

NEWS CLOUD

: 카테고리 기반으로 한 네이버 기사를 분석하여
기사 속 핵심단어를 파악하자

- 중간 발표, 9주차 발표 -

발표날짜 : 2019.05.09

과목명 : 산학 캡스톤 디자인1

교수명 : 정현숙

팀 : 5조

팀원 : 전연지 양진이 김소원

발표자 : 양진이

1

News Cloud 개요

- News Cloud 소개
- News Cloud 동기
- News Cloud 기대효과
- 유사 제품 Some Trend 소개

2

News Cloud 개발 환경

- 개발 환경 및 설치 순서

3

News Cloud 구성도

- News Cloud 전체 구성도
- News Cloud 화면 구성도

4

News Cloud 기능

- News Cloud 기능 흐름도
- S/W(Web 개발) 주요 기능
- S/W(Data 분석) 주요 기능
- N2H4 설명

5

News Cloud 수행 내용

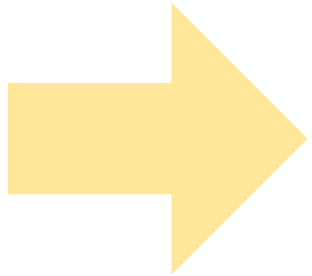
- News Cloud 수행 순서
- News Cloud 9주차 진행상황
- 팀원 소개 및 담당 업무
- 소감

6

News Cloud 참고 문헌

- 참고 사이트

NEWS CLOUD



실시간으로 특정 주제의 네이버 뉴스를 분석 및 수집하여
전체 뉴스를 읽지 않고도 시각화 기법인 Word Cloud 로
언론사에 따른 주요 키워드를 파악할 수 있는
R기반의 네이버 뉴스 수집 및 분석 웹 서비스

Word Cloud란,
문서의 키워드, 개념 등을 직관적으로 파악할 수 있도록
핵심단어를 시각적으로 돋보이게 하는 기법



“News Cloud” 개발 기대효과

○ 개인 맞춤 뉴스 제공

- 사용자들의 사용 로그를 수집하고, **로그를 기반으로 선호하는 카테고리**에 대한 **키워드를 파악**하게 된다.
파악한 결과를 바탕으로 개인별 성향에 따라 개인 맞춤 뉴스를 제공하게 된다.

○ 세부 카테고리별 뉴스 분석

- 사회, 생활/문화, 세계, IT/과학 **6가지 카테고리**와 **3개 이상의 추가적 세부 카테고리**로 나누어 뉴스 분석을 진행하게 된다.

○ 언론사별 뉴스 분석

- 선택한 언론사의 카테고리에 관한 Word Cloud 이미지를 통하여 **각 언론사 별 시사점과 뉴스의 키워드**를 쉽게 파악할 수 있다.

SomeTrend (썸트렌드)

- SomeTrend는 국내 텍스트 마이닝 전문업체인 다음소프트 (<http://www.daumsoft.com>)가 개발한 **소셜메트릭스(자연어처리 기술과 텍스트 마이닝 기술) 도구**를 사용하여 블로그와 트위터 문서를 분석하고 모니터링 결과를 제공하는 서비스
- 실시간 검색 키워드를 분석하여 SNS 트렌드 이슈에 대한 정보를 제공하는 사이트

<http://www.some.co.kr/>

News Cloud 개발 환경

+

개발환경 및 설치 순서

구분	상세내용
운영체제	Windows 10 64bit
개발도구	②Eclipse 2019-03 ③Tomcat 8 ⑥Rstudio 1.1.463 ④MySQL 8.0.25
사용언어	⑤R-3.5.3(⑦N2H4) ①Java(jdk-12) HTML JSP CSS

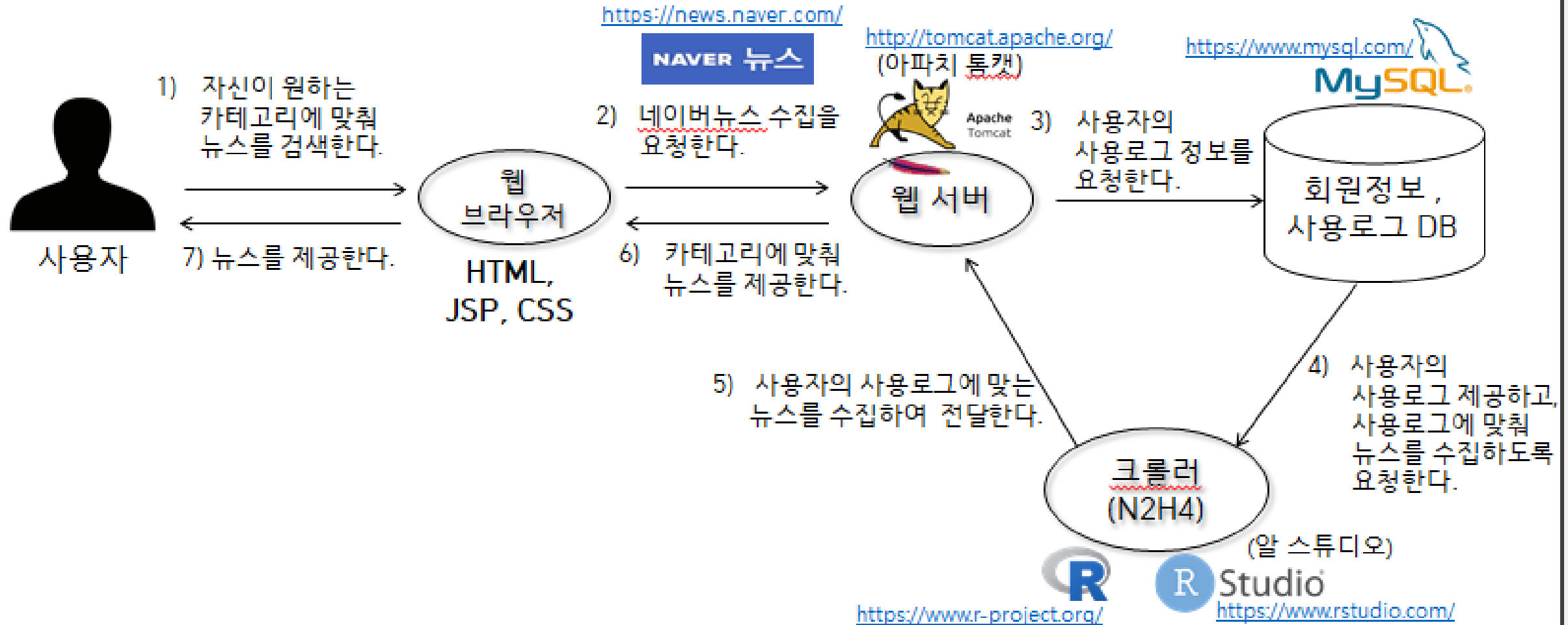
<설치 순서>

- ① Java를 사용하기 위해 **jdk를 설치**
- ② **Eclipse 설치**
- ③ 웹 프로그램은 구동시키기 위해선 네트워크 통신이 필요하게 됨 이것을 해결할 수 있는 **웹 서버 Tomcat을 설치**
- ④ 웹 서버안에서 데이터베이스를 사용하기 위해 **MySQL을 설치**
- ⑤ 데이터 분석을 위한 **R 설치**
- ⑥ R을 사용하여 데이터분석을 할 수도 있지만 R은 명령창 하나만 뜨기에 사용하는데에 조금 불편함 감이 있음. 반면, **Rstudio**는 네 분할(코딩 창, 명령창, 변수에 대한 정보 창, 파일/패키지/뷰어/도움말)을 실행시켜주기에 **사용하는데 더욱 편리하여 설치**
- ⑦ 뉴스 데이터를 수집하기 위해 R에서 지원하는 패키지 중 하나인 **N2H4 설치**

News Cloud 구성도

+

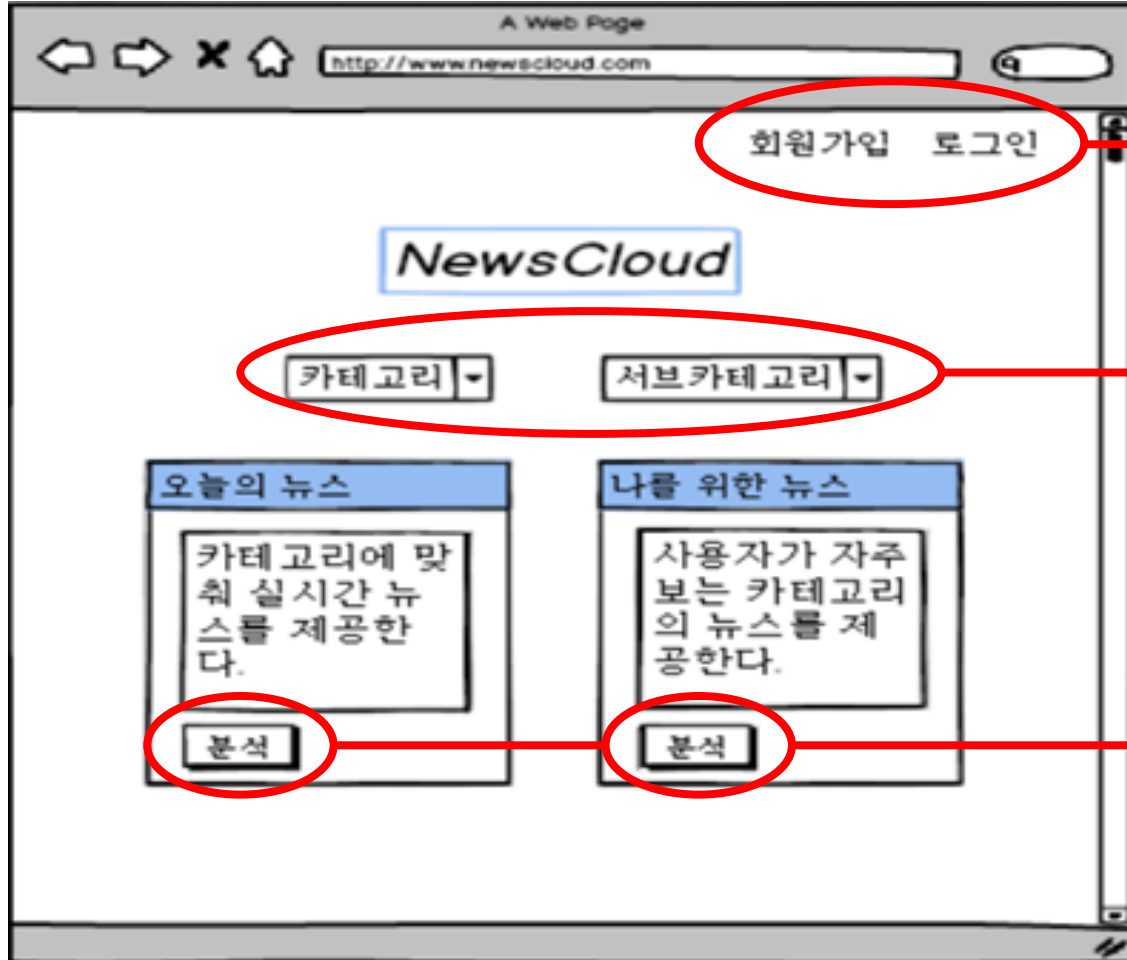
News Cloud 전체구성도



News Cloud 구성도

News Cloud 화면 구성도

<메인화면> <https://balsamiq.com/wireframes/>



맨 위 우측 화면에 회원가입, 로그인 기능을 넣어 사용자가 로그인 기능을 사용할 수 있도록 한다.

사용자는 자신이 원하는 카테고리 and 서브카테고리를 선택한 후

<오늘의 뉴스>기능이나, <나를 위한 뉴스>기능 분석을 선택할 수 있다.

News Cloud 구성도

News Cloud 화면 구성도

<오늘의 뉴스> <https://balsamiq.com/wireframes/>

The wireframe shows a web browser window titled 'A Web Page' with the URL 'http://www.newscloud.com'. The main content area is titled 'NewsCloud' and '오늘의 뉴스'. Below this is a navigation bar with '홈/오늘의 뉴스'. The main form has several sections: '카테고리' and '서브카테고리' dropdowns, '기간' with a date range '2000.00.00 ~ 2019.00.00', '언론사1' and '언론사2' dropdowns, a '분석' button, and a '워드 클라우드' section with two large gray boxes.

메인 화면에서 오늘의 뉴스 분석 기능을 선택하면,

사용자가 선택한 카테고리 와 서브카테고리에 관련된

기간, 언론사, 언론사2 목록이 생성된다.
언론사 2개를 선택해 분석 버튼을 누르게 되면

해당 기사들의 중요 핵심 키워드 빈도수에 따른
워드 클라우드 이미지를 볼 수 있게 된다.
이미지를 선택하게 되면
해당되는 실시간 뉴스를 볼 수 있다.

News Cloud 구성도

News Cloud 화면 구성도

<나를 위한 뉴스> <https://balsamiq.com/wireframes/>

The wireframe shows a web browser window titled 'A Web Page' with the URL 'http://www.newscloud.com'. The page content includes a header 'NewsCloud' and a main title '나를 위한 뉴스'. Below the title is a search bar with the placeholder text '홈/나를 위한 기사'. There are two dropdown menus for '카테고리' (Category) and '서브 카테고리' (Sub-category). A date range selector is set to '2000.00.00 ~ 2019.00.00'. There are two dropdown menus for '언론사1' (News Agency 1) and '언론사2' (News Agency 2). A button labeled '분석' (Analyze) is present. At the bottom, there is a section titled '워드 클라우드' (Word Cloud) with two large gray rectangular placeholders for word clouds.

메인 화면에서 나를 위한 뉴스 분석 기능을 선택하면,

→ 사용자가 가장 자주 본 카테고리로 지정된다.

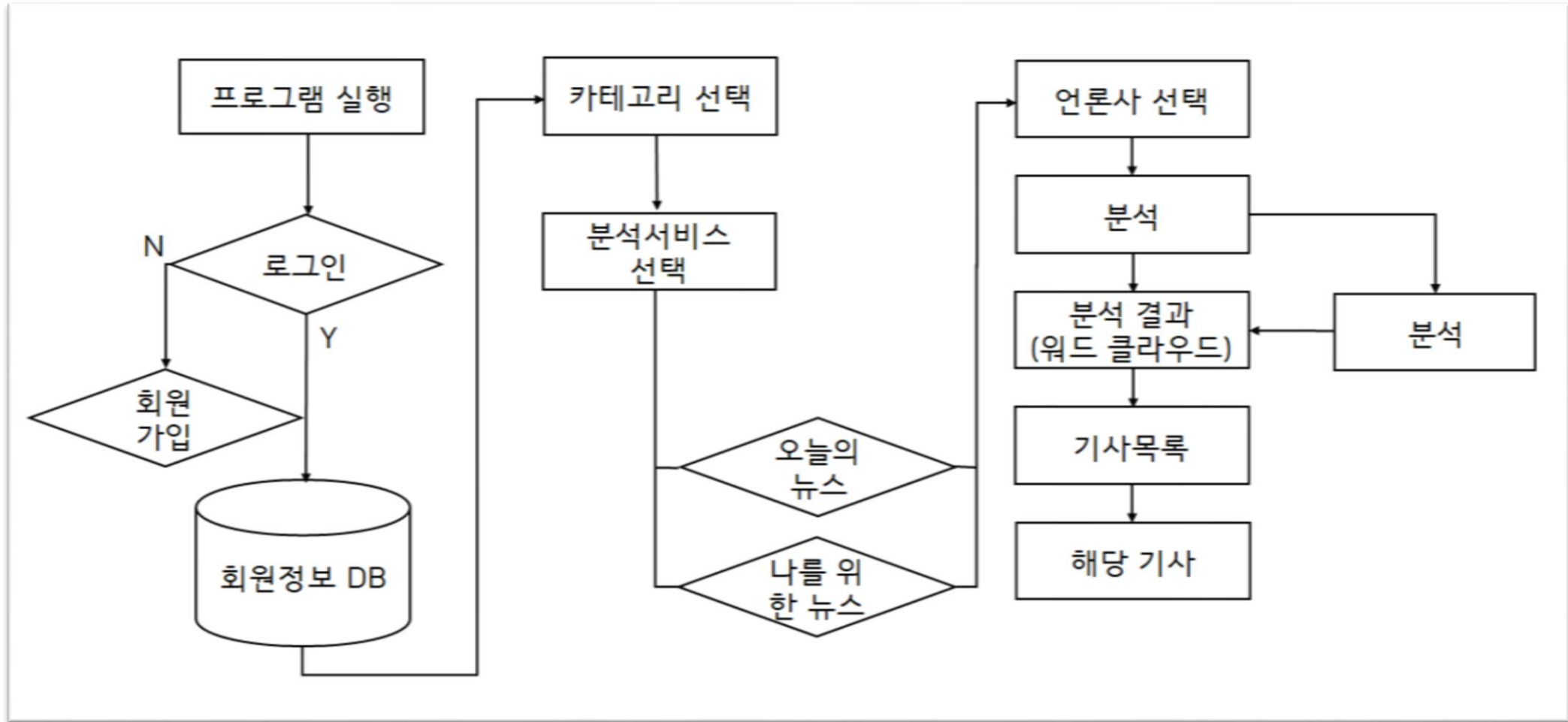
→ 사용자가 가장 자주 본 언론사 2개로 지정된다.

→ 해당 기사들의 빈도수에 따른 워드 클라우드 이미지를 볼 수 있게 된다. 이미지를 선택하게 되면 해당되는 뉴스를 볼 수 있다.

News Cloud 기능

+

News Cloud 기능 흐름도



News Cloud 기능

+

S/W(Web 개발) 주요 기능

기능	설명
로그인	데이터베이스에 저장된 회원 정보를 이용한 로그인 기능 제공
메인 화면	회원가입, 로그인, <오늘의 뉴스>, <나를 위한 뉴스> 분석 기능 제공
오늘의 뉴스	선택한 언론사와 카테고리의 당일 기사에 관한 워드 클라우드 이미지 제공
나를 위한 뉴스	사용자의 사용로그를 기준으로 분석하여 사용자가 선택한 기간, 언론사의 카테고리에 관한 워드 클라우드 이미지 제공
기사 제공	선택한 언론사의 카테고리에 관해 워드 클라우드 이미지를 선택하면 해당 기사 목록을 보여주고 링크를 걸어 기사를 볼 수 있도록 해줌

News Cloud 기능

+

S/W(Data 분석) 주요 기능

기능	설명
뉴스 수집기	<u>N2H4</u> 를 사용하여 네이버 뉴스에서 제공하는 6개의 카테고리(정치,경제,사회,생활/문화,세계,IT/과학)과 그에 따른 서브카테고리에 대한 뉴스를 수집
워드 클라우드(빈도분석-시각화)	뉴스에서 사용된 단어들의 빈도수를 파악하여 빈도수에 따라 크기와 색을 다르게 하여 쉽게 뉴스의 키워드를 파악할 수 있도록 이미지 제공

N2H4란,
네이버 뉴스 크롤링을 위한 도구

- R로 네이버 뉴스 데이터를 가져오는 방법
(R의 패키지 중 하나)

크롤링이란,
실시간으로 웹 데이터를
가져오는 방법

네이버 오픈 API 목록

  트윗  공유하기 6개

네이버 오픈API 목록 및 안내입니다.

API명	설명	호출제한
검색	네이버 블로그, 이미지, 뉴스 , 교과사전, 책, 카페, 지식iN 등 검색	25,000회/일
지도 (Web, Mobile)	네이버 지도 표시 및 주소 좌표 변환	20만/일
네이버 아이디로 로그인	외부 사이트에서 네이버 아이디로 로그인 기능 구현	없음
네이버 회원 프로필 조회	네이버 회원 이름, 닉네임, 이메일, 성별, 연령대, 프로필 조회	없음
Papago NMT 번역	인공신경망 기반 기계 번역 (영,중)	10,000글자/일
Papago SMT 번역	통계 기반 기계 번역 (영,일,중)	10,000글자/일
Clova Face Recognition	입력된 사진을 입력받아 얼굴윤곽/부위/표정/유명인 닮음도를 리턴	1,000건/일
데이터랩 (검색어트렌드)	통합검색어 트렌드 조회	1,000회/일

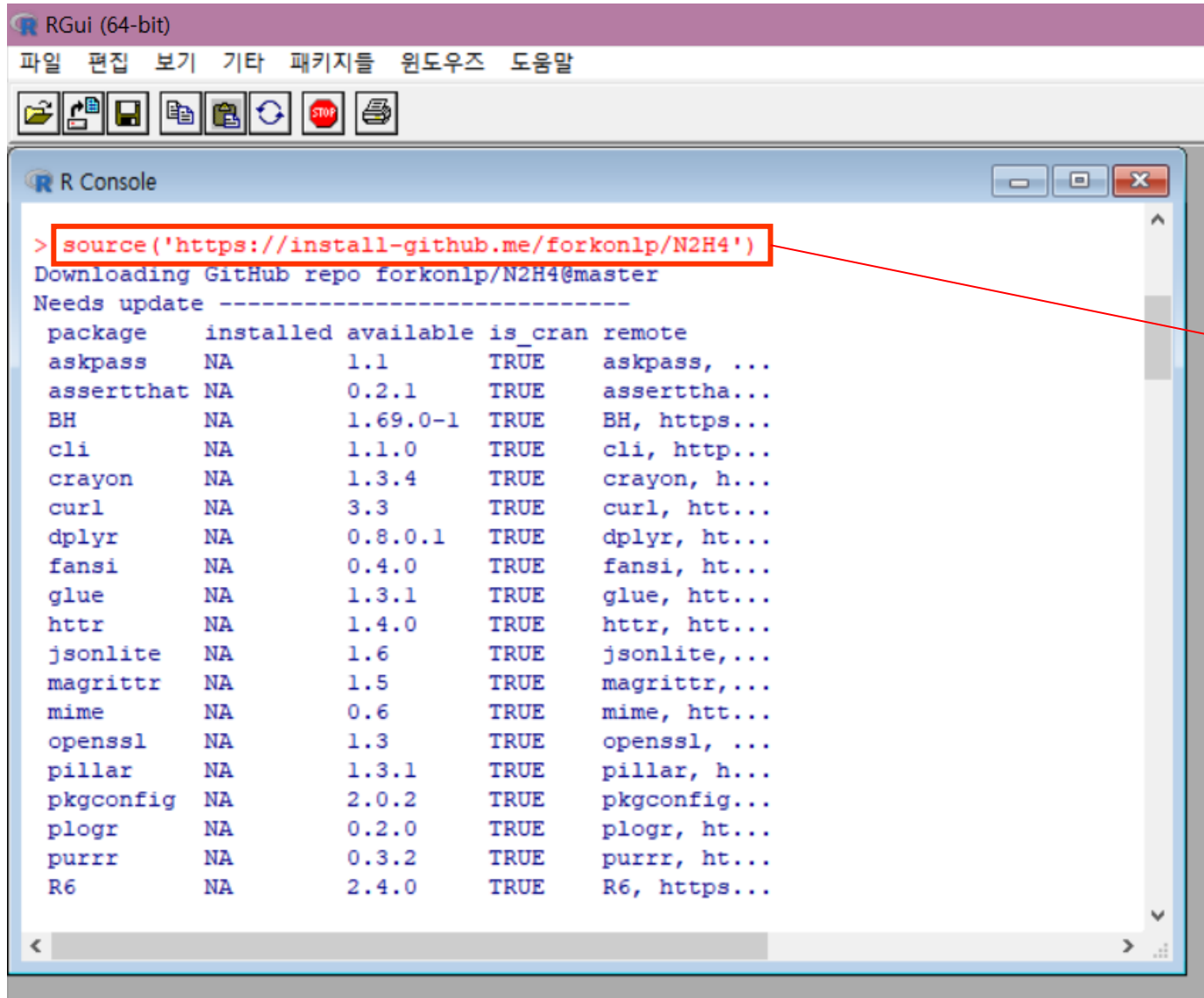
<https://developers.naver.com/products/intro/plan/>

네이버에서 제공하는 API 중 검색 API는 **검색**에 의해 뉴스를 제공하는 API이다.
하지만, News Cloud는 **카테고리 중심**으로 뉴스를 제공하기에 적절하지 않으며,
API 사용보다 크롤링 도구를 이용한 뉴스 데이터 수집이 더 편하고 빠르기 때문에
크롤링 도구를 이용한다.

News Cloud 기능

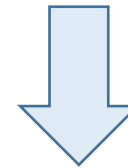
+

N2H4 설치 방법



```
> source('https://install-github.me/forkonlp/N2H4')
Downloading GitHub repo forkonlp/N2H4@master
Needs update -----
package      installed available is_cran remote
askpass      NA          1.1      TRUE  askpass, ...
assertthat   NA          0.2.1    TRUE  asserttha...
BH            NA          1.69.0-1  TRUE  BH, https...
cli           NA          1.1.0    TRUE  cli, http...
crayon        NA          1.3.4    TRUE  crayon, h...
curl          NA          3.3      TRUE  curl, htt...
dplyr         NA          0.8.0.1  TRUE  dplyr, ht...
fanssi        NA          0.4.0    TRUE  fanssi, ht...
glue          NA          1.3.1    TRUE  glue, htt...
httr          NA          1.4.0    TRUE  httr, htt...
jsonlite      NA          1.6      TRUE  jsonlite,...
magrittr      NA          1.5      TRUE  magrittr,...
mime          NA          0.6      TRUE  mime, htt...
openssl       NA          1.3      TRUE  openssl, ...
pillar        NA          1.3.1    TRUE  pillar, h...
pkgconfig     NA          2.0.2    TRUE  pkgconfig...
plogr         NA          0.2.0    TRUE  plogr, ht...
purrr         NA          0.3.2    TRUE  purrr, ht...
R6            NA          2.4.0    TRUE  R6, https...
```

`source('http://install-github.me/forkonlp/H2H4')`



`library(N2H4)`

News Cloud 기능



N2H4 기능 설명

<https://github.com/forkonlp/N2H4>

forkonlp / N2H4

Watch 16

★ Unstar 129

Fork 63

Code

Issues 5

Pull requests 0

Projects 1

Wiki

Insights

Branch: master N2H4 / R /

Create new file

Upload files

Find file

History



mrchypark fix datetime parse

Latest commit 23ef3c9 3 days ago

..

.searchNews.R	for poll	2 years ago
N2H4-package.R	update tibble	2 months ago
getComment.R	update tibble	2 months ago
getContent.R	fix datetime parse	3 days ago
getMainCategory.R	update tibble	2 months ago
getMaxPageNum.R	refactor code remove dep pack	a year ago
getNewsTrend.R	update tibble	2 months ago
getQueryUrl.R	fix getQueryUrl	2 months ago
getSubCategory.R	update tibble	2 months ago
getUrlListByCategory.R	update tibble	2 months ago
getUrlListByQuery.R	update tibble	2 months ago
getVideoUrl.R	refactor code remove dep pack	a year ago
setUrls.R	refactor code remove dep pack	a year ago

getComment : 네이버 뉴스 페이지의 관련 댓글 정보를 가져오는 기능

getContent : 네이버 뉴스 페이지 내에 url, 기사입력시간, 수정시간, 신문사, 제목, 내용 정보를 가져오는 기능

getMainCategory : 네이버 뉴스의 메인 카테고리를 가져오는 기능(정치, 경제, 사회 등등)

getMaxPageNum : 메인 카테고리의, 서브 카테고리 페이지에서 마지막 페이지 수를 가져오는 기능

getNewsTrend : 네이버 뉴스에서 검색시 검색 결과에 나오는 총 검색량을 가져오는 기능

getSubCategory : 네이버 뉴스의 서브 카테고리를 가져오는 기능

getUrlListByCategory : 뉴스 페이지의 제목과 url을 가져오는 기능

getUrlListQuery : 네이버 뉴스가 있는 기사들의 url을 가져오는 기능

N2H4 기능 예시

getContent 명령어 사용 : 네이버 뉴스 페이지 내에 url, 기사입력시간, 수정시간, 언론사, 제목, 뉴스 본문을 가져오는 기능

```
> library(N2H4)
> url<-"http://news.naver.com/main/read.nhn?mode=LS2D&mid=shm&sid1=105&sid2=731&oid=031&aid=0000393444"
> getContent(url)
# A tibble: 1 x 6
  url          datetime      edittime      press title      body
  <chr>        <dtm>        <dtm>        <chr>   <chr>   <chr>
1 http://ne~ 2016-11-17 14:39:00 2016-11-17 14:39:00 아이뉴스2~ 네이버-카카~ "[성상훈기자]~
> |
```

R에서의 결과값과
실제 뉴스의 정보가
같이 나온다는 것을
확인할 수 있음.

아이뉴스24

네이버-카카오, 新 성장동력 키워드는 '콘텐츠'

기사입력 2016.11.17. 오후 2:39 기사원문 스크랩 본문듣기 · 설정

2 댓글

요약봇 가 [공유]

<아이뉴스24>

[성상훈기자] 네이버와 카카오가 '변신'을 선언했다. 네이버는 포털 서비스를 넘어 '콘텐츠 플랫폼'으로 옷을 갈아 입고 있는 가운데 카카오도 카카오 페이지를 선봉에 내세

<https://news.naver.com/main/read.nhn?mode=LS2D&mid=shm&sid1=105&sid2=731&oid=031&aid=0000393444>

News Cloud 수행 내용

+

News Cloud 수행 순서

빅데이터
사용 도구 설치

R에서 패키지 N2H4
(네이버 뉴스 수집 크롤러)
설치

빅데이터
수집

날짜, 메인 카테고리, 서브카테고리 별로
해당하는 기사 및 세부사항 수집
(url, 기사 입력 시간, 기사 수정 시간,
언론사, 기사 제목,
기사 본문)

빅데이터
저장

수집한 기사들을
언론사 별로 모아 csv로 저장

기사 본문을
자연어 처리 기반 텍스트 마이닝 수행
(명사 추출, 본문 속 불필요한 단어 제거
(예: '무단전재 및 재배포 금지', '기자' 등),
사용 빈도 확인 등)

빅데이터
처리 및
분석

워드 클라우드 구현

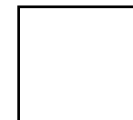
빅데이터
시각화



: 수행 완료



: 다음주
수행 예정



: 수행 예정

News Cloud 수행 내용

+

News Cloud 9주차 진행상황

<예시>

사용자가 5월 8일 "IT/과학"(메인카테고리)의
"모바일"(서브카테고리)을 선택 했을 경우

```
setwd("C:/RPrj") #경로설정
```

```
library(N2H4) #패키지 N2H4 사용
```

```
#변수 cate에 카테고리 메인카테고리 목록 저장
```

```
cate<-getMainCategory()
```

```
#변수 tcate에 선택한 메인카테고리의 sid값을 저장
```

```
tcate<-cate$sid1[6]
```

```
#변수 subCate에 선택한 메인카테고리의 서브카테고리의 목록 저장
```

```
subCate<-cbind(sid1=tcate,getSubCategory(sid1=tcate))
```

```
#변수 tscate에 선택한 서브카테고리의 sid값을 저장
```

```
tscate<-subCate$sid2[1]
```

```
strDate<-"20190508" #날짜설정
```

NAVER 뉴스 | TV연예 | 스포츠 | 뉴스스탠드 | 날씨

뉴스홈 속보 정치 경제 사회 생활/문화 세계 IT/과학
① ② ③ ④ ⑤ ⑥

↑ 메인 카테고리

IT/과학

← 서브 카테고리

모바일

①

인터넷/SNS

②

통신/뉴미디어

③

IT 일반

④

보안/해킹

⑤

컴퓨터

⑥

게임/리뷰

⑦

과학 일반

⑧

<https://news.naver.com/main/list.nhn?mode=LS2D&mid=shm&sid1=105&sid2=731>

News Cloud 수행 내용

+

News Cloud 9주차 진행상황

<예시>

사용자가 5월 8일 "IT/과학"(메인카테고리)의
"모바일"(서브카테고리)을 선택 했을 경우

```
setwd("C:/RPrj") #경로설정  
library(N2H4) #패키지 N2H4 사용
```

```
#변수 cate에 카테고리 메인카테고리 목록 저장  
cate<-getMainCategory()
```

```
#변수 tcate에 선택한 메인카테고리의 sid값을 저장  
tcate<-cate$sid1[6]
```

```
#변수 subCate에 선택한 메인카테고리의 서브카테고리의 목록 저장  
subCate<-cbind(sid1=tcate,getSubCategory(sid1=tcate))
```

```
#변수 tscate에 선택한 서브카테고리의 sid값을 저장  
tscate<-subCate$sid2[1]
```

```
strDate<-"20190508" #날짜설정
```

<sid >

"segment identifier"의 약자로,
각 카테고리를 식별할 수 있게 해주는 고유 값

-URL-

[https://news.naver.com/main/list.nhn?mode=LS2D
&mid=shm&sid1=105&sid2=731](https://news.naver.com/main/list.nhn?mode=LS2D&mid=shm&sid1=105&sid2=731)

```
> cate  
# A tibble: 6 x 2  
  cate_name sid1  
  <chr>      <chr>  
1 정치      100  
2 경제      101  
3 사회      102  
4 생활/문화 103  
5 세계      104  
6 IT/과학   105
```

```
> subCate  
  sid1 sub_cate_name sid2  
1  105      모바일   731  
2  105 인터넷/SNS   226  
3  105 통신/뉴미디어 227  
4  105      IT 일반  230  
5  105      보안/해킹 732  
6  105      컴퓨터  283  
7  105      게임/리뷰 229  
8  105      과학 일반 228
```

News Cloud 개요

+

News Cloud 9주차 진행상황

For문을 활용하여 뉴스본문 csv파일로 저장하기

뉴스 리스트 페이지의 url을 sid1(메인카테고리), sid2(서브카테고리), date(날짜)로 생성합니다.

pageUrl<-

```
paste0("http://news.naver.com/main/list.nhn?sid2=",sid2,"&sid1=",sid1,"&mid=shm&mode=LS2D&date="
,strDate)
```

리스트 페이지의 마지막 페이지수를 가져옵니다.

max<-getMaxPageNum(pageUrl)

for (pageNum in 1:max){

페이지번호를 포함한 뉴스 리스트 페이지의 url을 생성합니다.

pageUrl<-

```
paste0("http://news.naver.com/main/list.nhn?sid2=",sid2,"&sid1=",sid1,"&mid=shm&mode=LS2D&date="
,strDate,"&page=",pageNum)
```

뉴스 리스트 페이지 내의 개별 네이버 뉴스 url을 가져옵니다.

newsList<-getUrlListByCategory(pageUrl)

News Cloud 개요

+

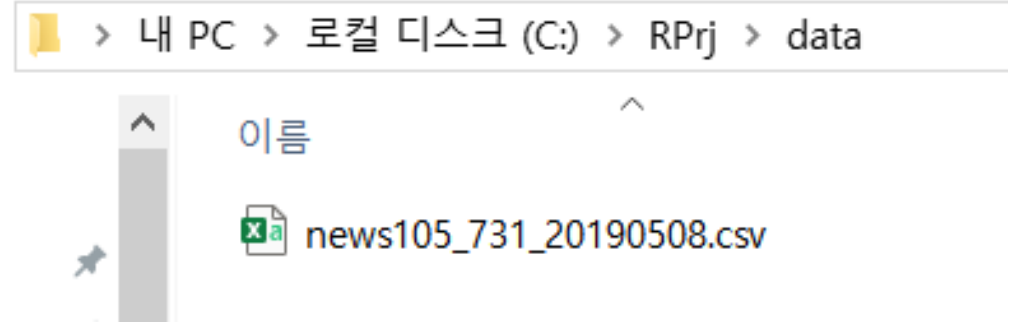
News Cloud 9주차 진행상황

For문을 활용하여 뉴스본문 csv파일로 저장하기

```
# url들의 정보를 가져옵니다.
for (newslink in newsList$links){

  # 불러오기에 성공할 때까지 반복합니다.
  tryi<-0
  tem<-try(getContent(newslink), silent = TRUE)
  # 성공할 때까지 반복하면 못나오는 문제가 있어서 5회로 제한합니다.
  while(tryi<=5&&class(tem)=="try-error"){
    tem<-try(getContent(newslink), silent = TRUE)
    tryi<-tryi+1
  }

  # 가져온 뉴스들을 press(언론사), body(기사본문)만 csv 형태로 저장한다.
  newsData1<-newsData[c("press","body")]
  # 언론사 종류가 같은 것들끼리 정렬하여 모은다.
  write.csv(newsData1[order(newsData$press),],
    file=paste0"tcate","_("./data/news",tcate,"_",tscate,"_",strDate,".csv"),row.names = F)
```



-> 지정된 경로에 csv파일이 저장된다.

News Cloud 수행 내용

+

팀원소개 및 담당 업무

구분	전연지	양진이	김소원
학번	20164244	20164316	20164226
연락처	010-3746-7697	010-3763-9216	010-7167-5324
이메일	jeonyeonji1028@gmail.com	zcxsad@naver.com	kksw5324@naver.com
깃허브	https://github.com/szduswldz	https://github.com/zcxsad	https://github.com/ksonwon
역할	팀장	팀원	팀원
담당업무	프로젝트 진행상황 관리, 데이터 분석/통합 개발 담당	데이터 분석 개발 담당	웹 개발 담당

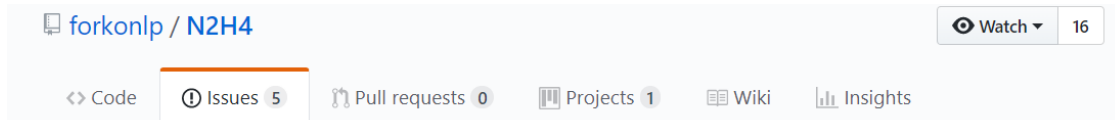
News Cloud 수행 내용

+

News Cloud 9주차 소감

N2H4의 기능인 'getContent' 명령어를 실행하던 중, 원하는 결과 값이 나오지 않아
이 점에서 힘이 들었습니다.

하지만 많은 시도에도 불구하고, 변화가 없어
아래 사진과 같이 N2H4 개발자의 깃허브를 통해 질문해보았습니다.
감사하게도 빠른 답변을 통해 N2H4 내의 문제였다는 점을 알았고,
개발자의 디버깅을 통해 원하는 결과를 얻을 수 있어 뿌듯하였습니다.



<https://github.com/forkonlp/N2H4/issues/85>

안녕하세요. 잘 되지 않아 여쭙봅니다. #85

Closed szduswldz opened this issue 15 days ago · 1 comment



szduswldz commented 15 days ago

안녕하세요. N2H4를 이용한 프로젝트를 만들고 있는 학생입니다.
다름이 아니라, 기능 중 getContent 가 잘 되지 않아 글을 쓰게 되었습니다.

getContent

```
getContent(turl = url, col=c("url","datetime","press","title","content"))
```

위와 같이 하였지만 url을 제외한 나머지는 모두 page is not news section 이라고 나옵니다.
다른 url로도 시도해보았지만 같은 결과가 나오는데
어떻게 해야할지 궁금합니다.

답 부탁드립니다. 감사합니다.



mrchypark commented 14 days ago

Member + 😊 ...

안녕하세요! 알려주신 덕분에 문제를 파악하고 수정하였습니다.
0.5.1 버전으로 설치하시면 기대하시는 대로 동작할 것입니다. 감사합니다.

<10주차 개발 계획>

기사들을 수집한 csv파일을 이용하여,
워드 클라우드 구현을 위해
기사 본문 속 불필요한 단어 (예 : '무단전재 및 재배포 금지', '기자' 등)를
제거하는 함수를 구현할 예정

News Cloud 수행 내용

+

News Cloud 참고 사이트

<http://www.some.co.kr/>

<https://news.naver.com/>

<http://tomcat.apache.org/>

<https://www.mysql.com/>

<https://www.r-project.org/>

<https://www.rstudio.com/>

<https://balsamiq.com/wireframes/>

<https://developers.naver.com/products/intro/plan/>

<https://github.com/forkonlp/N2H4>

<https://news.naver.com/main/list.nhn?mode=LS2D&mid=shm&sid1=105&sid2=731>

<https://github.com/forkonlp/N2H4/issues/85>

+

Q&A

+

감사합니다