Web Scraping: Open Forum

Pilsung Kang

School of Industrial Management Engineering

Korea University

Web Scraping: Open Forum

- Web scraping: Collect data from an open forum
  - ✓ http://www.ppomppu.co.kr/zboard/zboard.php?id=insurance
  - ✓ Date, Title, and Contents

Web Scraping: Open Forum

- Step 1: Check the structure of the URL
  - ✓ Check the part that changes with regard to the pages
    - http://www.ppomppu.co.kr/zboard/zboard.php?id=insurance&page=1&divpage=13
    - http://www.ppomppu.co.kr/zboard/zboard.php?id=insurance&page=2&divpage=13
    - http://www.ppomppu.co.kr/zboard/zboard.php?id=insurance&page=3&divpage=13
    - …

Web Scraping: Open Forum

- Step 1: Check the structure of the URL
  - ✓ Check the URL to each page

```
<tr align="center" class="list1" onMouseOver="this.style.backgroundColor='#F5F5F5'" onMouseOut="this.style.backgroundColor=''" style="height:16;word-break:break-all;" valign='middle'>
    <td class='eng list_vspace' colspan=2>45876</td>    <td class='han4 list_vspace' nowrap colspan=2><nobr class='han4 list_vspace'>일반</nobr></td>
    <!--<td nowrap colspan=2 style='padding:0'><input type=checkbox name=cart value="61484"></td>-->
    <td colspan=2 class='list_vspace' align="left"><div style='width:80px;overflow:hidden;text-overflow:ellipsis' class='list_name'><nobr class='list_vspace'> [* 익명 *]</nobr>
</div></td> <td align=left class=list_vspace>

    <img src=/images/icon_03.png border=0 align=absmiddle title='익명글'> <a href="view.php?id=insurance&page=1&divpage=10&no=61484"  ><font class=list_title>암보험 생활비 받는 암보험 괜찮은가요?</font></a>
        <td nowrap class='eng list_vspace' colspan=2  title="17.08.21 11:38:13" ><nobr class='eng list_vspace'>11:38:13</nobr></td> <td nowrap class='eng list_vspace' colspan=2></td>
<td nowrap class='eng list_vspace' colspan=2>5</td></tr>


        <tr><td colspan=13 class=line_separator height=1></td></tr>


<tr align="center" class="list0" onMouseOver="this.style.backgroundColor='#F5F5F5'" onMouseOut="this.style.backgroundColor=''" style="height:16;word-break:break-all;" valign='middle'>
    <td class='eng list_vspace' colspan=2>45875</td>    <td class='han4 list_vspace' nowrap colspan=2><nobr class='han4 list_vspace'>질문</nobr></td>
    <!--<td nowrap colspan=2 style='padding:0'><input type=checkbox name=cart value="61483"></td>-->
    <td colspan=2 class='list_vspace' align="left"><div style='width:80px;overflow:hidden;text-overflow:ellipsis' class='list_name'><nobr class='list_vspace'> [* 익명 *]</nobr>
</div></td> <td align=left class=list_vspace>

    <img src=/images/icon_06.png border=0 align=absmiddle title='모바일+익명글'> <a href="view.php?id=insurance&page=1&divpage=10&no=61483"  ><font class=list_title>태아보험 질문이요.</font></a> <span class=list_comment2> <span style='cursor:pointer;' onclick='win_comment("popup_comment.php?id=insurance&no=61483");'>3</span> </span>
        <td nowrap class='eng list_vspace' colspan=2  title="17.08.21 11:23:59" ><nobr class='eng list_vspace'>11:23:59</nobr></td> <td nowrap class='eng list_vspace' colspan=2></td>
<td nowrap class='eng list_vspace' colspan=2>13</td></tr>


        <tr><td colspan=13 class=line_separator height=1></td></tr>


<tr align="center" class="list1" onMouseOver="this.style.backgroundColor='#F5F5F5'" onMouseOut="this.style.backgroundColor=''" style="height:16;word-break:break-all;" valign='middle'>
    <td class='eng list_vspace' colspan=2>45874</td>    <td class='han4 list_vspace' nowrap colspan=2><nobr class='han4 list_vspace'>질문</nobr></td>
    <!--<td nowrap colspan=2 style='padding:0'><input type=checkbox name=cart value="61480"></td>-->
    <td colspan=2 class='list_vspace' align="left"><div style='width:80px;overflow:hidden;text-overflow:ellipsis' class='list_name'><nobr class='list_vspace'> [* 익명 *]</nobr>
</div></td> <td align=left class=list_vspace>

    <img src=/images/icon_06.png border=0 align=absmiddle title='모바일+익명글'> <a href="view.php?id=insurance&page=1&divpage=10&no=61480"  ><font class=list_title>비갱신형 암보험..40대 후반</font></a> <span class=list_comment2> <span style='cursor:pointer;' onclick='win_comment("popup_comment.php?id=insurance&no=61480");'>1</span> </span>
        <td nowrap class='eng list_vspace' colspan=2  title="17.08.21 05:19:07" ><nobr class='eng list_vspace'>05:19:07</nobr></td> <td nowrap class='eng list_vspace' colspan=2></td>
<td nowrap class='eng list_vspace' colspan=2>43</td></tr>
```

http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=13&no=61484

http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=13&no=61483

http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=13&no=61480

Web Scraping: Open Forum

- Step 1: Check the structure of the URL
  - ✓ Collect the URLs to the individual posts

```
# Extract the link of each post (for first 10 pages)
for( i in c(1:10)){
     tryCatch({ tmp_url <- paste(url, i, '&divpage=10', sep="")
     tmp_list <- read_html(tmp_url) %>% html_nodes('tr.list1') %>% html_nodes('a') %>%
                 html_attr('href')
tmp_list <- paste0('http://www.ppomppu.co.kr/zboard/',tmp_list)
```

```
> tmp_list
 [1] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61484"
 [2] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61480"
 [3] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61476"
 [4] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61473"
 [5] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61470"
 [6] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61468"
 [7] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61465"
 [8] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61463"
 [9] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61461"
[10] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61458"
[11] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61454"
[12] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61452"
[13] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61450"
[14] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61448"
[15] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61445"
```

Web Scraping: Open Forum

- Step 2: Collect the information
  - ✓ Title, date, and content from each post

Web Scraping: Open Forum

- Step 2: Collect the information
  - ✓ Title, date, and content from each post
    - Title

```
653                     <td valign=top nowrap style="padding-left:6px;line-height:140%;" class=han>
654                 <font class=view_title2><!--DCM_TITLE-->태아보험 질문이요.<!--/DCM_TITLE--></font> <sup><span class=list_comment>3</span></sup><br>
655 분류: <font class=view_cate>질문</font><br>
656 이름: <span title="">[* 익명 *] </b></span><br><img src=skin/DQ_Revolution_BBS/t.gif height=5 width=5 border=0><br>
657 <img src=skin/DQ_Revolution_BBS/t.gif height=5 width=5 border=0><br>
```

    - Date

```
649                     <td valign=top class=han width=100 align=right><img src='/images/no_face.jpg'"></td>
650                 <td width="6"></td>
651                 <td class="separator2" width="3"></td>
652                 <td width=3></td>
653                     <td valign=top nowrap style="padding-left:6px;line-height:140%;" class=han>
654                 <font class=view_title2><!--DCM_TITLE-->태아보험 질문이요.<!--/DCM_TITLE--></font> <sup><span class=list_comment>3</span></sup><br>
655 분류: <font class=view_cate>질문</font><br>
656 이름: <span title="">[* 익명 *] </b></span><br><img src=skin/DQ_Revolution_BBS/t.gif height=5 width=5 border=0><br>
657 <img src=skin/DQ_Revolution_BBS/t.gif height=5 width=5 border=0><br>
658 등록일: 2017-08-21 11:23<br>
```

Web Scraping: Open Forum

- Step 2: Collect the information
  - ✓ Title, date, and content from each post
    - Content

```
</script>

    <table border="0" cellspacing="0" cellpadding="0" width="900" class="pic_bg">
    <tr>
      <td style="padding:0 8 0 8;" align="left">
        <table width="100%" style="word-break:break-all;"><tbody><tr><td><!--DCM_BODY--><table border=0 cellspacing=0 cellpadding=0 width=100%  style="table-layout:fixed;" align="left">
<tr>
<td class='board-contents' align="left" valign=top class=han>
내년에 출생하는 아이가 있어서 태아보험을 가입하려는데 몇 가지 질문을 드릴게요. <br />
여기 게시판에서 보고 약간의 정보를 얻었는데 궁금한 점 입니다. <br />
<br />
1. 태아 보험 다이렉트로 가입하면 더 저렴한지. <br />
<br />
2. 태아 보험 가입 후 출생 이후에는 해지 하고 단독실비로 갈아타라는데 그렇게 하는 장점이 무엇인지 궁금합니다. <br />
그렇게 단독실비로 가입하면 가격적으로 더 저렴해지는지 혹은 커버가 더 많이 되는지 알고 싶어요. <br />
<br />
도움 좀 부탁드리겠습니다. <br />
좋은 하루 되세요~<br />
<br />
```

Web Scraping: Open Forum

- Step 2: Collect the information
  - ✓ Visit each page and collect the title, date, and content
  - ✓ To skip unexpected errors, use tryCatch function
  - ✓ Ex:
    - Do the instruction inside the tryCatch
    - If there is an error, store NULL to the title

```
# title
tryCatch({
     tmp_title <- repair_encoding(tmp_paragraph %>% html_nodes('font.view_title2') %>% html_text(T))
}, error = function(e){tmp_title <- NULL})
```

```
Best guess: UTF-8 (100% confident)
> tmp_title
[1] "태아보험 질문이요."
```

Web Scraping: Open Forum

- Step 2: Collect the information

  ✓ Visit each page and collect the title, date, and content

```
# date
tryCatch({
      tmp_date <- repair_encoding(tmp_paragraph %>% html_nodes('td.han') %>% html_text(T))[2]
      date_start_idx <- gregexpr(pattern ='등록일', tmp_date)[[1]][1] tmp_date <- substr(tmp_date,
      date_start_idx+5, date_start_idx+14)
}, error = function(e){tmp_date <- NULL})
```

  ✓ Before preprocessing

```
> tmp_date
[1] "태아보험 질문이요.?3\r\n분류: 질문\r\n이름: [* 익명 *] \r\n등록일: 2017-08-21 11:23\r\n\r\n조회수: 21 / 추천수: 0"
```

  ✓ After preprocessing

```
> tmp_date
[1] "2017-08-21"
```

Web Scraping: Open Forum

- Step 2: Collect the information

  ✓ Visit each page and collect the title, date, and content

```
# contents
tryCatch({
      tmp_contents <- repair_encoding(tmp_paragraph %>% html_nodes('td.board-contents') %>%
      html_text(T))
      tmp_contents <- gsub("[[:punct:]]", " ", tmp_contents)
      tmp_contents <- gsub("[[:space:]]", " ", tmp_contents)
      tmp_contents <- gsub("\\s+", " ", tmp_contents)
      tmp_contents <- str_trim(tmp_contents, side = "both")
}, error = function(e){tmp_contents <- NULL})
```

  ✓ Before preprocessing

```
> tmp_contents
[1] "내년에 출생하는 아이가 있어서 태아보험을 가입하려는데 몇 가지 질문을 드릴게요. \n여기 게시판에서 보고 약간의 정보를 얻었는데 궁금한 점 입니다. \n1. 태아
 보험 다이렉트로 가입하면 더 저렴한지. \n2. 태아 보험 가입 후 출생 이후에는 해지 하고 단독실비로 갈아타라는데 그렇게 하는 장점이 무엇인지 궁금합니다. \n그렇게
 단독실비로 가입하면 가격적으로 더 저렴해지는지 혹은 커버가 더 많이 되는지 알고 싶어요. \n도움 좀 부탁드리겠습니다. \n좋은 하루 되세요~"
```

  ✓ After preprocessing

```
> tmp_contents
[1] "내년에 출생하는 아이가 있어서 태아보험을 가입하려는데 몇 가지 질문을 드릴게요 여기 게시판에서 보고 약간의 정보를 얻었는데 궁금한 점 입니다 1 태아 보험 다
이렉트로 가입하면 더 저렴한지 2 태아 보험 가입 후 출생 이후에는 해지 하고 단독실비로 갈아타라는데 그렇게 하는 장점이 무엇인지 궁금합니다 그렇게 단독실비로 가입
하면 가격적으로 더 저렴해지는지 혹은 커버가 더 많이 되는지 알고 싶어요 도움 좀 부탁드리겠습니다 좋은 하루 되세요"
```

Web Scraping: Open Forum

- Step 2: Collect the information
  - ✓ Store the information in the dataframe and export it to a CSV file

```
ppomppu_insurance[Npost,1] <- tmp_title
ppomppu_insurance[Npost,2] <- tmp_date
ppomppu_insurance[Npost,3] <- tmp_contents
Npost <- Npost + 1

# Export the result
write.csv(ppomppu_insurance, file = "ppomppu_insurance.csv")
```

	V1	V2	V3
1	태아보험 질문이요.	2017-08-21	내년에 출생하는 아이가 있어서 태아보험을 가입하려는데 ...
2	단독실비 가입하려고 하는데 보험사 추천좀 부탁드리겠습...	2017-08-21	단독실비 인터넷으로 가입하려고 하는데요 피드백 바르고 ...
3	일상생활배상책임 질문이요	2017-08-20	휴대폰액정을 깨트려서 제 실비에 있는 일상배상책임으로 ...
4	보험가입하려합니다.	2017-08-20	87년생 남자 직장인 실비보험 암보험 갱신형 가입금액 최...
5	30대중반 가장 생명보험 가입 문의	2017-08-20	안녕하세요 아침에 보험증권을 정리하다보니 제가 사망시 ...
6	주택화재보험 가입 시 보험사의 좋고 나쁨이 있나요?(가입...	2017-08-19	안녕하세요 주택화재보험을 알아 보고 있습니다 이제 시작 ...
7	치아보험 문의 드립니다	2017-08-19	치아보험 보장 중에 충치치료도 받을 수 있는게 있나요 저...
8	보험 추천 부탁드립니다.	2017-08-18	만 30세 남 실비있음 만 31세 여 두명 암 심장 뇌 같은 보험 ...
9	여행자 보험 질문	2017-08-18	교환학생으로 인도네시아에 6개월 가량 체류할 예정입니다...
10	교보 생명 보험에 관해 몇 자 여쭙습니다.	2017-08-18	10년정도 납입한 교보생명 보험이 있습니다 지인분 추천으...


ANY
questions?