| 카카오톡 봇 강좌 | >

[중급] [장문복 주의] 파싱. 가장 효율적으로 하기.

인디벨 챗봇 마스터 1:1 채팅
2019.07.26. 01:03 조회 495

(모바일이라 가독성은 죄송함미다)

제가 여기에 올라오는 파싱 관련 글들을 날마다 쭉 봐보고 있는데, 10에 5~6 정도는 다 노가다성 파싱을 하고 계시더군요..안습

반복문을 잘 모르셔서 그렇다고들 하는데 반복문은 알고보면 그냥 누워서 배그 치킨먹기 입니다.

반복문 강좌는

여기 마른얼음님 강좌를 참고하시면 될 것 같습니다.

저는 반복문 중에서 for문을 사용하도록 하겠습니다.

일단 보통 파싱 하시는 분들의 소스를 보면 대부분 Utils.getWebText 메소드를 사용해 해당 URL의 HTML소스를 가져오고, String 형태인 그 소스를 바탕으로 문자열 자르기(split)을 이용하여 원하는 부분만 가져오시더라구요

저도 물론 카톡봇에 대해 잘 모를 시기에는 split 방식을 썼었습니다. 근데 여기서 아예 모든 HTML 태그들(<body>, , 등)과 속성들(class, id, href 등)을 정규식을 사용해 모두 없애고 오직 텍스트만 남은 상태에서 자르시는 분들도 계시더군요..

태그삭제식 파싱은 오히려 효율성을 떠나서 눈도 아프고(?) 코딩하기도 힘듭니다. 차라리 태그를 삭제하지 않고 그냥 split 방식을 이용하는 게 훨씬 효율적이라고 생각합니다.

요즘은 Jsoup라는 파싱 라이브러리(API)를 이용해 많이들 파싱 하시는데 단순히 split 방식으로도 충분히 효율 좋은 코딩을 하실 수 있습니다. (더군다나 Jsoup로 파싱할 때의 CSS 선택자를 좀 어려워 하시는 분들도 계시고...)

바로 앞서 말씀드린 "반복문"을 사용 한다면 가능합니다. 결론 부터 말씀드리면 일단 줄 수가 훨씬 줄어듭니다. 그래서 오히려 더 안 복잡해지고, 보기에도 좋습니다.

보통 한 개가 아닌, 일정한 구조를 가진 여러 개의 데이터(대표적으로 목록 태그인 ul과 li)를 가져올 때 하나하나 일일히 손수로 코딩해서 작업 하시더라구요.

그런데 이걸 반복문으로 한다?

```
var doc = Utils.getWebText("https://뭐시기저시기.com", null, false, false);
var li = doc.split("<ul class=₩"list₩">")[1].split("</ul>")[0].split("<li>"); //대괄호 없이 사용하면 모든 데이터를 가져옴
var result = [];

for(var i = 0; i < li.length; i++) {
```

```
result.push(li[i].split("<span>")[1].split("</span>")[0]);
}

replier.reply(result.join("₩n"));
```

엥? 10줄 정도 만에 끝났네요?

그럼 이번엔 사용이 좀 더 간편한 Jsoup를 이용해 봅시다.

```
var doc = org.jsoup.Jsoup.connect("https://뭐시기저시기.com").get();
var li = doc.select("ul.list > li");
var result = [];

for(var i = 0; i < li.size(); i++) {
result.push(li.get(i).select("span").text());
}

replier.reply(result.join("₩n"));
```

역시 약 10줄 만에 끝났습니다.

.
.
.
그러나 노가다 방식으로 하면?

```
var doc = Utils.getWebText("https://뭐시기저시기.com", null, false, false);
var li = doc.split("<ul class=₩"list₩">")[1].split("</ul>")[0].split("<li>");
var result = [];

result.push(li[1].split("<span>")[1].split("</span>")[0]);
result.push(li[2].split("<span>")[1].split("</span>")[0]);
result.push(li[3].split("<span>")[1].split("</span>")[0]);
result.push(li[4].split("<span>")[1].split("</span>")[0]);
result.push(li[5].split("<span>")[1].split("</span>")[0]);
result.push(li[6].split("<span>")[1].split("</span>")[0]);
result.push(li[7].split("<span>")[1].split("</span>")[0]);
```

.
.
.
ㅗㅜㅑ...벌써 10줄이 넘어버렸네요
확실히 뭐가 더 좋은지 같은 코딩러로써 이 강좌로 조금이나마 아셨으면 합니다.

더군다나 .length나 .size()를 이용해 데이터의 길이가 달라도 별 다른 제약 없이 가져올 수 있습니다. 와! 이보다 더 편리하고 간편할 순 없다!

인정?

어 인정.

인디벨님의 게시글 더보기 >

좋아요 1 댓글 9 공유 | 신고

댓글 등록순 최신순

댓글알림

삭제된 댓글입니다.

인디벨 작성자

2019.07.26. 01:12 답글쓰기

엘지
뭐라는건지 모르겠쿤
2019.07.26. 01:21 답글쓰기

사로로
저도 반복문으로 처리하죠.
jsoup 재밌다
2019.07.26. 03:23 답글쓰기

doami2005
고인물만 알아듣는 강의
2019.07.26. 04:06 답글쓰기

육덕몸매
장문복주의..
헥 암더 코리안 탑클래스!
2019.07.26. 05:50 답글쓰기

마른얼음 BOT
오예 내 강좌다(?)
2019.07.26. 11:25 답글쓰기

별명을 입력해 주세요
Html 태그 제거 안하면 가끔 너무 길어서 메시지가 안옴
2019.07.26. 14:56 답글쓰기

인디벨 작성자
그럼 필요없는 부분을 좀 짜르면 되죠!
2019.07.26. 14:57 답글쓰기

OnTheWay
REGEX...
2019.07.26. 16:41 답글쓰기

Hibot
댓글을 남겨보세요

등록

글쓰기 답글 목록 ▲ TOP

'| 카카오톡 봇 강좌 |' 게시판 글

전체 [중급] 말머리 글

이 게시판 새글 구독하기

[초급] 객체 생성하기 강좌 [7]	럭스봇	2019.07.27.
카카오톡봇 시작부터 지하철 실시간 정보까지 3편 자동응답 만들기(조건문) [11]	사로로	2019.07.26.
[중급] [장문복 주의] 파싱. 가장 효율적으로 하기. [9]	인디벨	2019.07.26.
카카오톡봇 시작부터 지하철 실시간 정보까지 2편 자동응답 만들기(연산자) [14]	사로로	2019.07.25.
카카오톡봇 시작부터 지하철 실시간 정보까지 1편 리스폰스 분석 [2]	사로로	2019.07.25.

1 2 3

전체보기

이 카페 인기글

태양, 달 정보 구현 완료

중수가 초보에게 하는 강좌 [2]

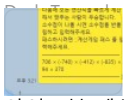
하하하하

♡1 💬11

[치***, 한**] 님 강제 탈퇴

AlphaDo

♡1 💬2



아이스봇 계산문제

JPG 개발 일지 #7

재승

♡0 💬4

안녕하세요

tomohong

♡1 💬3

자동응답봇으로 RPG봇 만들었읍니다

반가워요.

천방지축하연

♡0 💬4

도배하는 법 (잡글)

BennyK

♡0 💬5

1 2 3 4