



웹 데이터 크롤링

-Web Data Crawling-

Contents

I. 웹 크롤링 이해

II. R을 이용한 네이버 영화 댓글 크롤링 실습

III. Selenium



1. 웹 크롤링 이해

텍스트 크롤링 (Crawling)

- 텍스트 분석의 대상으로부터 분석에 사용할 수 있는 형태로 텍스트를 가져오는 것
- 크롤링의 대상: PDF파일, HWP파일, 웹 사이트, Social Media, 신문기사, 블로그 등
- 다양한 크롤링 도구(Tools)가 존재
- 무료 통계 패키지 R에도 웹 크롤링을 위한 패키지가 존재 (현재로는 완벽하지는 않음)
- 크롤링에 관해 다음과 같이 다양한 issue가 존재함
 - 이미지가 포함된 텍스트에서 이미지 크롤링
 - 로그인이 필요한 사이트에서의 웹 크롤링
 - 인코딩(Encoding) 문제로 인한 한글 깨짐 문제
 - 형식이 없는 텍스트를 형식이 있는 텍스트로 변환



웹 페이지의 구성

F12 또는 우클릭 후 소스보기
text(HTML, CSS, Javascript), Image, video 등

기상청

정보공개 참여와 소통 지식과 배움 행정과 정책 기상청소개 날씨누리 바로가기

청/차장 소개 미션·비전 조직·직원 주요업무 소속·산하기관소개 홍보실 찾아오시는길 관련사이트

1일 기상정보

대형 제주도 부근 - 남해상 - 부산 부근

5일-7일

남해상 중상층 300m, 일본 해상 300m

남해상 중상층 300m, 일본 해상 300m

제25호타

남해상 중상층 300m, 일본 해상 300m

기상청 "날씨누리"

http://www.weather.go.kr

새로운 도메인으로 만나보세요

기상특보 > 동네예보 > 날씨영상 >

알림판

2018 생활기상정보

F12 DOM 탐색기 콘솔 디버거 네트워크 성능 메모리 에뮬레이션

컨텐츠 형식

이름 / 경로	프로토콜	방법	결과 / 설명	컨텐츠 형식	받은 날짜	시간	초기화 형식
PopupSlider3.js?20171227 http://www.kma.go.kr/share/js/	HTTP	GET	200 OK	text/plain	(시작 캐시)	0 초	
gm_fontplus.gif http://www.kma.go.kr/images/main/	HTTP	GET	200 OK	image/gif	(시작 캐시)	0 초	
btn_global_search.gif http://www.kma.go.kr/images/main/	HTTP	GET	200 OK	image/gif	(시작 캐시)	0 초	
session.js http://www.kma.go.kr/share/js/	HTTP	GET	200 OK	text/plain	(시작 캐시)	0 초	
ATC201810081457251_9fd554c4-338e-4f74-a252-3... http://www.kma.go.kr/upload/focus/2018/10/08/	HTTP	GET	200 OK	image/jpeg	(시작 캐시)	0 초	
ATC201810041725441_f68c24d2-bb0b-432f-948f-9... http://www.kma.go.kr/upload/focus/2018/10/04/	HTTP	GET	200 OK	image/jpeg	(시작 캐시)	0 초	
ATC201810011452271_94b38719-de92-46de-a121-... http://www.kma.go.kr/upload/focus/2018/10/01/	HTTP	GET	200 OK	image/jpeg	(시작 캐시)	0 초	
num_bannerzone1_5.gif http://www.kma.go.kr/images/main/	HTTP	GET	200 OK	image/gif	(시작 캐시)	0 초	
nanumgothic.css http://www.kma.go.kr/share/css/fonts/	HTTP	GET	200 OK	text/css	(시작 캐시)	0 초	

찾기 (Ctrl+F)

머리글 본문 매개 변수 쿠키 타이밍

응답 본문 요청 본문

44th Session of the IPCC and
Session of Working Groups I, II and III

ber 2018 | Incheon, Republic of Korea

ipcc

Abdalah Mok...
Secretary

Hoesung Lee
Chairman

Jim Skea
Co-Chairman

이름: ATC201810081457251_9fd554c4-338e-4f74-a252-30aa266f6e67.jpg

1.09 초 소요됨 (DOMContentLoaded: 1.85 초, 로드: 1.87 초)

웹 페이지의 구성

콘텐츠 형식 중 text 관련

text/html - 웹 페이지상에서 문단, 제목, 표, 이미지, 동영상 등을 정의하고 그 구조와 의미를 부여하는 마크업 언어 like 사람

text/css - 배경색, 폰트, 콘텐츠의 레이아웃 등을 지정하여, HTML 콘텐츠를 꾸며주는 스타일 규칙 언어 like 패션

Text/plain – JavaScript – 동적으로 콘텐츠를 바꾸고, 멀티미디어를 다루고, 움직이는 이미지 등 웹 페이지를 꾸며주도록 하는 프로그래밍 언어 like 근육

웹 페이지의 구성

1. Html로 뼈대 만들기

```
1 | <p>Player 1: Chris</p>
```

Player 1: Chris

2. Css를 통해 꾸며주기

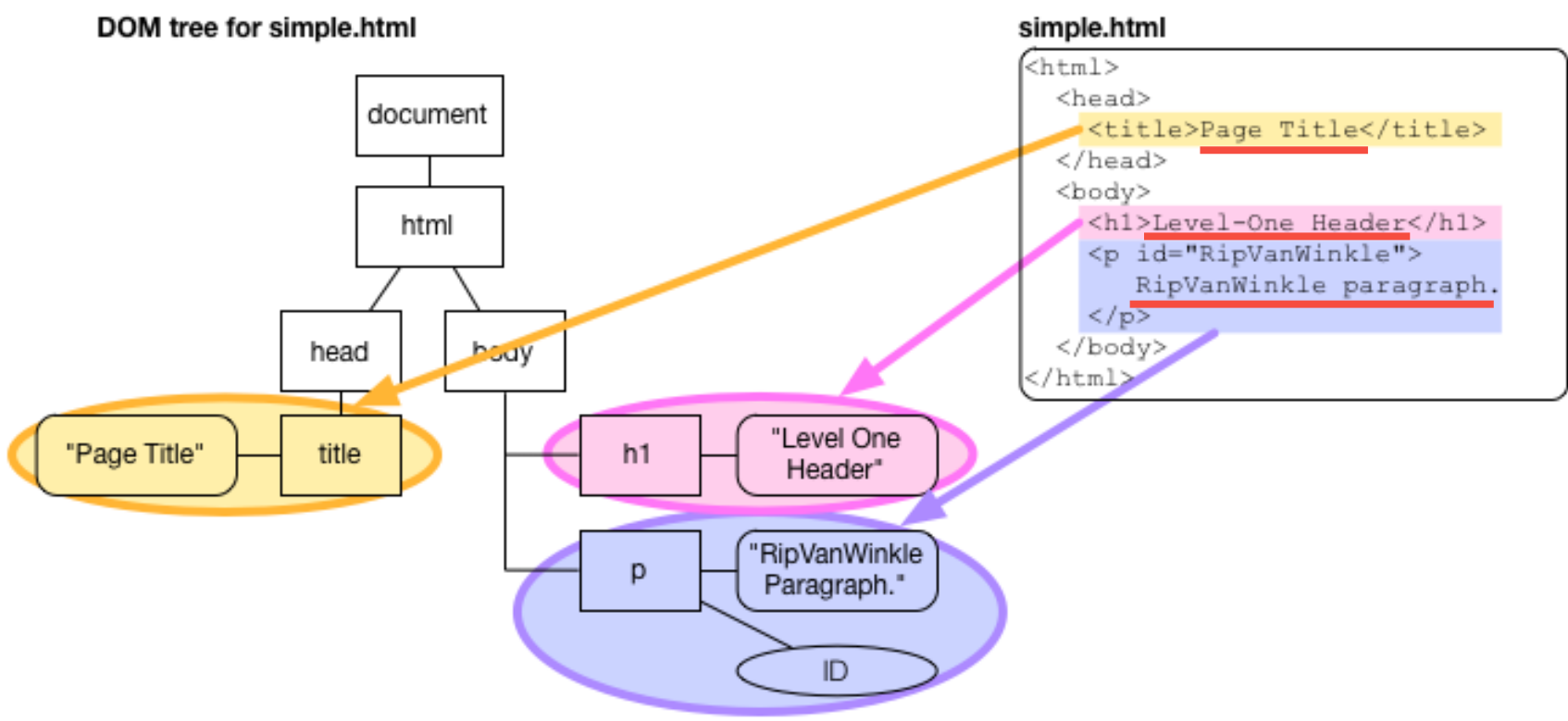
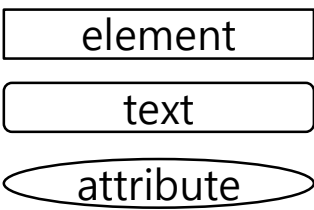
```
1 | p {  
2 |   font-family: 'helvetica neue', helvetica, sans-serif;  
3 |   letter-spacing: 1px;  
4 |   text-transform: uppercase;  
5 |   text-align: center;  
6 |   border: 2px solid rgba(0,0,200,0.6);  
7 |   background: rgba(0,0,200,0.3);  
8 |   color: rgba(0,0,200,0.6);  
9 |   box-shadow: 1px 1px 2px rgba(0,0,200,0.4);  
10 |   border-radius: 10px;  
11 |   padding: 3px 10px;  
12 |   display: inline-block;  
13 |   cursor: pointer;  
14 | }
```

PLAYER 1: CHRIS

3. JavaScript를 이용하면 클릭을 통해 이동하는 효과 등 동적으로 구현이 가능함

크롤링의 목적 : html 뼈대 사이의 text를 얻고자 함

HTML의 구조



HTML의 구성요소 및 예시

element : HTML에서 시작 태그와 종료태그로 이루어진 모든 명령어들을 의미

tag : element의 일부로 시작태그와 종료태그 두 종류가 있다. ex) `<script>`, `<dl>`, `</script>`, `</dl>`

attribute : 요소의 시작 tag 안에서 사용되는 것으로 좀 더 구체화된 명령어 체계를 의미 ex) `class`

arguments : attribute와 관련된 값 ex) `"boardViewSkin9_title"`

```
▶ <script>//<!-- var arrResizeImage = ...</script>
▲ <dl class="boardViewSkin9_title">
  <dt>기상업무 종사자 등의 교육훈련사업 위탁기관 지정 공고</dt>
  <dd>2018/10/01</dd>
</dl>
```

CSS 선택자

원하는 정보를 뽑아내기 위해선 CSS 선택자를 알아야 함

1. 타입 선택자 – 특정 element를 선택 ex) `title`, `div`, `article`

```
<title>9월 기상특성 &gt; 본청 &gt; 보도자료 &gt; 행정과 정책 &gt; 기상청 </title>
```

2. 클래스(class) 선택자 – 특정 값을 class 속성(attribute)의 값으로 갖는 element를 선택 ex) `li.sitemap` : " ." 이 class를 의미

```
<li class="sitemap"><label>골자크기</label></li>
```

3. 아이디(id) 선택자 – 특정 값을 id 속성(attribute)의 값으로 갖는 element를 선택 ex) `div#title` : "#" 이 id를 의미

```
<div id="title">
  <h4>공지사항</h4>
```

4. 속성 선택자 – 특정 속성을 갖고 있거나 특정 속성이 특정 값을 갖고 있는 element를 선택 ex) `label[for=sitelink5]`

```
<label class="blind" for="sitelink5">기상관련단체 바로가기</label>
```

rvest 패키지의 용어

node - html에서 tag라고 불리는 것

attr - html의 attribute

text - 시작 태그와 종료 태그 사이에 있는 글자

ex) `<dl class="arg"> 안녕하세요 </dl>`

rvest의 동작 순서

1. html 문서 데이터 가져오기
2. 필요한 노드 선택하기
3. 노드 내의 text 가져오기(attribute 가져오기)

ex) `read_html(url) %>% html_nodes("dl.arg") %>% html_text`



II. R을 이용한 네이버 영화 댓글 크롤링 실습

네이버 영화 ⓘ

다른 사이트를 보시려면 클릭하세요. [다른 사이트 더보기](#)



아이언맨 3 (Iron Man 3, 2013)

네티즌 ★★★★★ 8.86 (15,447) | 기자평론가 ★★★★★ 7.53 (9) 평점주기▶

SF, 액션, 모험 | 2013.04.25. 개봉 | 129분 | 미국 외 | 12세 관람가

감독 세인 블랙

관객수 9,001,679명

내용 <어벤져스> 뉴욕 사건의 트라우마로 인해 영웅으로서의 삶에 회의... [더보기](#)

관련정보 [명대사 보기](#)

↓ 다운로드

♡ 3,666

출연

관람객 평점

포토/동영상

시리즈 작품

AiTEMS 추천영화

편성표

★★★★★ 10 | chld**** | 👍 1,281

최고다. 후속작은 재미없을 거라는 편견을 깨준영화.

★★★★★ 10 | trex**** | 👍 753

자, 이제 어벤져스2가 기대된다.

★★★★★ 10 | hana**** | 👍 604

아이언맨 기대를 저버리지 않는군

14,452개 평점 전체보기

‘아이언맨3’ 검색 후

‘더보기’ 클릭

[정보모류 수정요청](#)

영화

영화상

상영작 · 예정작

- 현재 상영영화
- 개봉 예정영화
- 예고편

영화랭킹

예매

평점 · 리뷰

다운로드

인디극장

보기옵션

기본보기 넓게보기

원하시는 영화를 선택해 보세요.
 화제가 되고 있는 최신영화 정보를 한 눈에 확인하실 수 있습니다.

아이언맨 3

Iron Man 3, 2013

관람객 ★★★★★
기자·평론가 ★★★★★ 7.53

네트즌 ★★★★★ 8.86
내 평점 ★★★★★ 등록 >

개요 SF, 모험, 액션 | 미국, 중국 | 129분 | 2013.04.25 개봉
감독 셰인 블랙
출연 로버트 다우니 주니어(토니 스타크/아이언맨), 기네스 팰트로(페... [더보기](#) ▶
등급 [국내] 12세 관람가

[다운로드](#)
 3,666

주요정보

배우/제작진

포토

동영상

평점

리뷰

명대사/연관영화

줄거리

21세기 가장 매력적인 히어로의 귀환
 지금까지의 아이언맨은 잊어라!

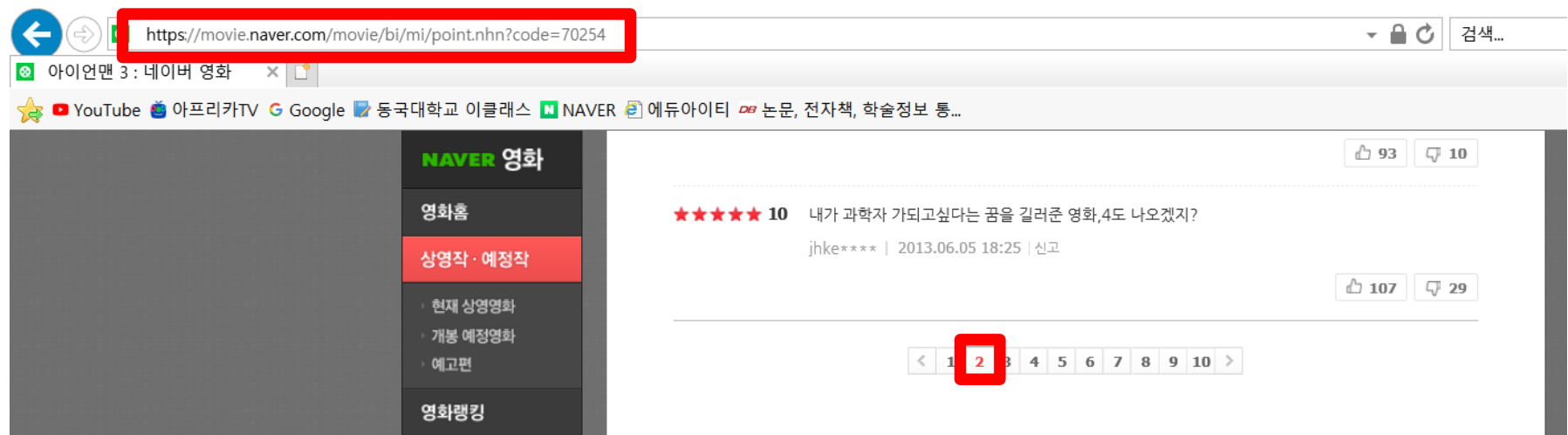
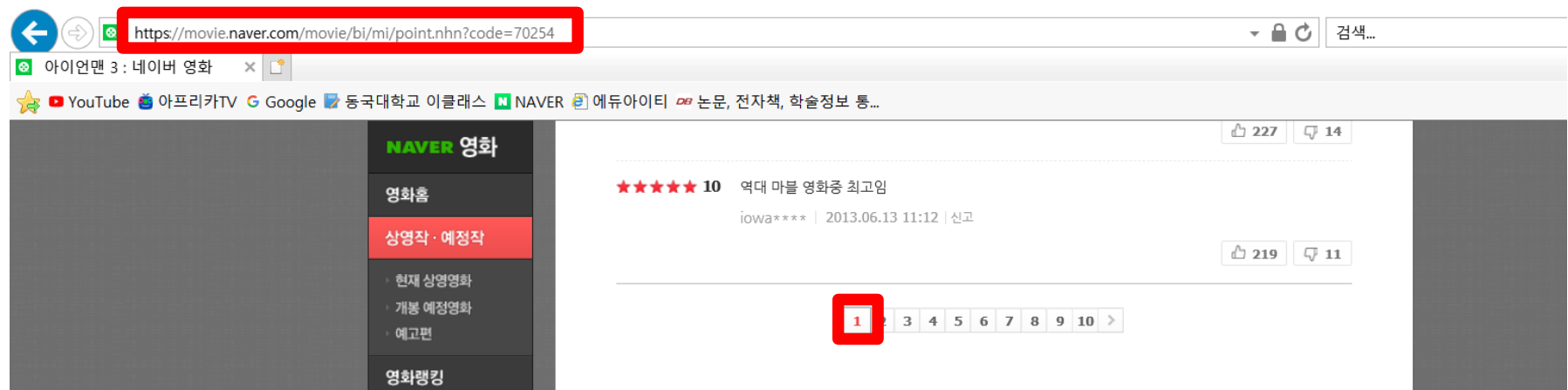
<어벤져스> 뉴욕 사건의 트라우마로 인해 영웅으로서의 삶에 회의를 느끼는 토니 스타크(로버트 다우니 주니어). 그가 혼란을 겪는 사이 최악의 테러리스트 만다린(벤 킹슬리)을 내세운 엑스트림이스 집단 AIM이 스타크 저택에 공격을 퍼붓는다. 이 공격으로 그에게 남은 건 망가진 슈트 한벌 뿐.

모든 것을 잃어버린 그는 다시 테러의 위협으로부터 세계와 사랑하는 여인(기네스 팰트로)를 지켜내야 하는 동시에 머릿속을 떠나지 않던 한가지 물음의 해답도 찾아야만 한다.

과연 그가 아이언맨인가? 슈트가 아이언맨인가?

'평점' 클릭

R을 이용한 네이버 영화 댓글 크롤링 실습



URL 즉, 주소가 페이지에따라 바뀌지 않음

R을 이용한 네이버 영화 댓글 크롤링 실습



하단의 '번호' 우클릭 후
'새 탭에서 열기' 클릭

R을 이용한 네이버 영화 댓글 크롤링 실습

네이버 영화

movie.naver.com/movie/bi/mi/pointWriteFormList.nhn?code=70254&type=after&isActualPointWriteExecute=false&isMileageSubscriptionAlready=false&isMileageSubscriptionReject=false&pag...

관람객 평점 14,452건

내 평점 등록

공감순 최신순 평점 높은 순 평점 낮은 순

스포일러 보기

관람객 평점만 보기

★★★★★ 10 최고다. 후속작은 재미없을 거라는 편견을 깨준 영화.
chld**** | 2013.07.10 11:03 | 신고

1281 94

★★★★★ 10 자, 이제 어벤져스2가 기대된다.
차세대별(trex****) | 2013.06.06 12:05 | 신고

753 78

★★★★★ 10 아이언맨 기대를 저버리지 않는군
hana**** | 2013.06.23 17:44 | 신고

604 54

★★★★★ 10 재미있음 코믹과 스토리를 잘 섞었는데 영화
스티브(terr****) | 2013.06.19 18:37 | 신고

507 43

★★★★★ 10 짱
sylv**** | 2013.06.23 14:26 | 신고

413 41

★★★★★ 10 지가 재미없게봤으면 죄다 알바인줄아네ㅜㅜ 괜히 이 영화가 900만 넘게 본줄아나ㅋㅋㅋㅋㅋ미친
김바니(bunn****) | 2013.07.02 18:43 | 신고

Elements

Console

Sources

Network

1

<!-- content -->

<input type="hidden" name="movieCode" id="movieCode" value="70254">

<input type="hidden" name="onlyActualPointYn" id="onlyActualPointYn" value="N">

<input type="hidden" name="includeSpoilerYn" id="includeSpoilerYn" value="N">

<input type="hidden" name="order" id="order" value="sympathyScore">

<input type="hidden" name="page" id="page" value="1">

<div class="ifn_area basic_ifn">

<div class="input_netizen">

<!-- [D] 관람객 평점 작성 완료 -->

<div id="actualPointWriteExecuteLayer" class="ly_viewer" style="display:none"></div>

<!-- //관람객 평점 작성 완료 -->

<!-- [D] 관람객 평점 작성 완료2 -->

<div id="pointWriteExecuteLayer" class="ly_viewer" style="display:none">

<!-- //관람객 평점 작성 완료2 -->

<div class="score_total"></div>

<div id="orderCheckbox" class="top_behavior"></div>

<div class="score_result">

<div class="star_score"></div>

<div class="score_reple"></div>

<div class="btn_area"></div>

::after

html body

Styles

Event Listeners

DOM Breakpoints

Properties

Accessibility

Filter

:hov .cls +

'F12' 누른 후

⑤ '  ' 를 클릭하여 '요소' 확인

R을 이용한 네이버 영화 댓글 크롤링 실습

네이버 영화

movie.naver.com/movie/bi/mi/pointWriteFormList.nhn?code=70254&type=after&isActualPointWriteExecute=false&isMileageSubscriptionAlready=false&isMileageSubscriptionReject=false&pag...

관람객 평점 14,452건

내 평점 등록

공감순

최신순

Color

Font 13px 나눔고딕, NanumGothic, 돋움, Dotu...

282.7 x 15.2

#333333

최고다. 후속작은 재미없을 거라는 편견을 깨준 영화

chld**** | 2013.07.10 11:03 | 신고

1281

94

자, 이제 어벤져스2가 기대된다.

차세대별(trex****) | 2013.06.06 12:05 | 신고

53

78

아이언맨 기대를 저버리지 않는군

hana**** | 2013.06.23 17:44 | 신고

604

5

재미있음 코믹과 스토리를 잘 섞었는 영화

스티브(terr****) | 2013.06.19 18:37 | 신고

507

43

짱

sylv**** | 2013.06.23 14:26 | 신고

413

41

지가 재미없게봤으면 죄다 알바인줄아네ㅋㅋ 관혀 이 영화가 900만 넘게 본줄아나ㅋㅋㅋㅋ미친

김바니(bunn****) | 2013.07.02 18:43 | 신고

413

41

Elements

Console

Sources

Network

>>

<?doctype html>

<html lang="ko">

><div style="display: none;"></div>

><head></head>

><body>

><!-- content -->

><input type="hidden" name="movieCode" id="movieCode" value="70254">

><input type="hidden" name="onlyActualPointYn" id="onlyActualPointYn" value="N">

><input type="hidden" name="includeSpoilerYn" id="includeSpoilerYn" value="N">

><input type="hidden" name="order" id="order" value="sympathyScore">

><input type="hidden" name="page" id="page" value="1">

>><div class="ifr_area basic_ifr">

>>><div class="input_netizen">

>>>><!-- [D] 관람객 평점 작성 완료 -->

>>>>><div id="actualPointWriteExecuteLayer" class="ly_viewer" style="display: none;"></div>

>>>>><!-- //관람객 평점 작성 완료 -->

>>>>><!-- [D] 관람객 평점 작성 완료2 -->

>>>>>><div id="pointWriteExecuteLayer" class="ly_viewer" style="display:none">

>>>>>>><!-- //관람객 평점 작성 완료2 -->

>>>>>>>><div class="score_total"></div>

>>>>>>>><div id="orderCheckbox" class="top_behavior"></div>

>>>>>>>>><div class="score_result">

>>>>>>>>>>

>>>>>>>>>>>

>>>>>>>>>>>><div class="stan_score"></div>

>>>>>>>>>>>><div class="score_reple">

>>>>>>>>>>>>><p>

>>>>>>>>>>>>>><!-- 스포일러 콘텐츠로 처리되는지 여부 -->

>>>>>>>>>>>>>>>>> == \$0

>>>>>>>>>>>>>>>>>></p>

>>>>>>>>>>>>>>>>>><dl></dl>

>>>>>>>>>>>>>>>>>></div>

>>>>>>>>>>>>>>>>>>><div class="btn_area"></div>

>>>>>>>>>>>>>>>>>>>><::after

html

body

div

div

div.score_result

ul

li

div.score_reple

p

span#_filtered_ment_0

Styles

Event Listeners

DOM Breakpoints

Properties

Accessibility

Filter

:hov

.cls

+

+

⑤ 첫 번째 댓글에 '🖱'를 올려 '요소' 확인
이 때, 두번째 댓글도 확인하여 어떤것이 바뀌는지 확인

R을 이용한 네이버 영화 댓글 크롤링 실습

네이버 영화

movie.naver.com/movie/bi/mi/pointWriteFormList.nhn?code=70254&type=after&isActualPointWriteExecute=false&isMileageSubscriptionAlready=false&isMileageSubscriptionReject=false&pag...

관람객 평점 14,440건 내 평점 등록

공감순 최신순 평점 높은 순 평점 낮은 순

스포일러 보기 관람객 평점만 보기

★★★★★ 10

최고다. 후속작은 재미없을 거라는 편견을 깨준 영화.

span#_filtered_ment_1

173.33 x 15.2

Color #333333

Font 13px 나눔고딕, NanumGothic, 돋움, Dotu...

1288 94

★★★★★ 10

라. 이제 어벤져스2가 기대된다

차세대별(trex****) | 2013.06.05 | 신고

756 78

★★★★★ 10

아이언맨 기대를 저버리지 않는군

hana**** | 2013.06.23 17:44 | 신고

610 54

★★★★★ 10

재미있음 코믹과 스토리를 잘 섞었는 영화

스티브(terr****) | 2013.06.19 18:37 | 신고

511 43

★★★★★ 10

짱

sylv**** | 2013.06.23 14:26 | 신고

418 41

★★★★★ 10

지가 재미없게봤으면 죄다 알바인줄아네ㅋㅋㅋ 괜히 이 영화가 900만 넘게 본줄아나ㅋㅋㅋㅋㅋ미친

김바니(bunn****) | 2013.07.02 18:43 | 신고

Elements Console Sources Network

```
none ></div>
<!-- //관람객 평점 작성 완료 -->
<!-- [D] 관람객 평점 작성 완료2 -->
<div id="pointWriteExecuteLayer" class="ly_viewer" style="display:none">
...</div>
<!-- //관람객 평점 작성 완료2 -->
<div class="score_total"></div>
<div id="orderCheckbox" class="top_behavior"></div>
<div class="score_result">
  <ul>
    <li></li>
    <li>
      <div class="star_score"></div>
      <div class="score_reple">
        <p>
          스포일러 컨텐츠로 처리되는지 여부 -->
          <span id="_filtered_ment_1"></span> == $0
        </p>
        <d1></d1>
      </div>
      <div class="btn_area"></div>
      ::after
    </li>
    <li></li>
    <li></li>
    <li></li>
  </ul>
</div>
```

html body div div div.score_result ul li div.score_reple p span#_filtered_ment_1

Styles Event Listeners DOM Breakpoints Properties Accessibility

Filter :hov .cls +

Console What's New

Highlights from the Chrome 80 update

Support for let and class redeclarations
When experimenting with new code in the Console, repeating let or class declarations no longer causes errors.

Improved WebAssembly debugging
The Sources panel has increased support for stepping over code, setting breakpoints, and resolving stack traces in source languages.

가장 뒤의 숫자가 1로 바뀜

R을 이용한 네이버 영화 댓글 크롤링 실습

← → ↺

movie.naver.com/movie/bi/mi/pointWriteFormList.nhn?code=70254&type=after&isActualPointWriteExecute=false&isMileageSubscriptionAlready=false&isMileageSubscriptionReject=false&pag...

☆

Y

⋮

관람객 평점

em

20.84 × 17.6

Color #333333

Font 14px tahoma

Padding 0px 5px 0px 5px

평점 남을 수

☐ 스포일러 보기

☐ 관람객 평점만 보기

★★★★★ 10

관고다. 후속작은 재미없을 거라는 편견을 깨준 영화.

**** | 2013.07.10 11:03 | 신고

1281 94

★★★★★ 10

자, 이제 어벤져스2가 기대된다.

차세대별(trex****) | 2013.06.06 12:05 | 신고

753 78

★★★★★ 10

아이언맨 기대를 저버리지 않는군

hana**** | 2013.06.23 17:44 | 신고

604 54

★★★★★ 10

재미있음 코믹과 스토리를 잘 섞었는데 영화

스티브(terr****) | 2013.06.19 18:37 | 신고

507 43

★★★★★ 10

짱

sylv**** | 2013.06.23 14:26 | 신고

413 41

★★★★★ 10

지가 재미있게봤으면 죄다 알바인줄아네ㅋㅋ 괜히 이 영화가 900만 넘게 본줄아나 ㅋㅋㅋㅋㅋ미친

김바니(bunn****) | 2013.07.02 18:43 | 신고

Elements

Console

Sources

Network

1

⋮

⌵

```
<!doctype html>
<html lang="ko">
  <div style="display: none;"></div>
  <head>...</head>
  <body>
    <!-- content -->
    <input type="hidden" name="movieCode" id="movieCode" value="70254">
    <input type="hidden" name="onlyActualPointYn" id="onlyActualPointYn" value="N">
    <input type="hidden" name="includeSpoilerYn" id="includeSpoilerYn" value="N">
    <input type="hidden" name="order" id="order" value="sympathyScore">
    <input type="hidden" name="page" id="page" value="1">
    <div class="ifr_area basic_ifr">
      <div class="input_netizen">
        <!-- [0] 관람객 평점 작성 완료 -->
        <div id="actualPointWriteExecuteLayer" class="ly_viewer" style="display: none;"></div>
        <!-- //관람객 평점 작성 완료 -->
        <!-- [0] 관람객 평점 작성 완료2 -->
        <div id="pointWriteExecuteLayer" class="ly_viewer" style="display:none">
          <!-- //관람객 평점 작성 완료2 -->
        </div>
        <div class="score_total"></div>
        <div id="orderCheckbox" class="top_behavior"></div>
        <div class="score_result">
          <ul>
            <li>
              <div class="star_score">
                <span class="st_off"></span>
                <em>10</em> == $0
              </div>
              <div class="score_reple">
                <p>
                  <!-- 스포일러 컨텐츠로 처리되는지 여부 -->
                  <span id="_filtered_ment_0"></span>
                </p>
              </div>
            </li>
          </ul>
        </div>
      </div>
    </div>
  </body>
</html>
```

html body div div.input_netizen div.score_result ul li div.star_score em

Styles

Event Listeners

DOM Breakpoints

Properties

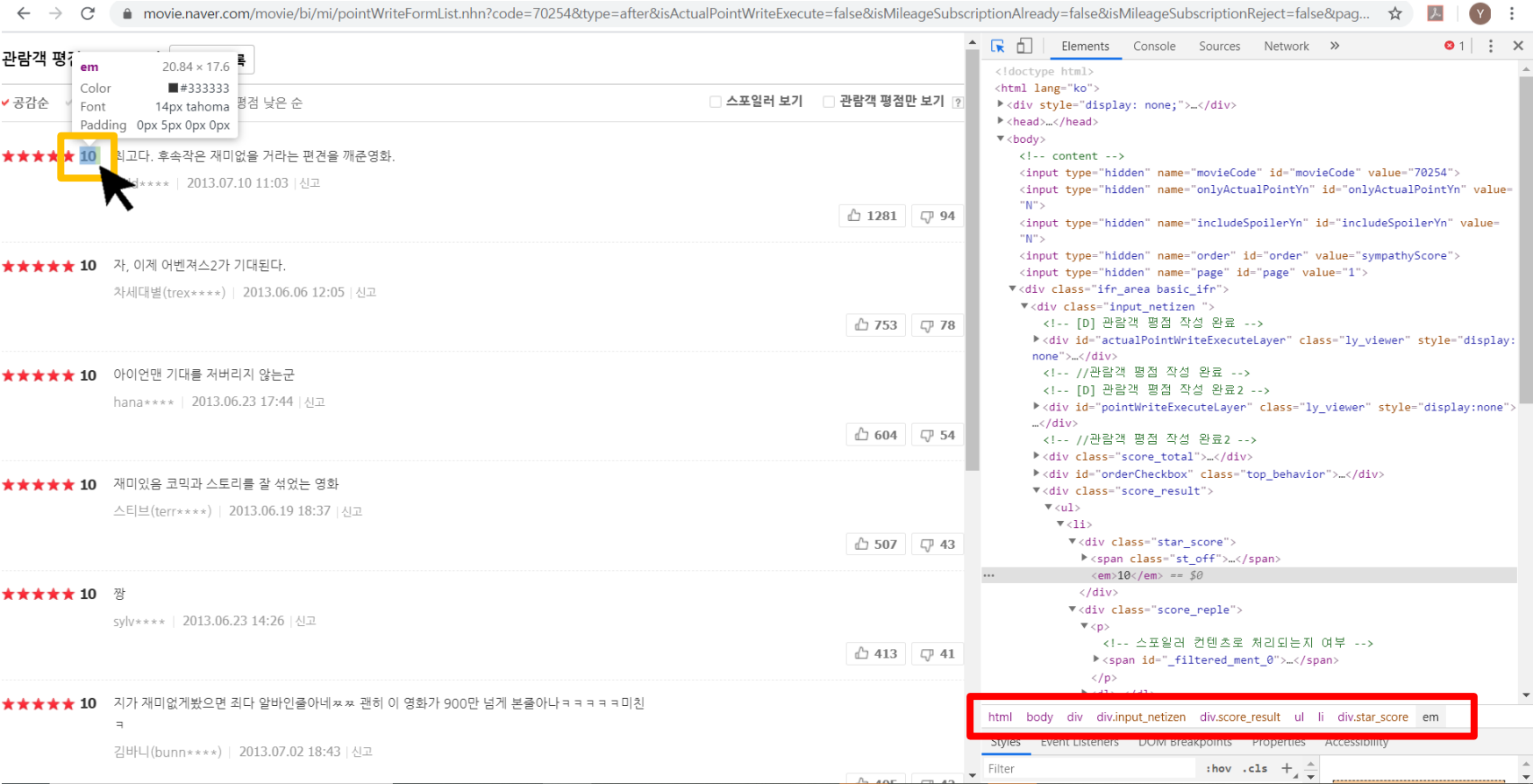
Accessibility

Filter

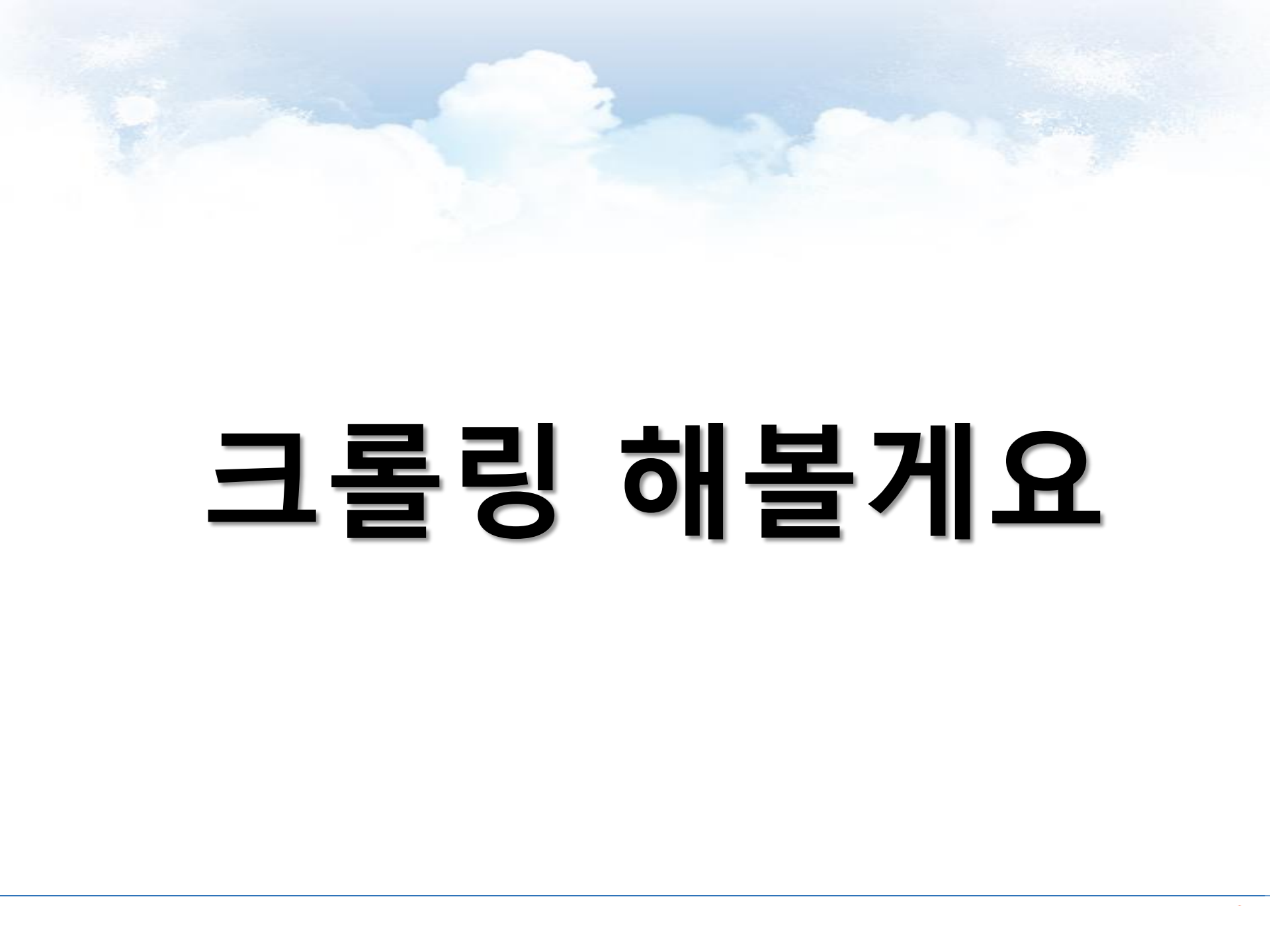
:hov .cls +

⑤ 첫 번째 평점에 '🖱'를 올려 '요소' 확인

R을 이용한 네이버 영화 댓글 크롤링 실습



⑥하단의 요소를 참고하면
쉽게 해당 요소를 찾을 수 있습니다.

A background image of a bright blue sky with soft, white, fluffy clouds. The clouds are scattered across the top half of the frame, with a denser cluster on the left and more wispy clouds on the right.

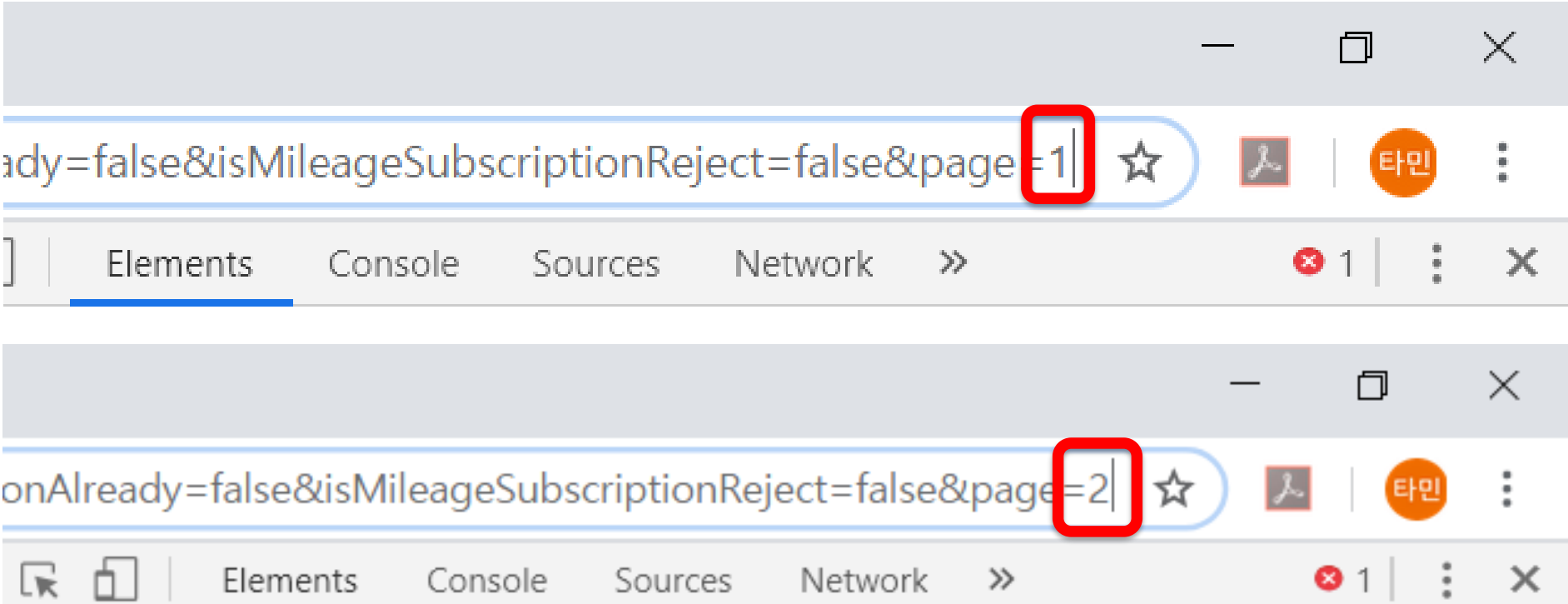
크롤링 해볼게요

- library(rvest)
 - 크롤링을 위한 패키지

크롤링에 필요한 함수들

- R에서 해당 URL의 html 소스코드를 가져오는 함수 **read_html()**
- 특정 태그가 포함하고 있는 소스코드 및 속성을 추출할 때 사용하는 함수 **html_nodes()**
- 해당 html에서 텍스트만 추출할 때 사용하는 함수 **html_text()**

1페이지와 2페이지의 주소 변화



```
url_base <-  
https://movie.naver.com/movie/bi/mi/pointWriteForm  
List.nhn?code=70254&type=after&isActualPointWriteE  
xecute=false&isMileageSubscriptionAlready=false&is  
MileageSubscriptionReject=false&page=
```

주소는 여러분 하고 싶은 영화 아무거나

**paste 함수와 for문을 이용해
페이지를 바꿀 겁니다.**

`paste(url_base, 1, sep = "")`

```
fileageSubscriptionReject=false&page=1'
```

붙었죠?

한 페이지 읽어오기

#주소설정

```
url<-paste(url_base,1,sep="")
```

#html 읽어오기

```
htxt<-read_html(url,encoding="UTF-8")
```

#node 읽기

```
table<-html_nodes(htxt,".score_result")
```

```
content<-html_nodes(table,".score_reple")
```

```
content2<-
```

```
html_nodes(content,paste("#_filtered_ment_",1,sep=""))
```

#text읽기

```
reviews<-html_text(content2) ; reviews
```

for문을 활용한 댓글 크롤링

```
all.reviews<-c()
for(page in 1:10){
  for(num in 1:9){
    url<-paste(url_base,page,sep="")
    htxt<-read_html(url,encoding="UTF-8")
    table<-html_nodes(htxt,".score_result")
    content<-html_nodes(table,".score_reple")
    content2<-html_nodes(content,paste("#_filtered_ment_",num,sep=""))
    reviews<-html_text(content2)
    if(length(reviews)==0){break}
    all.reviews<-c(all.reviews,reviews)
    print(page)
  }
}
```

각각 댓글 번호와 페이지를
바꿔가며 크롤링

R을 이용한 네이버 영화 댓글 크롤링 실습

```
> head(all.reviews)
```

[illegible][illegible][illegible][illegible]

```
> data<-gsub("[[:cntrl:]]","",all.reviews)
```

```
> head(data)
```

[1] "자, 이제 어벤져스2가 기대된다. "

[2] "아이언맨 기대를 저버리지 않는군 "

[3] "재미있음 코믹과 스토리를 잘 섞었는 영화 "

[4] "짱"

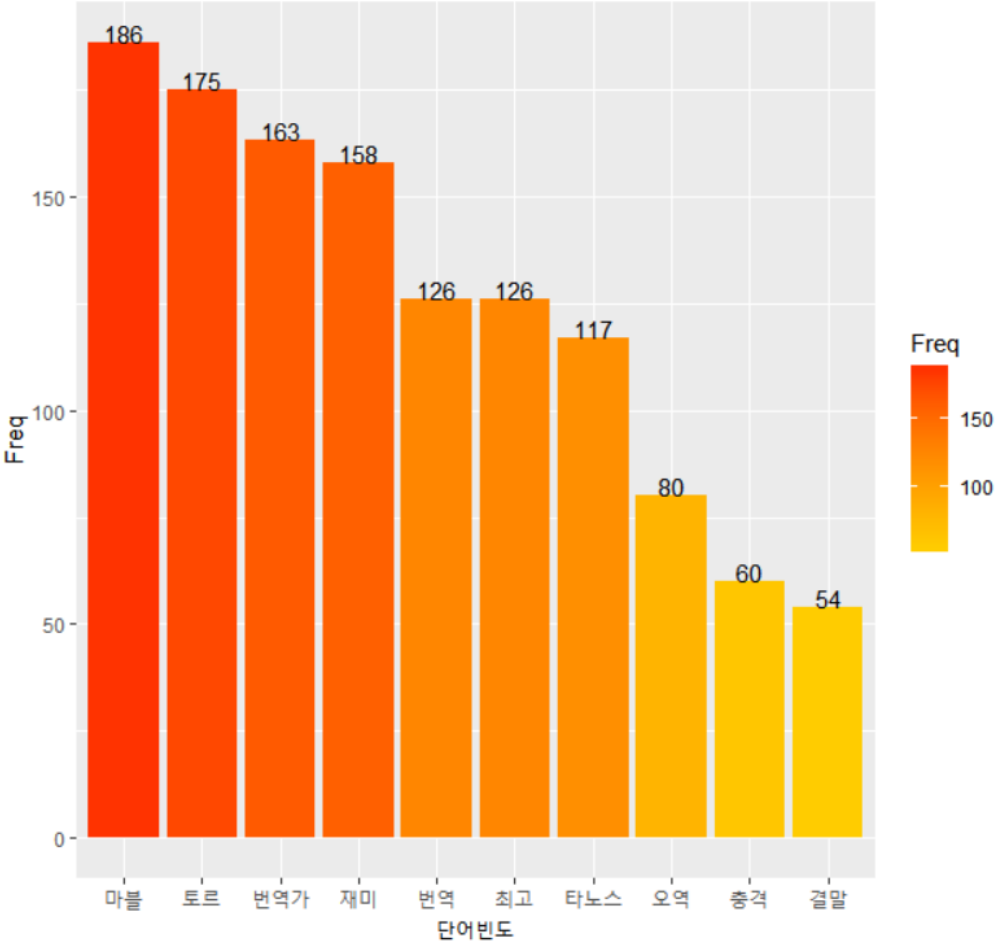
[5] "지가 재미없게봤으면 죄다 알바인줄아네 ㄱㄱ 관히 이 영화가 900만 넘게 본줄아나 ㄱㄱㄱㄱ미친 ㄱ "

[6] "1편, 2편은 3편을위해태어났다 "

결과 활용



어벤져스리뷰 단어빈도



**리뷰 크롤링 했던 코드를 가지고
평점을 크롤링 해보세요.**



III. Selenium

Selenium 이란?

여러 언어에서 웹드라이버를 통해 웹 자동화 테스트 혹은 웹 자동화를 도와주는 라이브러리.
즉, 여러 플랫폼의 브라우저 자동화를 지원하는 자동화도구!

예시)

- 1 세분류명 검색
- 2 기간입력
- 3 게시물 수 수집

NAVER 블로그

글

1

홍쇼핑 멸치

통합검색

블로그 홈 | 주제별 보기 | 이달의 블로그 | 공식블로그 | 파워블로그 | 챌린지 프로그램

글 | 블로그 | 별명·아이디

홍쇼핑 멸치에 대한 검색결과입니다. 57건

3

기간 입력

기간 전체

최근 1주

최근 1개월

기간 입력

2017-05-01


2017-05-31


적용

2

강원도 감자옹심이 감자전 공영홍쇼핑 아임쇼핑 2017. 5. 22.


감자옹심이 300g 10팩+감자전 240g 3팩+감자옹심이 소스 30g 10팩으로 구성 **홍쇼핑** 판매가 : 40,900원(자동 주문 시 1,000원 할인 39,900원) 모바일 구매... = 감자옹심이 = [재료] 감자옹심이 감자옹심이 소스 당근 애호박 양파 마늘 간장 다시 국물(다시 **멸치**, 북어 대가리, 다시마, 물) 당근, 양파...


 풀향기 | 풀향기의 맛있는 이야기



강원도 감자 진짜 옹심이와 강원도 감자 전통 감자전 공영홍쇼핑 판매~ 2017. 5. 21.

공영**홍쇼핑**에서 방송예정인... 강원도 감자 진짜 옹심이와 전통 감자전을... 받아보게 되었어요... 아이스 박스에 아이스팩을 넣어 냉동상태로... 만들어둔 **멸치**육수가 있어서... 끓고 있는 육수에 냉동상태의 옹 심이를 넣고... 애호박, 양파, 당근, 대파, 다진마늘을 넣어주고... 간은 소금으로...

 아과마린 | 아과마린의 쉬운 일상요리~



2

사용되는 함수- library(RSelenium)

remoteDriver(port, browserName) : 처음 킬 때 포트 지정 및 어떤 플랫폼 사용할지 지정. 이 함수를 지정한 객체로 뒤에 함수를 구성하게 됩니다.

예) `remDr <- remoteDriver(port=4445L, browserName="chrome")`

`remDr$open()` : 창 열기

`remDr$navigate(주소)` : 입력 주소로 이동

`$findElement(using, value)` : 요소를 찾는 함수

(using : "xpath", "css selector" 등, value :앞에서 했던 node 값)

`$clickElement()` : 마우스로 클릭

`$sendKeysToElement(list("검색어") 혹은 list(key="enter") 등등)` : 키보드 입력

그 외 스크롤 내리기 등 다양하게 있습니다.

Css selector와 Xpath의 차이점

Css selector : 대량으로 최적화되어 있으며 브라우저에 내장되어 있음.
속도가 매우 빠름.

Xpath : 모든 브라우저에 내장되어 있는 것은 아니며, 특히 IE에서는
Xpath를 이용하려면 먼저 JavaScript-Xpath와 같은 도구를 사
용 해야한다

하지만 Xpath에는 두드러지는 장점이 있다.

1. `//table/tr/td[contains(., "foo")]/../td[2]/input`. 과 같은 텍스트 내용을 기반으로 한 요소를 찾기 쉽다.
2. `../div[2]`. 과 같은 구조에서 관련된 상위 요소를 찾거나 반복문을 사용하기 좋다.

예시

```
remDr <-  
remoteDriver(port=4445L, browser  
Name="chrome")  
remDr$open()  
remDr$navigate("http://www.naver  
.com")
```

NAVER

naver.com


Chrome이 자동화된 테스트 소프트웨어에 의해 제어되고 있습니다.

NAVER whale

인터넷의 새로운 시작! 네이버 웨일로 차원이 다른 웹서핑을 경험해보세요!


다운로드

네이버를 시작페이지로 > | 주니어




메일 카페 블로그 지식iN 쇼핑 Pay TV 사전 뉴스 증권 부동산 지도 영화 뮤직 책 웹툰 더보기

1 유랑마켓



당신의 고려는 성공이 된다




고려사이버대학교
THE CYBER UNIVERSITY OF KOREA


2020학년도 신·편입 2차 모집
20.01.28(화) ~ 02.18(화) [지원하기 >](#)


연합뉴스 > 29번 환자, 감염경로 '아리송'...감시망 밖 첫 확진자 가능성


네이버뉴스 연예 스포츠 경제


뉴스스탠드 > 전체 언론사 MY 뉴스


스포츠서울


MBN


ChosunBiz


이데일리


한국일보


아이뉴스24


국민일보


경향신문


파이낸셜뉴스


NEWSIS


스포츠동아


JILJI.COM


프라이م경제

독서신문

서울파이낸스

쿠키뉴스

K 국방일보

텐아시아

네이버를 더 안전하고 편리하게 이용하세요.

NAVER 로그인

아이디·비밀번호 찾기

02.16. (일) | 이슈 1/5

'코로나19' 주요 증상과 예방 수칙

주요 증상

예방법

팩트체크

정부 정례 브리핑

N 쇼핑 x 아트원도

주머니 도둑전

GRA FOLIO 작가굿즈

1도인쇄 당선작 시리즈

예시

```
blogButton <-  
remDr$findElement(using="xpath",  
value='노드에서 우클릭 -> Copy  
-> Copy XPath')
```

이렇게 하면 블로그 버튼에 마우스가 올라갔다고
생각하시면 됩니다.

예시

```
blogButton$clickElement()
```

블로그 버튼이 눌러질겁니다.

예시

이제 검색을 해야겠죠? 먼저 검색창에 접근할게요.

이번엔 `css selector`로 해볼게요

```
webElemButton <-  
remDr$findElement(using="css  
selector",value= ' 검색창 노드를  
찾아서 우클릭 -> Copy  
-> Copy selector')
```

예시

반복문을 쓸거니까 앞에 뭔가 쓰여져 있는 것을 지우겠습니다. Shift + home + delete 하면 모든 글자가 지워져요. `sendKeysToElement`의 경우 list 형식의 값만 받습니다

```
webElemButton.sendKeysToElement(list(key='shift',key='home',key='delete'))
```

예시

이제 검색창에 검색어를 입력해야겠죠?

```
webElemButton$.sendKeysTo  
Element(list('자기이름'))
```

지금까지 코드를 잘 생각해보면 접근한
노드를 지정한 객체를 앞에 써주고 \$ 하고
어떤 행동을 할건지 적어주면 됩니다.

정리해보면

1. 내가 필요한 요소로 접근한다.
2. 내가 하고 싶은 행동을 취한다.

이 두 가지만 반복적으로 코딩해주면 됩니다.

내가 직접 할 일을 코딩한다고 생각하고 하세요.

쉽죠?

이제 직접 해봅시다.

1. 검색한다.
2. 날짜 지정란을 누른다.
3. 날짜를 지정한다.
4. 적용하기를 누른다.

예시

이제 검색이 되었으니 영화리뷰에서 했던 것처럼 크롤링 할건데 read_html만 조금 달라요.

```
html<-read_html(remDr$getPageSource()[[1]])
```

처음 창 열때 지정한 객체 remDr를 이용해 html의 정보를 가져옴.

그 뒤는 영화 리뷰 한 것과 똑같습니다.

NAVER블로그

블로그

김영석

통합검색

블로그 홈

주제별 보기

이달의 블로그

공식블로그

파워블로그

챌린지 프로그램

아이템 팩토리

블로그

별명·아이디

em.search_number 59.55 x 15.2

김영석에 대한 검색결과입니다. 28,476건


정확도

최신순

기간 전체


뇌물수수혐의 김영석 전 영천시장 가까스로 구속 모면(기각) 2018. 9. 18.

▲ 18일 오후 2시 뇌물수수혐의로 구속 전 피의자 심문(영장실질심사)을 받기위해 대구지방법원을 들어 서고있는 김영석 전 영천시장(사진= 뉴스1 캡처) [강지수 기자] 뇌물수수 혐의로 경찰수사를 받아오던 김영석 전 영천시장이 가까스로 구속을 면했다. 대구지방법원 영장전담 판사(이준규...




20대 청년 4인방...<2000여 전통주 중 인생술 찾아드립니다>〈술담화 박준형·김영석 이사〉 2019. 12. 13.

술담화 이사들 @더농부 (왼쪽부터) 이재욱 박준형 김태영 김영석 이사 전통주와 전통주 패키지 인식을 접한 청년 4명은 안타까움을 느꼈다.... 26-28세 청년들 4명 중 술담화의 디자인 업무를 담당하는 김영석 이사와 전통주 큐레이션을 담당하는 박준형 이사를 만나 상세 이야기를 나눴다. 20대 후반...



유지되는 유형과 무형의 산물을 포함하고 있는 이탈리아... 김영석, <이탈리아 이탈리아> 2019. 8. 11.

이르는, 그곳에서 벌어졌던 사건들과 그 사건들을 겪으며 그 자리에서 여태 유지되고 있는 유형과 무형



네이버 고지서
페이 포인트

#100만원 당첨 기
#재산세 #주민세

연관 검색어

김영석 기자

김영석 교수

김형석

이영석

김형석 작곡

블로그 앱 간편설

개발 가이드

블로그 글 권리보

Elements Console Sources Network Performance

<div id="content" class="content">

<!-- uiView: -->

<section class="wrap_search_list" ui-view>

<div class="category_search">

<!-- [D] .item이 활성화된 경우 aria-selected="true" / 비활성화

aria-selected="false" -->

<div class="navigator_category" role="tablist">...</div>

<!-- ngIf: postSearchCtrl.loaded -->

<div class="search_information" ng-if="postSearchCtrl.loaded">

<strong class="search_keyword">김영석

에 대한 검색결과입니다.

<em class="search_number">28,476건 == \$0

</div>

<div class="search_option"></div>

::after

</div>

<!-- end ngIf: postSearchCtrl.loaded -->

</div>

<bg-search-result-info search-display-info=

"postSearchCtrl.searchDisplayInfo" empty-result=

"postSearchCtrl.emptyResult" result-loaded="postSearchCtrl.loaded"

is-adult-user="postSearchCtrl.isAdultUser">...</bg-search-result-

info>

... #wrap #container div #content section div div span span em.search_number

Styles Event listeners DOM breakpoints Properties Accessibility em.search_number

Filter :hov .cls +

Console What's New

Highlights from the Chrome 80 update

Support for let and class redeclarations

When experimenting with new code in the Console, repeating let or class declarations no longer causes errors.

Improved WebAssembly debugging

The Sources panel has increased support for stepping over code, setting breakpoints, and resolving stack traces in source languages.

예시

```
html<-read_html(remoteDr$getPageSource()[[1]]);Sys.sleep(1)
content<-html_nodes(html, ".search_number")
num<-html_text(content)
num
```

이제 반복문만 사용하면 완성입니다.

**만약 건수가 아니라 블로그 제목을
크롤링 하고 싶다면??**