터 크롤링

트위터 API 인증

```
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
        auth.set_access_token(access_token, access_token_secret)
        api = tweepy.API(auth)
```

커서를 통한 특정 키워드에 대한 검색

```
tweepy.Cursor(api.search, q=keyword, since=sinceD, until=untilD, tweet_moede=mode, count=count).items()
```

T-SA: Demonstrate

Twitter Keyword Search API based Tweet Analysis

DBModule.py : 크롤링한 데이터 DB에 저장

MariaDB 연결 부분

```
conn = pymysql.connect(host=self.host, user=self.user, password=self.pswd, db=self.db, charset = self.charset)
curs = conn.cursor()
```

DB 데이터 삽입 부분

```
# Data삽입
sql = "insert into " + tableName.strip() + " values("
                    sql += "'" + values[0] + "',"
                    sql +=       str(values[2][0]) + ","
                    sql +=       str(values[2][1]) + ","
                    sql += "'" + values[1] + "');"
curs.execute(sql)
conn.commit()
```

DB 데이터 조회 부분

```
# 테이블 조회
sql = "select * from " + tableName.strip() + ";"
curs.execute(sql)
```

T-SA: Demonstrate

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: Demonstrate

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_MODEL기본 사전 사용 << 사전 정의 가능
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("NP","NNP","NNG");
// 요소들을 읽어오기 위한 Iterator 생성 후, tokens 내용 적재.
Ite rator<String> itrs = tokens.iterator();
// ----------------------------------------------------------------------------------------------
```

T-SA: Demonstrate

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: Demonstrate

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()
```

Thank you