창의코딩파이썬 프로젝트

네이버 실시간검색어 분석 및 시각화 프로그램

한림대학교 전자정보공과대학교 컴퓨터공학과

2017 June 17

작성자: 김기혁

GitHub : <https://github.com/rlgur41>



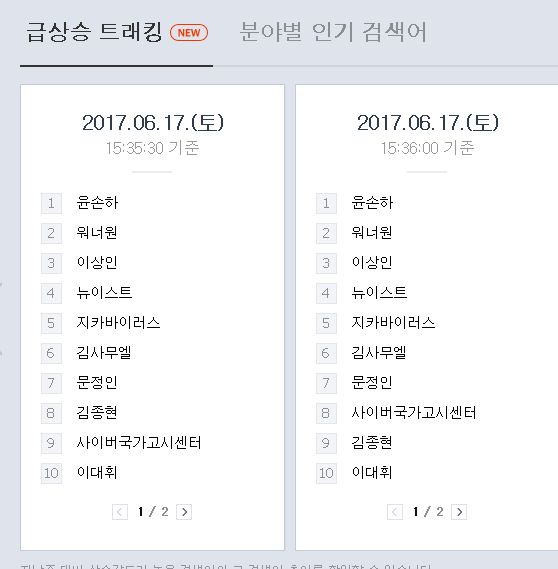
Index

1. 개발 배경
2. 개발 환경
3. 데이터 수집 처리, 방법 및 시각화 개요
4. 전체 프로그램 Flow Chart
5. 데이터 수집 프로그램 Flow chart
6. 데이터 시각화 프로그램 Flow chart

개발 배경

한국에서 높은 웹사이트 점유율을 가지고 있는 네이버의 핵심 모듈인 실시간 검색어는 그 날의 Hot issue를 나타내는 지표라고 생각했다. 이를 분석한다면 하루 동안 어떤 사건이 가장 이슈였는지 알 수 있을 텐데, 네이버 DataLab에서는 시간대별, 날짜별 실시간 검색어 순위만 제공해줄 뿐(A), 전체적인 검색어 빈도에 대해서는 제공해주지 않았다. 그래서 하루 동안 네이버 실시간 검색어 결과를 저장하고, 검색어 출현 빈도를 계산해서 높은 빈도의 검색어 중 사회적, 문화적으로 의미있는 검색어를 제공해주면 재미있을 것 같아서 이 프로젝트를 시작하게 되었다.

(A) : 네이버 Data-lab에서 제공해주는 검색어 정보



개발 환경

이 프로그램을 개발한 환경은 다음과 같다.

운영체제 : Windows10

사용 언어 : Python3.6

사용 라이브러리 : re, request, time, Turtle

개발툴 : Jupyter notebook

데이터 처리 플랫폼 : Hadoop, Spark

전체 소스코드 : https://github.com/rlgur41/Naver\_hack

re, request 라이브러리는 네이버 실시간 검색어를 얻는 데 필요한 모듈들이다. time은 네이버에서 많은 요청을 하는 IP에 Block을 걸기 때문에, 요청 시간을 조절하기 위해서 사용하였다. Turtle은 처리결과를 시각화하기 위해서 사용하였다. 데이터 처리를 위해 사용한 플랫폼으로 Hadoop과 Spark가 있는데, 이는 파이썬을 사용하지 않았으므로 구체적인 설명은 하지 않겠다. Hadoop과 Spark의 소개 및 처리에 사용된 프로그램 설명은 아래 자료링크로 대체한다.

Spark : https://github.com/rlgur41/Seminar\_presentation/blob/master/Apache%20Spark.pptx

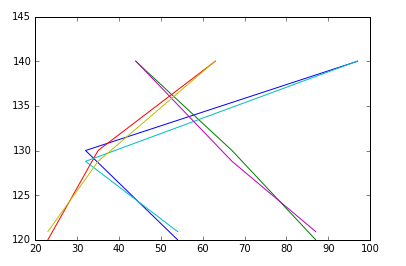
Hadoop

https://github.com/rlgur41/Seminar\_presentation/blob/master/Hadoop\_mapreduce.pptx

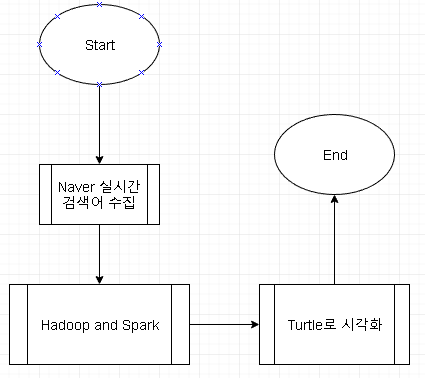
데이터 수집 처리, 방법 및 시각화 개요

분석 데이터는 2017년 5월 6일 네이버 실시간 검색어 1위부터 20위의 출현 빈도이고, 수집 시간은 00:00 ~ 23:00 이다. 데이터 분석에는 빅데이터 분산처리 플랫폼 Hadoop의 HDFS와 In memory기반 데이터처리 플랫폼 Spark를 사용하였고, Map Reduce를 이용해서 키워드의 빈도수를 계산하고, 출현빈도가 1000 이상인 키워드 중 중요한 키워드라고 생각하는 데이터를 파일로 저장하였다. 저장한 파일은 Text형태 포맷(키워드, 검색빈도)과 같이 나타나는데(Ex : 김기혁, 232) 처리 결과가 Text 형태로만 제공되는 것 이 재미없고 딱딱하다고 생각했다. Python에서는 이를 해결해주는 matplotlib 라는 그래프 라이브러리가 있는데, matplotlib는 입력 값을 그래프로 그려주는 라이브러리이다.(B) 하지만 이 라이브러리는 그래프가 실시간으로 그려지는 과정은 볼 수 없어서 조금 심심하다고 생각했다. 그러던 중 Python의 Turtle이라는 라이브러리를 알게 되었고, Turtle을 이용해서데이터를 그래프로 나타내고, 그래프가 그려지는 과정을 실시간으로 볼 수 있으면 재미있을 것 같아서 그래픽툴로 Turtle을 선택하였다. 데이터 처리 후 저장한 파일의 Text 데이터를 Turtle을 이용해 그래프로 그려주고, 그려지는 과정도 실시간으로 볼 수 있다.

(B) : matplotlib 그래프



전체 프로젝트 Flow Chart

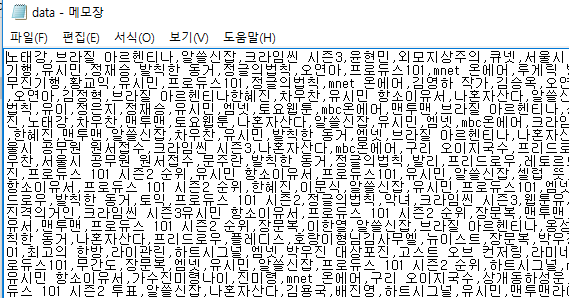


프로젝트의 전체적인 진행순서는 위와 같다.

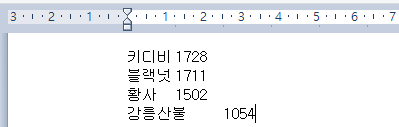
먼저 Naver 실시간 검색어를 하루 동안 수집하여 파일 형태로 만든다.

이 파일은 콤마(,)를 구분자로 하여 (키워드1, 키워드2, 키워드3... 키워드20) 과 같이 저장되어 있다(C). 이 파일을 Hadoop에 HDFS에 저장하여 분산처리가 가능하게끔 하고, Spark에서 HDFS에 해당 파일에 접근해 실시간 검색어 출현 빈도를 MapReduce를 이용해서 계산한다. 이 작업을 마치면 파일의 내부는 (키워드1 560, 키워드2 780) (D) 같이 특정 키워드에 대한 검색빈도가 저장되어있다. 이 파일의 항목을 가지고 Turtle를 이용해 그래프를 그리게 된다.

(C) : 분석할 데이터(네이버 실시간 검색어)

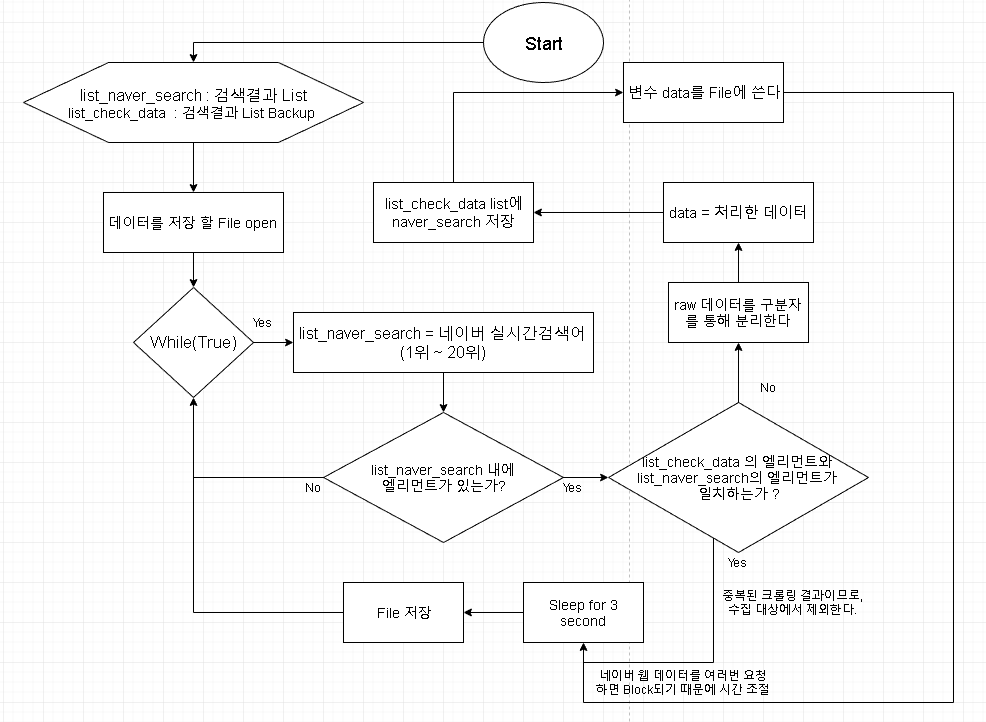


(D) : 처리한 데이터(네이버 실시간 검색어 빈도 수)



데이터 수집 프로그램Flow chart

File : data-lab.py

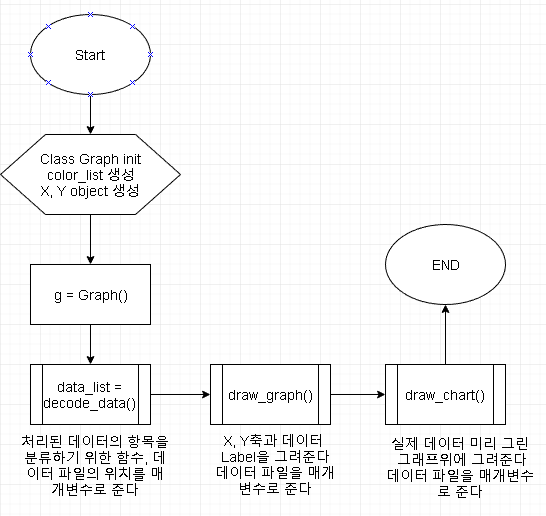




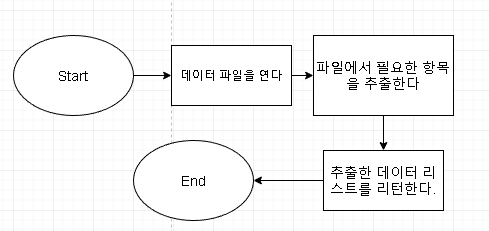
이 프로그램은 네이버 실시간 검색어 1위부터 20위를 수집하여 파일로 저장하는 프로그램이다. list\_naver\_search 라는 list type 변수에 re.findall() 함수를 이용해 현재 네이버 실시간 검색어를 저장하고, 중복되는 실시간 검색어를 제거하기 위해 list\_check\_data 리스트를 생성하여, 실시간 검색어를 새로 받으면 이전의 검색어들을 list\_check\_data에 저장하여 현재 받아온 실시간 검색어와 list\_check\_data를 비교하여 실시간검색어가 갱신되었는지 확인한 후 , 갱신되었다면 그 정보를 file에 쓴다. 네이버는 웹사이트에 많은 요청을 하는 IP를 Block 시키기 때문에, 3초정도 수집행위를 멈추고, 변경한 File을 저장하고, 다시 같은 작업을 반복한다.

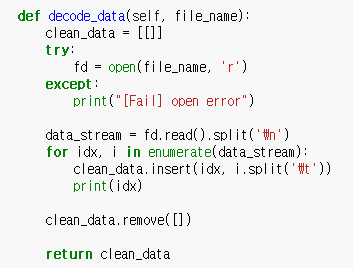
데이터 시각화 프로그램

File : turtle\_graph.py



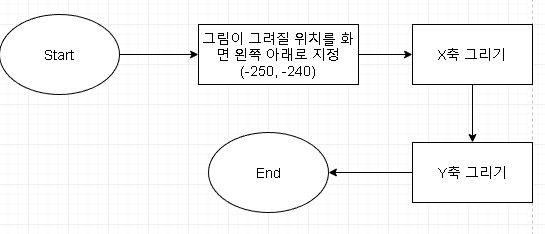
시각화 프로그램의 전체 흐름이다. 그래프를 그려주기 위해 Graph Class를 생성하고, 클래스 안에 있는 decode\_data()와 draw\_grah(), draw\_chart()를 순서대로 호출한다.

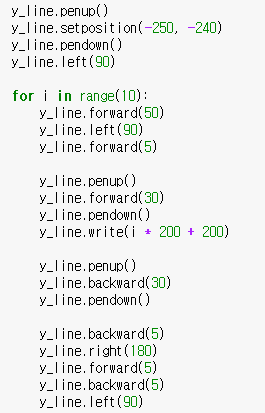
decode\_data() 



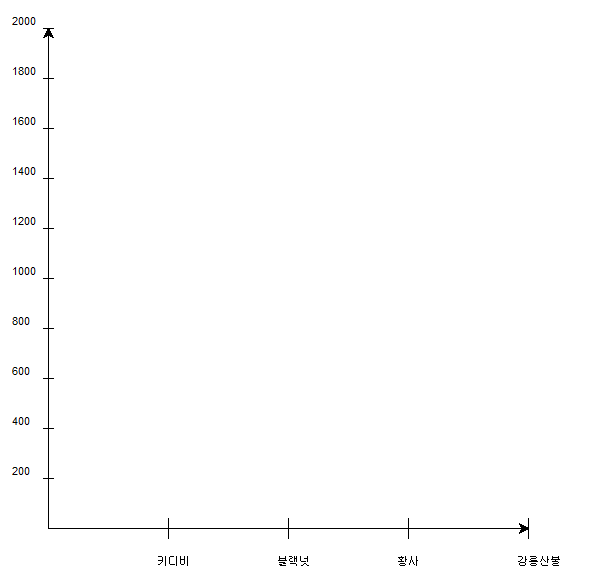
매개변수로 오는 file\_name은 저장된 파일의 절대 경로이다. 이 file\_name위치의 파일은 (키워드 \t 빈도수 \n) 과 같은 형태로 저장되어있다. 여기서 내가 필요한 정보는 키워드와 빈도수이므로, 해당 정보만 얻을 수 있도록 읽은 파일을 \n 을 기준으로 먼저 나누고, 그 안에서 \t를 기준으로 레코드를 나누어 얻은 키워드와 빈도수를 clean\_data 리스트에 넣고, 이를 return 한다.

draw\_graph()

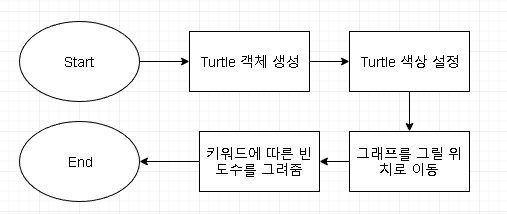


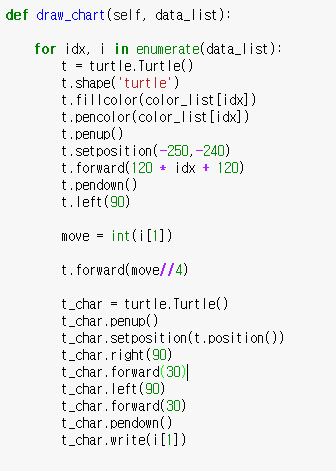
 

draw\_graph함수는 매개변수로 data\_list를 받는다. data\_list는 decode\_data()에서 필요한 데이터만 추출해서 넘겨준 리스트이다. 이 리스트에 있는 키워드와 빈도수를 이용해 그래프를 그린다. 먼저 그림을 그릴 위치를 화면 아래 왼쪽(-250, 240)으로 하고, X축과 Y축을 그린다. X축은 키워드를 나타내고, Y축은 빈도수를 나타낸다. X축을 그릴 때 키워드의 내용도 같이 그려지고, Y축은 200 간격으로 하고 최대 빈도수 비율에 맞게 그려진다. 아래는 draw\_graph()를 실행하면 나오는 결과다.

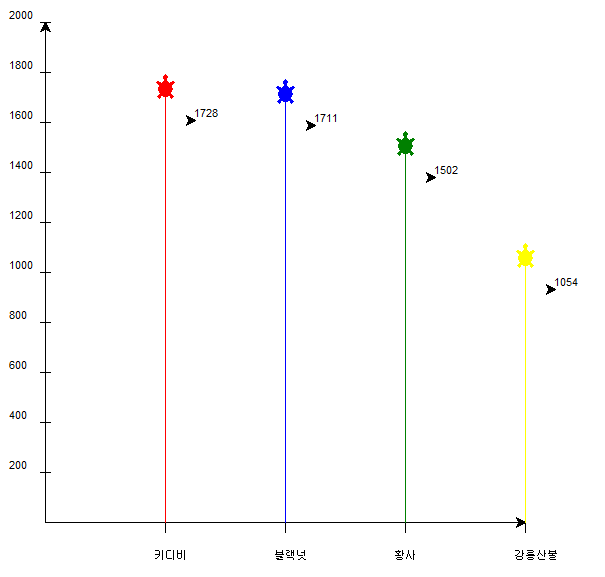


draw\_chart()





draw\_chart()는 매개변수로 data\_list를 받는다. data\_list는 decode\_data()에서 필요한 데이터만 추출해서 넘겨준 리스트이다. draw\_chart() 함수가 시작되면, 먼저 Turtle 객체를 생성하고, 그 객체에 색상과 모형을 정해주고, draw\_graph()에서 그려준 키워드 항목 위에 data\_list에 있는 키워드에 따른 빈도수를 그려주고, 빈도수를 그래프 옆에 출력해준다. 아래는 draw\_chart()를 실행했을때 나오는 결과다.



결과 그래프를 보면 파일에서 확인할 수 있었던 키디비, 블랙넛, 황사, 강릉산불이 X축의 Lable로 지정되어있다. 5월 6일에 키디비와 블랙넛이 고소 문제로 인터넷을 시끄럽게 달군 적이 있었다. 실시간검색어에서도 그 여파가 반영되어 많은 검색빈도를 나타냈다. 또 중국발 황사로 황사와 더불어 황사마스크, 공기청정기 등 환경과 관련된 검색어도 많은 빈도를 나타냈고, 강릉에서 발생한 산불로 강릉산불이라는 키워드가 많이 검색되었음을 확인 할 수 있었다.

Thank you

reference :

https://financedata.github.io/posts/naver\_realtime\_search\_crawling.html

http://datalab.naver.com/