



Web Scrapping: Open Forum

Pilsung Kang

School of Industrial Management Engineering

Korea University

Web Scrapping: Open Forum

- Web scraping: Collect data from an open forum

✓ <http://www.ppomppu.co.kr/zboard/zboard.php?id=insurance>

✓ Date, Title, and Contents

PPOMPPU 사람이 좋아 함께하는 곳.. 뽀뿌!

사기에는 비싸고.. 사용은 하고 싶고.. 고민될 때! 렌탈상담실!

뽀뿌	이벤트	정보	커뮤니티	갤러리	장터	포럼	뉴스	상담실	
게시글검색	휴대폰/가전	스포츠/레저	경제/지역	생활	게임	문화	취미	그룹	기타
휴대폰포럼	골프포럼	재테크포럼	결혼포럼	게임/오락	TV/드라마	DIY포럼	30+ 포럼	공포포럼	
구입개통수령	낚시포럼	소셜포럼	고민포럼	게임기포럼	독서/e-book	드론포럼	40+ 포럼	과학포럼	
휴대폰질문	농구포럼	보험포럼	대중교통	보드게임	만화/애니	사진/카메라	개발자포럼	문구포럼	
아이폰포럼	등산포럼	증권포럼	금연포럼	모바일게임	문화포럼	시계포럼	군대포럼	문서/서식	
아이패드포럼	바이크포럼	창업/자영업	전자담배포럼	플래시게임	애니툰서	인테리어	대학생포럼	뷰티/케어	
안드로이드	생활스포츠	비트코인	동식물포럼	GTA포럼	연예인포럼	전동휠포럼	미즈포럼	역사포럼	
안드로이드앱	수영포럼	로또포럼	마트/편의점	LOL포럼	영화포럼	주류포럼	봉사포럼	스타일포럼	
윈도우태블릿	스키/보드	토토/프로토	맛집포럼	디아블로포럼	음악/악기	취미포럼	성인포럼	전/현/무포럼	
기타스마트폰	스킨/스쿠버	부동산포럼	메디컬포럼	오버워치		커피포럼	직장인포럼	일바포럼	
가전포럼	스포츠포럼	지역별포럼	배달음식	클래시로알				취업포럼	
음향기기	야구포럼	국가별포럼	여행포럼	포켓몬GO				학습포럼	
컴퓨터포럼	사회인야구	해외포럼	연애포럼	피파온라인					
NAS포럼	자동차포럼	중국포럼	요리/레시피						
맥포럼	자전거포럼		운세포럼						
	축구포럼		육마포럼						
	캠핑포럼		자취포럼						
	테니스포럼		종교포럼						
	건강/헬스								

(B: 베타포럼)
 (R: 리뷰얼)
 Q: 포럼검색

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 Scrapping: Open Forum

• Step I: Check the structure of the URL

✓ Check the URL to each page

```
915 <tr align="center" class="list1" onMouseOver="this.style.backgroundColor='#F5F5F5'" onMouseOut="this.style.backgroundColor=''" style="height:16px;word-break:break-all;" valign="middle">
916 <td class="eng list_vspace" colspan=2>45876</td> <td class="han4 list_vspace" nowrap colspan=2>no<td class="han4 list_vspace">일반</td>
917 <!--<td nowrap colspan=2 style="padding:0"><input type="checkbox" name="cart" value="61484"></td-->
918 <td colspan=2 class="list_vspace" align="left"><div style="width:80px;overflow:hidden;text-overflow:ellipsis" class="list_name">no<td class="list_vspace">&nbsp;[+ 익명 +]</td>
</div></td> <td align="left" class="list_vspace">
919
920 &nbsp;<a href="view.php?id=insurance&page=1&divpage=10&no=61484"><font class="list_title">알보험 생활비 받는 알
보험 편함은가요?</font></a>
921 <td nowrap class="eng list_vspace" colspan=2 title="17.08.21 11:38:13">no<td class="eng list_vspace">11:38:13</td> <td nowrap class="eng list_vspace" colspan=2></td>
922 <td nowrap class="eng list_vspace" colspan=2>5</td></tr>
923
924 <tr><td colspan=13 class="line_separator" height=1></td></tr>
925
926
927 <tr align="center" class="list0" onMouseOver="this.style.backgroundColor='#F5F5F5'" onMouseOut="this.style.backgroundColor=''" style="height:16px;word-break:break-all;" valign="middle">
928 <td class="eng list_vspace" colspan=2>45875</td> <td class="han4 list_vspace" nowrap colspan=2>no<td class="han4 list_vspace">질문</td>
929 <!--<td nowrap colspan=2 style="padding:0"><input type="checkbox" name="cart" value="61483"></td-->
930 <td colspan=2 class="list_vspace" align="left"><div style="width:80px;overflow:hidden;text-overflow:ellipsis" class="list_name">no<td class="list_vspace">&nbsp;[+ 익명 +]</td>
</div></td> <td align="left" class="list_vspace">
931
932 &nbsp;<a href="view.php?id=insurance&page=1&divpage=10&no=61483"><font class="list_title">태아보험 질문이
요.</font></a>&nbsp;<span class="list_comment2"><span style="cursor:pointer;" onclick="win_comment('popup_comment.php?id=insurance&no=61483');">3</span> </span>
933 <td nowrap class="eng list_vspace" colspan=2 title="17.08.21 11:23:59">no<td class="eng list_vspace">11:23:59</td> <td nowrap class="eng list_vspace" colspan=2></td>
934 <td nowrap class="eng list_vspace" colspan=2>13</td></tr>
935
936 <tr><td colspan=13 class="line_separator" height=1></td></tr>
937
938
939 <tr align="center" class="list1" onMouseOver="this.style.backgroundColor='#F5F5F5'" onMouseOut="this.style.backgroundColor=''" style="height:16px;word-break:break-all;" valign="middle">
940 <td class="eng list_vspace" colspan=2>45874</td> <td class="han4 list_vspace" nowrap colspan=2>no<td class="han4 list_vspace">질문</td>
941 <!--<td nowrap colspan=2 style="padding:0"><input type="checkbox" name="cart" value="61480"></td-->
942 <td colspan=2 class="list_vspace" align="left"><div style="width:80px;overflow:hidden;text-overflow:ellipsis" class="list_name">no<td class="list_vspace">&nbsp;[+ 익명 +]</td>
</div></td> <td align="left" class="list_vspace">
943
944 &nbsp;<a href="view.php?id=insurance&page=1&divpage=10&no=61480"><font class="list_title">비갱신형 맘보
험 .40대 후반</font></a>&nbsp;<span class="list_comment2"><span style="cursor:pointer;" onclick="win_comment('popup_comment.php?id=insurance&no=61480');">1</span> </span>
945 <td nowrap class="eng list_vspace" colspan=2 title="17.08.21 05:19:07">no<td class="eng list_vspace">05:19:07</td> <td nowrap class="eng list_vspace" colspan=2></td>
946 <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 I: 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

PPOMPPU 포럼

홈 이벤트 정보 커뮤니티 갤러리 장터 포럼 뉴스 상담실

릴리안

회원가입 | 아이디 · 비밀번호 찾기 | 로그인

▶ 보험 포럼 입니다. [포럼지원센터](#)

각종 보험에 대한 정보를 공유하는 공간입니다.
보험 비교설계, 증권분석 요청은 [보험상담실]을 이용하시면 전문적으로 상담을 받으실 수 있습니다.

[관련메뉴](#) [재테크포럼](#) [증권포럼](#) [창업/자영업](#) [보험상담실](#)

 **태아보험 질문이요.** 3
분류: 질문
이름: [* 익명 *]
등록일: 2017-08-21 11:23
조회수: 13 / 추천수: 0

내년에 출생하는 아이가 있어서 태아보험을 가입하려는데 몇 가지 질문을 드릴게요.
여기 게시판에서 보고 약간의 정보를 얻었는데 궁금한 점 입니다.

1. 태아 보험 다이렉트로 가입하면 더 저렴한지.
2. 태아 보험 가입 후 출생 이후에는 해지 하고 단독실비로 갈아타라는데 그렇게 하는 장점이 무엇인지 궁금합니다.
그렇게 단독실비로 가입하면 가격적으로 더 저렴해지는지 혹은 커버가 더 많이 되는지 알고 싶어요.

도움 좀 부탁드립니다.
좋은 하루 되세요~

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>&nbsp;<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>&nbsp;<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

```
862 </script>
863
864 <table border="0" cellspacing="0" cellpadding="0" width="900" class="pic_bg">
865 <tr>
866 <td style="padding:0 8 0 8;" align="left">
867 <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">
868 <tr>
869 <td class='board-contents' align="left" valign=top class=han>
870 내년예 출생하는 아이가 있어서 태아보험을 가입하려는데 몇 가지 질문을 드릴게요. <br />
871 여기 게시판에서 보고 약간의 정보를 얻었는데 궁금한 점 입니다. <br />
872 <br />
873 1. 태아 보험 다이렉트로 가입하면 더 저렴한지. <br />
874 <br />
875 2. 태아 보험 가입 후 출생 이후에는 해지 하고 단독실비로 갈아타라는데 그렇게 하는 장점이 무엇인지 궁금합니다. <br />
876 그렇게 단독실비로 가입하면 가격적으로 더 저렴해지는지 혹은 커버가 더 많이 되는지 알고 싶어요. <br />
877 <br />
878 도움 좀 부탁드립니다. <br />
879 좋은 하루 되세요~<br />
880 <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 부탁드립니다. \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 다이렉트로 가입하면 더 저렴한지 2 태아 보험 가입 후 출생 이후에는 해지 하고 단독실비로 갈아타라는데 그렇게 하는 장점이 무엇인지 궁금합니다 그렇게 단독실비로 가입하면 가격적으로 더 저렴해지는지 혹은 커버가 더 많이 되는지 알고 싶어요 도움 좀 부탁드립니다 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 부탁드립니다...	2017-08-21	단독실비 인터넷으로 가입하려고 하는데요 피드백 빠르고 ...
3	일상생활배상책임 질문이요	2017-08-20	휴대폰액정을 패트려서 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 보험증권들 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 실비있을 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...

