주제

NEWS CLOUD

: 네이버에서 제공하는 기사들 속 다양한 분야의 이슈를 쉽게 알아내자

발표날짜 : 2019.05.23
과목명 : 산학 캡스톤 디자인1
교수명 : 정현숙
팀 : 5조
팀원 : 전연지 양진이 김소원
발표자 : 김소원

NEWS CLOUD는 무엇인가?

네이버에서 제공하는 기사들을 크롤링하여 분석한 후,
각 카테고리 별 화제가 되고 있는 뉴스 키워드를
시각화를 통해 쉽게 확인할 수 있는 웹 서비스

왜 만들게 되었는가?

대다수 젊은층들의 이슈를 접하는 경로
-> 유명 포털사이트의 메인 화면
SNS(페이스북, 트위터 등)
커뮤니티
자극적인 이슈를 메인으로 게시, 접근성 ↑

반면, 뉴스 사이트는 접근성이 상대적으로 떨어진다

왜 만들게 되었는가?

따라서 특정 분야에서 너무나 큰 이슈가 생기게 되면,
타 분야에서도 충분히 관심을 가질 수 있는 이슈가 나왔음에도 불구하고,
뉴스사이트를 들어가서 확인하지 않아 모르고 넘어가는 경우가 많았다.

정치와 연예인들 열애설... 2015.05.16.
열애설을 정치로 연관시켜서 사건을 파헤치는걸 문제삼는 게 아니야. 열애설이 잠잠해지면 그 사건에 대한 비난여론도 같이 잠잠해지는게 문제인 것이지. 이건 옛날부터...
영화같은 삶!! blog.naver.com/leedg1234/220360970624 | 블로그 내 검색

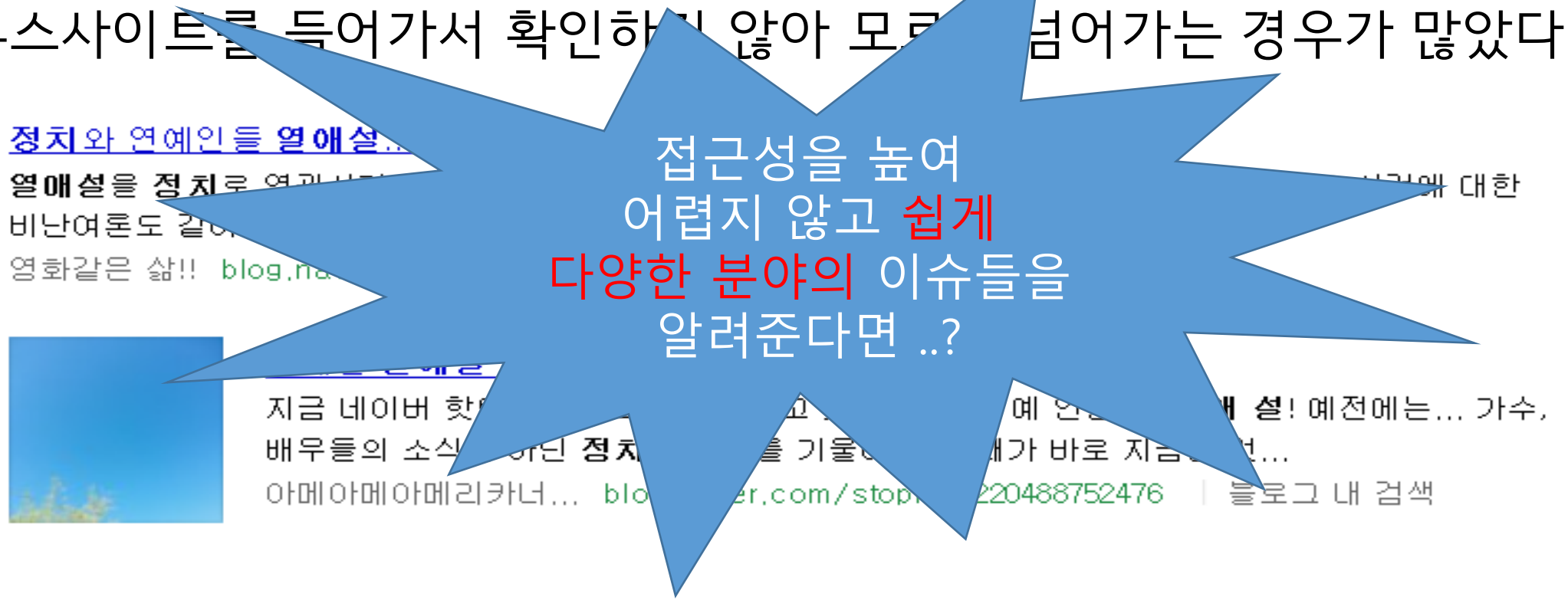
연예인 열애설 정치와도 연관이 있다? 2015.09.22.
지금 네이버 핫토픽 키워드에 떠오르고 있는 몇몇 연 예 인들의 열 애 설! 예전에는... 가수, 배우들의 소식이 아닌 정치에도 귀를 기울여야 할 때가 바로 지금인 것...
아메아메아메리카너... blog.naver.com/stopitfo/220488752476 | 블로그 내 검색

News Cloud 개요

News Cloud 소개

왜 만들게 되었는가?

따라서 특정 분야에서 너무나 큰 이슈가 생기게 되면,
타 분야에서도 충분히 관심을 가질 수 있는 이슈가 나왔음에도 불구하고,
뉴스사이트를 들어가서 확인하지 않아 모르고 넘어가는 경우가 많았다.



어떻게 쉽게 이슈를 알려줄 것인가?

'실시간 검색어' = 실시간으로 가장 많이 검색한 단어
>> 현재 이슈의 척도
이처럼, **실시간 기사 중 가장 많이 언급한 단어로 이슈의 키워드를 알 수 있음.**

'워드 클라우드'를 통해 시각화
-> 사용자가 관심이 생겨 기사를 보고 싶을 경우,
해당 기사로 넘어가는 기능 제공
키워드의 연관 분석도 진행 할 예정

왜 기사 수집을 네이버 뉴스로 선택하였는가?

1. 각종 언론사들의 기사를 집합한 대표적인 사이트는 네이버, 다음, 구글 등이 존재하지만, 이들은 사이트만 다를 뿐, **동일한 언론사이면 기사 내용도 동일**
   >> 따라서 하나의 뉴스 사이트로 가능

News Cloud 개요

News Cloud 소개

왜 기사 수집을 네이버 뉴스로 선택하였는가?

1. 각종 언론사들의 기사를 집합한 대표적인 사이트는 네이버, 다음, 구글 등이 존재하지만, 이들은 사이트만 다를 뿐, 동일한 언론사이면 기사 내용도 동일
   >> 따라서 하나의 뉴스 사이트로 가능

2. 세 사이트들 모두 카테고리로 기사를 나누지만
   네이버가 가장 세부적으로 카테고리를 나누어져 있음.

왜 기사 수집을 네이버 뉴스로 선택하였는가?

1. 각종 언론사들의 기사를 집합한 대표적인 사이트는 네이버, 다음, 구글 등이 존재하지만, 이들은 사이트만 다를 뿐, 동일한 언론사이면 기사 내용도 동일
   >> 따라서 하나의 뉴스 사이트로 가능

2. 세 사이트들 모두 카테고리로 기사를 나누지만
   네이버가 가장 세부적으로 카테고리를 나누어져 있음.

3. 네이버 기사를 쉽게 크롤링 할 수 있는 도구(N2H4)가 있음.

-> 네이버 뉴스로 선택

네이버의 모든 기사들을 수집하는 것인가?

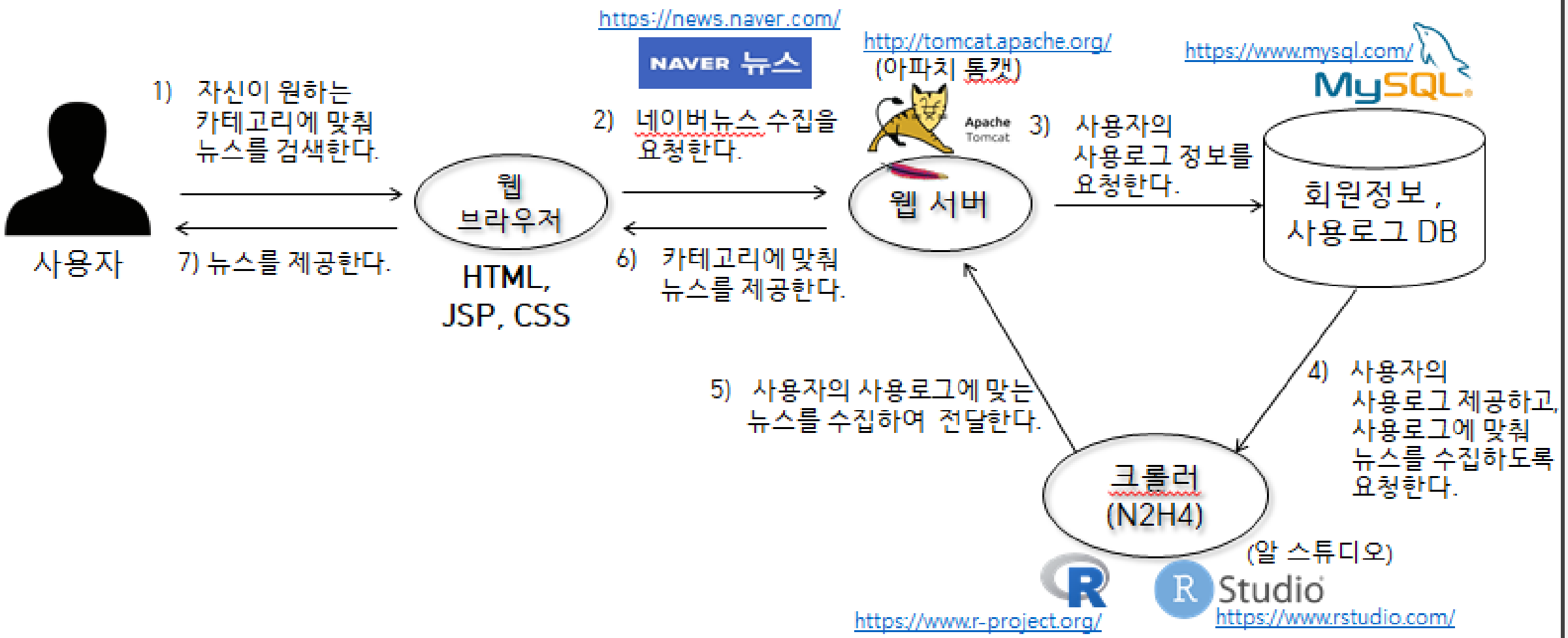
아니다.

사용자는 카테고리와 기간(최대 7일, 최소 1일)을 정할 수 있다.

>> 따라서 전체 기사가 아닌 부분만 빠르게 수집한다

News Cloud 구성도

News Cloud 전체구성도

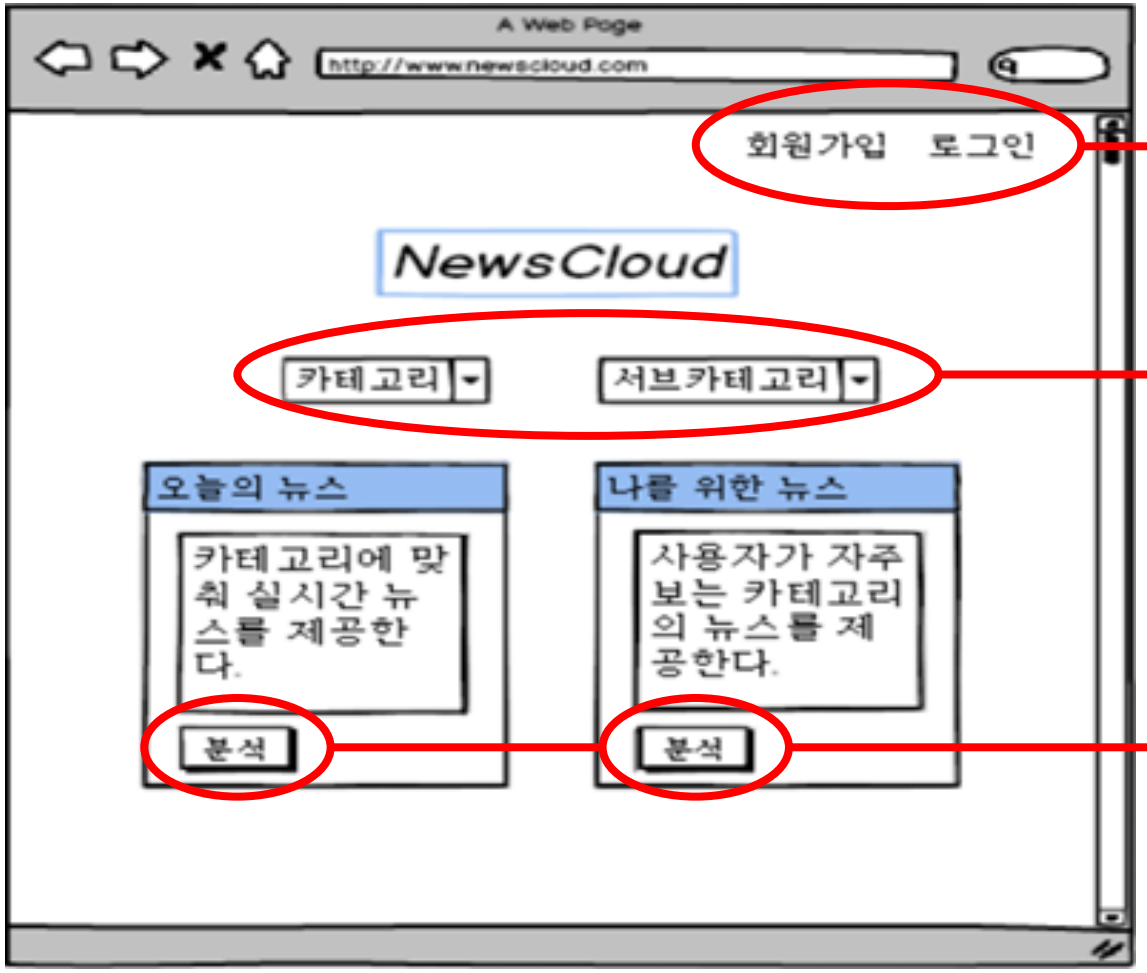


News Cloud 구성도

News Cloud 화면 구성도

<메인화면>

https://balsamiq.com/wireframes/



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

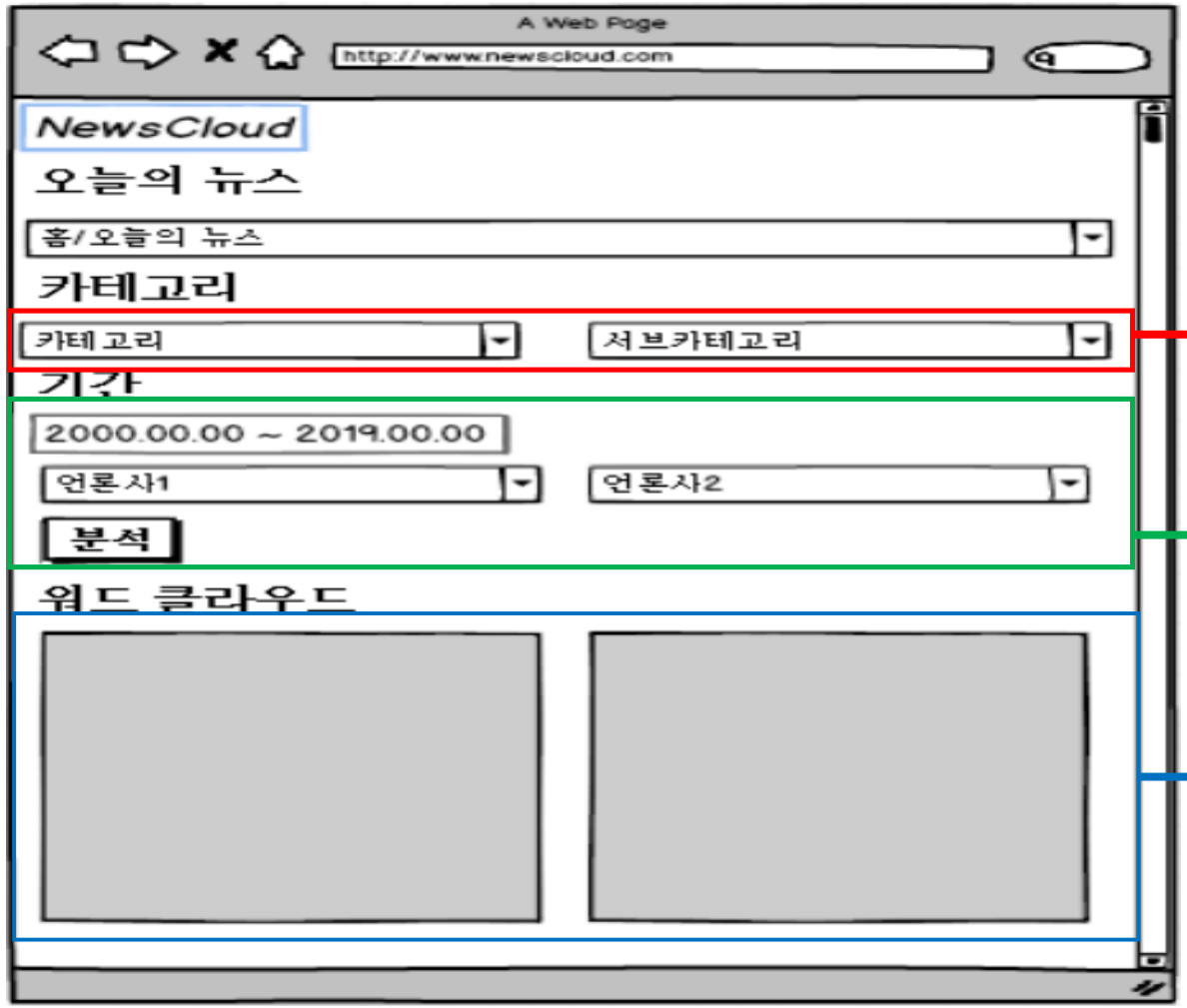
사용자는 자신이 원하는 카테고리와 서브카테고리를 선택한 후

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

News Cloud 구성도

News Cloud 화면 구성도

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



메인 화면에서 오늘의 뉴스 분석 기능을 선택하면,

사용자가 선택한 카테고리와 서브카테고리에 관련된

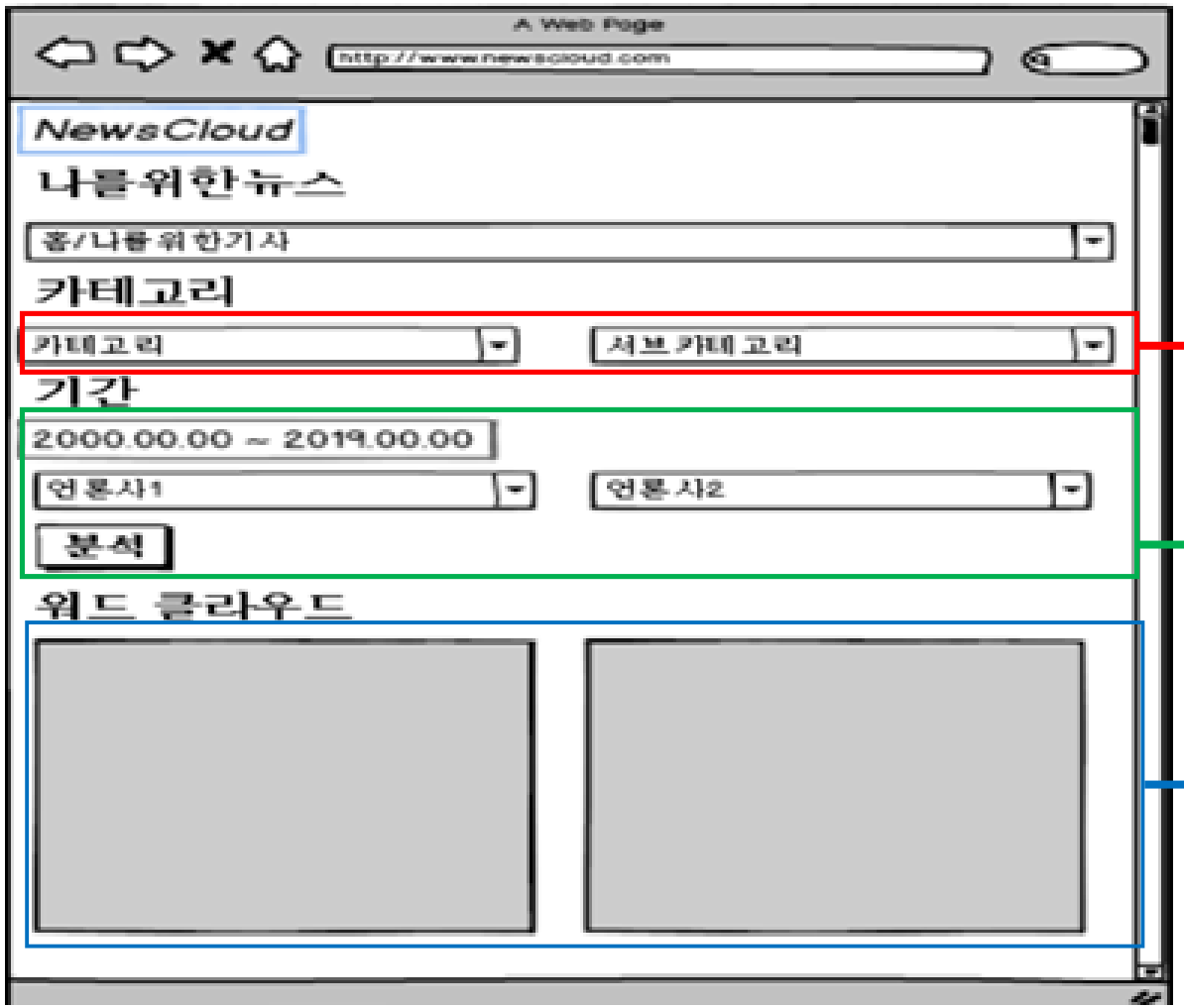
기간, 언론사, 언론사2 목록이 생성된다.
언론사 2개를 선택해 분석 버튼을 누르게 되면

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/



메인 화면에서 나를 위한 뉴스 분석 기능을 선택하면,

사용자가 선택한 카테고리와 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 개인 맞춤 뉴스를 볼 수 있다.

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 기간, 언론사의 카테고리에 관한 워드 클라우드 이미지 제공
나를 위한 뉴스	사용자의 사용로그를 기준으로 분석하여 사용자가 선택한 기간, 언론사의 카테고리에 관한 워드 클라우드 이미지 제공
연관 분석	사용자가 선택한 카테고리에 관한 키워드들의 연관분석 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	N2H4를 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
워드 네트워크(연관 분석-시각화)	뉴스 간에 등장하는 단어의 동시성을 파악하여 특정 단어가 들어간 뉴스에서 어떠한 단어가 동시에 출현하며, 연관성이 있는지 네트워크 분석결과 제공

N2H4란,
네이버 뉴스 크롤링을 위한 도구

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

크롤링이란,
실시간으로 웹 데이터를
가져오는 방법

News Cloud 기능

N2H4 설명

네이버 오픈 API 목록

네이버 오픈API 목록 및 안내입니다.

API명	설명	호출제한
검색	네이버 블로그, 이미지, 뉴스, 백과사전, 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 뉴스 API는 검색에 의해 뉴스를 제공하는 API가 존재함.
News Cloud는 카테고리 중심의 뉴스제공이며,
API사용보다 크롤링 도구를 이용한 뉴스 데이터 수집이 더 편하고 빠르기 때문에
크롤링 도구를 이용함.

News Cloud 기능

N2H4 기능 설명

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

News Cloud 기능

N2H4 기능 설명

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

https://github.com/forkonlp/N2H4

News Cloud 기능

N2H4 기능 예시

getContent 명령어 사용 : 네이버 뉴스 페이지 내에 url, 기사입력시간, 수정시간, 신문사, 제목, 내용 정보를 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>        <dttm>              <dttm>              <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 같게 나온다는 것을 확인할 수 있음.

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 사용할 수 있게 하는 환경인 **jdk를 설치**

② java프로그램을 만들기 위해 **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, **Rstudio는** 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 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

빅데이터 처리 및 분석

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 시각화

워드 클라우드 구현

: 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 서브카에고리의 목록 저장
subCate<-cbind(sid1=tcate,getSubCategory(sid1=tcate))

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

strDate<-"20190508"  #날짜설정
```

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

News Cloud 수행 내용

News Cloud 9주차

<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 서브카에고리의 목록 저
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**

```
> 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

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){

      # 페이지넘버를 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

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 body(기사본문)만 csv 형태로 저장한다.
write.csv(newsData$body, file=paste0",tcate,"_("./data/news",tscate,"_",strDate,".csv"),row.names = F)
```

-10주차-

```
write.csv(newsData, file=paste0("./data/news",tcate,"_",tscate,"_",strDate,".csv"),row.names = F)
# csv파일로 뉴스 기사 저장
cleaning_text<-function(dat){
    char<-gsub("[[:cntrl:]]","",dat)
    char<-gsub("[A-z]","",char)
    char<-gsub("₩₩▶","",char)
    char<-gsub("무단전재 및 재배포 금지","",char)
    char<-gsub("금지","",char)
    char<-gsub("재배포","",char)
    char<-gsub("년","",char)
    char<-gsub("무단","",char)
    char<-gsub("전재","",char)
    char<-gsub("바로가기","",char)
    char<-gsub("기자","",char)
}
# gsub: 필요없는 글 삭제
```

-10주차-

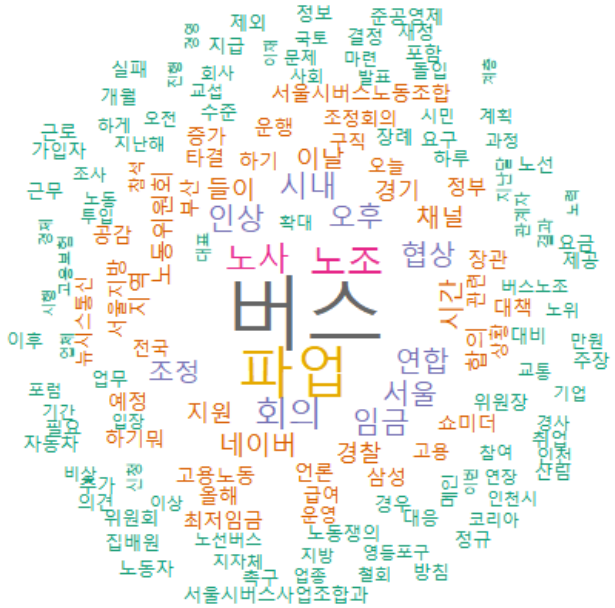
```
press_body<-cleaning_text(newsData$body)%>%sapply(extractNoun,USE.NAMES=FALSE)
# extractNoun: 문장을 단어로 만든 후 명사 추출
press_body1<-Filter(function(x){nchar(x)>=2},press_body1)
# 한 글자는 제외
write(unlist(press_body1),"word.txt")
wordcount<-table(press_body2)
# 리스트구조를 백터로 변환하여 txt파일로 저장 후 변수에 저장
pal<-brewer.pal(8,"Dark2")
wordcloud(words=names(wordcount), freq=wordcount, min.freq=2, max.words = 150,random.order=F,
rot.per=0.10,scale=c(4,.5), colors=pal)

#words:출력할 단어          #freq: 언급될 빈도수
#min.freq: 최소 빈도수     #max.words: 최대 빈도수
#random.order: 출력되는 순서 지정          #random.color: 글자 색상을 임의로 지정
#rot.per: 단어 배치를 90도 각도
```

News Cloud 수행 내용

News Cloud 10주차

-10주차-

사회 – 사건사고	사회 – 노동

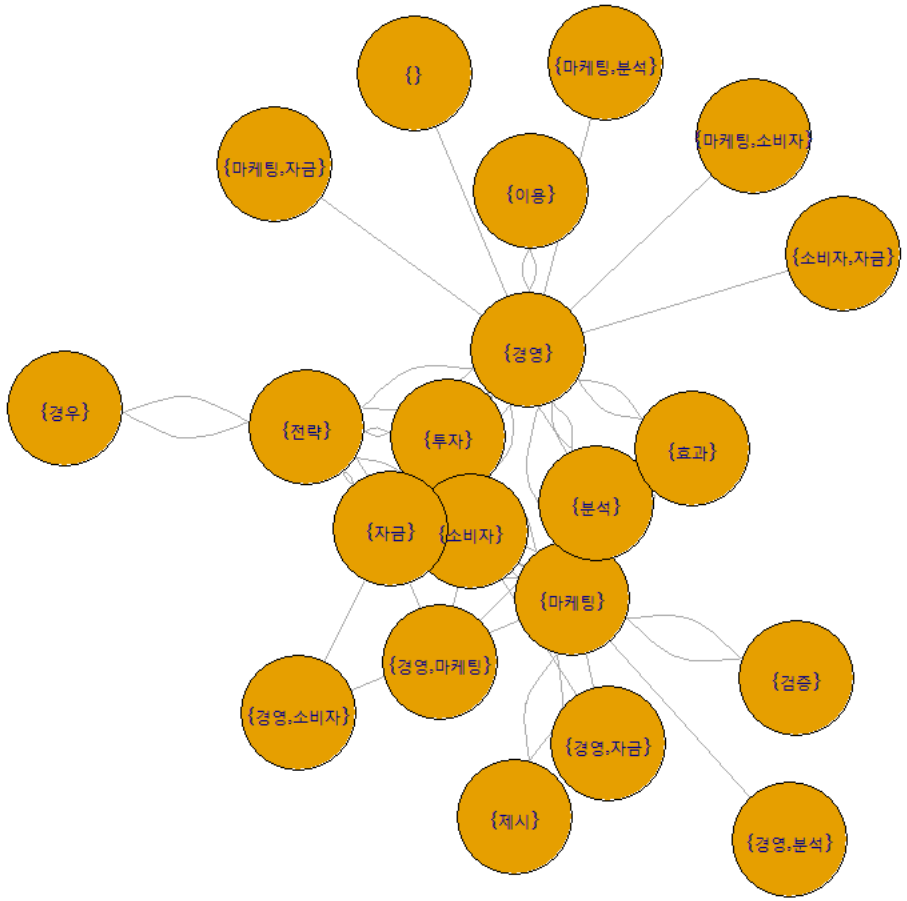


11주차 : 연관분석 진행중

```
# 단어 추출
tran<-Map(extractNoun, rules)
tran<-unique(tran)
tran<-sapply(tran, unique)
# 두 글자 이상인 문자(데이터) 벡터, 필터 후 tran에 저장
tran<-sapply(tran, function(x) {Filter(function(y){nchar(y) <=4 && nchar(y) >1 && is.hangul(y)},x)})
tran <- Filter(function(x) {length(x) >= 2}, tran)
# apriori 함수를 이용하여 데이터를 연관분석
ares <- apriori(wordtran, parameter = list(supp=0.1, conf=0.2))
# inspect 함수를 이용하여 연관 분석 결과 확인
inspect(ares)
# strsplit 함수를 이용하여 문자를 나눔
rules <- sapply(rules, strsplit, " ", USE.NAMES=F)
```

News Cloud 수행 내용

News Cloud 11주차



실행 결과

>> 연관분석은 나오게 되었지만 단어 사이 간에 연관성이 명확하게 확인되지 않았음.
단어 추출도 잘 되지 않았음.

News Cloud 수행 내용

News Cloud 12주차 개발예정

-12주차-

⊙단어 간에 연관성이 더 명확하게 보이게 보완

⊙노드와 노드 사이에 연관이 더 깊을수록
선의 굵기가 더 굵어지도록 각 노드에 가중치 부여

Q&A

감사합니다