

Data Analysis with Java

데이터 분석 프로그래밍02

Objective of Today's Class

Data Crawling

- ▶ Collect unstructured data with Selenium

Collect Unstructured Data(Cont'd)

Selenium

- ▶ A portable framework for testing web applications



Collect Unstructured Data(Cont'd)

Download a Selenium client and add it to the project

► <https://www.selenium.dev/downloads/>

Selenium Client & WebDriver Language Bindings

In order to create scripts that interact with the Selenium Server (Remote WebDriver) or create local Selenium WebDriver scripts, you need to make use of language-specific client drivers.

While language bindings for [other languages exist](#), these are the core ones that are supported by the main project hosted on GitHub.

LANGUAGE	STABLE VERSION	RELEASE DATE	BETA VERSION	BETA RELEASE DATE	LINKS
Ruby	3.142.6	October 04, 2019	4.0.0beta3	April 13, 2021	Download Beta Download Changelog API Docs
Java	3.141.59	November 14, 2018	4.0.0-beta-3	April 13, 2021	Download Beta Download Changelog API Docs
Python	3.141.0	November 01, 2018	4.0.0.b3	April 13, 2021	Download Beta Download Changelog API Docs
C#	3.14.0	August 02, 2018	4.0.0-beta2	March 17, 2021	Download Beta Download Changelog API Docs
JavaScript	3.6.0	October 06, 2017	4.0.0-beta.3	April 13, 2021	Download Beta Download Changelog API Docs

Collect Unstructured Data(Cont'd)

Download a ChromeDriver

► <https://chromedriver.chromium.org/downloads>

Current Releases

- If you are using Chrome version 91, please download [ChromeDriver 91.0.4472.19](#)
- If you are using Chrome version 90, please download [ChromeDriver 90.0.4430.24](#)
- If you are using Chrome version 89, please download [ChromeDriver 89.0.4389.23](#)
- If you are using Chrome version 88, please download [ChromeDriver 88.0.4324.96](#)
- For older version of Chrome, please see below for the version of ChromeDriver that supports it.

If you are using Chrome from Dev or Canary channel, please following instructions on the [ChromeDriver Canary](#) page.

For more information on selecting the right version of ChromeDriver, please see the [Version Selection](#) page.

The versions of the ChromeDriver and Chrome browser should be the same

Collect Unstructured Data(Cont'd)

Open a URL and Get the Page Source

```
6 public class Main {
7     public static WebDriver driver;
8     public static String base_url = "https://www.naver.com";
9     public static final String WEB_DRIVER_ID = "webdriver.chrome.driver";
10    public static final String WEB_DRIVER_PATH = "C:/Users/CTC/Downloads/chromedriver_win32/chromedriver.exe";
11
12    public static void main(String[] args) {
13        System.setProperty(WEB_DRIVER_ID, WEB_DRIVER_PATH);
14        driver = new ChromeDriver();
15        crawl();
16    }
17
18    public static void crawl() {
19        try {
20            driver.get(base_url);
21            System.out.println(driver.getPageSource());
22        } catch (Exception e) {
23            e.printStackTrace();
24        } finally {
25            driver.close();
26        }
27    }
28 }
```

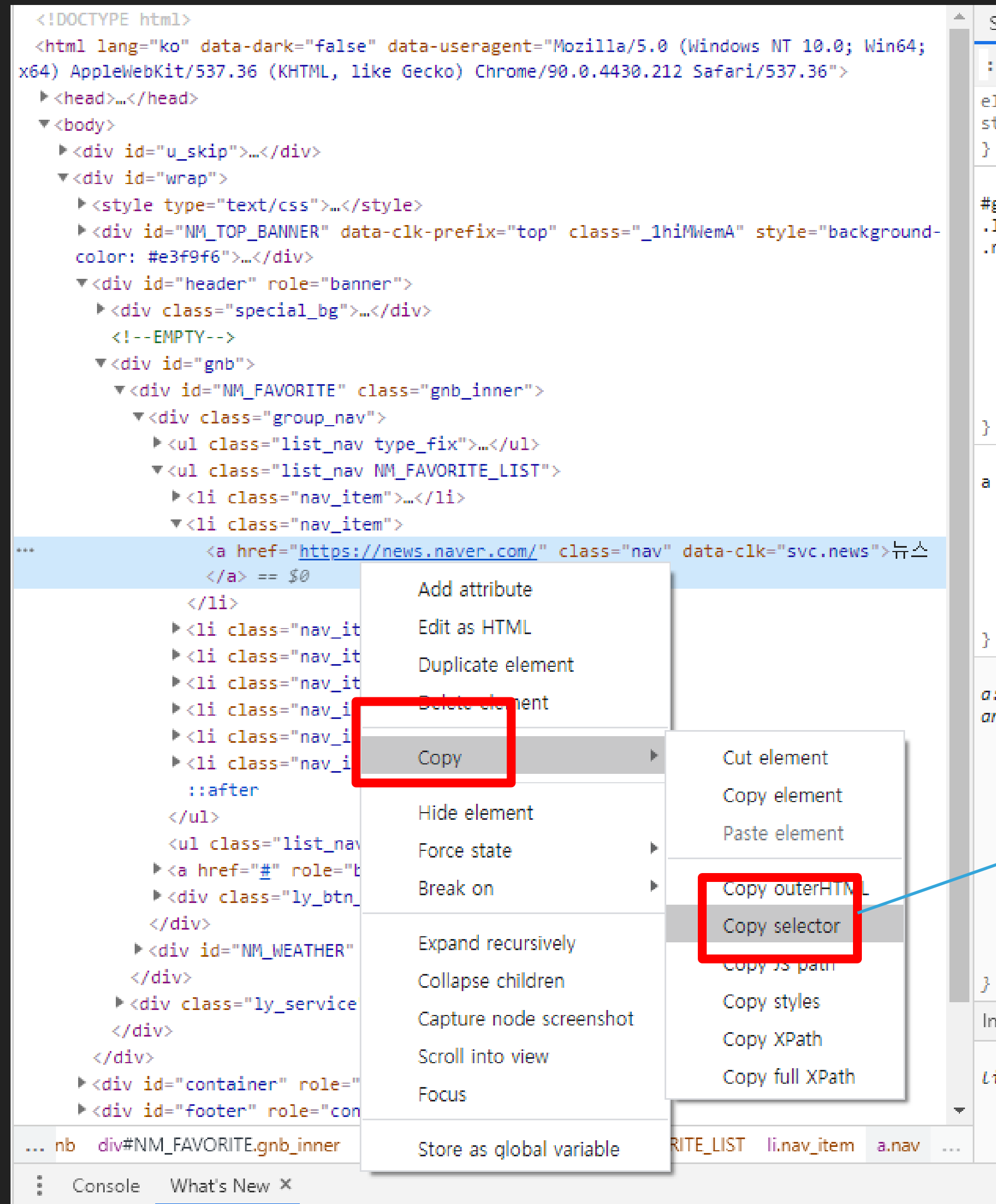
Collect Unstructured Data(Cont'd)

Click on a Button



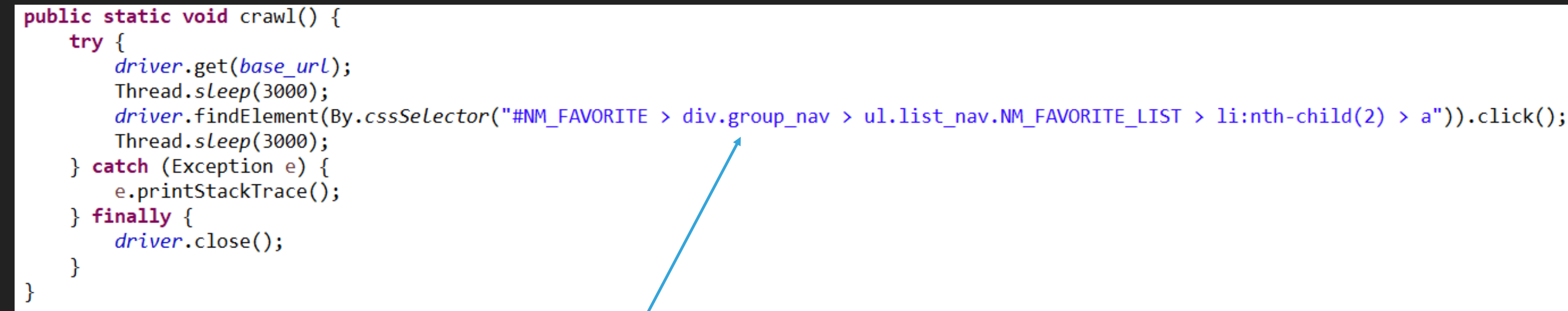
Collect Unstructured Data(Cont'd)

Click on a Button



#NM_FAVORITE > div.gnb_inner > ul.list_nav > li.nav_item > a.nav

Click on a Button



```
#NM_FAVORITE > div.group_nav > ul.list_nav.NM_FAVORITE_LIST > li:nth-child(2) > a
```

Collect Unstructured Data

► Crawl the text below

NAVER 뉴스 | TV연예 | 스포츠 | 뉴스스탠드 | 날씨

뉴스홈 **속보** 정치 경제 사회 생활/문화 세계 IT/과학 오피니언 포토 TV 랭킹뉴스

05.24 (월)

속보

정치

경제

사회

생활/문화

세계

IT/과학

오피니언

연합뉴스 속보

모바일 메인에서
보고싶은 뉴스
구독하세요!
바로가기 >

전체

신문게재기사만 | 제목형 | 요약형 | 포토만



불안정한 대기 영향...“전북 올여름도 집중호우 예상”
[KBS 전주] [앵커] 전주기상지청이 올여름 전북지역 기상 전망을 발표했습니다. 저기압과...
KBS | 1분전



정의선 현대차 회장 “경영 전 과정에서 탄소중립 달성할 것”
P4G 정상회의의 탄소중립 세션 연설...탄소중립 실현 의지 파력 [더팩트 | 서재근 기자] 정...
더팩트 | 1분전



[제보] 벤츠 E클래스 배터리 불량 속출...판매사도 “구매는 권장 안 해요”
[앵커] 한국에서 가장 많이 팔리는 수입차, 바로 메르세데스 벤츠입니다. 그런데 벤츠 E클...
KBS | 1분전



“그저 ‘몸짓’이 할 수 있는 걸로 이주민들을 그리고 싶었다”
[경향신문] 2010년 초연 후 11년 만의 공연 유럽 이민노동자 다룬 책서 영감 공연은 곳의...
경향신문 | A14면 TOP | 1분전

Collect Titles of My Favorite Idol News Articles

▶ Search for the news titles which were published for the last 1 week

▶ Restore the data into a CSV file (No., Title, Date)

The screenshot shows a Naver search results page for the keyword 'bts'. The page layout includes a top navigation bar with the Naver logo and 'bts' search term, followed by a secondary navigation bar with categories like '통합', '이미지', '뉴스' (selected), 'VIEW', '지식IN', '동영상', '쇼핑', '어학사전', '지도', and '책'. Below this is a filter bar with options for '관련도순', '최신순', and '오래된순', along with a '응답' button. The main content area displays three news articles, each with a 'PiCK' badge indicating a selected article. The first article is titled 'BTS, 빌보드 뮤직 어워즈 4관왕 기염...자체 최다 기록(종합)' and includes a summary about their record-breaking performance at the Billboard Music Awards. The second article is titled '英 인기 토크쇼 출연 BTS "최근 곡, 딱 맞는 옷 입은 듯"' and discusses their appearance on a British talk show. The third article is titled '빌보드에서 뽐뽐 터진 '다이너마이트'...BTS 4관왕' and highlights their success with the song 'Dynamite'. Each article snippet includes a small thumbnail image and a list of related news links.

N | bts

통합 이미지 뉴스 VIEW 지식IN 동영상 쇼핑 어학사전 지도 책 ...

• 관련도순 • 최신순 • 오래된순 응답

N 원하는 기사를 빨리찾는 TIP! 뉴스검색 가이드

PiCK 해당 언론사가 주요기사로 직접 선정한 기사입니다. >

연합뉴스 PiCK | 1일 전 | 네이버뉴스

BTS, 빌보드 뮤직 어워즈 4관왕 기염...자체 최다 기록(종합)

방탄소년단(BTS)이 미국 3대 음악시상식 중 하나인 '빌보드 뮤직 어워즈'(BBMA)에서 4관왕에 오르며 자체 최다 수상 기록을 경신했다. BTS는 24일(한국시간) 열린...

2년 만에 기록 경신...BTS, 빌보드 뮤직 어... SBS PiCK | 23시간 전 | 네이버뉴스
BTS, '톱 셀링 송' 수상...빌보드 4관왕, 자체 ... JTBC PiCK | 1일 전 | 네이버뉴스
BTS, 빌보드어워즈 4관왕..."다이너... 연합뉴스 PiCK | 23시간 전 | 네이버뉴스
BTS, 빌보드 뮤직 어워즈서 '톱 셀링 송'... 서울경제 PiCK | 1일 전 | 네이버뉴스

관련뉴스 20건 전체보기 >

SBS PiCK | 2시간 전 | 네이버뉴스

英 인기 토크쇼 출연 BTS "최근 곡, 딱 맞는 옷 입은 듯"

BTS 리더 RM은 최근 발매된 '버터'를 소개해달라는 말에 "무거운 메시지가 없는 여름 댄스 팝 트랙"이라며 "가사에 나오듯 사람들을 버터처럼 부드럽게 춤추게 ...

BTS "아미와 우리 서로 행복... 이제는 하나... 세계일보 | 1시간 전 | 네이버뉴스
BTS "힙합으로 시작한 우리, 이제야 딱 맞... 연합뉴스 | 2시간 전 | 네이버뉴스

MBC PiCK | 14시간 전 | 네이버뉴스

빌보드에서 뽐뽐 터진 '다이너마이트'...BTS 4관왕

그룹 방탄 소년단 BTS가 미국 3대 음악상 가운데 하나인 '빌보드 뮤직 어워즈'에서 4개 부문 수상을... 지난 1년간 전 세계인들이 가장 많이 들은 곡, '톱 셀링 송' 부...

BTS, 빌보드 뮤직 어워즈서 4관왕 올라... 조선비즈 PiCK | 1일 전 | 네이버뉴스
BTS, 빌보드 뮤직 어워즈 4관왕...'톱 셀... 부산일보 PiCK | 1일 전 | 네이버뉴스
BTS, 빌보드 뮤직 어워즈 4관왕 기염...자... TV조선 PiCK | 1일 전 | 네이버뉴스
BTS, 빌보드어워즈 4관왕..."다이너마이트... 연합뉴스 | 23시간 전 | 네이버뉴스

관련뉴스 10건 전체보기 >

Compose an Automated Testing Tool

- ▶ Target site : a bulletin board which was designed and composed in class
- ▶ Function
 - Click on the “write” button
 - Insert random values into the fields
 - Click on the “submit” button
 - The processes above should be done 1,000 times

Keywords

What we need to understand are ...

- ▶ Correlation coefficient r
- ▶ R^2
- ▶ Adj. R^2
- ▶ Coefficient
- ▶ Significance Level