자동화된 트위터 데이터 수집 시스템 설계 및 구현 : 환경 데이터를 중심으로

김도형*, 구자환**, 김응모***
*성균관대학교 소프트웨어대학
** 성균관대학교 사회과학대학
*** 성균관대학교 사회과학대학 소비자가족학과 / 소프트웨어대학
shape1248@skku.edu, jhkoo@skku.edu, ukim@skku.edu

Design and Implementation of Automated Twitter Data Collecting System : Focus on Environmental Data

Do-Hyung Kim* Jahwan Koo**, Ung-Mo Kim***
*College of Software, Sungkyunkwan University
** College of Social Sciences, Sungkyunkwan University
*** Dept. of Consumer and Family Sciences, College of Social Sciences / College of Software, Sungkyunkwan University

요 약

소셜 네트워크 서비스의 사용자가 늘어나면서, 소셜 네트워크 서비스상에서 발생하는 빅데이터를 활용한 서비스가 늘어나고 있다. 소셜 네트워크 서비스 데이터는 실시간으로 생성되며, 따라서 데이터 수집 시스템 역시 자동화하여 준 실시간으로 데이터를 수집할 필요가 있다. 본 논문에서는 대표적인 소셜 네트워크 서비스인 트위터의 데이터를 지속적으로 수집하기 위한 자동 수집 시스템을 제안한다. 수집 시스템은 Twitter API 를 활용하는 Python 라이브러리를 통해 내용 및 메타데이터를 수집하며, 수집된 데이터를 재 검증한 뒤 저장한다. 또한 구현된 시스템에 환경 데이터를 주제로 하는 쿼리를 입력하여 실제 트위터 데이터를 수집하며 구현된 시스템을 검증해보았다.

1. 서론

최근 트위터, 페이스북과 같은 SNS(Social Network Service) 사용자가 늘어나면서, SNS 에서 발생하는 대량의 데이터를 정치, 경제, 문화 등 다양한 분야에 활용하려는 노력이 계속되고 있다. 이러한 SNS 에서 생성되는 데이터는 빅데이터의 정의에 정확히 부합한다. 즉, 정형화되지 않고, 대용량이며, 실시간으로 생성된다[1]. 따라서 이러한 SNS 빅데이터를 특성에 맞게 활용하기 위해서는 안정적이고 지속적으로 데이터를 수집할 수 있는 시스템이 요구된다.

많은 연구자들이 단기적인 데이터 분석을 위해 트위터 데이터 수집 시스템을 설계, 구현해왔다. 트위터는 트위터 데이터에 접근하기 위한 API 를 공개하고 있으며, 웹페이지를 통한 검색도 가능하다. 또한 트위터 데이터를 수집하는 다양한 Third party API 및 라이브러리도 만들어져 있다.

우리 연구에서는 정해진 쿼리에 따라 트위터 데이터를 수집하며, 불필요한 데이터와 중복 데이터를 제거한 후 날짜 별로 나누어 저장하는 과정을 지속적으로 수행하는 시스템을 제안한다.

이를 통해 준 실시간으로, 지속적으로 트위터 데이터를 수집할 때 발생할 수 있는 데이터 중복 및 각종 예외 상황에 대한 관리를 사용자의 조작 없이도 시스템 자체적으로 지원할 수 있다.

본 논문의 구조는 다음과 같다. 2 장에서는 트위터 데이터를 수집 시스템을 설계 및 구현하였던 다양한 선행 연구를 소개하고, 트위터 데이터를 수집하는 다양한 방법을 서술한다. 3 장에서는 본 연구에서 제안하는 자동화된 트위터 데이터 수집 시스템의 요구조건과 구조를 제안한다. 이어 4 장에서 환경을 주제로 실제 트위터 데이터를 수집하여 결과를 확인한 후, 5 장에서 결론과 향후 계획에 대해 밝힌다.

2. 관련 연구

2.1. SNS 데이터 수집 선행연구

Byun, Kim, et al(2012)은 Twitter API 와 Java 를 이용하여 자동화된 트위터 데이터 수집 시스템을 제안하였다. 제안된 시스템은 트위터 계정간 팔로우 관계를

통해 데이터를 수집한 후, 수집된 데이터 내에서 키워드 검색을 통해 원하는 데이터를 데이터베이스에 저장한다.[2]

최민석(2015)은 Twitter API 를 이용하여 유저가 입력한 관심 키워드에 대한 트위터 데이터를 수집한 후, 분석 및 시각화하는 시스템을 제안하였다. 제안된 시스템은 트윗 데이터를 하루 분량만 MySQL DB 에 저장하며, 그 외에는 분석 완료된 데이터만 저장한다. 또한 관심 키워드에 대한 급격한 트윗 증가를 감지하여 유저가 이를 파악할 수 있도록 하였다.[3]

하승도 등(2016)은 트위터로부터 대화형 말뭉치 데이터를 수집하기 위하여 웹 크롤링 방식을 제안하였으며, 기존의 트위터 API 기반의 데이터 수집기와 비교하여 같은 시간동안 더 많고 완전한 말뭉치 데이터를 얻을 수 있었다.[4]

2.2. 트위터 데이터의 수집 방법

트위터는 Tweet 데이터에 접근하기 위한 API[5]를 공개하고있다. 트위터 계정을 통해 API key 를 지급받아 사용할 수 있으며, 무료 라이센스의 경우 최근 7일 이내의 데이터에 대한 검색이 가능하다.

트위터 웹 페이지의 고급 검색 기능을 활용하여 웹 크롤링 방법으로 트위터 데이터를 수집할 수도 있다. https://twitter.com/search?q=<keyword>&src=typed_query&f=live 주소의 <keyword>에 검색 쿼리를 입력하여 URL 을 생성할 수 있으며, 한 번에 모든 정보를 표시하지 않기 때문에 같은 웹 페이지에 대해 스크롤 정보를 지속적으로 갱신하며 요청해야 완전한 정보를 얻을 수 있다.

Python3 라이브러리인 GetOldTweets3[6]은 트위터의 웹 API 를 사용하여 트위터 데이터를 수집한다. 트위터가 제공하는 프로그래밍 언어 수준의 API 가 가진 단점인 날짜 제한이 없고, 웹 크롤링 방식과 마찬가지로 쿼리를 작성하여 데이터를 수집할 수 있다. 따라서 본 연구에서는 해당 라이브러리를 통해 트위터의 데이터를 수집한다.

3. 트위터 크롤러 시스템 분석 및 구성

3.1. 요구조건

본 연구에서 설계한 Tweet 데이터 수집 시스템은 다음과 같은 요구조건을 충족시켜야 한다.

시스템을 구축한 후에는 추가적인 유저의 조작 없이 자동적이고 지속적으로 최신 데이터를 수집하여 기존 데이터에 추가해야 한다.

수집할 Tweet 데이터는 여러 종류이며, 각각 사전에 작성된 검색식과 제외식이 존재한다. Tweet 의 내용이 검색식에 포함되면서 제외식에는 포함되지 않는 데이터를 수집, 저장해야 한다.

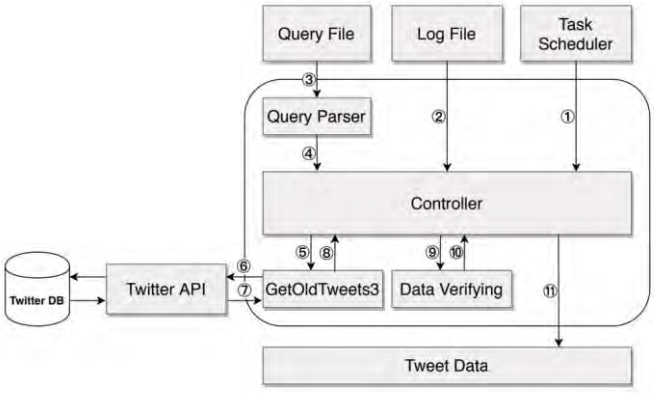
3.2. 시스템 아키텍쳐

GetOldTweets3 라이브러리는 Python 환경에서 사용가능하다. GetOldTweets3.manager.TweetCriteria 함수를 사용하여 검색 쿼리 및 검색 기간 등을 설정한 후, GetOldTweets3.manager.TweetManager.getTweets 함수를 사용하면 Twitter API 를 활용하여 데이터를 수집한 후, 수집 결과를 되돌려준다.

Twitter API 는 검색 쿼리를 통해 데이터를 요청 시 쿼리의 검색어를 Tweet 의 내용뿐 아니라 작성자명까지 포함하여 검색하는 특징이 있다. Twitter API 를 사용하는 GetOldTweets3 라이브러리를 통해 수집한 데이터 역시 같은 특성을 가진다.

이렇게 수집된 데이터는 본 시스템이 원하는 데이터가 아니므로, 추가적인 데이터 검증 작업을 통해 내용에 검색어가 포함되지 않는 데이터를 제거하는 작업이 필요하다.

제외어 역시 Twitter API 에 포함하여 요청할 경우, 수집되어야 할 데이터가 제외식이 포함된 작성자명으로 인해 수집되지 않을 수 있다. 따라서 검색어만으로 이루어진 쿼리를 작성하여 데이터를 수집하고, 이후 제외어가 포함된 데이터를 제거하는 작업이 필요하다.



(그림 1) Tweet 수집 시스템의 activity diagram

1. Task Scheduler 가 크롤러 프로그램을 실행시킨다.

2. 이전에 작성된 로그 파일을 통해 검색 시작시점을 파악한다.

3,4. 쿼리 파일을 해석하여 검색어와 제외어 키워드 뭉치를 만든다.

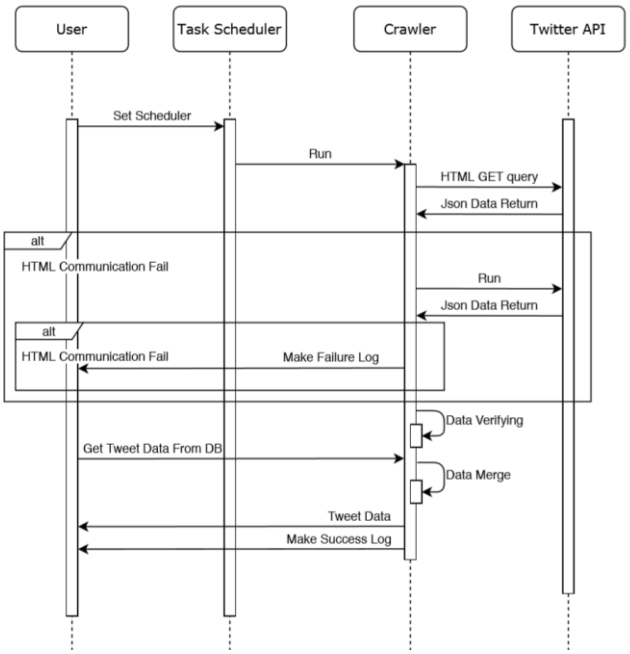
5. GetOldTweets3 라이브러리에 검색어로 이루어진 쿼리와 검색 기간을 지정하여 데이터를 수집 요청한다.

6,7. GetOldTweets3 라이브러리는 Twitter API 에 해당 쿼리를 전송하여 데이터를 수집한다.

8. 수집된 데이터를 Controller 에 되돌려준다.

9,10. 데이터 검증 작업을 통해 내용에 검색어가 포함되지 않거나, 제외어가 포함된 데이터를 제거한다.

11. 최종적으로 얻은 데이터를 저장한다.



(그림 2) Tweet 수집 시스템의 sequence diagram

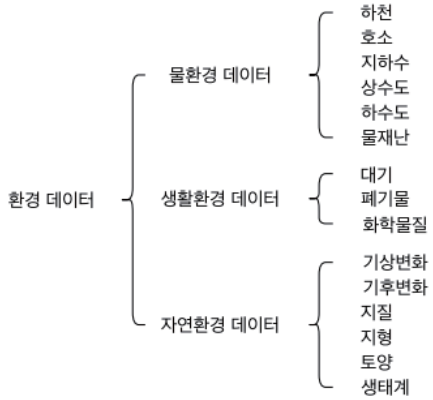
(그림 2)에 명시된 Crawler 의 내부 구조는 (그림 1)의 activity diagram 에서 표현된 내용과 같다.

유저가 OS 의 Task Scheduler 를 설정하면, Crawler 는 오직 Task Scheduler 에 의해서만 실행되며, 완성된 데이터는 User 에게 파일로써 제공된다. 이 과정에서 Twitter API 가 응답하지 않을 경우에 대해 최대 두 번까지 수집을 시도한다.

4. 구현 방법

우리는 본 시스템을 활용하여 Twitter 에서 환경 관련 언급이 포함된 Tweet 을 수집하였다.

수집할 환경 데이터는 (그림 3)과 같이 3 종류의 대분류로 나눌 수 있으며, 3 종의 데이터는 다시 15 종의 소분류로 나뉜다.



(그림 3) 환경데이터 분류

각 소분류 데이터 수집을 위한 검색어 및 제외어는 아래 <표 1>과 같다.

수집한 데이터는 대한민국 표준시에 따라 Tweet 이 작성된 날짜 별로 나뉘어 csv 파일 형태로 저장된다.

4.1. 로그 파일

시스템 재부팅 및 프로그램 실행 실패 등의 상황에서도 이전 수집 시점에 이어서 수집할 수 있도록, 데이터 수집 성공 시 로그 파일에 데이터 수집 시점과 수집 데이터 종류, 수집 기간 정보를 남긴다. 크롤링 시스템은 프로그램 실행 시점마다 로그 파일을 통해 이전 수집 시점으로부터 이어서 데이터를 수집한다.

```
Collecting tweeter data of 05-Jan-2020...
  Keyword     | Response data   | 09:00~23:59 | 00:00~09:00 | invalid data | running time
지하수        |     26 tweets   |        23   |         3   |          0   |  0:00:03
하천          |   1286 tweets   |       979   |       304   |          3   |  0:02:03
하천          |      3 tweets   |         1   |         2   |          0   |  0:02:05
하천          |      7 tweets   |         6   |         1   |          0   |  0:02:07
호소          |    843 tweets   |       565   |       271   |          7   |  0:03:23
상수도        |     40 tweets   |        29   |        11   |          0   |  0:03:27
하수도        |     21 tweets   |        16   |         5   |          0   |  0:03:30
물재난        |    297 tweets   |       231   |        65   |          1   |  0:03:58
대기          |    852 tweets   |       645   |       204   |          3   |  0:05:18
폐기물        |      6 tweets   |         4   |         2   |          0   |  0:05:20
폐기물        |     78 tweets   |        47   |        31   |          0   |  0:05:27
화학물질      |     22 tweets   |        15   |         7   |          0   |  0:05:31
기상변화      |    236 tweets   |       156   |        80   |          0   |  0:05:53
기후변화      |    573 tweets   |       395   |       177   |          1   |  0:06:46
지질          |    960 tweets   |       677   |       280   |          3   |  0:08:15
지형          |    121 tweets   |        88   |        33   |          0   |  0:08:27
토양          |     19 tweets   |        15   |         4   |          0   |  0:08:29
생태계        |     17 tweets   |        11   |         4   |          2   |  0:08:32
Collecting tweeter data of 06-Jan-2020...
  Keyword     | Response data   | 09:00~23:59 | 00:00~09:00 | invalid data | running time
지하수        |     27 tweets   |        16   |        11   |          0   |  0:00:03
하천          |   1161 tweets   |       810   |       350   |          1   |  0:01:50
하천          |      5 tweets   |         3   |         2   |          0   |  0:01:52
하천          |      9 tweets   |         7   |         2   |          0   |  0:01:54
호소          |    849 tweets   |       579   |       263   |          7   |  0:03:11
```

(그림 4) 데이터 수집중인 시스템의 UI

date	time	permalink	text
2020-01-06	8:44:44	https://twi	[실시간 2위 소한] "소한맞아?" 겨울에 눈대신 비...때아닌 100㎜ 이상
2020-01-06	8:38:03	https://twi	ㅠㅠㅠ마음속으로는 이미 홍수야ㅠㅠ
2020-01-06	8:34:05	https://twi	그래도 폭우는아니니 다행이에여^^
2020-01-06	8:25:50	https://twi	홍철: 인천대 총장 "4대강사업은 이 정부의 대표적인 지역정책. 홍수방
2020-01-06	8:18:58	https://twi	리안한테 오가게 될 것 같네요(비서실: 눈물바다 홍수)
2020-01-06	8:18:24	https://twi	베트남의 수도 Ha N?i[하 노이]는 물(河) 안쪽(內)이라는 이름답게 홍
2020-01-06	8:12:18	https://twi	중드 특 : A작품 끝났다고 덕질끝나는거 아님. 그작품 최애캐의 배우
2020-01-06	8:08:34	https://twi	청계천 정비를 했음에도 홍수로 인해서 청계천이 범람했다.범람을 막
2020-01-06	8:00:48	https://twi	떡밥 늘 홍수지만 연말연초엔 미친듯이 터지네여... 컴백전까지는...다...
2020-01-06	7:52:09	https://twi	" ..... " 푹ㅡ. (적막을 깨더니 보이는 것은 자신의 두 손과 다리에 칼을
2020-01-06	7:27:24	https://twi	폭우
2020-01-06	7:18:22	https://twi	SCP-3637	무호화	이 사랑은 많은 물이 꺼치지 못하겠고 홍수라도
2020-01-06	7:17:35	https://twi	폭우가 눈 앞을 가리는 어느 밤 그 어떤 불빛조차 없던 그날 라디오도
2020-01-06	7:15:49	https://twi	폭우를 대비해서 지하에 건설된 대규모 집수조 등을 둘러보는 도시 속
2020-01-06	7:14:47	https://twi	대덕구 여름철 집중호우대비 물막이벽 보급
2020-01-06	7:01:23	https://twi	또 비가 많이 내리는 홍수철엔 나무가 최대한 물을 흡수했다가 나무
2020-01-06	6:08:23	https://twi	한양성내 홍수가 잦아서 개천을 정비하기 위해서 개천도감을 만들었

(그림 5) 수집된 데이터(물재난)

5. 결론

본 논문에서는 가장 대표적인 소셜 네트워크 서비스인 트위터의 데이터를 자동화되어 지속적으로 수집하기 위한 시스템을 설계하였다. 그 후, 환경 데이터를 수집하는 Case study 를 통해 트위터 데이터가 정상적으로 수집됨을 확인하였다.

추후에 데이터 수집 속도를 향상시키기 위한 병렬 수집 시스템과 같은 방법에 대한 연구와, 수집된 데이터를 Hadoop 과 같은 대용량 데이터 분석 시스템에 저장하여 데이터 관리, 분석, 시각화에 쉽게 활용할 수 있도록 추가 연구가 수행된다면 시스템의 완성도가 높아질 것으로 기대된다.

<표 1> 15 종 환경 데이터 수집을 위한 검색어와 제외어

키워드	검색어	제외어
하천	하천수질 OR 수질오염 OR 유역오염 OR 물오염 OR 하천유량 OR 하천정비 OR 하천 OR 개천 OR 도랑 OR 개울 OR 한강 OR 영산강 OR 낙동강 OR 금강 OR4 대강 OR 이포보 OR 여주보 OR 강천보 OR 함안창녕보 OR 창녕합천보 OR 달성보 OR 강정고령보 OR 칠곡보 OR 충주댐 OR 횡성댐 OR 안동댐 OR 임하댐 OR 합천댐 OR 남강댐 OR 밀양댐 OR 군위댐 OR 부항댐 OR 대청댐 OR 용담댐 OR 섬진강댐 OR 주암댐 OR 부안댐 OR 보령댐 OR 장흥댐 OR 영주댐 OR 청평댐 OR 화천댐 OR 괴산댐 OR 춘천댐 OR 의암댐 OR 팔당댐 OR 도암댐 OR 보성강댐 OR 승촌보 OR 소양강댐 OR 하천환경 OR 하천관리 OR 구미보 OR 낙단보 OR 상주보 OR 백제보 OR 공주보 OR 세종보 OR 죽산보	매매 OR 투어 OR 여행 OR 관광 OR 드라이브 OR 맛집 OR 펜션 OR 전원주택 OR 업체 OR 가볼만한곳
호소	호수 OR 늪 OR 저수지 OR 소택 OR 습원 OR 담수호 OR 민물호수 OR 담호	매매 OR 투어 OR 여행 OR 관광 OR 드라이브 OR 맛집 OR 펜션 OR 전원주택 OR 업체 OR 가볼만한곳
상수도	상수도 OR 수돗물 OR 상수관 OR 상하수도	제품 OR 필터 OR 미국 OR 매매 OR 투어 OR 여행 OR 관광 OR 드라이브 OR 맛집 OR 펜션
하수도	하수도 OR 하수처리 OR 하수관 OR 상하수도	매매 OR 투어 OR 여행 OR 관광 OR 드라이브 OR 맛집 OR 펜션 OR 전원주택 OR 업체 OR 고객
지하수	지하수 OR 암반수	생수 OR 천연 OR 매매 OR 투어 OR 여행 OR 관광 OR 드라이브 OR 맛집 OR 펜션 OR 전원주택
물재난	홍수 OR 폭우 OR 집중호우 OR 하천범람	매매 OR 투어 OR 여행 OR 관광 OR 드라이브 OR 맛집 OR 펜션 OR 전원주택 OR 업체 OR 가볼만한곳
대기	대기환경 OR 미세먼지 OR 미먼 OR 황사 OR 대기오염 OR 대기 배출 OR 도로 재비산 먼지 OR 자동차 배출가스 OR 실내공기질 OR 공기오염 OR 매연 OR 스모그 OR 오존 OR 부유먼지 OR 초미세먼지 OR 황산화물 OR 질소산화물	매매 OR 투어 OR 여행 OR 관광 OR 드라이브 OR 맛집 OR 펜션 OR 전원주택 OR 업체 OR 가볼만한곳
폐기물	폐기물 (음식물 OR 분뇨 OR 사업장 OR 생활 OR 건설 OR 자동차 OR 지정 OR 전기 OR 전자 OR 매립 OR 소각 OR 재활용 OR 해역배출 OR 가연성 OR 불연성 OR 비가연성 OR 플라스틱 OR 생활 OR 처리) OR 재활용쓰레기 OR 음식물쓰레기	매매 OR 투어 OR 여행 OR 관광 OR 드라이브 OR 맛집 OR 펜션 OR 전원주택 OR 업체 OR 가볼만한곳
화학물질	화학물질 OR 환경호르몬 OR 화학사고 OR 유해화학물질 OR 잔류성유기오염물질 OR 화학제품성분 OR 유독물질 OR 유해물질	매매 OR 투어 OR 여행 OR 관광 OR 드라이브 OR 맛집 OR 펜션 OR 전원주택 OR 업체 OR 가볼만한곳
기상변화	일사량 OR 일조량 OR 강우량 OR 습도 OR 운량 OR 기압 OR 풍향 OR 풍속 OR 기상변화 OR 강수량	매매 OR 투어 OR 여행 OR 관광 OR 드라이브 OR 맛집 OR 펜션 OR 전원주택 OR 업체 OR 가볼만한곳
기후변화	한파 OR 폭염 OR 가뭄 OR 집중호우 OR 지구온도 OR 해수면 OR 온난화 OR 대기혼탁도 OR 오존층 OR 엘니뇨 OR 라니냐 OR 풍후 OR 기온변화 OR 강수량변화 OR 기후변화 OR 폭설 OR 온실가스 OR 온실효과	매매 OR 투어 OR 여행 OR 관광 OR 드라이브 OR 맛집 OR 펜션 OR 전원주택 OR 업체 OR 가볼만한곳
지질	지질 OR 지층 OR 지질재해 OR 지질학 OR 광물자원 OR 해저지질 OR 해저광물 OR 석유자원 OR 산사태 OR 지열 OR 지질자원 OR 지질도 OR 지진 OR 국토지질 OR 석유해저 OR 지반	매매 OR 투어 OR 여행 OR 관광 OR 드라이브 OR 맛집 OR 펜션 OR 전원주택 OR 업체 OR 가볼만한곳
지형	급경사지 OR 지표면 OR 지형굴곡 OR 지형경사도 OR 지형표고 OR 지형 OR 씽크홀 OR 지형도	매매 OR 투어 OR 여행 OR 관광 OR 드라이브 OR 맛집 OR 펜션 OR 전원주택 OR 업체 OR 가볼만한곳
토양	토양 OR 토양오염 OR 토양환경 OR 성숙토	매매 OR 투어 OR 여행 OR 관광 OR 드라이브 OR 맛집 OR 펜션 OR 전원주택 OR 업체 OR 가볼만한곳
생태계	생태환경 OR 생태관광 OR 자연복원 OR 생태서비스 OR 생태계파괴 OR 국립습지 OR 환경영향평가 OR 생물계	매매 OR 투어 OR 여행 OR 관광 OR 드라이브 OR 맛집 OR 펜션 OR 전원주택 OR 업체 OR 가볼만한곳

참고문헌

[1] A. Bhardwaj, R. Singh, V. Deep and P. Sharma, BDT3V — A Technique for Big Data Testing considering 3V's, 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT), Bangalore, India, 2018, pp. 222-225.

[2] Byun, C., Kim, Y., Lee, H., & Kim, K. K. Automated Twitter data collecting tool and case study with rule-based analysis. In Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services, Bali Indonesia, December 2012, pp. 196-204.

[3] 최민석. 효율적인 트윗 분석 시스템 설계 및 구현 방법. Journal of Digital Convergence, 13(2), pp. 43-50, 2015

[4] 하승도, 선동성, 이해준, 이상구. 대화 말뭉치 구축을 위한 웹 크롤러 기반 대화 수집기. 한국정보과학회 학술발표논문집, pp. 334-336, 2016

[5] Twitter, Inc. "Overview — Twitter Developers", 2020, https://developer.twitter.com/en/docs/tweets/search/overview

[6] PyPI, "GetOldTweets3 · PyPI". 2020, https://pypi.org/project/GetOldTweets3/