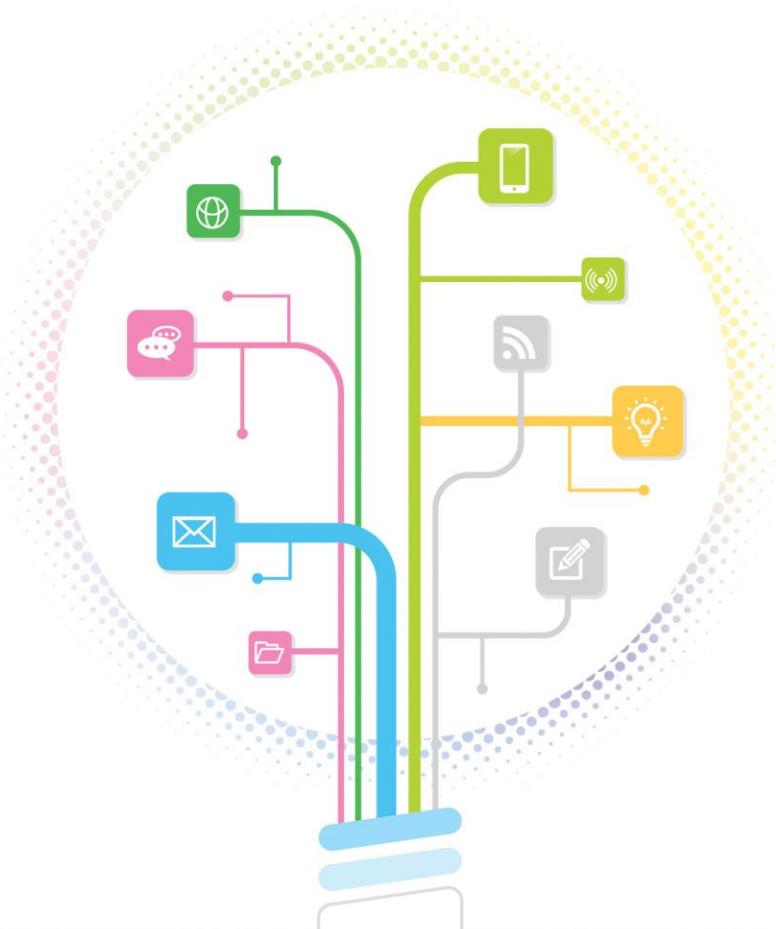


데이터 분석 콘텐츠 활용 매뉴얼

3 소셜



미래창조과학부



한국정보화진흥원



KBIG
빅데이터
전략센터

CONTENTS

Beginning Level 초급과정

I 개요

개요	9
----	---

II 수집

개요	13
수집 데이터	14
데이터 수집	16
데이터 작업 영역 이동 스크립트	19

III 가공

개요	23
데이터 가공 R 스크립트	28

IV 저장

개요	33
R Studio 활용 저장	34



V 분석

개요	39
R Studio 활용 분석	40
R Studio 저장	43

VI 시각화

개요	47
분석 데이터 시각화	49
데이터 분석	51

VII 예제 문제

예제 문제1. 트위터 데이터에서 일별 키워드 등장 건수를 계산하라.	55
예제 문제2. 트위터 사용자별 이용 건수를 계산하라.	56

CONTENTS

Intermediate Level **중급과정**

I 개요

개요	61
----	----

II 수집

개요	65
수집 데이터	66
데이터 수집	70
데이터 작업 영역 이동 스크립트	73

III 가공

개요	77
데이터 가공 R 스크립트	84

IV 저장

개요	89
R Studio 활용 저장	90



V 분석

개요	95
R Studio 활용 분석	96
R Studio 저장	106

VI 시각화

개요	111
분석 데이터 시각화	113
데이터 분석	115

VII 예제 문제

예제 문제1. 뉴스 데이터에서 월별 키워드 빈도를 계산하라.	119
예제 문제2. 특정 주간의 소셜 키워드 빈도를 계산하라.	120



소셜 

Beginning Level

초급과정







I 개요

개요

9

I

개요

> 개요

소셜 미디어 데이터는 솔트룩스에서 제공해 준 트위터를 바탕으로, 소셜 네트워크에서 생산되어지는 비정형 구조의 의미 정보를 형태소 분석으로 의미가 있는 키워드 만을 추출하여 특정 기간 또는 소셜 미디어 데이터 전체에 대한 이슈 키워드를 알아보고자 한다. 이러한 방법으로 이슈 키워드 분석을 통하여 사회적 이슈와 관심 정보를 확인할 수 있다.

> 활용 데이터

- **sample_201101.json** : 트위터 데이터(2011년 1월 중)

> 선행학습

- **리눅스** – 파일시스템 구조, 쉘 명령어, 쉘 스크립트 실행 방법
- **R 프로그래밍 언어** – 파일 불러오기, 라이브러리 등록, 데이터 함수(프레임, 테이블), 그래프 함수, 제어문(함수 호출, 외부 함수, 함수 정의) 사용방법
- **데이터 구조** – CSV, JSON 데이터 구조, Text 파일 저장 구조 이해
- **R 차트** – 내부 차트(막대, 바, 원 등), 외부 차트(막대, 클라우드, 3D, D3 차트) 사용방법

▶ 요구사항

- 비정형 데이터 분석에서 불필요한 정보를 리스트로 정리하고, 이를 제거한다.
- 중요도가 높은 키워드 또는 특수 문자 키워드를 단어 사전에 추가하고, 단어 사전을 이용하여 형태소 분석을 실시한다.
- 추출된 명사형 데이터를 이용하여 단어의 빈도를 계산하라.

▶ 분석 절차

- 수집된 소셜 데이터 셋을 분석 저장소로 복사한다.
- 비정형 데이터 분석을 위해 데이터 구조인 JSON 포맷을 이해하고, 이 중에서 문장 분석을 위하여 메시지 정보의 키에 대응하는 값 정보만을 추출한다.
- 트위터 정보는 @RT 정보, URL 정보, 숫자 등과 같은 불필요한 정보를 포함하고 있기 때문에 추출된 트위터 메시지 값 정보에서 이를 정보를 제거하는 정제 처리를 실시한다.
- 분석 중에 관심있는 키워드나 중요도가 높은 키워드를 단어 사전에 추가하여 핵심 키워드에 대한 단어 사전의 가중치 정보를 보완한다.
- 정제된 트위터의 메시지 값에서 명사형 키워드만을 추출하고, 키워드 출현 빈도를 계산한다.
- 계산된 키워드 출현 빈도를 WordCloud 그래프로 출력하고, 결과 데이터를 저장한다.
- 출력된 그래프에서 소셜 네트워크 상에서 참여자들이 거론하였던 이슈 키워드를 확인할 수 있으며, 이를 통하여 사회적 관심 분야에 대한 패턴 또는 흐름을 판단할 수 있다.



II 수집

개요	13
수집 데이터	14
데이터 수집	16
데이터 작업 영역 이동 스크립트	19



수집



개요

소셜 데이터는 소셜 네트워크 서비스, 소셜 미디어 서비스, 마이크로 소셜 서비스, 뉴스 미디어 서비스 등을 통하여 거론된 사회적 정보들로 구성되며, 실시간으로 사회적 이슈를 분석하기 위해서는 실시간으로 데이터를 수집하고 관리되어야 한다. 이를 위해 소셜 데이터의 수집은 소셜 미디어 서비스(트위터, 페이스북, 블로그, 뉴스 등)에서 제공하는 API 또는 수집기를 통해 실시간으로 수집할 수 있으며, 사회적 이슈나 비즈니스 분석 등과 같은 다양한 사회적 분석 모델에 적용할 수 있다.

▶ 수집 방법

- **API 데이터 수집** : 소셜 데이터의 수집은 API를 이용하여 일정량의 데이터를 실시간으로 수집하고 있으며, 각 소셜 서비스 회사의 정책에 따라 수집량을 제한하고 있다.
- **데이터 제공** : 소셜 데이터는 OpenAPI, 자료수집기 (Crawler), 데이터 구매 등으로 데이터를 수집 할 수 있으며, 실습용 자료는 빅데이터 분석활용센터에 접속하여 소셜 초급 데이터 셋을 다운로드 받을 수 있도록 원시데이터를 제공하고 있다.

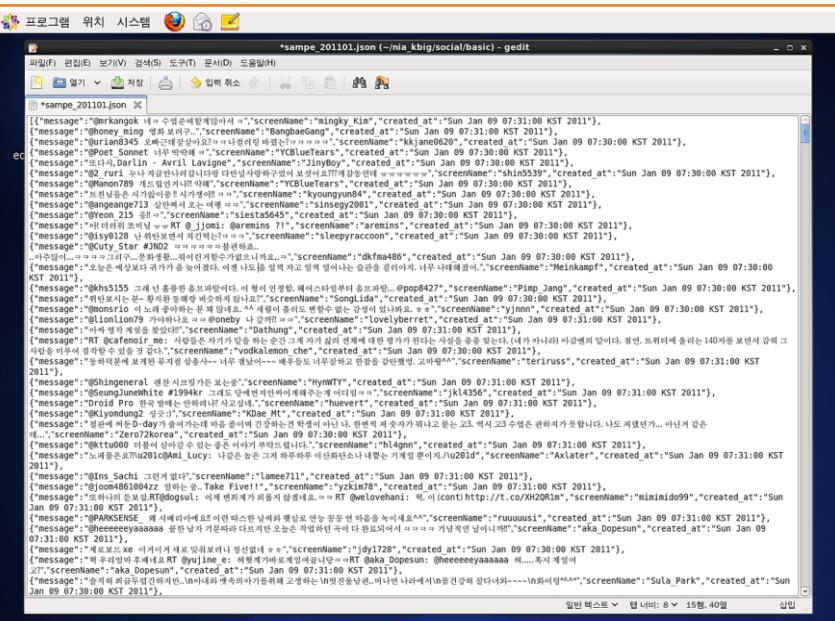


용어정리

- **비정형 데이터(Unstructured Data)** : 일정한 규격이나 형태를 지닌 숫자 데이터와 달리 그림, 영상, 문서 등과 같이 서로 다른 형태의 구조화되지 않은 데이터

> 수집 데이터

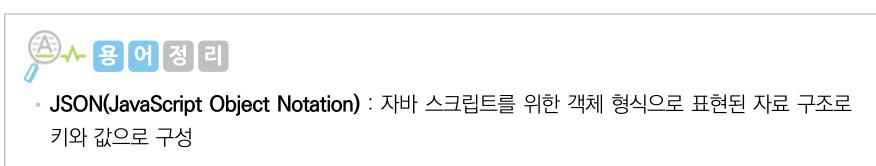
▶ 트위터 데이터(sample_201101.json)



▶ 트위터의 데이터 구조

- **JSON 구조** : 트위터에서 제공하는 데이터는 비정형 데이터로 빠르게 표현할 수 있는 JSON 구조를 가진다. 다양한 종류의 키와 값으로 구성되지만 분석에 필요한 데이터만을 수집하기 때문에 핵심 키로 재구성하였다.

- **message** : 소셜미디어 참여자들이 자신의 계정으로 업로드한 데이터 정보
 - **screenName** : 소셜미디어 참여자의 닉네임 정보
 - **create_at** : 트위터에 업로드한 시간 정보



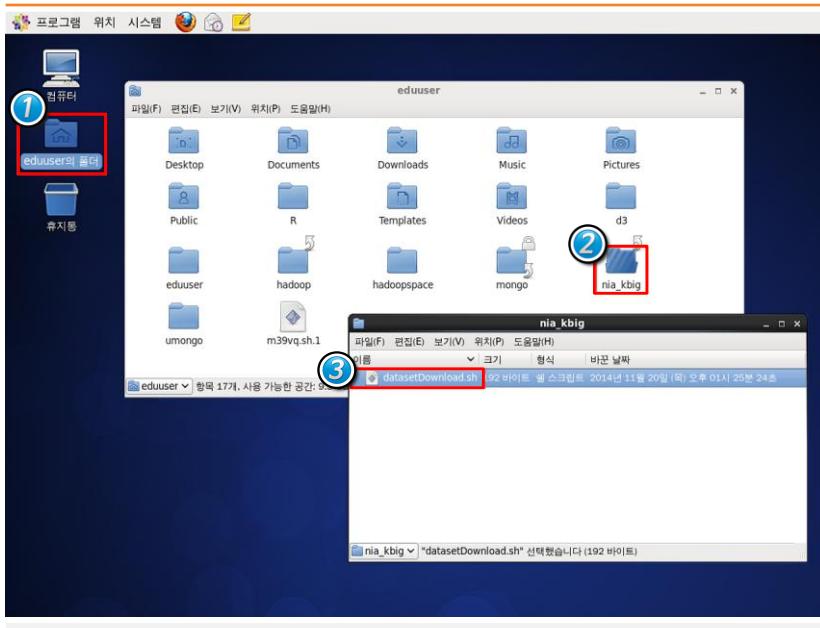
▶ 뉴스 데이터 JSON 구조의 예

```
01. [
02. {
03.   "message": "@mrkangok 네ㅋ 수업준비할게많아서 ㅋ",
04.   "screenName": "mingky_Kim",
05.   "created_at": "Sun Jan 09 07:31:00 KST 2011"
06. },
07. {
08.   "message": "@honey_ming 영화 보려구..",
09.   "screenName": "BangbaeGang",
10.   "created_at": "Sun Jan 09 07:31:00 KST 2011"
11. },
12. ~ 중간 생략 ~
13. {
14.   "message": "@urian8345 오빠근데잘살아요?ㅋㅋ나컬러링 바꿨는?ㅋㅋㅋㅋ",
15.   "screenName": "kkjane0620",
16.   "created_at": "Sun Jan 09 07:30:00 KST 2011"
17. },
18. {
19.   "message": "@luckyjhs ㅊㅋㅊㅋ",
20.   "screenName": "kyokyoc",
21.   "created_at": "Mon Jan 10 03:15:00 KST 2011"
22. }
23. ]
```

> 데이터 수집

- 데이터 저장소에서 서버 로컬로 데이터 셋을 복사해 온다.
- **sample_201101.json** : 소셜 데이터

> 실습코드 디렉토리로 이동

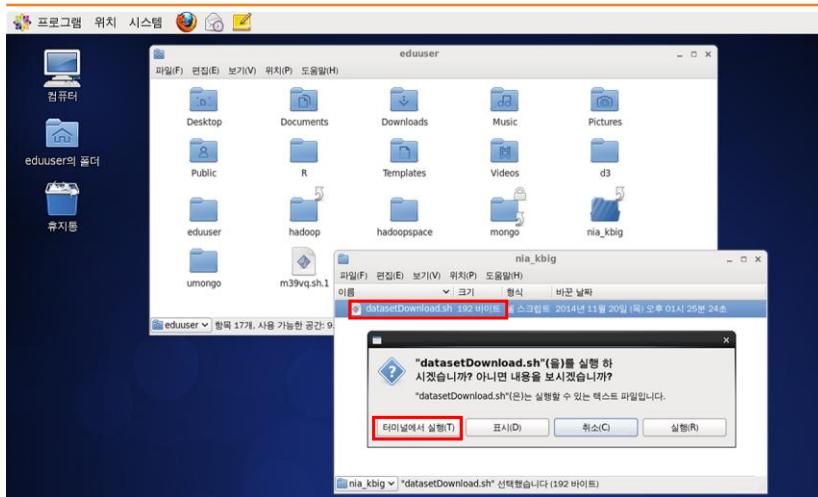


- ① 로그인 후 바탕화면에서 eduuser 폴더를 오픈한다.
- ② nia_kbig 폴더를 오픈한다.
- ③ datasetDownload.sh를 더블클릭하여 실행한다.

II. 수집

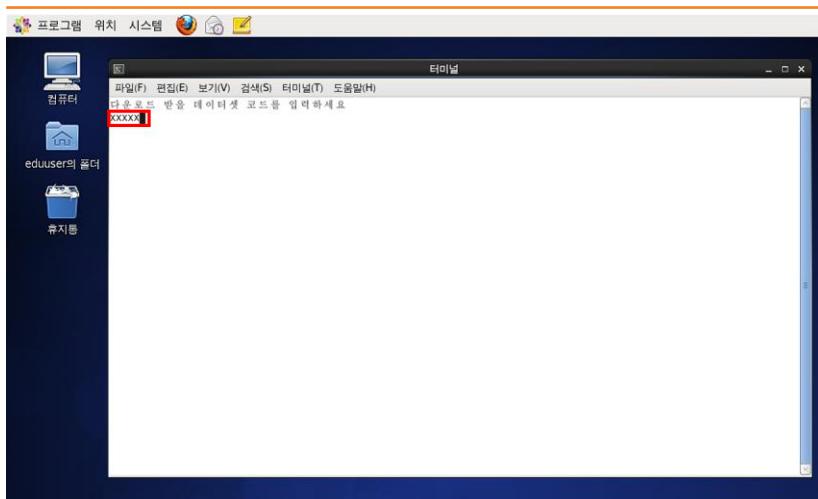
▶ 레파지토리에서 데이터 수집

datasetDownload.sh (원시데이터로 컬서버로 복사)



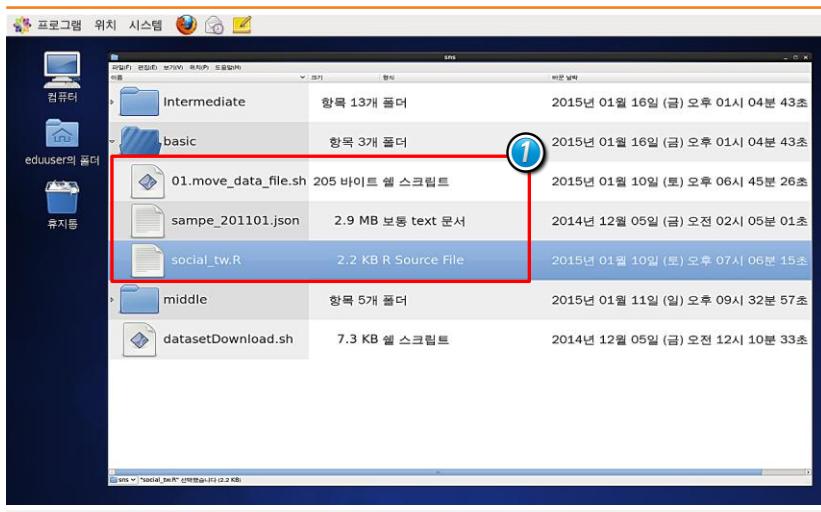
- '터미널에서 실행' 버튼을 클릭한다.

▶ 데이터셋 코드 입력



- 다운로드 받을 데이터셋 코드를 입력 후 엔터

> 데이터셋과 실습용 쉘 스크립트



- 실습용 데이터셋과 실습용 스크립트를 확인한다.

> ① 데이터 및 스크립트

- **01.move_data_file.sh** : 작업 영역 Data 폴더로 자료 이동하는 스크립트
- **social_tw.R** : R 분석 스크립트
- **sample_201101.json** : 소셜 트위터 분석 샘플 데이터

II. 수집

▶ 데이터 작업 영역 이동 스크립트(01.move_data_file.sh)

▶ 데이터 작업 공간으로 이동

- 로컬로 수집해 온 데이터를 작업 영역 Data 폴더로 자료를 이동하는 스크립트

01.move_data_file.sh

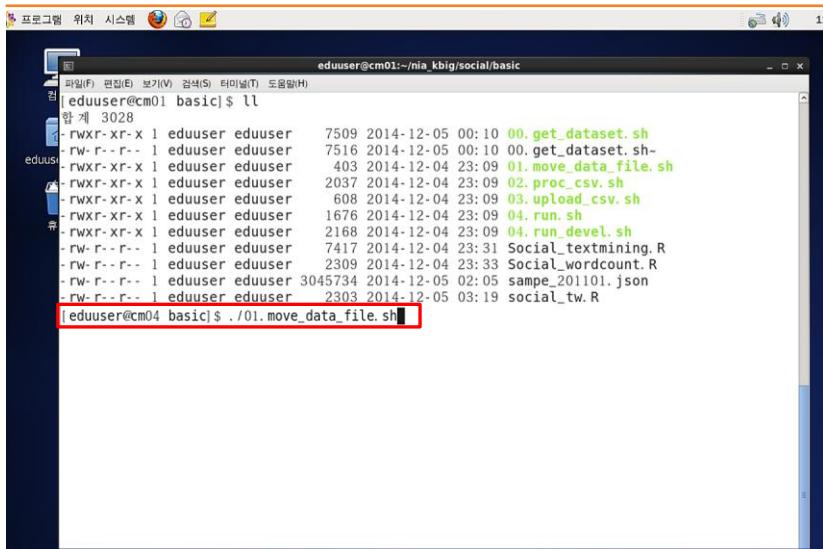
```
01. #!/bin/bash
02. #Social Data file define
03. TARGET_SOCIAL=/home/eduuser/nia_kbig/sns/basic/sample_*.json
04.
05. # 작업 디렉토리 정의
06. LOCAL_DIR=/home/eduuser/nia_kbig/data/
07. mv $TARGET_SOCIAL $LOCAL_DIR
08.
```



- 분석 원시 데이터 이동 스크립트 소스(01.move_data_file.sh)
- 라인 01~03 : 이동시킬 데이터 파일과 파일의 위치를 “TARGET_SOCIAL”로 정의하며, 기호 “#”은 주석을 의미한다.
- 라인 05~06 : 데이터를 이동시킬 위치 정보를 “LOCAL_DIR”이라는 이름으로 기록한다.
- 라인 07 : mv 명령을 이용하여 분석할 데이터를 소셜 폴더에서 분석 폴더로 이동시킨다.

▶ 수집 데이터셋 작업 영역 폴더 이동

- 로컬로 수집해 온 데이터를 작업 영역 Data 폴더로 자료를 이동하는 스크립트



The screenshot shows a terminal window titled 'eduuser@cm01:~/nia_kbig/social/basic'. The user has run the command 'll' to list files in the current directory. The output shows several files with their modification times and names. The file '01.move_data_file.sh' is highlighted with a red rectangle. Below the list, the user has typed the command './01.move_data_file.sh' and is waiting for input.

```
eduuser@cm01:~/nia_kbig/social/basic
[eduuser@cm01 basic]$ ll
합계 3028
drwxr-xr-x 1 eduuser eduuser 7509 2014-12-05 00:10 00.get_dataset.sh
-rw-r--r-- 1 eduuser eduuser 7516 2014-12-05 00:10 00.get_dataset.sh-
-rw-r--r-- 1 eduuser eduuser 403 2014-12-04 23:09 01.move_data_file.sh
drwxr-xr-x 1 eduuser eduuser 2037 2014-12-04 23:09 02.proc_csv.sh
drwxr-xr-x 1 eduuser eduuser 608 2014-12-04 23:09 03.upload_csv.sh
drwxr-xr-x 1 eduuser eduuser 1676 2014-12-04 23:09 04.run_sh
drwxr-xr-x 1 eduuser eduuser 2168 2014-12-04 23:09 04.run-devel.sh
-rw-r--r-- 1 eduuser eduuser 7417 2014-12-04 23:31 Social_textmining.R
-rw-r--r-- 1 eduuser eduuser 2309 2014-12-04 23:33 Social_wordcount.R
-rw-r--r-- 1 eduuser eduuser 3045734 2014-12-05 02:05 sampe_201101.json
-rw-r--r-- 1 eduuser eduuser 2303 2014-12-05 03:19 social_tw.R
[eduuser@cm04 basic]$ ./01.move_data_file.sh
```

- 로컬에 원시데이터를 작업 영역 폴더로 이동(/home/eduuser/nia_kbig/data/) 시킨다. `./01.move_data_file.sh` 입력 후 엔터





III 가공

개요	23
데이터 가공 R 스크립트	28



가공

> 개요

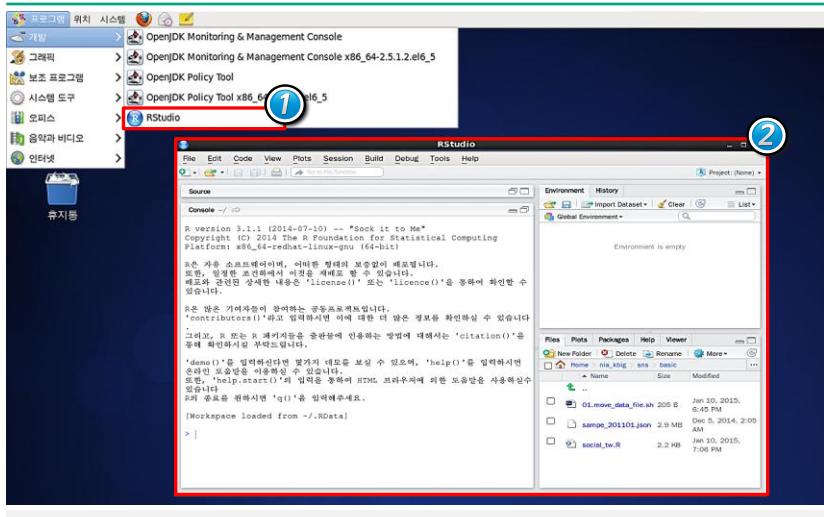
소셜 미디어 데이터의 가공은 수집된 데이터에서 구조에 따라 분석에 필요한 데이터만을 분류하고, 문장에서 불필요한 문자를 제거한다. 그리고 데이터에서 분석에 필요한 데이터만을 선별하고, 불필요한 데이터와 문장을 제거하여 분석의 효율성을 높인다. 이 과정에서 키워드 추출과 정제, 숫자와 문자를 분별, 사전 관리 등이 이루어진다.



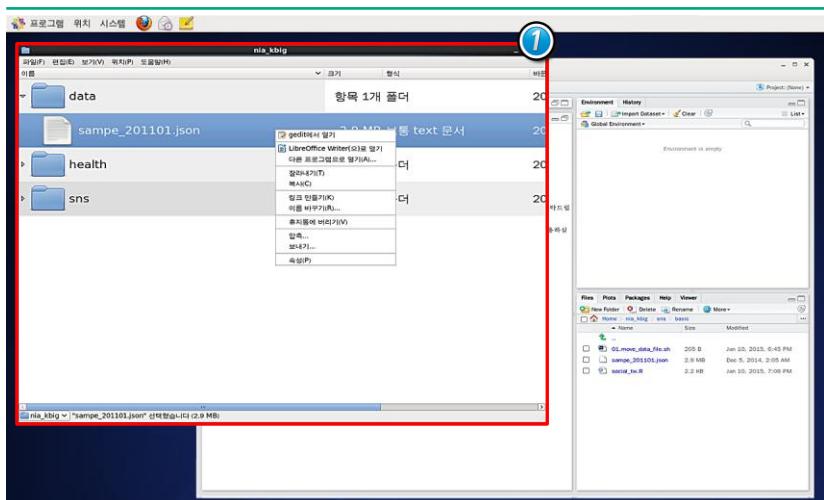
> 가공 방법

- **데이터 구조 가공** : 소셜 데이터는 한글과 영어로 구성되어 있기 때문에 Encoding 구조와 데이터 구조인 JSON 형태를 확인하고, 필요한 Encoding 구조인 “UTF-8”과 JSON 구조에서 필요한 Key와 Value를 추출한다.
- **데이터 가공 준비** : R에서 데이터 가공을 위한 라이브러리 리스트를 확인하고, 해당 라이브러리를 설치한다. 또한 한글 형태소 분석을 위해 데이터 상태를 확인한다.
- **트위터 문장 구조 가공** : 트위터는 RT(ReTwit)에서 발생된 데이터, URL 정보는 분석에서 사용되지 않기 때문에 1차적으로 해당 데이터를 삭제한다.
- **가공 분석을 위해**, 프로그래밍 도구인 R을 실행한다. R은 10,000 줄 이상의 데이터 처리 제약이 있기 때문에 대용량의 소셜 데이터 처리를 위해서는 Map Reduce와 결합하여 처리한다.

▶ 데이터 가공



1. ① 왼쪽 상단의 [“프로그램” 클릭] > [“개발” 클릭] > [“RStudio” 클릭]으로 분석 도구인 R을 실행한다. ② 는 R이 실행된 결과이다.

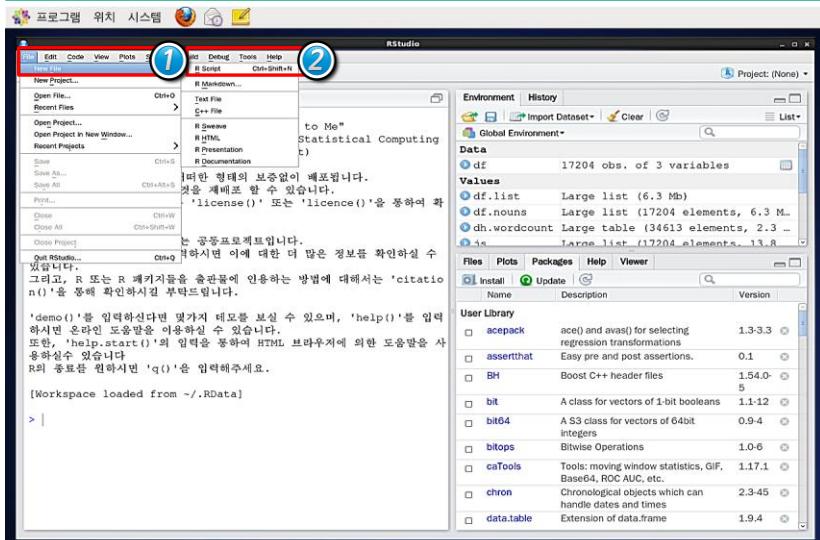


2. R 프로그램이 실행된 상태에서 트위터 샘플 데이터의 분석에 필요한 Key와 Value를 JSON 구조에서 확인한다. 작업 영역(①)인 “data” 폴더에서 이전 작업에서 이동 시킨 “sample_201101.json” 파일을 선택한 후, 더블클릭 또는 단축메뉴의 “gedit에서 열기”를 선택하여 확인할 수 있다.

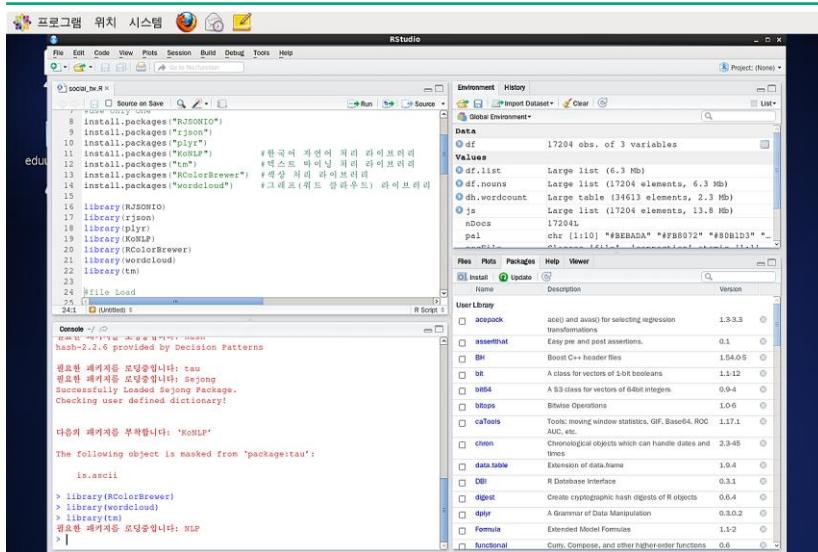
III. 가공

```
01. [
02. {
03.   "message": "@mrkangok 네ㅋ 수업준비할게많아서 ㅋ",
04.   "screenName": "mingky_Kim",
05.   "created_at": "Sun Jan 09 07:31:00 KST 2011"
06. },
07. {
08.   "message": "@honey_ming 영화 보려구..",
09.   "screenName": "BangbaeGang",
10.   "created_at": "Sun Jan 09 07:31:00 KST 2011"
11. },
12. {
13.   "message": "@urian8345 오빠근데잘살아요?ㅋㅋ나컬러링 바꿨는?ㅋㅋㅋㅋ",
14.   "screenName": "kkjane0620",
15.   "created_at": "Sun Jan 09 07:30:00 KST 2011"
```

- 전체 구조는 배열 구조(Array)로 구성되어 있다. 그리고 분석을 위한 문장에 대한 JSON key는 “message”이며, 해당 값은 비정형 구조로 이루어져 있다. 또한 “screenName”은 사용자의 이름으로 사용자 분석에 활용할 수 있으며, 작성된 날짜에 해당하는 JSON key는 “created_at”이다.



- ① ② 소셜 트위터 데이터에서 JSON 데이터 인식과 분리, 한글 데이터 분석에 필요한 기능을 위해 프로그램 작업 파일("New File" 클릭 > "R_Script" 클릭)을 선택한다.



4. 분석에 필요한 라이브러리 파일을 설치하기 위해, 필요한 라이브러리를 작성하고 실행한다.

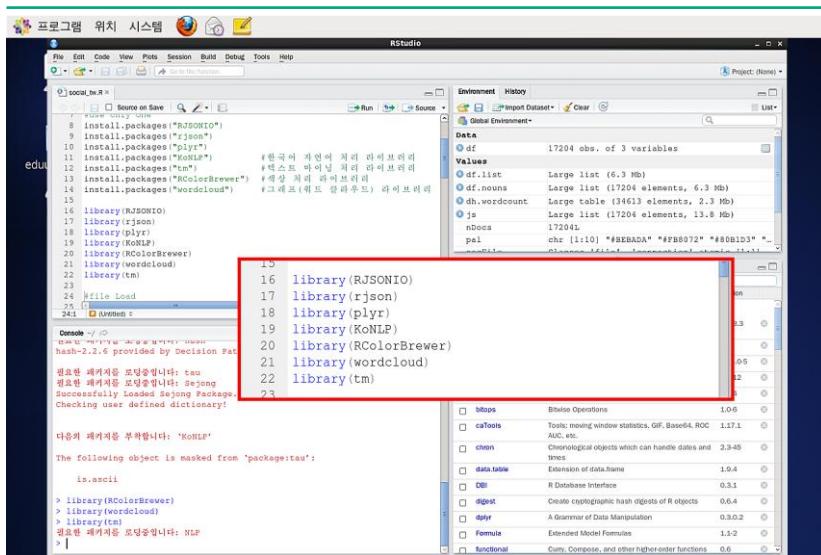
- #주) R 프로그램 분석을 위한 사전 라이브러리 설치는 install.package("라이브 러리 이름")으로 설치하거나 오른쪽 하단의 패널을 이용하여 설치한다. 설치할 패키지 리스트는 아래와 같다. 작성된 줄의 끝에서 “**Ctrl+ Enter**”를 입력하여 실행한다.

```

01. #라이브러리 리스트
02. install.packages("RJSONIO")      #JSON처리를 위한 라이브러리
03. install.packages("rjson")        #JSON처리를 위한 라이브러리
04. install.packages("plyr")
05. install.packages("KoNLP")        #한국어 자연어 처리 라이브러리
06. install.packages("tm")          #텍스트 마이닝 처리 라이브러리
07. install.packages("RColorBrewer")  #색상 처리 라이브러리
08. install.packages("wordcloud")    #그래프(워드 클라우드) 라이브러리

```

III. 가공

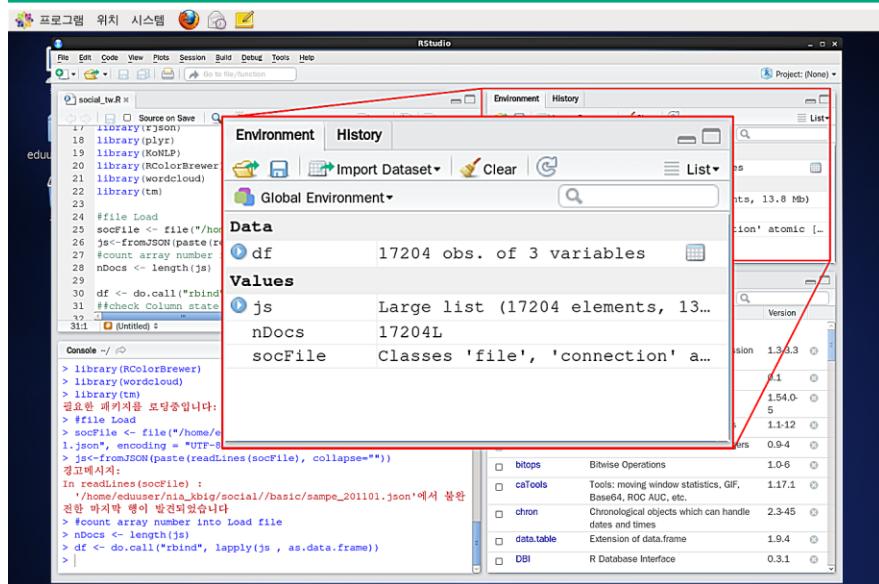


5. 설치된 라이브러리를 프로그램에 이용하기 위해 “library(‘이름’)”을 이용하여 불러온다.

- #주) 설치된 라이브러리들을 불러오는 리스트는 아래와 같다.

01. library(RJSONIO)
02. library(rjson)
03. library(plyr)
04. library(KoNLP)
05. library(RColorBrewer)
06. library(wordcloud)
07. library(tm)

▶ 데이터 가공 R 스크립트(social_tw.R)



- 분석에 필요한 파일과 Encoding 구조에 맞게 지정된 위치에서 파일을 불러온다.
- #주) 실행 후, 각각의 변수에 입력된 값은 오른쪽 상단의 패널을 통해 확인할 수 있다.

```

01. #수집된 파일 불러온다. 저장된 파일의 위치와 Encoding 정보를 확인하여 입력한다.
02. socFile<-file("/home/eduuser/nia_kbig/data/sample_201101.json", encoding =
  ↪ "utf-8")
03. js<-fromJSON(paste(readLines(socFile), collapse=""))
04.
05. #소설 데이터의 크기(라인수)를 계산하여 불러온 데이터를 프레임 구조로 변환한다.
06. nDocs <- length(js)
07. df <- do.call("rbind", lapply(js , as.data.frame))

```



- 데이터 가공/분석 R 스크립트 소스(social_tw.R)
- 라인 02 : 분석할 데이터 파일을 불러오기 위해 파일의 위치와 인코딩 타입을 함께 사용한다.
- 라인 03 : 정의된 파일에서 라인단위로 읽어 JSON 구조에 맞게 데이터를 기록한다.
- 라인 06~07 : 분석할 데이터 파일의 크기와 데이터 프레임으로 변환한다.

III. 가공

```
08. #데이터 프레임으로 변환한 구조를 확인한다.  
09. colnames(df)  
10. #수집 데이터에서 @RW정보와 URL정보를 제거한다.  
11. removeTwit <- function(x) { gsub("@[:graph:]*", "", x) }  
12. removeURL <- function(x) { gsub("http://[:graph:]*", "", x) }  
13. df$message <- sapply(df$message , removeTwit)  
14. df$message <- sapply(df$message, removeURL)
```



- 28페이지 데이터 가공/분석 R 스크립트 소스(social_tw.R)
- 라인 09 : 데이터 프레임의 컬럼 이름을 확인한다.
- 라인 11~12 : 트위터 데이터에서 RT 정보와 URL 정보를 제거하기 위해, 데이터 파싱 정보를 함수로 정의한다.
- 라인 13~14 : RT 정보와 URL 정보를 메시지에서 찾아서 삭제하기 위해, 앞에서 지정한 삭제 함수를 이용하여 불필요한 정보를 삭제한다.

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화



IV 저 장

개요

33

R Studio 활용 저장

34

> 개요

비정형 데이터는 수집 데이터와 분석 데이터를 구분하여 JSON 구조를 갖는 데이터 모델링을 통해 분리하여 저장한다. 이를 위해, NoSQL과 같은 비정형 데이터베이스를 이용하여 기록하고 관리한다. NoSQL은 잘 설계된 데이터 모델링을 통해 대용량 데이터를 안정적으로 관리할 수 있다. 단, 데이터 모델이 완성되지 않은 상태에서는 텍스트 파일이나 CSV와 같은 구조로 저장하고 관리한다.



> 저장 방법

- **가공된 데이터 임시 저장** : 소셜 데이터인 트위터에서 분석을 위해 추출한 데이터를 임시 파일로 저장한다.
- **저장 파일의 구조 정의** : 문장에서 빈 공간을 가진 데이터를 제거한 데이터만을 저장한다.
- **소스 저장** : 작성 중인 소셜 트위터 분석 프로그램을 저장한다.

```

> #temp data save
> write.table(df, file="/home/eduuser/nia_kbig/social/basic/re_message.csv",
  ↪ append=FALSE, quote=FALSE, sep="", row.names=FALSE)

```

1. 안정적인 분석을 위해, 가공된 데이터를 파일로 저장한다. 저장된 파일은 다른 분석 프로그램으로도 사용할 수 있도록 “CSV” 파일로 저장한다.

- #주) 저장된 파일은 실행시켜 놓은 터미널을 통해 확인할 수 있다. ①은 저장된 파일과 위치에 해당하는 파일이고, ②는 R에서 실행된 CSV로, 임시 저장되는 명령어의 실행 정보이다.

```

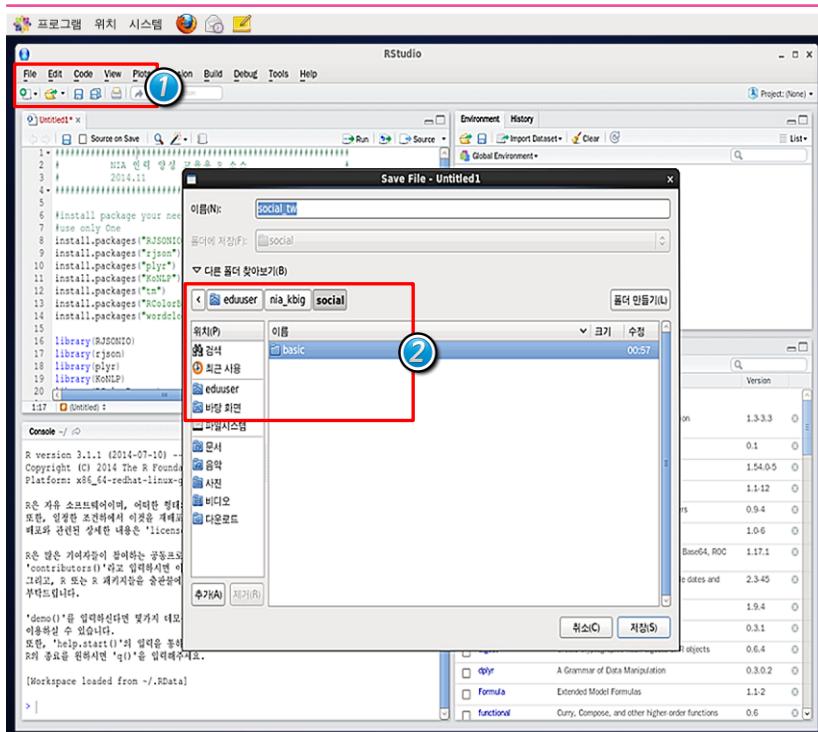
01. #temp data save
02. write.table(df, file="/home/eduuser/nia_kbig/social/basic/re_message.csv",
  ↪ append=FALSE, quote=FALSE, sep="", row.names=FALSE)

```



- 데이터 가공/분석 R 스크립트 소스(social_tw.R)
- 라인 02 : 가공된 데이터를 임시로 저장하기 위한 위치, 파일명을 지정하여 테이블 구조로 저장한다.

IV. 저장



2. 소셜 트위터 분석을 위해 작성 중인 프로그램 소스를 저장한다.

- #주) 작성 중인 프로그램 소스를 저장하는 방법은 메뉴의 “File” > “Save”를 이용하거나 도구상자의 저장 아이콘을 이용한다. 저장 시 저장 위치는 eduuser라는 폴더를 선택하여 하위 폴더를 따라 저장할 위치를 이동하여 최종적으로 “basic”으로 선택한다. 파일명은 본인이 작성한다.

I.개요

II.수집

III.기공

IV.저장

V.분석

VI.시각화

W



V 분석

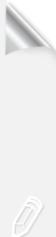
개요	39
R Studio 활용 분석	40
R Studio 저장	43

V

분석

> 개요

소셜 데이터 분석은 한국어 또는 영어와 같이 언어에 따른 데이터 사전을 이용하여 명사형 또는 연관 키워드를 추출하는 형태소 분석을 실시한다. 그리고, 추출된 형태소 분석 데이터를 이용하여 키워드 빈도 분석, 연관 분석, 토픽 분석 등과 같은 텍스트 마이닝 및 분류 분석을 처리할 수 있다.



> 데이터 분석 방법

- **가중치 빈도 계산** : 키워드 빈도에 가중치를 부여하기 위한 핵심 키워드를 사전에 추가하고, tm 라이브러리 안에 있는 가중치 빈도 계산을 이용한다.
- **명사형 단어 추출** : 단어를 추출하여 사전에 동사, 형용사 등의 단어를 제거하고 명사 구조만을 추출한다. 단, 한국어 사전이 불완전하기 때문에 명사형 형태로 단어를 추출한다.
- **키워드 빈도수 계산** : 명사형 단어를 인덱싱하고, 해당 단어가 등장하는 키워드의 빈도를 계산한다. 이때 tm 라이브러리와 KoNLP 라이브러리를 이용하여 계산한다.



용어정리

- **형태소 분석** : 한국어 형태소 분석을 위해서는 공개용 한국어 사전을 이용하지만, 일반적으로 정확한 분석을 위해서는 상업용 분석 사전을 이용한다. 개방형 형태소 분석은 설치된 라이브러리 중 한국어 자연어 처리 라이브러리인 “KoNLP”에서 제공한다.

> R Studio 활용 분석(social_tw.R)

> 데이터 불러오기

소셜 트위터 샘플 데이터 (re_message.csv)

```
프로그램 위치 시스템 파일(F) 최근IE 보기(V) 검색(S) 히어로ID 도움말(H) eduser@cm01:~/mia_khig/social/basic

[eduser@cm01 basic]$ cat re_message.csv | more
1
messagesScreenNamecreated at

네가 수령한 메시지를 받아서 mingky_Kimsun Jan 09 07:31:00 KST 2011
영화 보러 가는 GangSun Jan 09 07:31:00 KST 2011
오빠근데 잘 살아요? 나 걸려질 바에는? kkjane0620Sun Jan 09 07:30:00 KST 2011
너무 멍때리 YCloudBearSun Jan 09 07:30:00 KST 2011
다시, Darlin Avril LavigneJnyboSun Jan 09 07:30:00 KST 2011
누나 거울만 끌라 갑니다 라 디번님께서 하구 있어 보셨어요?? 깨광운데 ㅠㅠㅠㅠㅠ shin5539Sun Jan 09 07:30:00 KST 2011
개드림이거나? 악당YCloudBearSun Jan 09 07:30:00 KST 2011
전진님은 잘해요? kyongyoun84Sun Jan 09 07:30:00 KST 2011
산수께 오는 사랑 Sinsop001Sun Jan 09 07:30:00 KST 2011
여기에서 siesas5645Sun Jan 09 07:30:00 KST 2011
여기에서 더러워 험금 꾸꾸架子 armeniusSun Jan 09 07:30:00 KST 2011
단, 위안부 당시 죄 칸찍는구요 SleepypracoSun Jan 09 07:30:00 KST 2011
#JND2 각자각색 불편하죠... 이 우습잖아... 쿄 쿄 쿄 그리고... 문화생활... 워이런거 할수가 없으니까요.. dkfma486Sun Jan 09 07:30:00 KST 2011
오늘은 예상보다 귀가가 좀 늦어졌다. 이젠 나도 좀 일찍 자고 일찍 일어나는 습관을 길러야지. 너무 나태해졌어. MeinkampfSun Jan 09 07:30:00 KST 2011
그래 넌 흠한- 흠한- 등등이다. 이 형이 진정해. 에스티일부터 웜파파... PimpGangSun Jan 09 07:30:00 KST 2011
위안부로 본- 환경지구를 둘째로 비벼지 겁나요? SonglidaSun Jan 09 07:30:00 KST 2011
이도래 좋아하는 분께 많네요. 새 월이 터울로 면할수 있는 감성이 있나봐요. * ejynnnsun Jan 09 07:30:00 KST 2011
가하나요? 나 같아? wylovelibertySun Jan 09 07:30:00 KST 2011
아직 맹계 게임을 찾았다! DathungSun Jan 09 07:31:00 KST 2011
ATM 사용자는 자기가 맘을 하는 순간 그간 고개 차기 삶의 전제에 대해 한 번도 평가가 된다는 사실을 증명 잇는다. (내가 아니라) 아감벌의 철학. 전문, 트위터에 유통되는 140자 글을 보면서 감히 그 사람을 미루어 짐작할 수 있을 것 같다. vodkalemon_chesun Jan 09 07:30:00 KST 2011
도덕덕분에 보게 된 듯자 커질 삼총사--- 너무 끊놨어--- 배우들도 너무 잘하고 한참을 길دان했어. 고마웡^^terirusSun Jan 09 07:31:00 KST 2011
광장 시크릿카드를 보는 중HNTYSun Jan 09 07:31:00 KST 2011
#1994크 그대도 당에 먼저 암자씨에게 해주는 게 어디일까? jkl4356Sun Jan 09 07:31:00 KST 2011
Droid Pro 한국 블랙해도 좋아? 나? 사고 심네. huevertsun Jan 09 07:31:00 KST 2011
설국 :) KDaE_MtSun Jan 09 07:31:00 KST 2011
질책에 팔레트 D-day가 줄어 가는데 패션 줄이며 긴장을하는 건 학생이 아닌 나. 한번씩 저 수자가 뛰어나고 물는 고3. 역시 고3 수업은 결혼하지가 못합니까? 나도 속 저렸던가... zero72koreaSun Jan 09 07:30:00 KST 2011
더불어 살아갈 수 있는 좋은 이기야 부탁드린답니다. hn4lgnnSun Jan 09 07:31:00 KST 2011
제 꽃을 봐요? 같은 날은 그저 하루 하루 이상한 소나내내는 기계일 뿐이지. //AxLaterSun Jan 09 07:31:00 KST 2011
그런 거 없던데! lameeee711Sun Jan 09 07:31:00 KST 2011
일하는 중. Take Five! yzmkm75Sun Jan 09 07:31:00 KST 2011
하나의 보물보강 RT 이게 면역력에 회복력이 돋보였어요. RT. 하. 이 (cont) mimimido99Sun Jan 09 07:31:00 KST 2011
```

- #주) 데이터 확인은 대용량 데이터로, Open Office를 이용하여 확인할 수 없기 때문에 Linux 명령어로 확인한다.
 - \$ cat re message.csv | more

V. 분석

▶ 데이터 분석

- #주) 앞의 작성 중인 R 프로그램 소스에 이어서 작업한다. 작업 내용은 아래와 같다.

```
30 #cleaning data cleansing
31 removeTwit <- function(x) { gsub("@[:graph:]*", "", x) }
32 #url data cleansing
33 removeURL <- function(x) { gsub("http://[:graph:]*", "", x) }
34
35 df$message <- sapply(df$message, removeTwit)
36 df$message <- sapply(df$message, removeURL)
37
38 #temp data save
39 write.table(df, file="/home/eduuser/nia_kbig/social/basic/re_message.csv", append=FALSE, quote=FALSE)
40
41
42
43
44
45
46
47
48 #Load seajong dictionary in KoNLP
49 useSejongDic()
50 #adding your word in dictionary file and Stemming but English
51 mergeUserDic(data.frame(c("사람","강남","트위터","콘텐츠"), c("ncn")))
52
53 df$message <- sapply(df$message, function(x) { paste(extractNoun(x), collapse = " ") })
54
55 #word extraction
56 df.nouns <- sapply(df$message, extractNoun, USE.NAMES=F)
57 dh.wordcount <- table(unlist(df.nouns))
58
59 df.list<-list(df.nouns)
60
61 #save result
62 write.csv(dh.wordcount, file="/home/eduuser/nia_kbig/social/basic/social_result.csv")
63
64
```

1. 사전 데이터를 불러온다.

- #주) 실행된 결과는 아래와 같으며, 한국어 명사는 87,007개의 단어로 구성되어 있다.

01. #Load seajong dictionary

02. useSejongDic()

```
> #Load seajong dictionary in KoNLP
> useSejongDic()
Backup was just finished!
87007 words were added to dic_user.txt.
> |
```

2. 읽어들인 사전에 사용자 또는 개발자가 필요로 하는 단어를 추가한다.

- #주) 추가된 사용자 키워드는 명사 또는 명사형으로 필요한 경우에 추가적으로 넣을 수 있다. 이곳에 자신만의 추가적인 단어를 구성한다. 실행된 결과는 아래와 같다.

```
01. # adding your word #Stemming for English
02. mergeUserDic(data.frame(c("사람","강남","트위터", "콘텐츠"), c("ncn")))
> #adding your word in dictionary file and Stemming but English
> mergeUserDic(data.frame(c("사람","강남","트위터", "콘텐츠"), c("ncn"))
))
4 words were added to dic_user.txt.
> |
```

3. 명사형 단어를 분리하여 단어 사전에 맞는 단어로 변환하고, 단어의 빈도수와 가중치에 따라 분류하여 계산한다.

- #주) 추출한 단어에서 빈 공간(" ")을 기준으로, 단어를 분리한다. 그리고 분리된 키워드에서 명사형 키워드의 빈도를 계산한다. 실행된 결과는 아래와 같다.

```
01. df$message <- sapply(df$message, function(x) { paste(extractNoun(x), collapse = " ") })
02.
03. #word extraction
04. df.nouns <- sapply(df$message, extractNoun, USE.NAMES=F)
05. dh.wordcount <- table(unlist(df.nouns))
06. df.list<-list(df.nouns)

> df$message <- sapply(df$message, function(x) { paste(extractNoun(x), collapse = " ") })
50건 이상의 경고들이 있습니다 (처음 50건의 경고들을 확인하기 위해서는 warnin gs()를 이용하세요)
> #word extraction
> df.nouns <- sapply(df$message, extractNoun, USE.NAMES=F)
50건 이상의 경고들이 있습니다 (처음 50건의 경고들을 확인하기 위해서는 warnin gs()를 이용하세요)
> dh.wordcount <- table(unlist(df.nouns))
> df.list<-list(df.nouns)
> #save result
> write.csv(dh.wordcount, file="/home/eduuser/nia_kbig/social/basic /social_result.csv")
> |
```

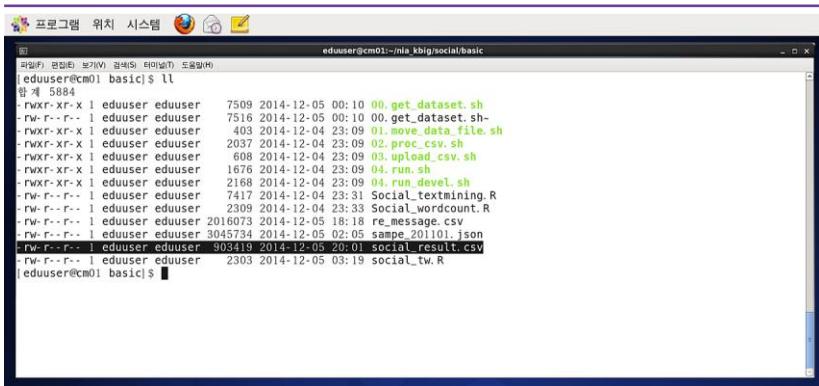


부연설명

- 데이터 가공/분석 R 스크립트 소스(social_tw.R)
- 라인 01 : 트위터의 메시지에서 명사형 단어를 추출하기 위해 단어를 추출한다.
- 라인 04 : 추출된 단어에서 명사형 키워드를 추출한다.
- 라인 05~06 : 명사형 키워드를 테이블 구조로 변환한 후, 키워드 수를 계산하기 위해 임시 변수에 기록한다. 그리고 명사형 단어를 리스트 구조로 변환하여 관리하는 명령을 실행한다.

> R Studio 저장

> 분석 결과 저장

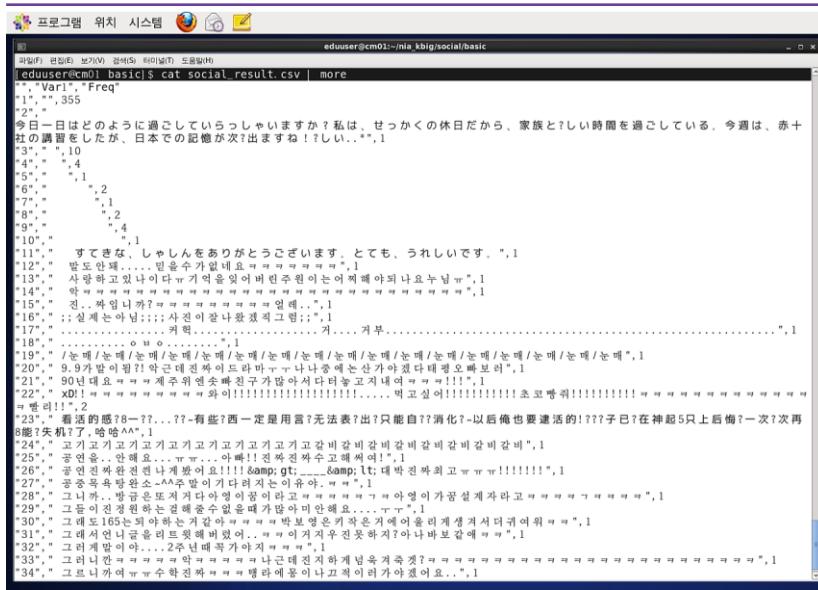


```
eduuser@cm01:~/nia_kbig/social/basic$ ll
total 5884
-rwxr-xr-x 1 eduuser eduuser 7509 2014-12-05 00:10 00.get_dataset.sh
-rw-r--r-- 1 eduuser eduuser 7516 2014-12-05 00:10 00.get_dataset.sh~
-rwxr-xr-x 1 eduuser eduuser 403 2014-12-04 23:09 01.move_data_file.sh
-rwxr-xr-x 1 eduuser eduuser 2037 2014-12-04 23:09 02.proc_csv.sh
-rwxr-xr-x 1 eduuser eduuser 608 2014-12-04 23:09 03.upload_csv.sh
-rwxr-xr-x 1 eduuser eduuser 1676 2014-12-04 23:09 04.run.sh
-rwxr-xr-x 1 eduuser eduuser 2168 2014-12-04 23:09 04.run-devel.sh
-rw-r--r-- 1 eduuser eduuser 7377 2014-12-04 23:31 Social_textmining.R
-rw-r--r-- 1 eduuser eduuser 2399 2014-12-04 23:33 Social_wordcount.R
-rw-r--r-- 1 eduuser eduuser 2016073 2014-12-05 18:45 re_message.csv
-rw-r--r-- 1 eduuser eduuser 3045734 2014-12-05 02:05 result_10.json
-rw-r--r-- 1 eduuser eduuser 302419 2014-12-05 20:01 social_result.csv
-rw-r--r-- 1 eduuser eduuser 2303 2014-12-05 03:18 social_tw.R
eduuser@cm01:~/nia_kbig/social/basic$
```

- 분석된 결과를 저장한다. 분석 결과는 다양한 방법으로 저장할 수 있으며, 앞의 정제 데이터 저장과 동일한 CSV 형태로 저장한다.
- #주) 저장된 파일의 확인은 실행되어 있는 터미널을 이용하여 확인한다.

```
01. #save final data
02. write.csv(dh.wordcount,
    ↪ file="/home/eduuser/nia_kbig/data/social_final_result.csv")
```

> 결과 데이터



I. 개요

II. 수집

III. 가공

IV. 저장

V.문서

VI.人각호



1

2



VI 시각화

개요	47
분석 데이터 시각화	49
데이터 분석	51

VI

시각화

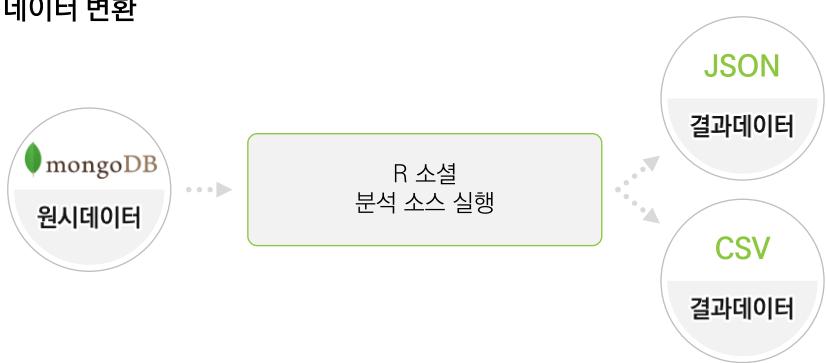
> 개요

분석 결과는 다양한 방법으로 시각화하여 분석할 수 있으며, 이를 통해 데이터의 변화 및 분포를 해석하고 데이터에 대한 분석 효과를 적용할 수 있다. 특히, 데이터의 출현 빈도는 이슈 키워드 중심으로 변화하기 때문에 이에 대한 기간 또는 분기별 키워드의 변화 정보를 판단할 수 있는 정보이다. 또한 최종적으로 획득된 분석 결과는 시각화 이후에 저장하여 전체 데이터에 대한 분석에 사용할 수 있다.

> 시각화 방법 및 활용 기술

- **분석 결과에 대한 그래프 설정** : 분석된 결과에 대하여 키워드 출현 빈도에 따라 구름 모양으로 분포시키는 WordCloud 그래프 출력으로 한다.
- **저장된 포맷에 맞는 그래프 도구 설정** : 저장된 데이터를 이용하여 시각화할 수 있는 도구도 다양하기 때문에 이를 위해 시각화 도구를 선택한다.
- **의미해석** : 시각화된 WordCloud 그래프에서 크기와 색상으로 표현된 키워드는 자주 또는 가장 많이 언급된 키워드이며, 검색엔진을 통해 사회적 이슈와 등장 기간을 확인한다.

▶ 데이터 변환



> 분석 데이터 시각화(social_tw.R)

> 데이터 시각화

1. 시각화를 위한 데이터 범위와 그래프를 설정한다.

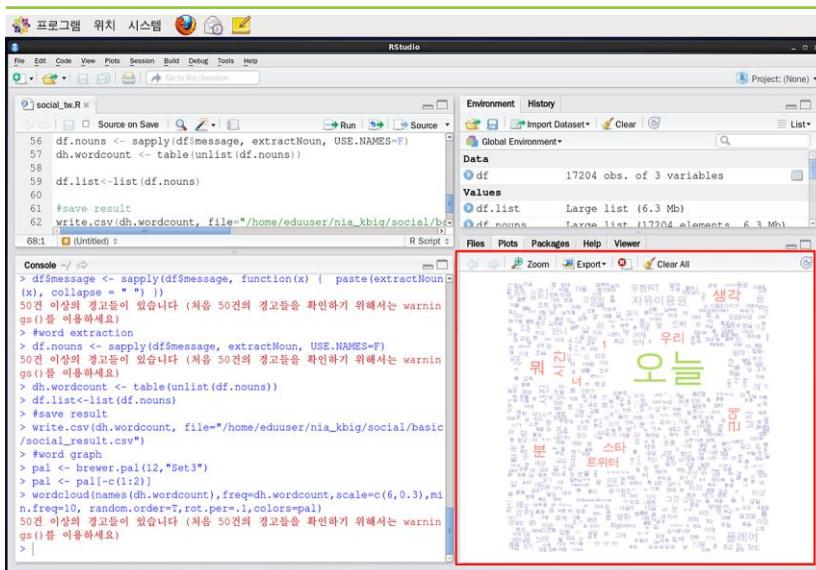
- #주) 그래프는 Word Cloud에 출력하기 위한 필요 데이터와 모델을 지정한다. 작성한 코드의 실행 결과는 아래와 같다. 데이터 크기가 크기 때문에 범위를 다 수용하지 못하는 경고 메시지가 동작된다.

```
01. #word graph  
02. pal <- brewer.pal(12, "Set3")  
03. pal <- pal[-c(1:2)]  
04. wordcloud(names(dh.wordcount), freq=dh.wordcount, scale=c(6,0.3), min.freq  
    ↲ =10, random.order=T, rot.per=.1, colors=pal)
```

```
> #word graph  
> pal <- brewer.pal(12, "Set3")  
> pal <- pal[-c(1:2)]  
> wordcloud(names(dh.wordcount), freq=dh.wordcount, scale=c(6,0.3), mi  
n.freq=10, random.order=T, rot.per=.1, colors=pal)  
50건 이상의 경고들이 있습니다 (처음 50건의 경고들을 확인하기 위해서는 warnin  
gs()를 이용하세요)  
> |
```

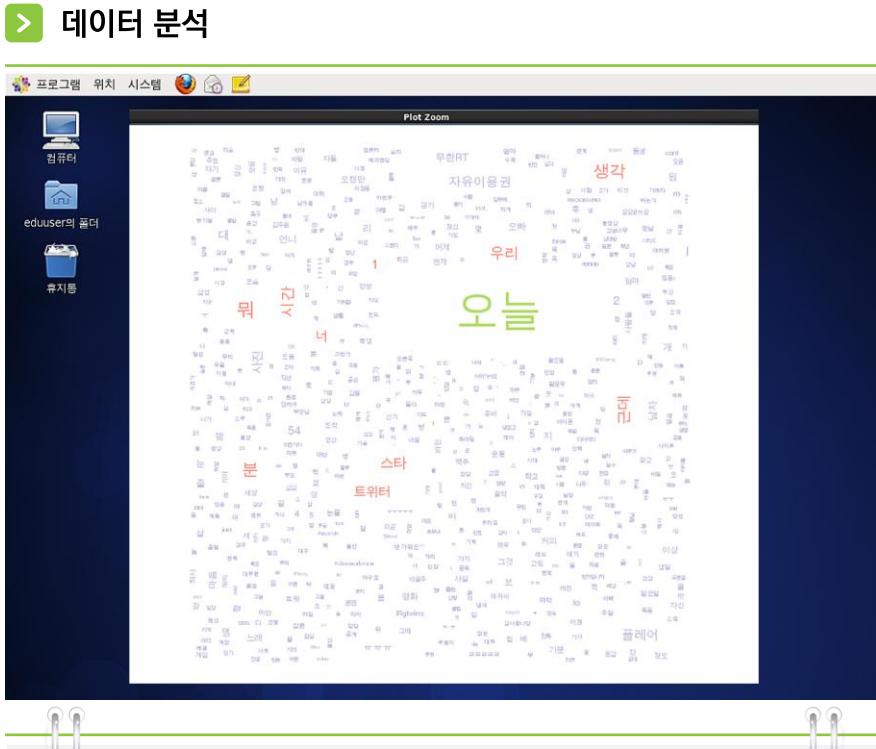


- 데이터 가공/분석 R 스크립트 소스(social_tw.R)
- 라인 02~03 : 그래프 출력을 위한 출력 타입을 정의한다.
- 라인 04 : 워드클라우드 그래프 출력을 위한 데이터, 크기, 색상, 표현 방식을 지정하고 출력시킨다.



2. 시각화 결과에 대한 그래프는 RStudio의 오른쪽 하단 패널에 “Plots”으로 확인할 수 있다.

- #주) 그래프를 Zoom으로 확인하면 위와 같다.



- 분석 결과는 해당 기간의 트위터 데이터에서 의미 있는 키워드로 “강남역”, “친구”, “사람”, “집” 등과 같은 키워드가 등장하였다.
- 특정 키워드 중에서 “RT”, “ㅋㅋㅋ”, “ㅠㅠ” 와 같은 키워드의 등장은 한국어 형태소 사전이 불완전하기 때문이다.
- 이러한 특수 키워드의 등장은 소셜에서 함축적인 키워드 사용으로 한글의 맞춤법이 어긋난 단어가 많다는 것을 알 수 있다.
- wordCloud의 형태에서 불필요한 키워드를 제거한 상태의 의미를 살펴보면, 사람, 오늘, 일, 스타, 강남역, 친구 등으로 지역적 의미를 가진 키워드로 “강남역”, 사회적 의미를 가진 키워드로 “친구”, “스타”, 시간적 의미를 가진 키워드로 “오늘” 등이다.
- 한국어 형태소 분석을 위한 단어 사전이 취약한 상태로, 영어와 혼합된 키워드, 특수 문자 등이 제거가 안된 상태이다. 따라서 정확한 분석을 위해서는 상업용 한글 형태소 분석기를 사용하는 것이 좋다.

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화



VII 예제문제

예제 문제1

55

예제 문제2

56

예 / 제 / 문 / 제

예제 1

트위터 데이터에서 일별 키워드 등장 건수를 계산하라.

- 트위터 키워드에서 일별로 등장하는 키워드 건수를 분석하고, 2011년 1월 9일의 키워드를 워드 클라우드 그래프로 출력하라.

- 트위터 데이터에서 날짜를 추출하기 위한 날짜 라이브러리(lubridate)를 추가한다.
- JSON의 날짜 필드를 이용하여 2011년 1월 9일 날짜 데이터를 추출한다.
- 추출한 데이터에서 형태소 분석을 한다.
- 키워드별 건수를 계산하고, 저장한다.
- 2011년 1월 9일 키워드에 대한 등장 변화를 워드 클라우드 그래프를 사용하여 시각화한다.

예제 2

트위터 사용자별 이용 건수를 계산하라.

- 일별 트위터를 이용하는 사용자 수를 계산하고 비교하라.

- 트위터 원시 데이터에서 날짜 데이터만을 추출하여 저장한다.
- 추출된 날짜 데이터를 일자별로 그룹핑하여 건수를 계산한다.
- 일자별 구한 사용자의 건수에 대하여 막대그래프를 이용하여 시각화하여 비교 분석한다.



소셜 

Intermediate Level

중급과정







I 개요

개요

61

I

개요



개요

소셜 미디어 데이터는 솔트룩스에서 제공해 준 트위터와 뉴스 데이터를 바탕으로, 두 종류 이상의 비정형 구조를 가진 데이터를 통합하고 의미 정보 분석을 위해 형태소 분석과 단어 출현 빈도를 계산한다. 그리고 이 정보에서 상위 키워드 만을 다시 추출하여 특정 기간 또는 소셜 미디어 데이터 전체에 대한 이슈 키워드를 알아보고자 한다. 이러한 방법으로 이슈 키워드 분석을 통하여 사회적 이슈와 관심 정보를 확인할 수 있다.

> 활용 데이터

- **sample_news_201301.json** : 뉴스 데이터(2013년 1월 데이터 일부)
- **sample_news_201302.json** : 뉴스 데이터(2013년 2월 데이터 일부)
- **sample_news_201303.json** : 뉴스 데이터(2013년 3월 데이터 일부)
- **sample_tw_201301_1.json** : 트위터 데이터(2013년 1월 데이터 일부)
- **sample_tw_201302_1.json** : 트위터 데이터(2013년 2월 데이터 일부)
- **sample_tw_201303_1.json** : 트위터 데이터(2013년 3월 데이터 일부)

> 선행학습

- **리눅스** – 파일시스템 구조, 쉘 명령어, 쉘 스크립트 실행 방법
- **R 프로그래밍 언어** – 파일 불러오기, 라이브러리 등록, 데이터 함수(프레임, 테이블), 그래프 함수, 제어문(함수 호출, 외부 함수, 함수 정의) 사용방법
- **데이터 구조** – CSV 데이터 구조, JSON 데이터 구조, 파일 저장 구조 이해
- **R 차트** – 내부 차트(막대, 바, 원 등), 외부 차트(막대, 클라우드, 3D, D3 차트) 사용방법

▶ 요구사항

- 뉴스와 트위터의 비정형 데이터 분석에서 불필요한 정보를 리스트로 정리하고, 이를 제거한다.
- 중요도가 높은 키워드 또는 특수 문자 키워드를 단어 사전에 추가하고, 단어 사전을 이용하여 형태소 분석을 실시한다.
- 추출된 명사형 단어 데이터의 빈도를 계산한다.
- 빈도 데이터에서 상위 정보만을 추출하고 막대 그래프로 시각화 한다.

▶ 분석 절차

- 수집된 소셜 데이터 셋을 분석 저장소로 복사한다.
- 비정형 데이터 분석을 위해 데이터 구조인 JSON 포맷을 이해하고, 이 중에서 문장 분석을 위하여 메시지 정보의 키에 대응하는 값 정보만을 추출한다.
- 트위터 정보는 @RT 정보, URL 정보, 숫자 등과 같은 불필요한 정보를 포함하고 있기 때문에 추출된 트위터 메시지 값 정보에서 이들 정보를 제거하는 정제 처리를 실시한다.
- 뉴스 정보는 숫자와 문자를 분리하여 비정형 데이터를 재구성한다.
- 분석 중에 관심있는 키워드나 중요도가 높은 키워드를 단어 사전에 추가하여 핵심 키워드에 대한 단어 사전의 가중치 정보를 보완한다.
- 정제된 트위터의 메시지와 뉴스 컨텐트 값에서 명사형 키워드만을 추출하고, 키워드 출현 빈도를 계산한다.
- 계산된 키워드 출현 빈도에서 중요 키워드 추출을 위해 출현빈도 재계산을 수행하여 중요 키워드만을 추출한다.
- 추출된 중요 키워드에 대한 정보를 막대 그래프로 출력하고, 결과 데이터를 저장한다.
- 출력된 그래프에서 소셜 네트워크 상에서 참여자들이 거론하였던 이슈 키워드의 가중치와 빈도를 결합한 사회적 관심 및 키워드 출현 빈도를 파악하여 특정 키워드에 대한 변화를 판단할 수 있다.



II 수집

개요	65
수집 데이터	66
데이터 수집	70
데이터 작업 영역 이동 스크립트	73



수집

▶ 개요

소셜 데이터는 소셜 네트워크 서비스, 소셜 미디어 서비스, 마이크로 소셜 서비스, 뉴스 미디어 서비스 등을 통하여 거론된 사회적 정보들로 구성되며, 실시간으로 사회적 이슈를 분석하기 위해서는 실시간으로 데이터를 수집하고 관리되어야 한다. 이를 위해 소셜 데이터의 수집은 소셜 미디어 서비스(트위터, 페이스북, 블로그, 뉴스 등)에서 제공하는 API 또는 수집기를 통해 실시간으로 수집할 수 있으며, 사회적 이슈나 비즈니스 분석 등과 같은 다양한 사회적 분석 모델에 적용할 수 있다.

▶ 수집 방법

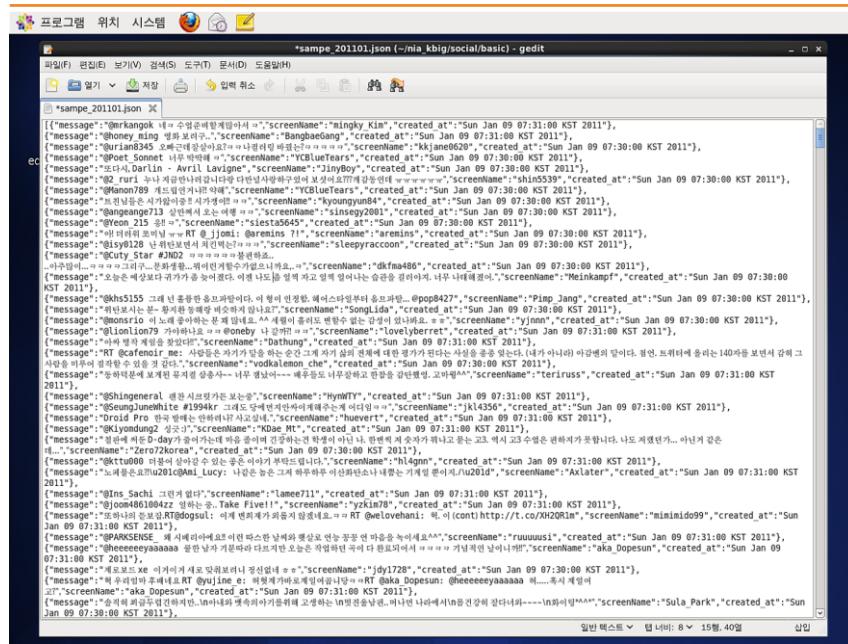
- **API 데이터 수집** : 소셜 데이터의 수집은 API를 이용하여 일정량의 데이터를 실시간으로 수집하고 있으며, 각 소셜 서비스 회사의 정책에 따라 수집량을 제한하고 있다.
- **데이터 제공** : 소셜 데이터는 OpenAPI, 자료수집기 (Crawler), 데이터 구매 등으로 데이터를 수집할 수 있으며, 실습용 자료는 빅데이터 분석 활용센터에 접속하여 소셜 초급 데이터 셋을 다운로드할 수 있도록 원시데이터를 제공하고 있다.



- **비정형 데이터(Unstructured Data)** : 일정한 규격이나 형태를 지닌 숫자 데이터와 달리 그림, 영상, 문서 등과 같이 서로 다른 형태의 구조화 되지 않은 데이터

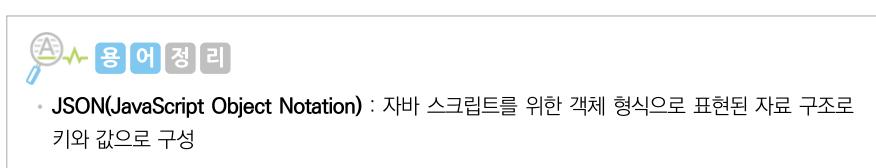
> 수집 데이터

▶ 트위터 데이터(sample_tw_201301.json)



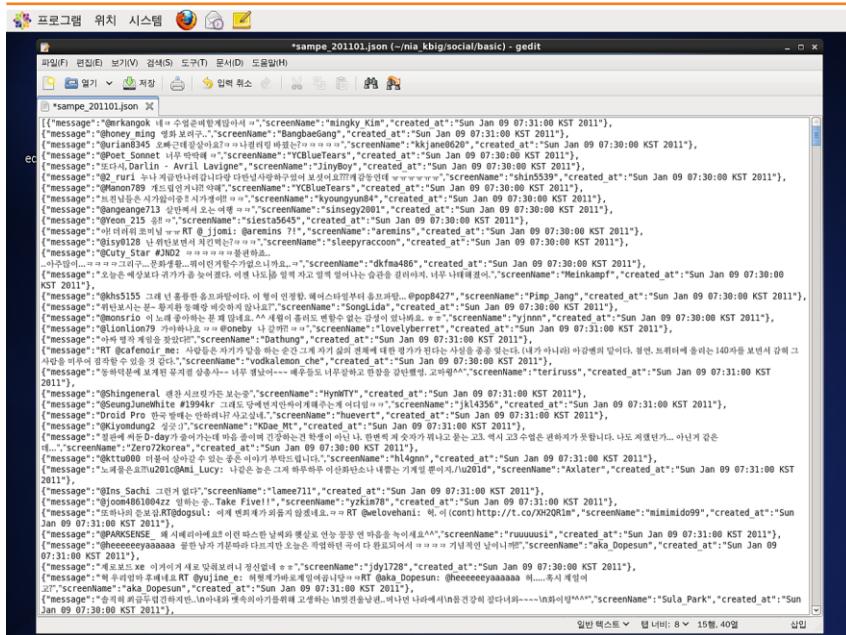
▶ 트위터의 데이터 구조

- **JSON 구조** : 트위터에서 제공하는 데이터는 비정형 데이터로 빠르게 표현 할 수 있는 JSON 구조를 가진다. 다양한 종류의 키와 값으로 구성되지만 분석에 필요한 데이터만을 수집하기 때문에 해싱 키로 재구성하였다.
 - **message** : 소셜미디어 참여자들이 자신의 계정으로 업로드한 데이터 정보
 - **screenName** : 소셜미디어 참여자의 닉네임 정보
 - **create_at** : 트위터에 업로드한 시각 정보



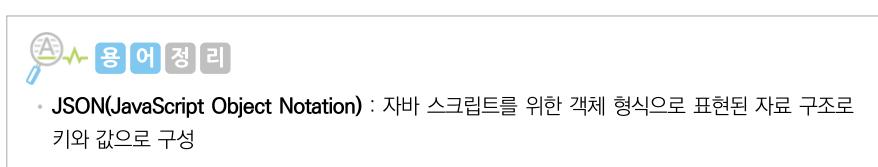
II. 수집

▶ 뉴스 데이터(sample_news_201301.json)



▶ 뉴스 데이터 구조

- **JSON 구조** : 뉴스에서 제공하는 데이터는 비정형 데이터로 빠르게 표현할 수 있는 JSON 구조를 가진다. 다양한 종류의 키와 값으로 구성되지만 분석에 필요한 데이터만을 수집하기 때문에 핵심 키로 재구성하였다.
 - **content** : 뉴스 데이터의 문장이며, 특수 문자 처리를 위해서는 “\특수기호”로 처리
 - **title** : 뉴스 기사의 제목(title) 정보
 - **date** : 기사를 업로드한 시간 정보



▶ 트위터 데이터 JSON 구조의 예

```
01. [
02. {
03.   "message": "@mrkangok 네ㅋ 수업준비할게많아서ㅋ",
04.   "screenName": "mingky_Kim",
05.   "created_at": "Sun Jan 09 07:31:00 KST 2011"
06. },
07. {
08.   "message": "@honey_ming 영화 보려구..",
09.   "screenName": "BangbaeGang",
10.   "created_at": "Sun Jan 09 07:31:00 KST 2011"
11. },
12. ~ 중간 생략 ~
13. {
14.   "message": "@urian8345 오빠근데잘살아요?ㅋㅋ나컬러링 바꿨는?ㅋㅋㅋㅋ",
15.   "screenName": "kkjane0620",
16.   "created_at": "Sun Jan 09 07:30:00 KST 2011"
17. },
18. {
19.   "message": "@luckyjhs ㅊㅋㅊㅋ",
20.   "screenName": "kyokyoc",
21.   "created_at": "Mon Jan 10 03:15:00 KST 2011"
22. }
23. ]
```

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화

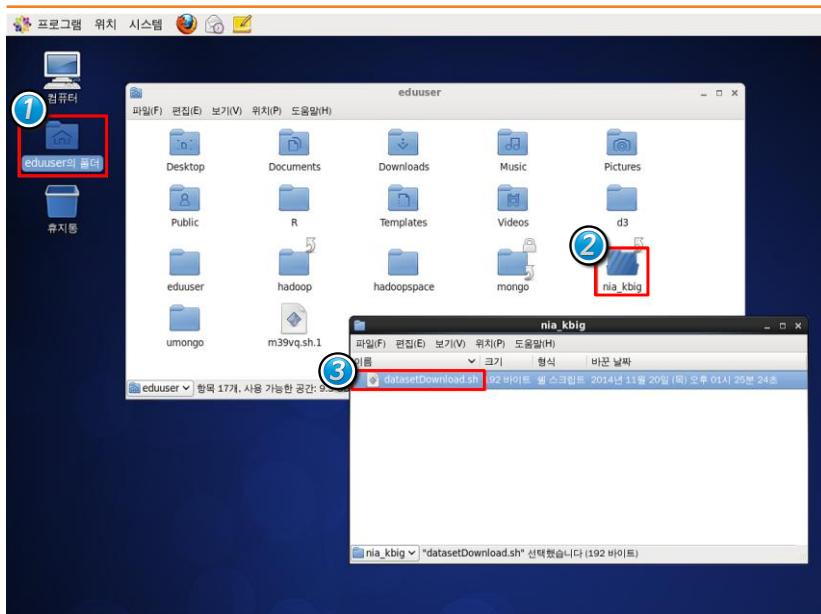
▶ 뉴스 데이터 JSON 구조의 예

```
01. [
02. {
03.   "content": "금융·증권 ▼\n상호금융 집중 관리, 은행 수준으로 규제 강화\n 입력 : 2
04.   "title": "상호금융 집중 관리, 은행 수준으로 규제 강화",
05.   "date": "20130131"
06. },
07. {
08.   "content": "쿠키 연예 배우 송혜교가 31일 오후 서울 한남동 블루스퀘어에서 열린
09.   "title": "쿠키 포토 송혜교, 시각장애인 연기 도전!",
10.   "date": "20130131"
11. },
12. {
13.   "content": "이재용기자\n      jylee@sed.co.kr\n경영정상화를 위한 수주활동에
14.   "title": "한진중공업 사측·노조, '시신투쟁 즉각 중단해야'",
15.   "date": "20130131"
16. },
17. ~ 중간 생략 ~
18. {
19.   "content": "경북 포항대의 홍보 교수들은 2007년 4월 포항·경주의 고교 3학년
20.   "title": "학생 1인당 20만원'…대학·교사간 돈거래",
21.   "date": "20130128"
22. },
23. {
24.   "content": "박시후 종영 소감 \"청앨\", 빨리 끝났으면 했는데…아쉽다\"\n
25.   "title": "박시후 종영 소감 \"청앨\", 빨리 끝났으면 했는데…아쉽다\"",
26.   "date": "20130128"
27. }
28. ]
```

> 데이터 수집

- 데이터 저장소에서 서버 로컬로 소셜 데이터 셋을 복사해 온다.
(트위터 2013년 1월 ~ 3월, 뉴스 2013년 1월 ~ 3월)

> 실습코드 디렉토리로 이동

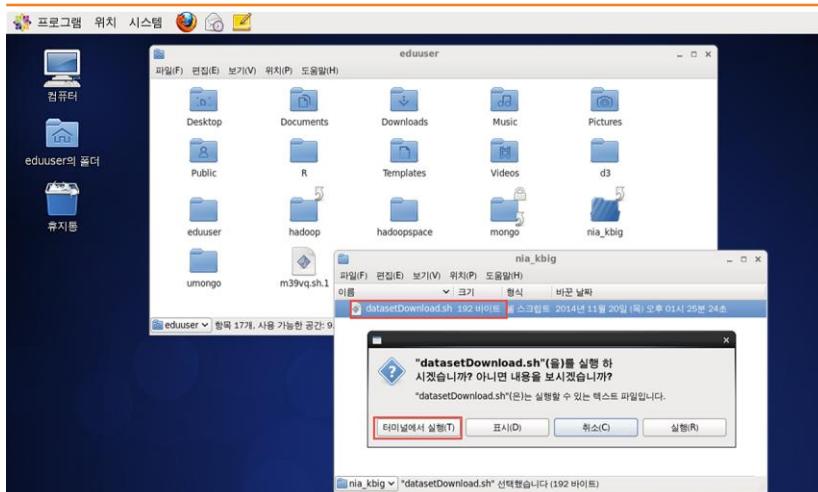


- 로그인 후 바탕화면에서 edouser 폴더를 오픈한다.
- nia_kbig 폴더를 오픈한다.
- datasetDownload.sh를 더블클릭하여 실행한다.

II. 수집

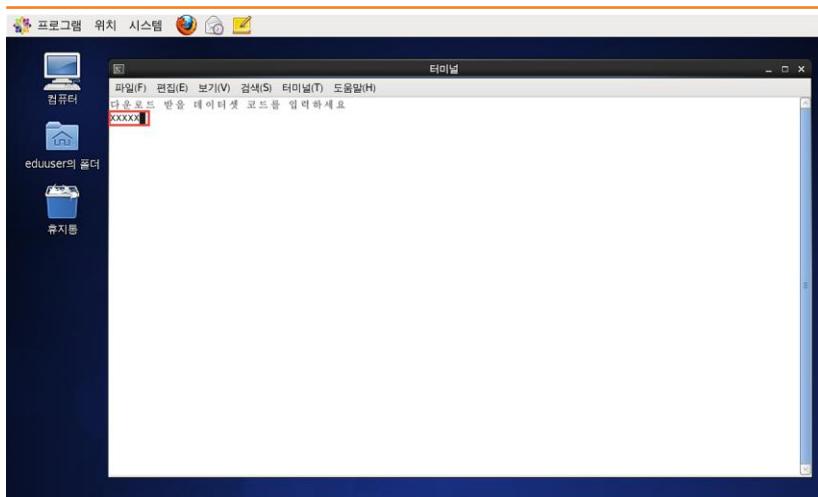
▶ 레파지토리에서 데이터 수집

datasetDownload.sh (원시데이터로 컬서버로 복사)



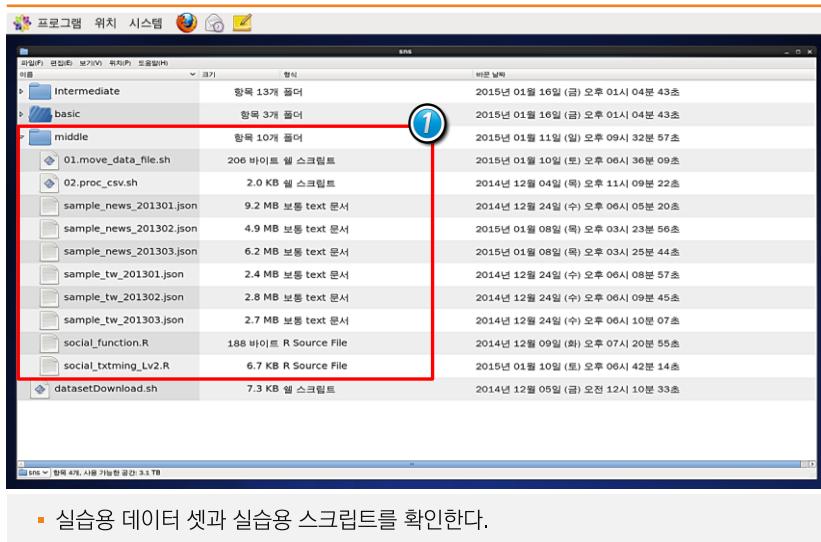
- '터미널에서 실행' 버튼을 클릭한다.

▶ 데이터셋 코드 입력



- 다운로드 받을 데이터셋 코드를 입력 후 엔터

▶ 데이터셋과 실습용 쉘 스크립트



▶ ① 데이터 및 스크립트

■ 01.move_data_file.sh :

작업 영역 Data 폴더로 자료 이동하는 스크립트.

- social_txtrmining_Lv2.R : R 분석 스크립트
- sample_tw_201301.json : 소셜 트위터 분석 샘플 데이터
- sample_tw_201302.json : 소셜 트위터 분석 샘플 데이터
- sample_tw_201303.json : 소셜 트위터 분석 샘플 데이터
- sample_news_201301.json : 뉴스 분석 샘플 데이터
- sample_news_201302.json : 뉴스 분석 샘플 데이터
- sample_news_201303.json : 뉴스 분석 샘플 데이터

II. 수집

▶ 데이터 작업 영역 이동 스크립트(01.move_data_file.sh)

▶ 데이터 작업 공간으로 이동

- 로컬로 수집해 온 데이터를 작업 영역 Data 폴더로 이동시키는 명령 스크립트

01.move_data_file.sh

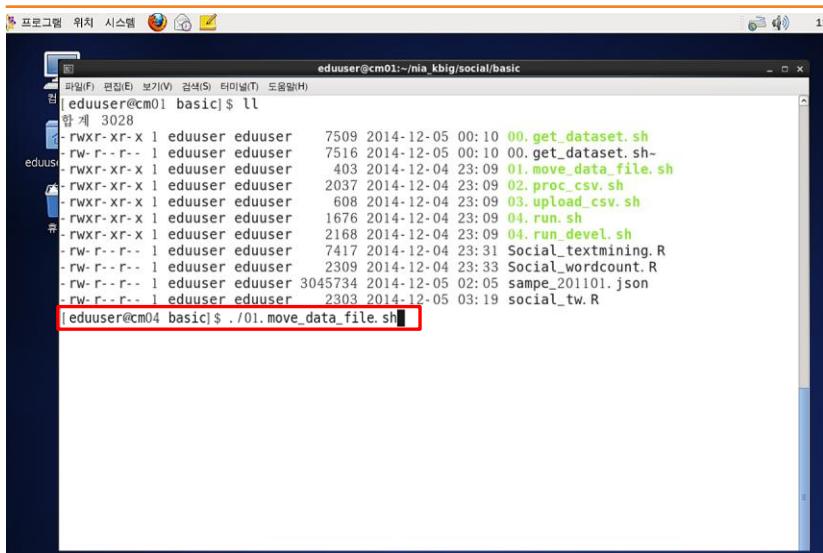
```
01.#!/bin/bash
02. #Social Data file define
03. TARGET_SOCIAL=/home/eduuser/nia_kbig/sns/middle/sample_*.json
04.
05. # 작업 디렉토리 정의
06. LOCAL_DIR=/home/eduuser/nia_kbig/data/
07. mv $TARGET_SOCIAL $LOCAL_DIR
08.
```



- 분석 원시 데이터 이동 스크립트 소스(01.move_data_file.sh)
- 라인 01~03 : 이동시킬 데이터 파일과 파일의 위치를 “TARGET_SOCIAL”로 정의하며, 기호 “#”은 주석을 의미한다.
- 라인 05~06 : 데이터를 이동시킬 위치 정보를 “LOCAL_DIR”이라는 이름으로 기록한다.
- 라인 07 : mv 명령을 이용하여 분석할 데이터를 소셜 폴더에서 분석 폴더로 이동시킨다.

> 수집 데이터셋 작업 영역 폴더 이동

- 로컬로 수집해 온 데이터를 작업 영역 Data 폴더로 자료를 이동하는 스크립트



The screenshot shows a terminal window titled 'eduuser@cm01:~/nia_kbig/social/basic'. The user has run the command 'll' to list files in the current directory, which contains 3028 files. The user then runs the command '/01.move_data_file.sh'.

```
eduuser@cm01 basic]$ ll
total 3028
drwxr-xr-x 1 eduuser eduuser 7509 2014-12-05 00:10 00.get_dataset.sh
-rw-r--r-- 1 eduuser eduuser 7516 2014-12-05 00:10 00.get_dataset.sh-
-rw-r--r-- 1 eduuser eduuser 403 2014-12-04 23:09 01.move_data_file.sh
drwxr-xr-x 1 eduuser eduuser 2037 2014-12-04 23:09 02.proc_csv.sh
-rw-r--r-- 1 eduuser eduuser 608 2014-12-04 23:09 03.upload_csv.sh
-rw-r--r-- 1 eduuser eduuser 1676 2014-12-04 23:09 04.run.sh
-rw-r--r-- 1 eduuser eduuser 2168 2014-12-04 23:09 04.run-devel.sh
-rw-r--r-- 1 eduuser eduuser 7417 2014-12-04 23:31 Social_textmining.R
-rw-r--r-- 1 eduuser eduuser 2309 2014-12-04 23:33 Social_wordcount.R
-rw-r--r-- 1 eduuser eduuser 3045734 2014-12-05 02:05 sampe_201101.json
-rw-r--r-- 1 eduuser eduuser 2303 2014-12-05 03:19 social_tw.R
|eduuser@cm04 basic]$ ./01.move_data_file.sh
```

- 로컬에 원시데이터를 작업 영역 폴더로 이동(/home/eduuser/nia_kbig/data/) 시킨다. ./01.move_data_file.sh 입력 후 엔터





III 가공

개요

77

데이터 가공 R 스크립트

84



가공

▶ 개요

소셜 미디어 데이터의 가공은 수집된 데이터에서 구조에 따라 분석에 필요한 데이터만을 분류하고, 문장에서 불필요한 문자를 제거한다. 그리고 데이터에서 분석에 필요한 데이터만을 선별하고, 불필요한 데이터와 문장을 제거하여 분석의 효율성을 높인다. 이 과정에서 키워드 추출과 정제, 숫자와 문자를 분별, 사전 관리 등이 이루어진다.

▶ 가공 방법

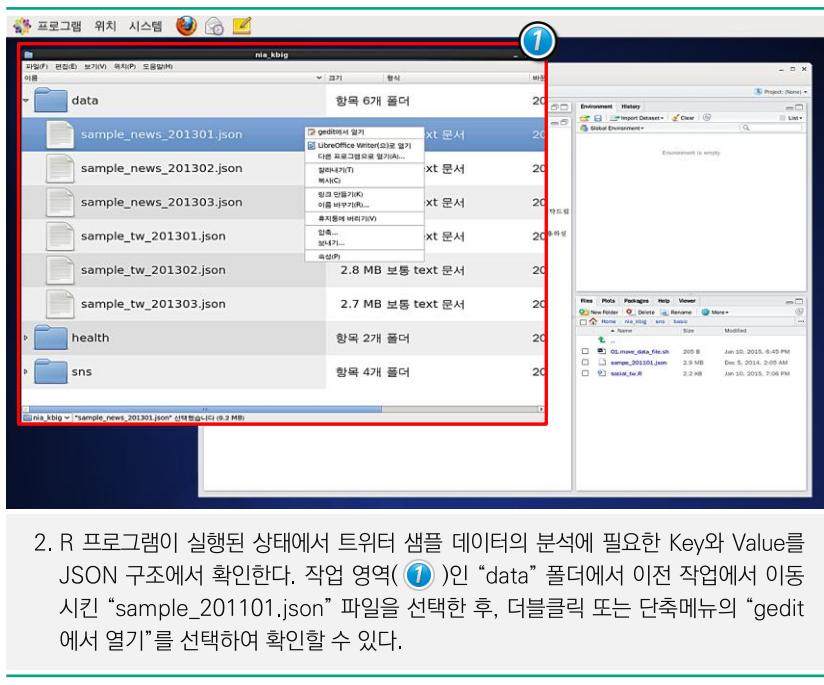
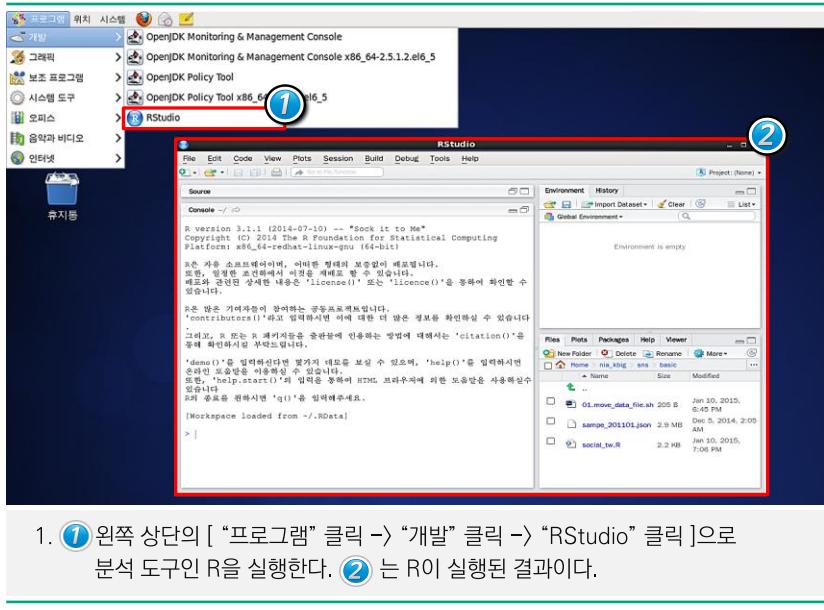
- **데이터 구조 가공** : 소셜 데이터는 한글과 영어로 구성되어 있기 때문에 Encoding 구조와 데이터 구조인 JSON 형태를 확인하고, 필요한 Encoding 구조인 “UTF-8”과 JSON 구조에서 필요한 Key와 Value를 추출한다.
- **데이터 가공 준비** : R에서 데이터 가공을 위한 라이브러리 리스트를 확인하고, 해당 라이브러리를 설치한다. 또한 한글 형태소 분석을 위해 데이터 상태를 확인한다.
- **트위터 문장 구조 가공** : 트위터는 RT(ReTwit)에서 발생된 데이터, URL 정보는 분석에서 사용되지 않기 때문에 1차적으로 해당 데이터를 삭제한다.
- **가공 분석을 위해, 프로그래밍 도구인 R을 실행한다.** R은 10,000 줄 이상의 데이터 처리 제약이 있기 때문에 대용량의 소셜 데이터 처리를 위해서는 Map Reduce와 결합하여 처리한다. 단, Rhadoop은 의존성이 높은 도구로 설치와 이용이 까다롭기 때문에 하둡 버전별로 다루는 것이 중요하다.



용 어 경 리

- **Rhadoop()** : R에서 대용량 데이터 처리 Map Reduce 이용을 위한 프로그램 보조 도구

▶ 데이터 가공



III. 가공

```
01. [
02. {
03.     "message": "@mrkangok 네ㅋ 수업준비할게많아서ㅋ",
04.     "screenName": "mingky_Kim",
05.     "created_at": "Sun Jan 09 07:31:00 KST 2011"
06. },
07. {
08.     "message": "@honey_ming 영화 보려구..",
09.     "screenName": "BangbaeGang",
10.     "created_at": "Sun Jan 09 07:31:00 KST 2011"
11. },
12. {
13.     "message": "@urian8345 오빠근데잘살아요?ㅋㅋ나컬러링 바꿨는?ㅋㅋㅋㅋ",
14.     "screenName": "kkjane0620",
15.     "created_at": "Sun Jan 09 07:30:00 KST 2011"
```

- 전체 구조는 배열 구조(Array)로 구성되어 있다. 그리고 분석을 위한 문장에 대한 JSON key는 “message”이며, 해당 값은 비정형 구조로 이루어져 있다. 또한 “screenName”은 사용자의 이름으로 사용자 분석에 활용할 수 있으며, 작성된 날짜에 해당하는 JSON key는 “created_at”이다.

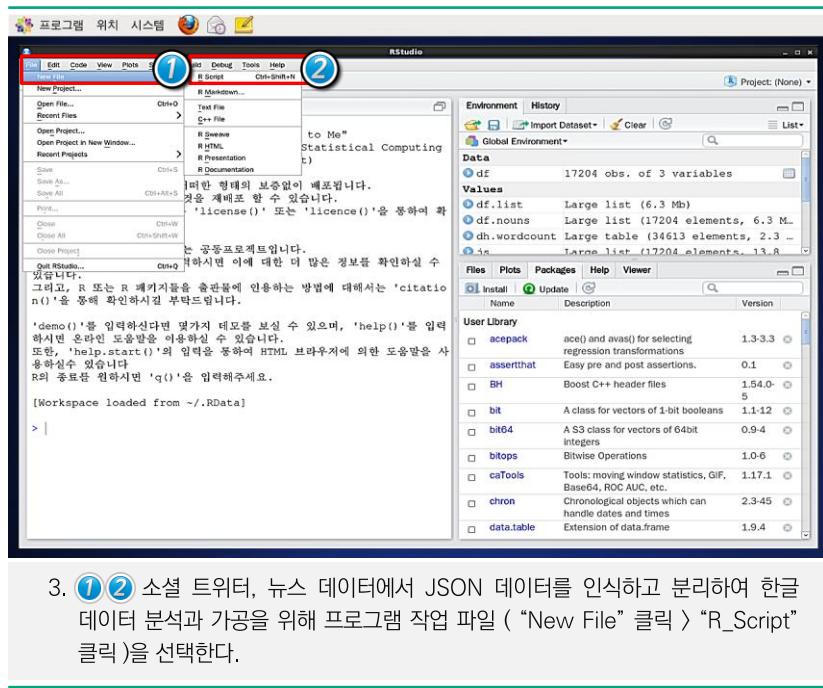
```

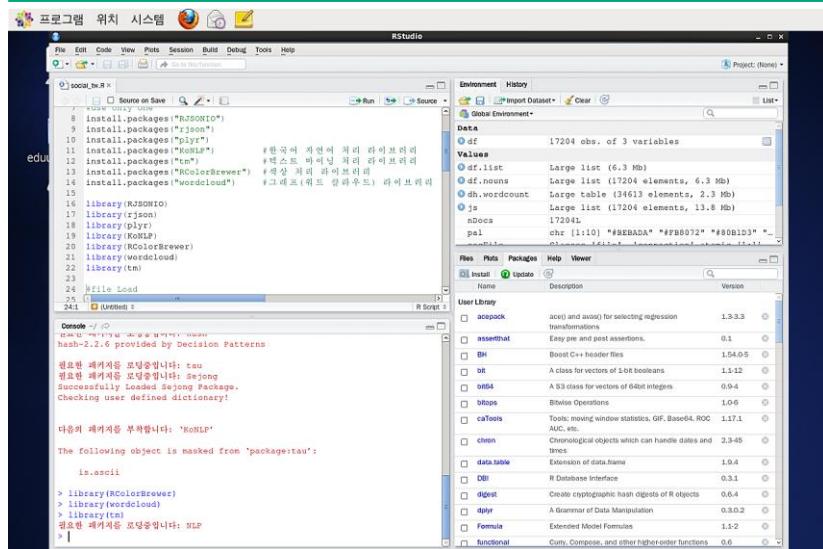
01. [
02. {
03.   "content": "금융·증권 ▼\n상호금융 집중 관리, 은행 수준으로 규제 강화\n 입력 : 2
04.   "title": "상호금융 집중 관리, 은행 수준으로 규제 강화",
05.   "date": "20130131"
06. },
07. {
08.   "content": "쿠키 연예 배우 송혜교가 31일 오후 서울 한남동 블루스퀘어에서 열린
09.   "title": "쿠키 포토 송혜교, 시각장애인 연기 도전!",
10.   "date": "20130131"
11. },
12. {
13.   "content": "이재용기자\n      jylee@sed.co.kr\n경영정상화를 위한 수주활동에
14.   "title": "한진중공업 사측 · 노조, \"시신투쟁 즉각 중단해야\"",
15.   "date": "20130131"
16. }

```

- 전체 구조는 배열 구조(Array)로 구성되어 있다. 그리고 분석을 위한 문장에 대한 JSON key는 “content”이며, 해당 값은 비정형 구조로 이루어져 있다. 또한 “title”은 기사제목으로 헤드라인 분석에 활용할 수 있으며, 작성된 날짜에 해당하는 “date”이다.

III. 가공





4. 분석에 필요한 라이브러리 파일을 설치하기 위해, 필요한 라이브러리를 작성하고 실행한다.

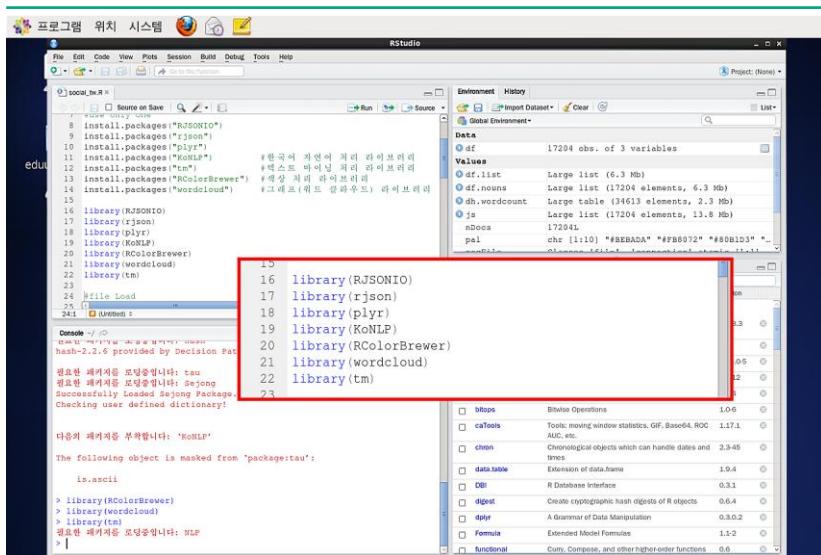
- #주) R 프로그램 분석을 위한 사전 라이브러리 설치는 install.package("라이브러리 이름")으로 설치하거나 오른쪽 하단의 패널을 이용하여 설치한다. 설치할 패키지 리스트는 아래와 같다. 작성된 줄의 끝에서 “**Ctrl+ Enter**”를 입력하여 실행한다.

```

01. #라이브러리 리스트
02. install.packages("rjson")      #JSON처리를 지원하는 라이브러리
03. install.packages("plyr")
04. install.packages("KoNLP")      #한국어 자연어 처리를 지원하는 라이브러리
05. install.packages("tm")         #텍스트 마이닝 처리를 지원하는 라이브러리
06. install.packages("RColorBrewer") #색상 처리를 지원하는 라이브러리
07. install.packages("wordcloud")   #그래프(워드 클라우드)를 지원하는 라이브러리
08. install.packages("gdata")       #단어의 빈도 분석을 지원하는 라이브러리
09. install.packages("data.table")  #데이터 구조 변경을 지원하는 라이브러리

```

III. 가공

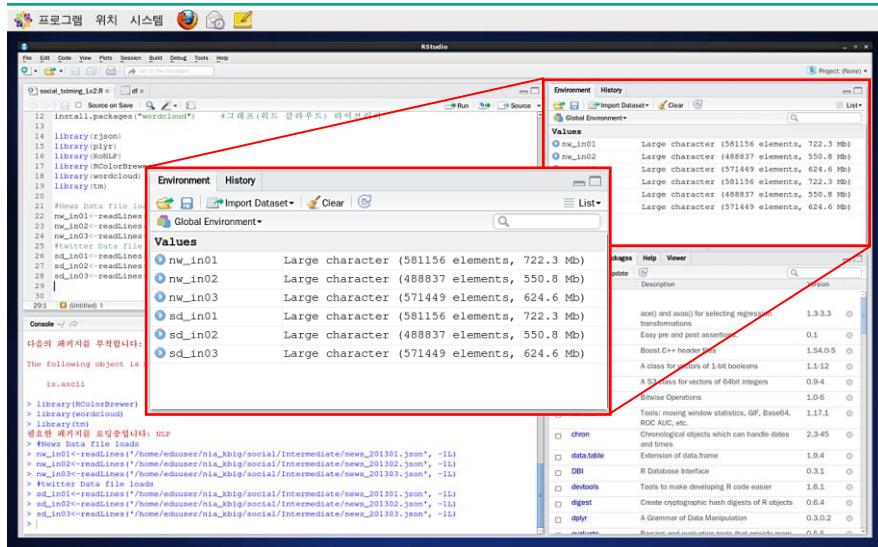


5. 설치된 라이브러리를 프로그램에 이용하기 위해 “library(‘이름’)”을 이용하여 불러온다.

- #주) 설치된 라이브러리들을 불러오는 리스트는 아래와 같다.

01. library(RJSONIO)
02. library(rjson)
03. library(plyr)
04. library(KoNLP)
05. library(RColorBrewer)
06. library(wordcloud)
07. library(tm)
08. library(data.table)
09. library(gdata)
10. library(ggplot2)

> 데이터 가공 R 스크립트(social_txmining_Lv2.R)



6. 분석에 필요한 파일과 JSON 구조에 맞게 지정된 변수로 파일을 불러온다.

- #주) 실행 후, 각각의 변수에 입력된 값은 오른쪽 상단의 패널을 통해 확인할 수 있다.

```

01. ##News Data file loads
  ↪ 수집된 파일 불러온다. 저장된 파일의 위치와 Large 데이터를 확인하여 입력한다.

02. nw_in01<-readLines('/home/eduuser/nia_kbig/data/sample_
  ↪ news_201301.json', -1L)
03. nw_in02<-readLines('/home/eduuser/nia_kbig/data/sample_
  ↪ news_201302.json', -1L)
04. nw_in03<-readLines('/home/eduuser/nia_kbig/data/sample_
  ↪ news_201303.json', -1L)

06. #twitter Data file loads
07. sd_in01<-readLines('/home/eduuser/nia_kbig/data/sample_
  ↪ news_201301.json', -1L)
08. sd_in02<-readLines('/home/eduuser/nia_kbig/data/sample_
  ↪ news_201302.json', -1L)
09. sd_in03<-readLines('/home/eduuser/nia_kbig/data/sample_
  ↪ news_201303.json', -1L)

11. #JSON converts

12. #News Data file loads

13. tr_nw01<- lapply(nw_in01, function(x) t(unlist(fromJSON(x))))
14. tr_nw02<- lapply(nw_in02, function(x) t(unlist(fromJSON(x))))

```

III. 가공

```
15. tr_nw03<- ldply(nw_in03, function(x) t(unlist(fromJSON(x))))  
16.  
17. #twitter Data file loads  
18. tr_sd01<- ldply(sd_in01, function(x) t(unlist(fromJSON(x))))  
19. tr_sd02<- ldply(sd_in02, function(x) t(unlist(fromJSON(x))))  
20. tr_sd03<- ldply(sd_in03, function(x) t(unlist(fromJSON(x))))
```



- 84페이지 데이터 가공/분석 R 스크립트 소스(social_txtrmining_Lv2.R)
- 라인 02~09 : 분석을 위한 데이터 파일을 트위터, 뉴스로 구분하여 각각 불러온다. 데이터의 크기가 크기 때문에 Large 데이터 표기로 공간을 확보한다.
- 라인 12~20 : 불러온 트위터와 뉴스 데이터를 JSON 구조에 맞춰 데이터 프레임으로 임시 변환하고, 분석 구조에 맞춰 컬럼 단위로 기록한다.

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화



IV 저 장

개요	89
R Studio 활용 저장	90

IV

저장

> 개요

비정형 데이터는 수집 데이터와 분석 데이터로 구분하여 JSON 구조를 갖는 데이터 모델링을 통해 분리하여 저장한다. 이를 위해, NoSQL과 같은 비정형 데이터베이스를 이용하여 기록하고 관리한다. NoSQL은 잘 설계된 데이터 모델링을 통해 대용량 데이터를 안정적으로 관리할 수 있다. 단, 데이터 모델이 완성되지 않은 상태에서는 텍스트 파일이나 CSV와 같은 구조로 저장하고 관리한다.



> 저장 방법

- **가공된 데이터 임시 저장** : 소셜 데이터 중 트위터에서 분석에 필요한 데이터만을 정제하여 파일로 데이터를 저장한다.
- **저장 파일의 구조 정의** : 문장에서 빈 공간을 가진 데이터를 제거한 데이터만을 저장한다.
- **소스 저장** : 작성 중인 소셜 트위터 분석 프로그램을 저장한다.

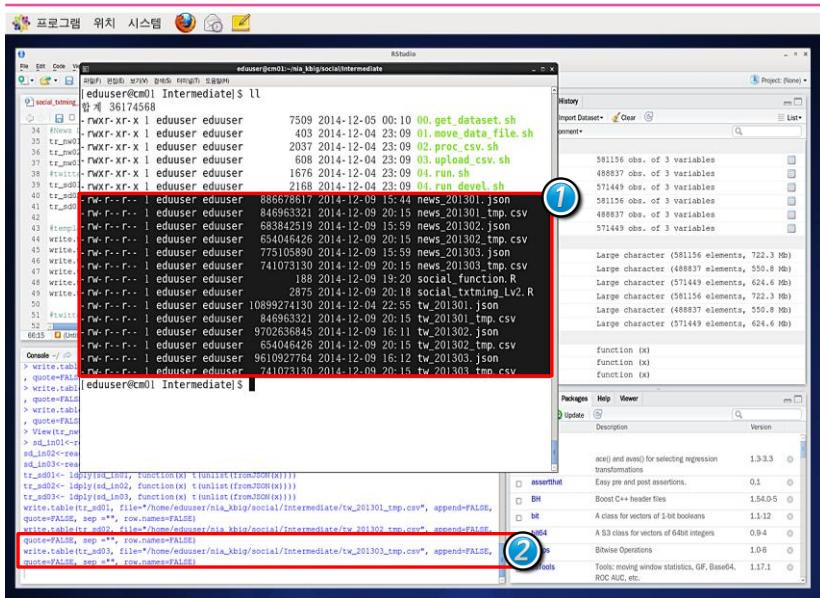


용어정리

- **NoSQL** : Not Only SQL의 약자로 기존 RDBMS 형태의 관계형 데이터베이스가 아닌 다른 형태의 데이터 저장 기술을 의미하며, 다른 형태의 데이터 저장 구조를 총칭하며, 제품에 따라 각기 그 특성이 매우 달라서 NoSQL을 하나의 제품군으로 정의할 수는 없음

> R Studio 활용 저장

> 데이터 저장



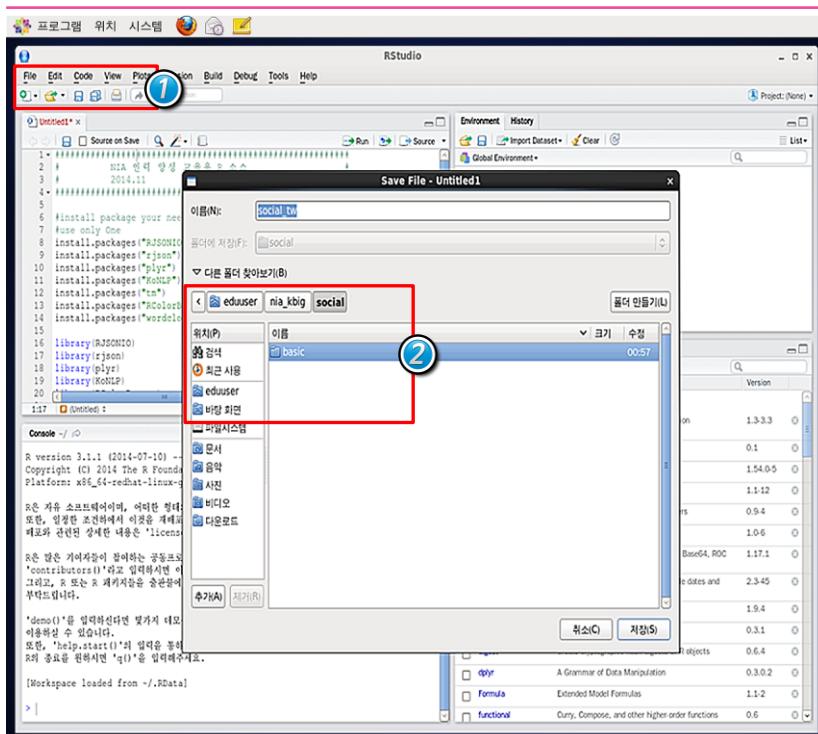
1. 안정적인 분석을 위해, 가공된 데이터를 파일로 저장한다. 저장된 파일은 다른 분석 프로그램으로도 사용할 수 있도록 “CSV” 파일로 저장한다.
- #주) 저장된 파일은 실행시켜 놓은 터미널을 통해 확인할 수 있다. ①은 저장된 파일과 위치에 해당하는 파일이고, ②는 R에서 실행된 CSV로, 임시 저장되는 명령어의 실행 정보이다.

```

01. #template data save frame structured data
02. write.table(tr_nw01, file="/home/eduuser/nia_kbig/data/news_
  ↪ 201301_tmp.csv", append=FALSE, quote=FALSE, sep = "", row.names=FALSE)
03. write.table(tr_nw02, file="/home/eduuser/nia_kbig/data/news_
  ↪ 201302_tmp.csv", append=FALSE, quote=FALSE, sep = "", row.names=FALSE)
04. write.table(tr_nw03, file="/home/eduuser/nia_kbig/data/news_
  ↪ 201303_tmp.csv", append=FALSE, quote=FALSE, sep = "", row.names=FALSE)
05. write.table(tr_sd01, file="/home/eduuser/nia_kbig/data/tw_
  ↪ 201301_tmp.csv", append=FALSE, quote=FALSE, sep = "", row.names=FALSE)
06. write.table(tr_sd02, file="/home/eduuser/nia_kbig/data/tw_
  ↪ 201302_tmp.csv", append=FALSE, quote=FALSE, sep = "", row.names=FALSE)
07. write.table(tr_sd03, file="/home/eduuser/nia_kbig/data/tw_
  ↪ 201303_tmp.csv", append=FALSE, quote=FALSE, sep = "", row.names=FALSE)

```

IV. 저장



2. 소셜 트위터, 뉴스 데이터 분석을 위해 작성 중인 프로그램 소스를 저장한다.

- #주) 작성 중인 프로그램 소스를 저장하는 방법은 메뉴의 “File” → “Save”를 이용하거나 도구상자의 저장 아이콘을 이용한다. 저장 시 저장 위치는 eduuser라는 폴더를 선택하여 하위 폴더를 따라 저장할 위치를 이동하여 최종적으로 “Intermediate”를 선택한다. 파일명은 본인이 작성한다.

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화

W





V 분석

개요	95
R Studio 활용 분석	96
R Studio 저장	106

V

분석

> 개요

소셜 데이터 분석은 한국어 또는 영어와 같이 언어에 따른 데이터 사전을 이용하여 명사형 또는 연관 키워드를 추출하는 형태소 분석을 실시한다. 그리고, 추출된 형태소 분석 데이터를 이용하여 키워드 빈도 분석, 연관 분석, 토픽 분석 등과 같은 텍스트 마이닝 및 분류 분석을 처리할 수 있다.

> 데이터 분석 방법

- **가중치 빈도 계산** : 키워드 빈도에 가중치를 부여하기 위한 핵심 키워드를 사전에 추가하고, 텍스트마이닝 라이브러리인 “tm” 안에 있는 가중치 빈도 계산을 이용한다.
- **명사형 단어 추출** : 단어를 추출하여 사전에 동사, 형용사 등의 단어를 제거하고 명사 구조만을 추출한다. 단, 한국어 사전이 불완전하기 때문에 명사형 형태로 단어를 추출한다.
- **키워드 빈도수 계산** : 명사형 단어를 인덱싱하고, 해당 단어가 등장하는 키워드의 빈도를 계산한다. 이때 “tm”, “KoNLP”, “NLP” 라이브러리를 이용하여 계산한다.



용어 정리

- **형태소 분석** : 한국어 형태소 분석을 위해서는 공개용 한국어 사전을 이용하지만, 일반적으로 정확한 분석을 위해서는 상업용 분석 사전을 이용한다. 개방형 형태소 분석은 설치된 라이브러리 중 한국어 자연어 처리 라이브러리인 “KoNLP”에서 제공한다.

> R Studio 활용 분석

> 데이터 불러오기

소셜 트위터 샘플 데이터 (tw_201301_tmp.csv)

```
eduuser@cm01:~/nia_kbkg/social/Intermediate
eduuser@cm01 Intermediate $ cat tw_201301_tmp.csv | more
[...]
[1] 네ㅋ 수업준비 할게 많아서ㅋㅋ mingky_KimSun Jan 09 07:31:00 KST 2011
영화 보려구.. BangbaeGangSun Jan 09 07:31:00 KST 2011
오빠근데 잘살아요?ㅋㅋ 나걸러링 바꿨는데ㅋㅋㅋ kkjane0620Sun Jan 09 07:30:00 KST 2011
너무 딱딱해ㅋ YCBBlueTearsSun Jan 09 07:30:00 KST 2011
또 다시, Darlin - Avril LavigneJinyBoySun Jan 09 07:30:00 KST 2011
누나 지금 만나러갑니다랑 다만날사랑해구 있어 보셨어요??? 캐감동인데ㅠㅠㅠㅠ sh
in5539Sun Jan 09 07:30:00 KST 2011
개드립인거냐?! 악해 YCBBlueTearsSun Jan 09 07:30:00 KST 2011
트친님들은 시가암이중!! 시가챙이!!ㅋㅋ kyoungyun84Sun Jan 09 07:30:00 KST 2011
살만께서 오는 여행ㅋㅋ sinsegyl2001Sun Jan 09 07:30:00 KST 2011
응!!ㅋ siesta5645Sun Jan 09 07:30:00 KST 2011
아! 더리워 죠미님ㅠㅠ RT?! areminsSun Jan 09 07:30:00 KST 2011
난 위판보면서 치킨먹는?ㅋㅋㅋ sleepyracoonSun Jan 09 07:30:00 KST 2011
#JND2ㅋㅋㅋㅋㅋ 불편하죠... 아주많이...ㅋㅋㅋ그리구...문화생활...뭐이런거 할
수가 없으니까요., dkfma486Sun Jan 09 07:30:00 KST 2011
오늘은 예상보다 귀가가 좀 늦어졌다. 이젠 나도 좀 일찍 자고 일찍 일어나는 습관을
길러야지. 너무 나태해졌어. MeinkampfSun Jan 09 07:30:00 KST 2011
그래 넌 훌륭한 움브파탈이다. 이 형이 인정함. 헤어스타일부터 움브파탈... Pimp_Ja
ngSun Jan 09 07:30:00 KST 2011
위탄보시는 분~ 황지환 동해랑 비슷하지 않나요? SongLidaSun Jan 09 07:30:00 KST 2011
이 노래 좋아하는 분 꽤 많네요. ^^ 세월이 흘러도 변할수 없는 감성이 있나봐요. Hong
yjnnnSun Jan 09 07:30:00 KST 2011
가야하나요ㅋㅋ 나갈까?!ㅋㅋ lovelyberretSun Jan 09 07:31:00 KST 2011
아싸 명작 게임을 찾았다!! DathungSun Jan 09 07:31:00 KST 2011
RT 사람들은 자기가 말을 하는 순간 그게 자기 삶의 전체에 대한 평가가 된다는 사실
```

- #주) 데이터 확인은 대용량 데이터로, Open Office를 이용하여 확인할 수 없기 때문에 Linux 명령어로 확인한다.
- \$ cat tw_201301_tmp.csv | more

V. 분석

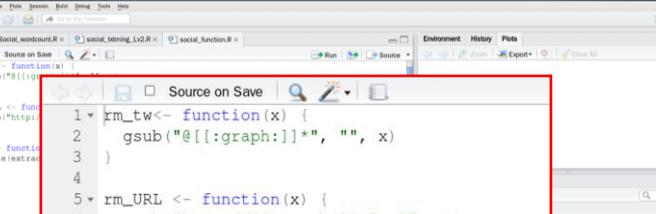
뉴스 샘플 데이터 (news_201301_tmp.csv)

```
프로그램 위치 시스템 브라우저(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)
eduuser@cm02:~/edu-source
[eduuser@cm02 edu-source]$ cat nw_201301_tmp.csv | more ①
Contenttitledate
전체 물량 81 청약 선방 21 3순위 126㎡형 서울 인천 최고 8. 33대1 기록 한라 간설
비발디' 아파트 순위 내 전체 모집 자수 81 선방했다 이 시 분양 한 김포한강신도시 내 아파트 소형 것
일 평가 22 2층 결제 원 한강신도시 한라비발디 3순위 청약 접수 21 결과 857 가구 모집 696명(특)
주택 평 전용면적 126㎡형은 60 가구 모집 126㎡형 2. 1대로 마감 3순위 서울 인천은 최고 8. 33대1 경
계자 청약 선방 한 이후 '뛰어난 입지여건' 와 '한강조망권을 극대화 한 설계특화'를 실제 한강신도시
태 공원 옆 위치 해 꽤적 한 주거환경 집안 한강 전경 수혜 멀티 조망권 중대 행 복구 실속 행 중대 행
줄 청약 유리하게 작용 평가 한 부동산 전문가 "김포한강신도시는 전체 인구 23 명 불과 청약 성
위에서 청약이 이 것 '한강신도시 한라비발디'는 지하2층~지상19층 12 개 동 857 세대 대단지로 전용면
적 126㎡형 60 세 구성 분양가 3.3㎡당 1,060 만 원 청약금 5 2 회 분납 중도금 60 전액, 무이자 적용
점자 대상 5 월 2 일 4 일 사용 간 청약 진행 견본 주택 김포시 고촌 읍 신곡리 1064번지에 위치 문 15
강신도시 한라비발디 #39; 3순위 서울 인천 최고 8.33대1, 20110422
내년 고등학교 입학생 한국사 수 5급 공무원 공채 시험 한국사 과목 내년 수 등 각종 공무원 시험 한
국 권장 교육과학기술부(장관 이주호 국사편찬위원회(위원장 이태진 역사교육과정 개발주선위원회(위
용·역사교과 강화방안)을 공동 발표 교과부 '우리 사회·문화·민족·글로벌적인 변화 진행 환경
'며 "독도 문제 등 주변국 지속 적 역사 교육 관련 한 주변 상황 학생들이 인식 필요 한국사 수 이수
방안 고교 선택 과목 한국사 2012 학년 고교 입학생 문과·이과·예체능계열 인문계고·특성화고 등
임 때 85 시 5 단 안팎 한국사 과목 현행 2009 개정 교육과정 포함 과목 가운데 수 과목 한국사 처음
학교 현장 한국사 능력 검정시험 성적 반영 방식 한국사 과목 수 사법시험·법원 5 금 시험·국회 9 금 시
원 공채 한국사 능력 검정시험 성적 반영 방식 한국사 과목 수 사법시험·법원 5 금 시험·국회 9 금 시
는 그동안 학생들이 역사교과서 지루하게 접 고려해 우리 역사 수교과서 내용 수정 그동안 초·
사적 방식 역사 기술 해 학생들 역사 공부 분량 압기 과목 정도 처부 한 편 교과부 이 탐구·체험·
정 주제 중심 서술 방식 김포 교과부 공표 교육 관련 단체 환영 목소리 내 참교육 학부모 회장은 숙
말 반면 전국교직원노동조합 '우리 역사 교육 강화 취지 자체 특정 상황 교육과정 운영지침 학교 현장
부처 치밀하게 대책 한다"고 주문 e 뉴스 팀 한국사, 내년부터 고등학교 필수과목으로 20110422
1990년대 '문화대통령'으로 가수 서태지(39·본명 김지아)가 위자료
원에 이씨는 올해 1 월 19 서씨를 상대 5 억 원 위자료 재산분할 청구 소송 두 사람 변호인 3 월 14 일
3 차 변론 준비기 일 다음달 23 일 양측 위자료 청구 소송 사건 이해적 3 4 명 변호인 단선임 두 사람
부 확인 사실 혼 관계 아이 둘 소송 이번 소송 양육권 부분 포함 일부 재산분할 청구 액수 50 억 원 주
태지 소속사 서태지컴퍼니는 "서태지는 음반 작업 차 해외 연락 있다"며 개인 사이 일 사실 생각 뿐 반
사람 혼인 이혼 위자료·재산분할 청구소송 사실 공식 인정 소속사 주장 이씨는 1993년 미국 유학
을 떠나 1997년 미국에서 둘 결혼식 결혼 후 애들 랜다 에리조나 등 결혼 생활 2000년 6월
```

- #주) 데이터 확인은 대용량 데이터로, Open Office를 이용하여 확인할 수 없기 때문에 Linux 명령어로 확인한다.
- \$ cat news_201301_tmp.csv | more

> 데이터 분석

- #주) 앞의 작성 중인 R 프로그램 소스에 이어서 작업한다. 작업 내용은 아래와 같다.



```
1 rm_tw<- function(x) {
2   gsub("@[:graph:]*", "", x)
3 }
4
5 rm_URL <- function(x) {
6   gsub("http://[:graph:]*", "", x)
7 }
8
9 msg <- function(x) {
10   paste(extractNoun(x) , collapse = " ")
11 }
12
```

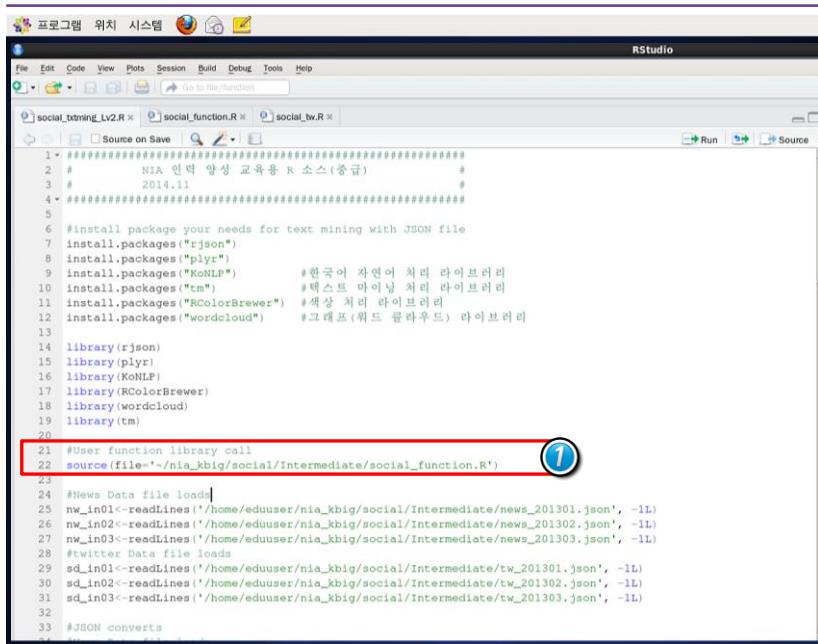
1. 형태소 분석을 위한 트위터 정제, 명사 추출 함수를 묶어서 사용자 라이브러리 파일로 작성한다.

```
01. rm_tw <- function(x) {  
02.   gsub("@[:graph:]*", "", x)  
03. }  
04. Rm_URL <- function(x) {  
05.   gsub("http://[:graph:]*", "", x)  
06. }  
07. msg <- function(x) {  
08.   paste(extractNoun(x), collapse = "")  
09. }
```



- 데이터 가공/분석 R 스크립트 소스(social_function.R)
 - 라인 01 ~ 03 : 트위터의 메시지 형식 중에서 RT 기호와 빈공간 등의 데이터를 찾아서 삭제한다.
 - 라인 04 ~ 06 : 트위터의 메시지 형식 중에서 URL 데이터를 찾아서 삭제한다.
 - 라인 07 ~ 09 : 데이터 중에서 명사형 데이터를 찾아서 분리한다.

V. 분석



```
1 * #####  
2 # NIA 인력 양성 교육용 R 소스(중급) #  
3 # 2014.11 #  
4 * #####  
5  
6 #install package your needs for text mining with JSON file  
7 install.packages("rjson")  
8 install.packages("plyr")  
9 install.packages("KoNLP") #한국어 자연어 처리 라이브러리  
10 install.packages("tm") #텍스트 마이닝 처리 라이브러리  
11 install.packages("RColorBrewer") #색상 처리 라이브러리  
12 install.packages("wordcloud") #그래프(워드 클라우드) 라이브러리  
13  
14 library(rjson)  
15 library(plyr)  
16 library(KoNLP)  
17 library(RColorBrewer)  
18 library(wordcloud)  
19 library(tm)  
20  
21 #User function library call  
22 source(file='~/nia_kbkg/social/Intermediate/social_function.R') 1  
23  
24 #News Data file loads  
25 nw_in01<-readLines('~/home/eduuser/nia_kbkg/social/Intermediate/news_201301.json', -1L)  
26 nw_in02<-readLines('~/home/eduuser/nia_kbkg/social/Intermediate/news_201302.json', -1L)  
27 nw_in03<-readLines('~/home/eduuser/nia_kbkg/social/Intermediate/news_201303.json', -1L)  
28 #Twitter Data file loads  
29 sd_in01<-readLines('~/home/eduuser/nia_kbkg/social/Intermediate/tw_201301.json', -1L)  
30 sd_in02<-readLines('~/home/eduuser/nia_kbkg/social/Intermediate/tw_201302.json', -1L)  
31 sd_in03<-readLines('~/home/eduuser/nia_kbkg/social/Intermediate/tw_201303.json', -1L)  
32  
33 #JSON converts
```

2. 작성된 사용자 라이브러리 파일은 ① 과 같이 Library 영역 다음에서 파일을 읽어온다.

```
01. #User function library call  
02. source(file="~/nia_kbkg/sns/middle/social_function.R")
```

```

# twitter data cleasing 201301 data
tr_sd01$message <- sapply(tr_sd01$message , rm_tw)
tr_sd01$message <- sapply(tr_sd01$message, rm_URL)
#twitter data cleasing 201302 data
tr_sd02$message <- sapply(tr_sd02$message , rm_tw)
tr_sd02$message <- sapply(tr_sd02$message, rm_URL)
#twitter data cleasing 201303 data
tr_sd03$message <- sapply(tr_sd03$message , rm_tw)
tr_sd03$message <- sapply(tr_sd04$message, rm_URL)

#Load seajong dictionary in KoNLP
useSeajongDic()
#adding your word in dictionary file and Stemming but English

```

The screenshot shows an RStudio interface with a script file open. The code is written in R, performing data cleaning on Twitter datasets from 2013. It uses the `sapply` function to remove specific characters and URLs from messages. Lines 51 through 59 are highlighted with a red box, indicating the part of the code responsible for writing the cleaned data to a CSV file. A red arrow points to the first line of the script.

3. 저장 코드 다음으로 트위터 데이터에서 불필요한 키워드를 정리한다.

▪ #주) 앞에서 작성한 사용자 라이브러리를 이곳에서 함수명으로 다시 호출한다.

```

01. #twitter data cleasing 201301 data
02. tr_sd01$message <- sapply(tr_sd01$message , rm_tw)
03. tr_sd01$message <- sapply(tr_sd01$message, rm_URL)
04. #twitter data cleasing 201302 data
05. tr_sd02$message <- sapply(tr_sd02$message , rm_tw)
06. tr_sd02$message <- sapply(tr_sd02$message, rm_URL)
07. #twitter data cleasing 201302 data
08. tr_sd03$message <- sapply(tr_sd03$message , rm_tw)
09. tr_sd03$message <- sapply(tr_sd04$message, rm_URL)

```



- 데이터 가공/분석 R 스크립트 소스(social_txtmining_Lv2.R)
- 라인 01~03 : 2013년 1월 트위터 샘플 데이터의 메시지 형식 중에서 RT 기호와 URL의 데이터를 찾아서 삭제한다.
- 라인 04~06 : 2013년 2월 트위터 샘플 데이터의 메시지 형식 중에서 RT 기호와 URL의 데이터를 찾아서 삭제한다.
- 라인 07~09 : 2013년 3월 트위터 샘플 데이터의 메시지 형식 중에서 RT 기호와 URL의 데이터를 찾아서 삭제한다.

V. 분석

```
49 write.table(tr_sd03, file="/home/eduuser/nia_kbig/social/Intermediate/tw_201303_tmp.csv", append=FALSE, quote=FALSE)
50
51 #twitter data cleasing 201301 data
52 tr_sd01$message <- sapply(tr_sd01$message, rm_tw)
53 tr_sd01$message <- sapply(tr_sd01$message, rm_URL)
54 #twitter data cleasing 201302 data
55 tr_sd02$message <- sapply(tr_sd02$message, rm_tw)
56 tr_sd02$message <- sapply(tr_sd02$message, rm_URL)
57 #twitter data cleasing 201302 data
58 tr_sd03$message <- sapply(tr_sd03$message, rm_tw)
59 tr_sd03$message <- sapply(tr_sd03$message, rm_URL)
60
61 #Load seajong dictionary in KoNLP
62 useSejongDic()
63 #adding your word in dictionary file and Stemming but English
64 mergeUserDic(data.frame(c("대한민국", "경제", "교육", "대학", "콘텐츠", "서비스"), c("ncn")))
65
66 #word extraction - news
67 tr_nw01$content <- sapply(tr_nw01$content, msg(x) )
68 tr_nw02$content <- sapply(tr_nw02$content, function(x) { paste(extractNoun(x), collapse = " ") })
69 tr_nw03$content <- sapply(tr_nw03$content, function(x) { paste(extractNoun(x), collapse = " ") })
70
71 #word extraction - twitter
72 tr_sd01$message <- sapply(tr_sd01$message, function(x) { paste(extractNoun(x), collapse = " ") })
73 tr_sd02$message <- sapply(tr_sd02$message, function(x) { paste(extractNoun(x), collapse = " ") })
74 tr_sd03$message <- sapply(tr_sd03$message, function(x) { paste(extractNoun(x), collapse = " ") })
75
76 #nouns extraction - news
77 tr_nw01.nouns <- sapply(tr_nw01$content, extractNoun, USE.NAMES=F)
78 tr_nw02.nouns <- sapply(tr_nw02$content, extractNoun, USE.NAMES=F)
79 tr_nw03.nouns <- sapply(tr_nw03$content, extractNoun, USE.NAMES=F)
80
81 #nouns extraction - twitter
82 tr_tw01.nouns <- sapply(tr_tw01$message, extractNoun, USE.NAMES=F)
83
```

4. 사전 데이터를 불러온다.

- #주) 실행된 결과는 아래와 같으며, 한국어 명사는 87,007개의 단어로 구성되어 있다.

```
01. #Load seajong dictionary
```

```
02. useSejongDic()
```

```
> #Load seajong dictionary in KoNLP
> useSejongDic()
Backup was just finished!
87007 words were added to dic_user.txt.
> |
```

5. 읽어들인 사전에 사용자 또는 개발자가 필요로 하는 단어를 추가한다.
- #주) 추가된 사용자 키워드는 명사 또는 명사형으로 필요한 경우에 사전에 추가하여 자신만의 사용자 사전을 구축할 수 있다. 실행된 결과의 예는 아래와 같다.

```
01. #adding your word #Stemming for English
02. mergeUserDic(data.frame(c("대한민국", "경제", "교육", "대학", "콘텐츠", "서비스"),
  ↪ c("ncn")))
```

```
> #Load seajong dictionary in KoNLP
> useSejongDic()
Backup was just finished!
87007 words were added to dic_user.txt.
> #adding your word in dictionary file and Stemming but English
> mergeUserDic(data.frame(c("대한민국", "경제", "교육", "대학", "콘텐츠",
+ c("ncn"))))
6 words were added to dic_user.txt.
> |
```



6. 명사형 단어를 분리하여 단어 사전에 맞는 단어로 변환하고, 단어의 빈도수와 가중치에 따라 분류하여 계산한다.

- #주) 추출한 단어에서 빈 공간(“ ”)을 기준으로, 단어를 분리한다. 그리고 분리된 키워드에서 명사형 키워드의 빈도를 계산한다. 실행된 결과는 아래와 같다.

```
01. #word extraction – news
02. tr_nw01$content <- sapply(tr_nw01$content, function(x) { paste(extractNoun(x),
  ↪ collapse = " ") })
03. tr_nw02$content <- sapply(tr_nw02$content, function(x) { paste(extractNoun(x),
  ↪ collapse = " ") })
04. tr_nw03$content <- sapply(tr_nw03$content, function(x) { paste(extractNoun(x),
  ↪ collapse = " ") })
05.
06. #word extraction – twitter
07. tr_sd01$message <- sapply(tr_sd01$message, function(x) { paste(extractNoun(x),
  ↪ collapse = " ") })
08. tr_sd02$message <- sapply(tr_sd02$message, function(x) { paste(extractNoun(x),
  ↪ collapse = " ") })
09. tr_sd03$message <- sapply(tr_sd03$message, function(x) { paste(extractNoun(x),
  ↪ collapse = " ") })
```

7. 소셜 데이터의 명사형 추출 결과는 각 문장에 대하여 명사 형태를 가지고 있는지를 사전을 이용하여 형태소를 판단한다.

- #주) 한글과 영어에 대한 형태소를 위한 함수가 다르며, 각각의 기능은 지원하는 형태소 분석 함수에 따라 다르다. 실행된 결과의 예는 아래와 같다.

	V1
1	c("전체", "물량", "81", "청약…증대형", "불구", "청약", "선별", "21", "3순위", "126m²형", "서울인천", "최고", "8.", "33대", "감인", "이래적", "일", "평가", "22", "금융", "결제", "원", "한강신도시", "한리", "비밀", "디", "3순위", "청약", "접수", "2", "실감캐", "한라", "건설", "관계자", "청약", "선방", "한", "이유", "뛰어난", "일지어건'와", "한강조망권을", "극대화", "한", "105m~106m²형을", "집중", "적", "배치", "해", "부담", "줄", "것", "청약", "유리", "하게", "작용", "평가", "한", "부동산", "진용면적", "기준", "105m²형", "513", "세", "대", "106m²형", "284", "세", "대", "126m²형", "60", "세", "구성", "분양가", "1064번지에", "위치", "문", "1599", "3737", "한경닷컴", "이유", "선", "기", "yuri", "hankyung", "com")
2	c("내년", "고등학교", "일학생", "한국사", "수", "5", "급", "공무원", "공채", "시험", "한국사", "과목", "내년", "수", "등", "급", "역사교육", "강화방안을", "공동", "발표", "교과", "부", "우리", "사회", "다문화·다민족·글로벌적인", "변화", "진행", "이날", "발표", "한", "방안", "고교", "선택", "과목", "한국사", "2012", "학년", "고교", "일학생", "문과·예체능계열", "한국사", "반영", "확대", "교과", "부", "학교현장", "한국사", "소양", "교사", "리", "2013", "년", "신규", "교원", "임용", "(한국사)", "과목", "포함", "방안", "관련", "부처", "사이", "논의", "그동안", "학생", "들이", "역사교과서", "지루", "하게", "탐구·체험·토론", "활동", "나용", "강화", "일회", "인물", "이야기", "특정", "주제", "중심", "서술", "방식", "검토", "교과", "과정", "운영지침", "학교", "현장", "교과목", "등", "수", "있다"며, "학교교과", "흔한", "교과", "부", "치밀", "하게", "대책", "1990", "년대", "문화대통령"으로, "가수", "서태지(39·본명, "이지아(33·본명, "김지아)가", "집", "진행", "3", "차", "번들", "준비기", "일", "다음달", "23", "일", "양측", "위자료", "청구", "소송사건", "이래적", "3", "4", "계시", "상태", "것", "이날", "서태지", "소속사", "서태지컴퍼니는", "서태지는", "음반작업", "차", "해외", "연락", "있다", "서서를", "연인", "서서기", "1996", "년", "초", "은퇴", "후", "미국으로", "1997", "년", "미국에서", "둘", "결혼식", "결혼", "재산분할", "청구소송", "소멸시효", "이상", "협의", "것", "판단", "해", "소", "제기", "둘", "사이", "자녀", "소문", "사실무", "마음", "이씨는", "2007", "년", "MBC", "태왕사신기", "여주인공", "드라마", "첫발", "이", "MBC", "베토벤", "비이어스"(결혼·출산", "이혼", "줄고…경기기", "▶", "명동은", "명품전쟁…롯데·신세계", "사이", "대형", "병행수입", "관", "등장", "&", "mك", "co", "kr", "무단", "전재", "재배", "포", "금지")
3	#nouns extraction – news tr_nw01.nouns <- data.table(sapply(tr_nw01\$content, extractNoun, USE.NAMES=F)) tr_nw02.nouns <- data.table(sapply(tr_nw02\$content, extractNoun, USE.NAMES=F)) tr_nw03.nouns <- data.table(sapply(tr_nw03\$content, extractNoun, USE.NAMES=F)) #nouns extraction – twitters tr_sd01.nouns <- data.table(sapply(tr_sd01\$message, extractNoun, USE.NAMES=F)) tr_sd02.nouns <- data.table(sapply(tr_sd02\$message, extractNoun, USE.NAMES=F)) tr_sd03.nouns <- data.table(sapply(tr_sd03\$message, extractNoun, USE.NAMES=F)) #명사형 keyword 빈도수 계산(tf-idf) – news dh_nw01.wordcount <- table(unlist(tr_nw01.nouns)) dh_nw02.wordcount <- table(unlist(tr_nw02.nouns)) dh_nw03.wordcount <- table(unlist(tr_nw03.nouns))



- 데이터 가공/분석 R 스크립트 소스(social_txmining_Lv2.R)
- 라인 01~04 : 각각의 뉴스 데이터에서 명사하여 단어를 추출한다.
- 라인 06~09 : 각각의 트위터 데이터에서 명사하여 단어를 추출한다.
- 라인 11~14 : 각 뉴스 명사형 데이터에서 키워드 빈도수를 TF/IDF 알고리즘을 이용하여 계산한다.

```

01. #명사형 keyword 빈도수 계산(tf-idf) – twitter
02. dh_sd01.wordcount <- table(unlist(tr_sd01.nouns))
03. dh_sd02.wordcount <- table(unlist(tr_sd02.nouns))
04. dh_sd03.wordcount <- table(unlist(tr_sd03.nouns))
05.
06. #Text Mining – news
07. doc_01 <- Corpus(VectorSource(tr_nw01.nouns))
08. doc_02 <- Corpus(VectorSource(tr_nw02.nouns))
09. doc_03 <- Corpus(VectorSource(tr_nw03.nouns))
10.
11. #Text Mining – twitter
12. doc_04 <- Corpus(VectorSource(tr_sd01.nouns))
13. doc_05 <- Corpus(VectorSource(tr_sd02.nouns))
14. doc_06 <- Corpus(VectorSource(tr_sd03.nouns))
15.
16. #숫자 제거 – integration compute – for news
17. doc_01 <- tm_map(doc_01, removeNumbers)
18. doc_02 <- tm_map(doc_02, removeNumbers)
19. doc_03 <- tm_map(doc_03, removeNumbers)
20.
21. #숫자 제거 – integration compute – for twitter
22. doc_04 <- tm_map(doc_04, removeNumbers)
23. doc_05 <- tm_map(doc_05, removeNumbers)
24. doc_06 <- tm_map(doc_06, removeNumbers)
25.
26. #Document Matrix 생성 – for integration data
27. doc_01 <- DocumentTermMatrix(doc_01)
28. doc_02 <- DocumentTermMatrix(doc_02)
29. doc_03 <- DocumentTermMatrix(doc_03)
30.

```



데이터 가공/분석 R 스크립트 소스(social_txtmining_Lv2.R)

- 라인 01~04 : 각 트위터 명사형 데이터에서 키워드 빈도수를 TF/IDF 알고리즘을 이용하여 계산한다.
- 라인 06~14 : 트위터와 뉴스 데이터를 텍스트 마이닝 알고리즘으로 키워드를 정리하고 분류한다.
- 라인 16~24 : 트위터와 뉴스 데이터에서 숫자를 제거하여 순수 명사형만을 보관한다.
- 라인 27~29 : 각 데이터를 하나로 통합하기 위해 데이터를 메트릭스 구조로 변환한다.

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화

V. 분석

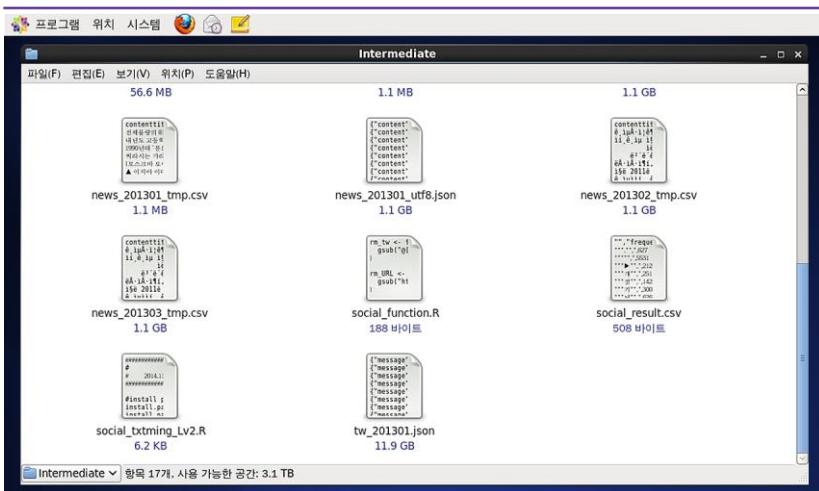
```
01. doc_04 <- DocumentTermMatrix(doc_04)
02. doc_05 <- DocumentTermMatrix(doc_05)
03. doc_06 <- DocumentTermMatrix(doc_06)
04.
05. #단어 빈도 분석 -> 각각의 문서의 단어를 통합하고, 변환
06. totalDoc <- as.matrix(cbind(doc_01))
07. totalDoc <- as.matrix(cbind(totalDoc,doc_02))
08. totalDoc <- as.matrix(cbind(totalDoc,doc_03))
09. totalDoc <- as.matrix(cbind(totalDoc,doc_04))
10. totalDoc <- as.matrix(cbind(totalDoc,doc_05))
11. totalDoc <- as.matrix(cbind(totalDoc,doc_06))
12.
13. #추출된 단어 문서를 저장, 키워드에 대한 리스트로 CSV보다는 text 파일로 저장이 우수
   ↳ write.csv(totalDoc, file="/home/eduuser/nia_kbig/sns/middle/term_doc.txt")
14.
15. ## Term matrix에서 단어 검사를 위한 Document Term Matrix로 재변환 처리
16. frequency <- colSums(totalDoc)
17. frequency <- subset(frequency, frequency >= 600)
18. frequency_result <- as.data.frame(x=frequency, row.names = names(frequency),
   ↳ optional = FALSE)
19.
20. #선정된 데이터 값을 임시 저장
21. write.csv(frequency_result, file="/home/eduuser/nia_kbig/sns/middle/
   ↳ term_frequency.csv")
22. #그래프 출력
23. barplot(frequency, las=1, legend = rownames(frequency))
```



- 데이터 가공/분석 R 스크립트 소스(social_txtmining_Lv2.R)
- 라인 01~03 : 각 데이터를 하나로 통합하기 위해 데이터를 메트릭스 구조로 변환한다.
- 라인 05~11 : 각 데이터 빈도 계산을 위해 하나의 데이터로 통합한다.
- 라인 13 : 통합된 데이터를 임시로 저장한다.
- 라인 15~18 : 각 컬럼 단위로 통합된 데이터의 빈도수를 계산하고, 그 중에서 특정 출현 빈도수 이상(600)의 단어만을 추출하여 데이터를 분리하여 관리한다.
- 라인 21 : 추출된 특정 수 이상의 단어를 임시 저장한다.
- 라인 23 : 추출된 단어에 대한 빈도수를 막대 그래프로 출력한다.

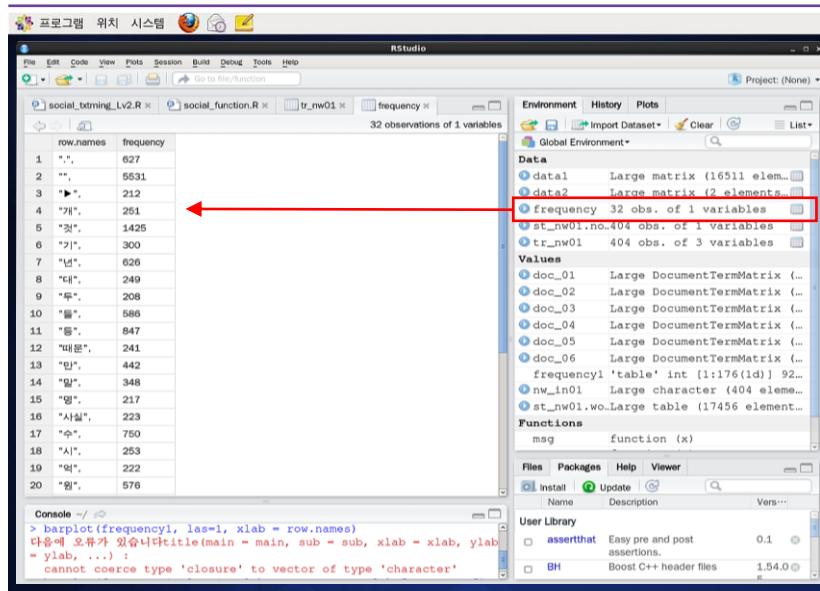
> R Studio 저장

> 분석 결과 저장



- 분석된 결과를 저장한다. 분석 결과는 다양한 방법으로 저장할 수 있으며, 앞의 정제 데이터 저장과 동일한 CSV 형태로 저장한다.
- #주) 저장된 파일의 확인은 실행되어 있는 터미널을 이용하여 확인한다.

➤ 결과 데이터



I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화



1

2



VI 시각화

개요	111
분석 데이터 시각화	113
데이터 분석	115

VI

시각화



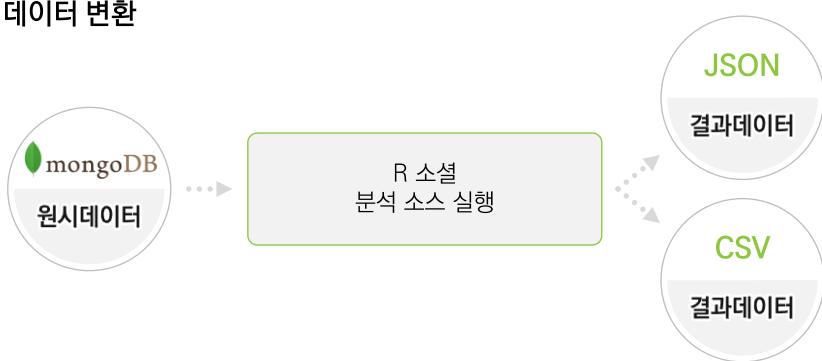
개요

분석 결과는 다양한 방법으로 시각화하여 분석할 수 있으며, 이를 통해 데이터의 변화 및 분포를 해석하고 데이터에 대한 분석 효과를 적용할 수 있다. 특히, 데이터의 출현 빈도는 이슈 키워드 중심으로 변화하기 때문에 이에 대한 기간 또는 분기별 키워드의 변화 정보를 판단할 수 있는 정보이다. 또한 최종적으로 획득된 분석 결과는 대량의 데이터로 상위 데이터의 출현에 따라 분리가 필요하며, 이를 시각화한 후에 키워드의 빈도 상태값에 따라 의미 해석을 검색엔진과 함께 진행할 수 있다.

> 시각화 방법

- **분석 결과에 대한 그래프 설정** : 분석된 결과에 대하여 키워드 출현 빈도를 200개 이상으로 제한된 키워드 출현 빈도에 대한 막대그래프 출력으로 한다.
- **저장된 포맷에 맞는 그래프 도구 설정** : 저장된 데이터를 이용하여 시각화할 수 있는 도구도 다양하기 때문에 이를 위해 시각화 도구를 선택한다.
- **의미 해석** : 시각화된 Bar 그래프에서 크기와 표현된 키워드는 자주 또는 가장 많이 언급된 키워드 정보 중에서도 상위 키워드이며, 검색엔진을 통해 사회적 이슈와 등장 빈도를 확인한다.

▶ 데이터 변환



> 분석 데이터 시각화

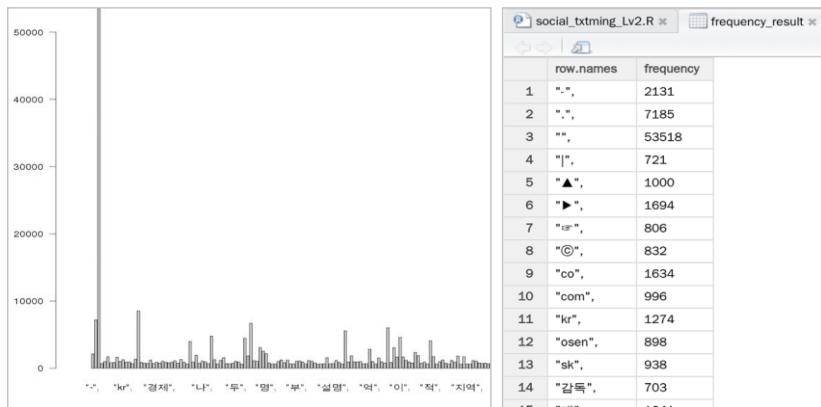
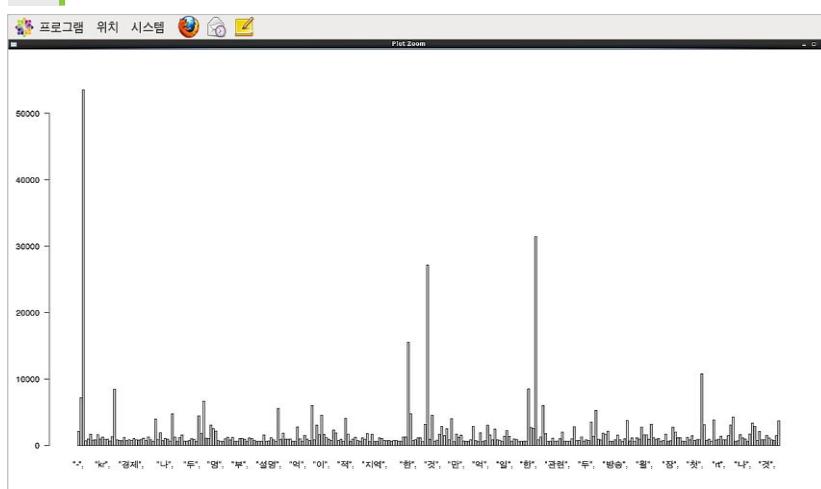
> 데이터 시각화

- 시각화를 위한 데이터 범위와 그래프를 설정한다.
- #주) 그래프는 막대그래프에 출력하기 위해 제한된 데이터를 출력하도록 설정하였지만, 다양한 출력을 위해 R과 함께 Open Office를 이용하도록 한다. 작성한 코드의 실행 결과는 아래와 같다. R의 데이터 구조상 X축 표현이 부족하다.

```

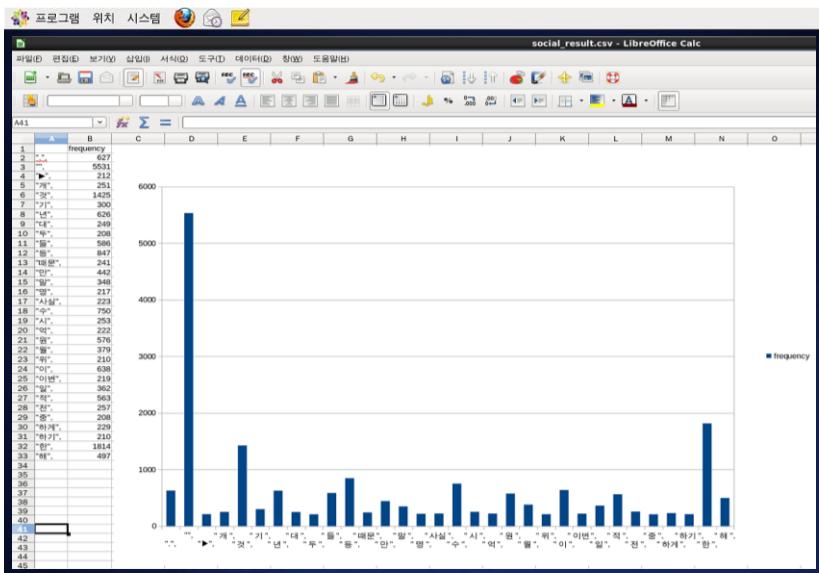
01. #package dependency clear and Visualization
02. #그래프 출력
03. barplot(frequency, las=1, legend = rownames(frequency))
04.

```

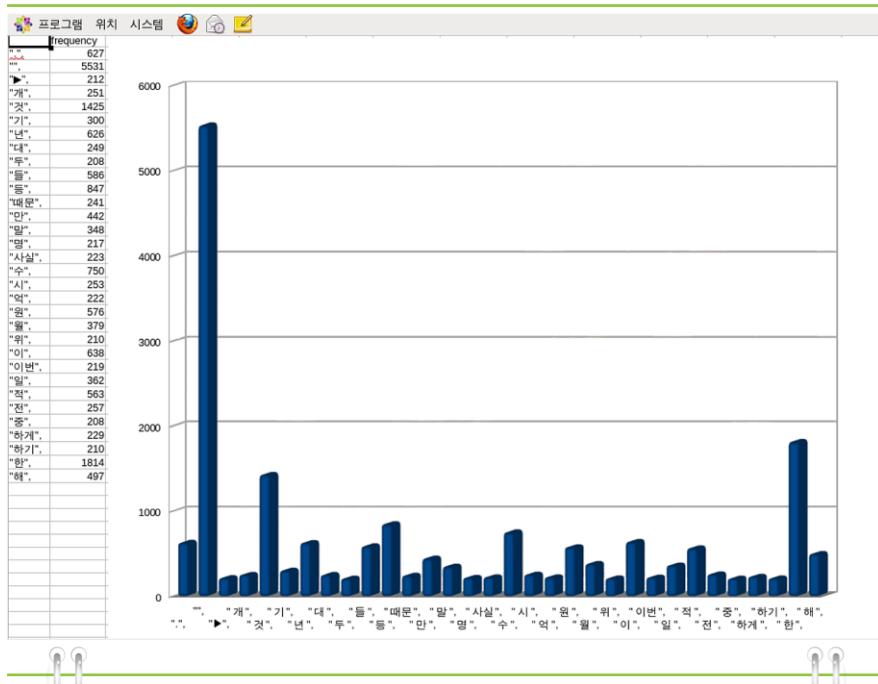


2. 시각화 결과에 대한 그래프는 RStudio의 오른쪽 하단 패널에 “Plots”으로 확인할 수 있다. 또한 Open Office를 이용하여 저장된 결과(CSV)를 읽어 아래와 같이 그래프로 처리할 수 있다.

- #주) RStudio에서는 그래프를 Zoom으로 확인하면 위와 같다.



▶ 데이터 분석



- 한국어 형태소와 단어 빈도의 계산 결과는 보이지 않는 스페이스가 가장 많은 빈도를 차지하고 있으나, 이는 문자의 인코딩(Encoding)에서 발생되는 문제로 제외할 수 있으며, 다음으로 많이 등장한 빈도의 글자는 “한”이다. 이는 “한국”과 같이 단어를 부연해 주는 단어로 많이 사용되고 있음을 알 수 있다. 또한 “것”도 마찬가지로 불용어지만 형태소 분석의 기능이 더욱 지원되어야 할 것이다.
- 특히, 트위터에서는 단어가 아니 인터넷 음어를 제거하는 기능으로 키워드의 발현 빈도를 조정하여 제거하거나 사전의 보강으로 지원할 수 있다.
- 소설 데이터에서 정제가 되어야 할 단어 중에서 “년”, “월”, “일”과 같은 명사형 처리 지원이 필요하며, 등장키워드에 대한 조합 분석이 추가적으로 수행되어야 한다.
- 한국어 형태소 분석의 정확한 분석을 위해서는 상업용 한글 형태소 분석기를 사용과 실시간 소설 분석 처리가 좋으며, 기능 분석상 공개용 버전으로 소설 데이터를 처리하는 것이 성능상 곤란하기 때문에 데이터에 대한 절제된 범위를 사용하였다.

I. 개요

II. 수집

III. 가공

IV. 저장

V. 분석

VI. 시각화



VII 예제문제

예제 문제1

119

예제 문제2

120

예 / 제 / 문 / 제

예제 1

뉴스 데이터에서 월별 키워드 빈도를 계산하라.

- 뉴스 데이터에서 3개월간 등장하는 키워드의 빈도를 계산하고, 워드 클라우드 그래프로 출력하라.

- 트위터 데이터에서 날짜를 추출하기 위한 날짜 라이브러리(lubridate)를 추가한다.
- JSON의 날짜 필드를 이용하여 2013년 10월부터 2013년 12월까지의 데이터를 추출한다.
- 추출한 데이터에서 형태소 분석을 한다.
- 키워드별 건수와 빈도를 계산하고, 저장한다.
- 키워드에 대한 등장 변화를 워드 클라우드 그래프를 사용하여 시각화한다.

예제 2

특정 주간의 소셜 키워드 빈도를 계산하라.

- 특정 주간의 소셜 키워드에 대한 이슈 키워드 Top 10을 산출하고, 흐름을 비교하라.

- 분석하고자 하는 소셜 데이터(트위터, 뉴스)에서 특정 주간의 날짜 데이터만을 추출하여 저장한다.
- 추출된 소셜 데이터를 일자별로 그룹핑하여 키워드 빈도를 계산한다.
- 일자별 키워드 빈도를 정렬하여 Top 10을 추출하고, 일자별 키워드에 대한 빈도수를 꺾은선 그래프로 시각화하여 이슈 키워드를 비교 분석한다.

데이터 분석 콘텐츠 활용 매뉴얼

2014년 12월 인쇄

2015년 1월 발행

발 행 처 한국정보화진흥원 빅데이터전략센터

집 필 신신애, 김성현, 박재원, 김현태, 김지홍, 정다운,
이승하, 신은비

주 소 서울시 중구 청계천로 14

연 락 처 (02) 2131-0114

인 쇄 HNJ Printing

〈비매품〉

[데 이 터 분 석 콘 텐 츠]

활용 매뉴얼

NIA  한국정보화진흥원

(100-775) 서울시 종구 청계천로 14 한국정보화진흥원
TEL 02-2131-0114 FAX 02-2131-0109
www.nia.or.kr

