



OPCUG



Users helping users
for over 40 years

Spiders, Scrapers, Crawlers, and Bots

Alan German
Ottawa PC Users' Group

September 14, 2022

Test Type

NEW CAR ASSESSMENT TEST

Configuration

VEHICLE INTO BARRIER

Make

FORD

Model

FOCUS

Year

2012

Impact Angle

0°

Closing Speed

56 (kph)

Offset Distance

0 (mm)

Performer

KARCO ENGINEERING

Contract/Study Title

35 MPH NCAP FRONTAL - 2012 FORD FOCUS 4-DOOR SEDAN



77 JPG photos.



2 PDF reports
1 ZIP data file
(from 7 formats)

Reports

v7478_vin1FAHP3F26CL215091_ACM.PDF
MC0200 Complete Report Ford Focus Frontal.pdf

Downloads

NHTSA UDS-1992
NHTSA EVS ASCII X-Y
Aftair Binary Format (ABF)
NHTSA ISO_MME
DIADEM TDMS
NHTSA XML
NHTSA JSON

Video Downloads

1	Camera No. 12 - Driver Windshield.wmv	(2.98 MB)
2	2012 Ford Focus 4-Door Sedan Frontal Impact FOR WEB.wmv	(959.45 KB)
3	Camera No. 9 - Right Front Half.wmv	(2.98 MB)
4	Camera No. 5 - Steering Column - Top.wmv	(3.03 MB)
5	Camera No. 7 - Right Overall.wmv	(3.03 MB)
6	Camera No. 10 - Right Angle.wmv	(2.98 MB)
7	Camera No. 4 - Left Angle.wmv	(2.98 MB)
8	Camera No. 6 - Steering Column - Bottom.wmv	(3.03 MB)
9	Complete Documentation Video.wmv	(81.7 MB)
10	Camera No. 18 - Real Time Left View of Impact.wmv	
11	Camera No. 16 - Driver Dummy On-Board View.wmv	
12	Camera No. 10 Real Time Left Overall.wmv	
13	Camera No. 11 - Windshield.wmv	
14	Camera No. 17 - Passenger Dummy On-Board View.wmv	
15	Camera No. 2 - Driver Close-Up.wmv	(3.03 MB)
16	Camera No. 3 - Left Front Half.wmv	(3.03 MB)
17	Camera No. 8 - Passenger Close-Up.wmv	(3.03 MB)
18	Camera No. 15 - Pit Row.wmv	(7.98 MB)

21 WMV videos

Test Type
NEW CAR ASSESSMENT TEST

Configuration
VEHICLE INTO BARRIER

Make
FORD

Model
FOCUS

Year
2012

Impact
2'

Closing
56 (kph)

Offset Dist
0 (mm)

Performer
KARCO ENGINEERING

Contract/Study Title
35 MPH NCAP FRONTAL - 2012 FORD FOCUS 4-DOOR SEDAN

9400 Crash Tests !



77 JPG photos.



2 PDF reports
1 ZIP data file
(from 7 formats)

Reports

v7478_vin1FAHP3F26CL215091_ACM.PDF
MC0200 Complete Report Ford Focus Frontal.pdf

Downloads

NHTSA UDS-1992
NHTSA EVS ASCII X-Y
Aftair Binary Format (ABF)
NHTSA ISO_MME
DIADEM TDM3
NHTSA XML
NHTSA JSON

Video Downloads

1	Camera No. 12 - Driver Windshield.wmv	(2.98 MB)
2	2012 Ford Focus 4-Door Sedan Frontal Impact FOR WEB.wmv	(959.45 KB)
3	Camera No. 9 - Right Front Half.wmv	(2.98 MB)
4	Camera No. 5 - Steering Column - Top.wmv	(3.03 MB)
5	Camera No. 7 - Right Overall.wmv	(3.03 MB)
6	Camera No. 10 - Right Angle.wmv	(2.98 MB)
7	Camera No. 4 - Left Angle.wmv	(2.98 MB)
8	Camera No. 6 - Steering Column - Bottom.wmv	(3.03 MB)
9	Complete Documentation Video.wmv	(81.7 MB)
10	Camera No. 10 - Real Time Left View of Impact.wmv	
11	Camera No. 16 - Driver Dummy On-Board View.wmv	
12	Camera No. 10 Real Time Left Overall.wmv	
13	Camera No. 11 - Windshield.wmv	
14	Camera No. 17 - Passenger Dummy On-Board View.wmv	
15	Camera No. 2 - Driver Close-Up.wmv	(3.03 MB)
16	Camera No. 3 - Left Front Half.wmv	(3.03 MB)
17	Camera No. 8 - Passenger Close-Up.wmv	(3.03 MB)
18	Camera No. 15 - Pit Room.wmv	(7.98 MB)

21 WMV videos



spiders scrapers crawlers bots



All

Images

Shopping

Videos

News

More

Tools

About 1,110,000 results (0.58 seconds)

<https://blogs.asarkar.com> › algorithms-curated › S...

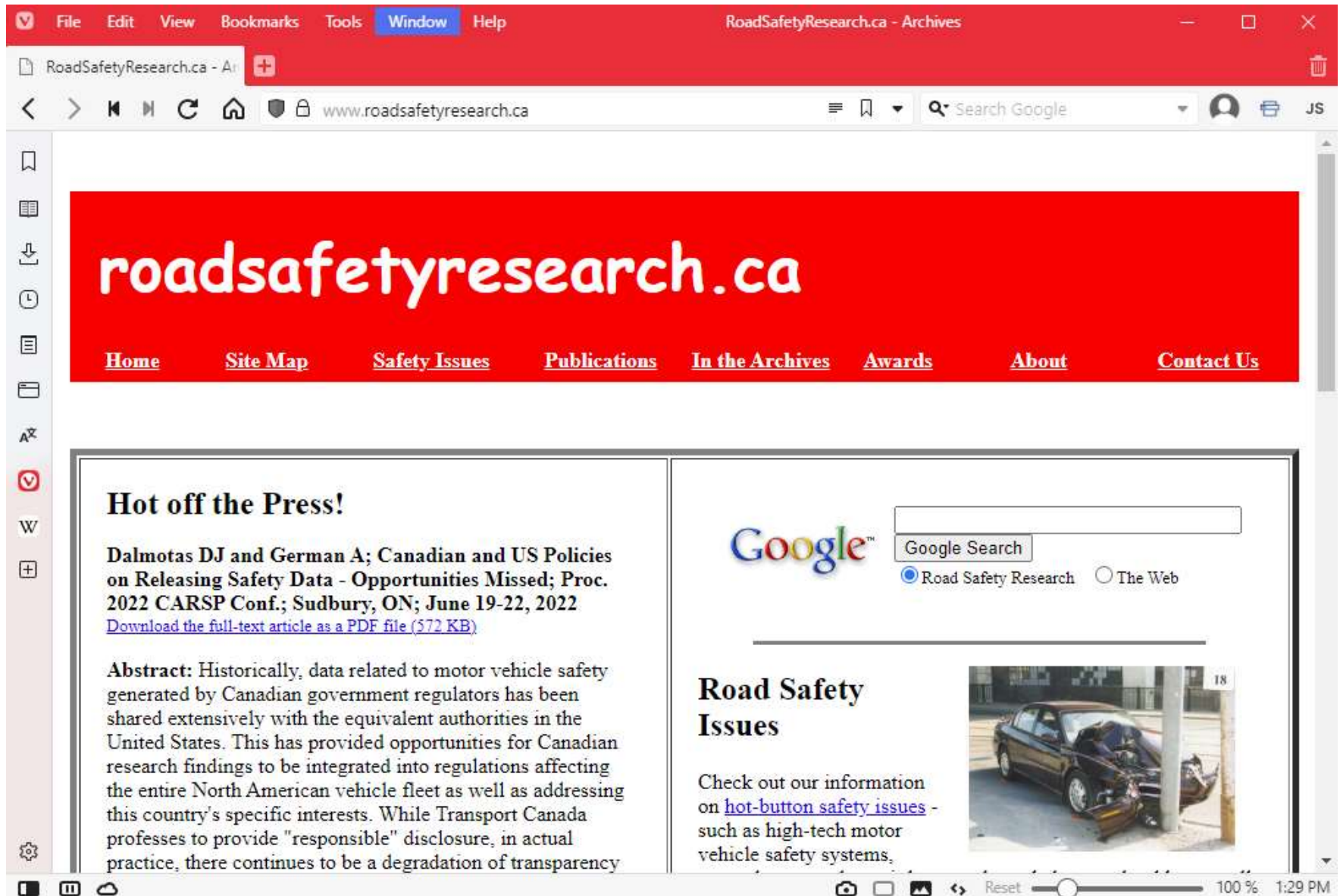
Spiders, crawlers, harvesters, bots

Spiders, crawlers, harvesters, bots ... Scrapers. • Crawlers automatically harvest all files on the web ... Also known as a web **spider, bot, harvester**.

79 pages

Crawlers automatically harvest all files on the web...

What you see...



What your browser sees...

[illegible]

Web page

Hot off the Press!

Dalmotas DJ and German A; Canadian and US Policies on Releasing Safety Data - Opportunities Missed; Proc. 2022 CARSP Conf.; Sudbury, ON; June 19-22, 2022

[Download the full-text article as a PDF file \(572 KB\)](#)

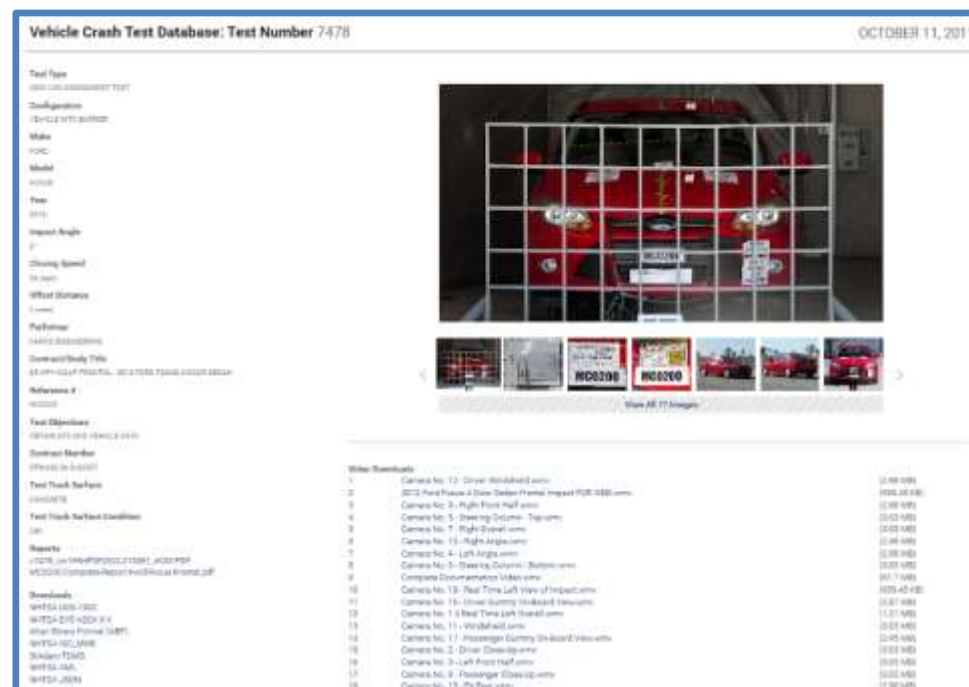
Abstract: Historically, data related to motor vehicle safety generated by Canadian government regulators has been shared extensively with the equivalent authorities in the United States. This has provided opportunities for Canadian research findings to be integrated into regulations affecting the entire North American vehicle fleet as well as addressing this country's specific interests. While Transport Canada has focused on

HTML code

```
Opportunities Missed; Proc. 2022 CARSP Conf.; Sudbury, ON; June 19-22, 2022</b><br>
<a href="https://roadsafetyresearch.ca/papers/DataPolicies.pdf"><font size="-1">
Download the full-text article as a PDF file (572 KB)</font></a>
<p>
```

URL

<https://roadsafetyresearch.ca/papers/DataPolicies.pdf>



D:\alan\djd\pdfminer\WebScraper\PageCode_7478.txt - Notepad++

File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?

PageCode_7478.txt

```
data-v-0ee9b3e3="" class="data-couple"><div data-v-0ee9b3e3=""><label data-v-0ee9b3e3="">Contract/Study Title
</label></div><div data-v-0ee9b3e3="" class="detail-text">35 MPH NCAP FRONTAL - 2012 FORD FOCUS 4-DOOR SEDAN
</div></div><div data-v-0ee9b3e3="" class="data-couple"><div data-v-0ee9b3e3=""><label data-v-0ee9b3e3="">
Reference #</label></div><div data-v-0ee9b3e3="" class="detail-text">MC0200</div></div><div data-v-0ee9b3e3=""
class="data-couple"><div data-v-0ee9b3e3=""><label data-v-0ee9b3e3="">Test Objectives</label></div><div
data-v-0ee9b3e3="" class="detail-text">OBTAIN ATD AND VEHICLE DATA</div></div><div data-v-0ee9b3e3="" class=
"data-couple"><div data-v-0ee9b3e3=""><label data-v-0ee9b3e3="">Contract Number</label></div><div
data-v-0ee9b3e3="" class="detail-text">DTNH22-06-D-00027</div></div><div data-v-0ee9b3e3="" class="data-couple"
><div data-v-0ee9b3e3=""><label data-v-0ee9b3e3="">Test Track Surface</label></div><div data-v-0ee9b3e3=""
class="detail-text">CONCRETE</div></div><div data-v-0ee9b3e3="" class="data-couple"><div data-v-0ee9b3e3=""
><label data-v-0ee9b3e3="">Test Track Surface Condition</label></div><div data-v-0ee9b3e3="" class=
"detail-text">DRY</div></div><div data-v-0ee9b3e3="" id="report-container" class="bottom-sm"><div
data-v-0ee9b3e3="" class="section-header">Reports</div><!--><div data-v-0ee9b3e3=""><a data-v-0ee9b3e3=""
href="https://nrd-static.nhtsa.dot.gov/reports/vehdb/v00000/v07400/v07478R002.pdf" target="_blank">
v7478_vin1FAHP3F26CL215691_ACM.PDF </a></div><div data-v-0ee9b3e3=""><a data-v-0ee9b3e3="" href=
"https://nrd-static.nhtsa.dot.gov/reports/vehdb/v00000/v07400/v07478R001.pdf" target="_blank">MC0200 Complete
Report Ford Focus Frontal.pdf </a></div></div><div data-v-0ee9b3e3="" id="downloads-container" class=
"bottom-sm"><div data-v-0ee9b3e3="" class="section-header">Downloads</div><div data-v-0ee9b3e3="" class=
"document-download"><div data-v-0ee9b3e3=""><a data-v-0ee9b3e3="" href=
"https://nrd-static.nhtsa.dot.gov/compress/uds/vehdb/v00000/v07400/v07478UDS.zip" target="_blank">NHTSA
UDS-1992</a></div><div data-v-0ee9b3e3=""><a data-v-0ee9b3e3="" href=
"https://nrd-static.nhtsa.dot.gov/compress/ev/vehdb/v00000/v07400/v07478.zip" target="_blank">NHTSA EV5 ASCII
X-Y</a></div><div data-v-0ee9b3e3=""><a data-v-0ee9b3e3="" href=
"https://nrd-static.nhtsa.dot.gov/compress/abf/vehdb/v00000/v07400/v07478.abf" target="_blank">Altair Binary
Format (ABF) </a></div><div data-v-0ee9b3e3=""><a data-v-0ee9b3e3="" href=
"https://nrd-static.nhtsa.dot.gov/compress/iso/vehdb/v00000/v07400/v07478ISO.zip" target="_blank">NHTSA ISO_MME
</a></div><div data-v-0ee9b3e3=""><a data-v-0ee9b3e3="" href=
"https://nrd-static.nhtsa.dot.gov/compress/tdms/vehdb/v00000/v07400/v07478tdms.zip" target="_blank">DIAdem TDMS
</a></div><div data-v-0ee9b3e3=""><a data-v-0ee9b3e3="" href=
"https://nrd-static.nhtsa.dot.gov/compress/xml/vehdb/v00000/v07400/v07478xml.zip" target="_blank">NHTSA XML
```

Hyper Text Markup Language file length : 232,517 lines : 636 Ln : 1 Col : 1 Pos : 1 Windows (CR LF) ANSI INS

D:\alan\djd\pdfminer\WebScraper\PageCode_7478.txt - Notepad++

File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?

PageCode_7478.txt

```
data-v-0ee9b3e3="" class="data-couple"><div data-v-0ee9b3e3=""><label data-v-0ee9b3e3="">Contract/Study Title
</label></div><div data-v-0ee9b3e3="" class="detail-text">35 MPH NCAP FRONTAL - 2012 FORD FOCUS 4-DOOR SEDAN
</div></div><div data-v-0ee9b3e3="" class="data-couple"><div data-v-0ee9b3e3=""><label data-v-0ee9b3e3="">
Reference #</label></div><div data-v-0ee9b3e3="" class="detail-text">MC0200</div></div><div data-v-0ee9b3e3=""
class="data-couple"><div data-v-0ee9b3e3=""><label data-v-0ee9b3e3="">Test Objectives</label></div><div
data-v-0ee9b3e3="" class="detail-text">OBTAIN ATD AND VEHICLE DATA</div></div><div data-v-0ee9b3e3="" class=
"data-couple"><div data-v-0ee9b3e3=""><label data-v-0ee9b3e3="">Contract Number</label></div><div
data-v-0ee9b3e3="" class="detail-text">DTNH22-06-D-00027</div></div><div data-v-0ee9b3e3="" class="data-couple"
><div data-v-0ee9b3e3=""><label data-v-0ee9b3e3="">Test Track Surface</label></div><div data-v-0ee9b3e3=""
class="detail-text">CONCRETE</div></div><div data-v-0ee9b3e3="" class="data-couple"><div data-v-0ee9b3e3=""
><label data-v-0ee9b3e3="">Test Track Surface Condition</label></div><div data-v-0ee9b3e3="" class=
"detail-text">DRY</div></div><div data-v-0ee9b3e3="" id="report-container" class="bottom-sm"><div
data-v-0ee9b3e3="" class="section-header">Reports</div><!--<div data-v-0ee9b3e3=""><a data-v-0ee9b3e3=""
href="https://nrd-static.nhtsa.dot.gov/reports/vehdb/v00000/v07400/v07478R002.pdf" target="_blank">
v7478_vin1FAHP3F26CL215691_ACM.PDF </a></div><div data-v-0ee9b3e3=""><a data-v-0ee9b3e3="" href=
"https://nrd-static.nhtsa.dot.gov/reports/vehdb/v00000/v07400/v07478R001.pdf" target="_blank">MC0200 Complete
Report Ford Focus Frontal.pdf </a></div></div><div data-v-0ee9b3e3="" id="downloads-container" class=
"bottom-sm"><div data-v-0ee9b3e3="" class="section-header">Downloads</div><div data-v-0ee9b3e3="" class=
"document-download"><div data-v-0ee9b3e3=""><a data-v-0ee9b3e3="" href=
"https://nrd-static.nhtsa.dot.gov/compress/uds/vehdb/v00000/v07400/v07478UDS.zip" target="_blank">NHTSA
UDS-1992</a></div><div data-v-0ee9b3e3=""><a data-v-0ee9b3e3="" href=
"https://nrd-static.nhtsa.dot.gov/compress/ev/vehdb/v00000/v07400/v07478.zip" target="_blank">NHTSA EV5 ASCII
X-Y</a></div><div data-v-0ee9b3e3=""><a data-v-0ee9b3e3="" href=
"https://nrd-static.nhtsa.dot.gov/compress/abf/vehdb/v00000/v07400/v07478.abf" target="_blank">Altair Binary
Format (ABF)</a></div><div data-v-0ee9b3e3=""><a data-v-0ee9b3e3="" href=
"https://nrd-static.nhtsa.dot.gov/compress/iso/vehdb/v00000/v07400/v07478ISO.zip" target="_blank">NHTSA ISO_MME
</a></div><div data-v-0ee9b3e3=""><a data-v-0ee9b3e3="" href=
"https://nrd-static.nhtsa.dot.gov/compress/tdms/vehdb/v00000/v07400/v07478tdms.zip" target="_blank">DIAdem TDMS
</a></div><div data-v-0ee9b3e3=""><a data-v-0ee9b3e3="" href=
"https://nrd-static.nhtsa.dot.gov/compress/xml/vehdb/v00000/v07400/v07478xml.zip" target="_blank">NHTSA XML
```

Hyper Text Markup Language file length : 232,517 lines : 636 Ln : 1 Col : 1 Pos : 1 Windows (CR LF) ANSI INS

```
<div data-v-0ee9b3e3=""><a data-v-0ee9b3e3="" href="https://nrd-static.nhtsa.dot.gov/reports/vehdb/v00000/v07400/v07478R001.pdf" target="_blank">MC0200 Complete Report Ford Focus Frontal.pdf</a></div>
```

HTML tags:

`<div>...</div>` *A “section” of code*

`<a>...</a>` *Anchor tag – link to a file*

MC0200 Complete Report Ford Focus Frontal.pdf

<https://nrd-static.nhtsa.dot.gov/reports/vehdb/v00000/v07400/v07478R001.pdf>

What do we need to do?

Build a web crawler that will:

- Browse to a web page
- Grab the underlying HTML code
- Search through the code and extract links to:
 - PDF files
 - JPG photos
 - WMV videos
- Download the files from the identified links



How are we going to do this?

- Give Google a call



How are we going to do this?

- Or... DIY Python Script
- Python modules
 - Selenium Web Driver
(automates browser operations)
 - ChromeDriverManager
(Selenium/Chrome interface)
 - re (Regular expression operations)
(text string search routines)



Two basic operations...

(1) Grab the HTML code

```
# Set the baseline URL for NHTSA's vehicle crash test database

NHTSA_URL = "https://www.nhtsa.gov/research-data/research-testing-databases#/vehicle"

# Set the URL for the crash test number
# e.g. https://www.nhtsa.gov/research-data/research-testing-databases#/vehicle/14050

crashTest = NHTSA_URL + "/" + testNumber

# Link to NHTSA's web page

driver.get(crashTest)
time.sleep(timeDelay)

# Create the file (WebScraper.src) to store the page source (HTML) code;
# get the page source code; store it as the variable content and write
# it to WebScraper.src

fo = open(cwd + "WebScraper.src", "w")
print(driver.title)
content = driver.page_source
fo.write(content)
```

Two basic operations...

(2) Extract link and download file

```
# =====
# EV5 ASCII X-Y (ZIP file)
# =====

# Search content for the link to the EV5 ASCII X-Y file

result = re.search('NHTSA UDS-1992</a></div><div data-v-0ee9b3e3=\"\">
<a data-v-0ee9b3e3=\"\" href=\"(.*)\"></div>', content)
if result:
    targetString = result.group(1)
# Extract the URL for the file download
result = re.search('https(.*)\" target', targetString)
if result:
    fileURL = "https" + result.group(1)

# Download the ZIP file in 1 MB chunks

r = requests.get(fileURL, stream=True)
with open(testFolder + "/" + fileName, "wb") as zip:
    for chunk in r.iter_content(chunk_size=1024):
        # writing one chunk at a time to pdf file
        if chunk:
            zip.write(chunk)
```

Two basic operations...

(2) Extract link and download file

```
# =====  
# EV5 ASCII X-Y (ZIP file)  
# =====  
  
# Search content for the link to the EV5 ASCII X-Y file  
  
result = re.search('NHTSA UDS-1992</a></div><div data-v-0ee9b3e3=\"\">  
<a data-v-0ee9b3e3=\"\" href=\"(.*)\"></div>', content)  
if result:  
    targetString = result.group(1)  
# Extract the URL for the file download  
result = re.search('https(.*)\" target', targetString)  
if result:  
    fileURL = "https" + result.group(1)  
  
# Download the ZIP file in 1 MB chunks  
  
r = requests.get(fileURL, stream=True)  
with open(testFolder + "/" + fileName, "wb") as zip:  
    for chunk in r.iter_content(chunk_size=1024):  
        # writing one chunk at a time to pdf file  
        if chunk:  
            zip.write(chunk)
```

```
blank">NHTSA UDS-1992</a></div><div data-v-0ee9b3e3=""><a data-v-0ee9b3e3="" href="
https://nrd-static.nhtsa.dot.gov/compress/ev/vehdb/v00000/v07400/v07478.zip" target
="_blank">NHTSA EV5 ASCII X-Y</a></div><div data-v-0ee9b3e3=""><a data-v-
```

result = re.search('

NHTSA UDS-1992</div><div data-v-0ee9b3e3=\" \">
<a data-v-0ee9b3e3=\" \" href=\" \"

(.*?)



**Get all the characters between the
First and second strings**

</div>'

, content)

```
blank">NHTSA UDS-1992</a></div><div data-v-0ee9b3e3=""><a data-v-0ee9b3e3="" href="
https://nrd-static.nhtsa.dot.gov/compress/ev/vehdb/v00000/v07400/v07478.zip" target
="_blank">NHTSA EV5 ASCII X-Y</a></div><div data-v-0ee9b3e3=""><a data-v-
```

result = re.search('

NHTSA UDS-1992</div><div data-v-0ee9b3e3=\" \">
<a data-v-0ee9b3e3=\" \" href=\" \"

(.*?)

</div>'

, content)

```
https://nrd-static.nhtsa.dot.gov/
compress/ev/vehdb/v00000/
v07400/v07478.zip
" target="_blank">
NHTSA EV5 ASCII X-Y</a>
```

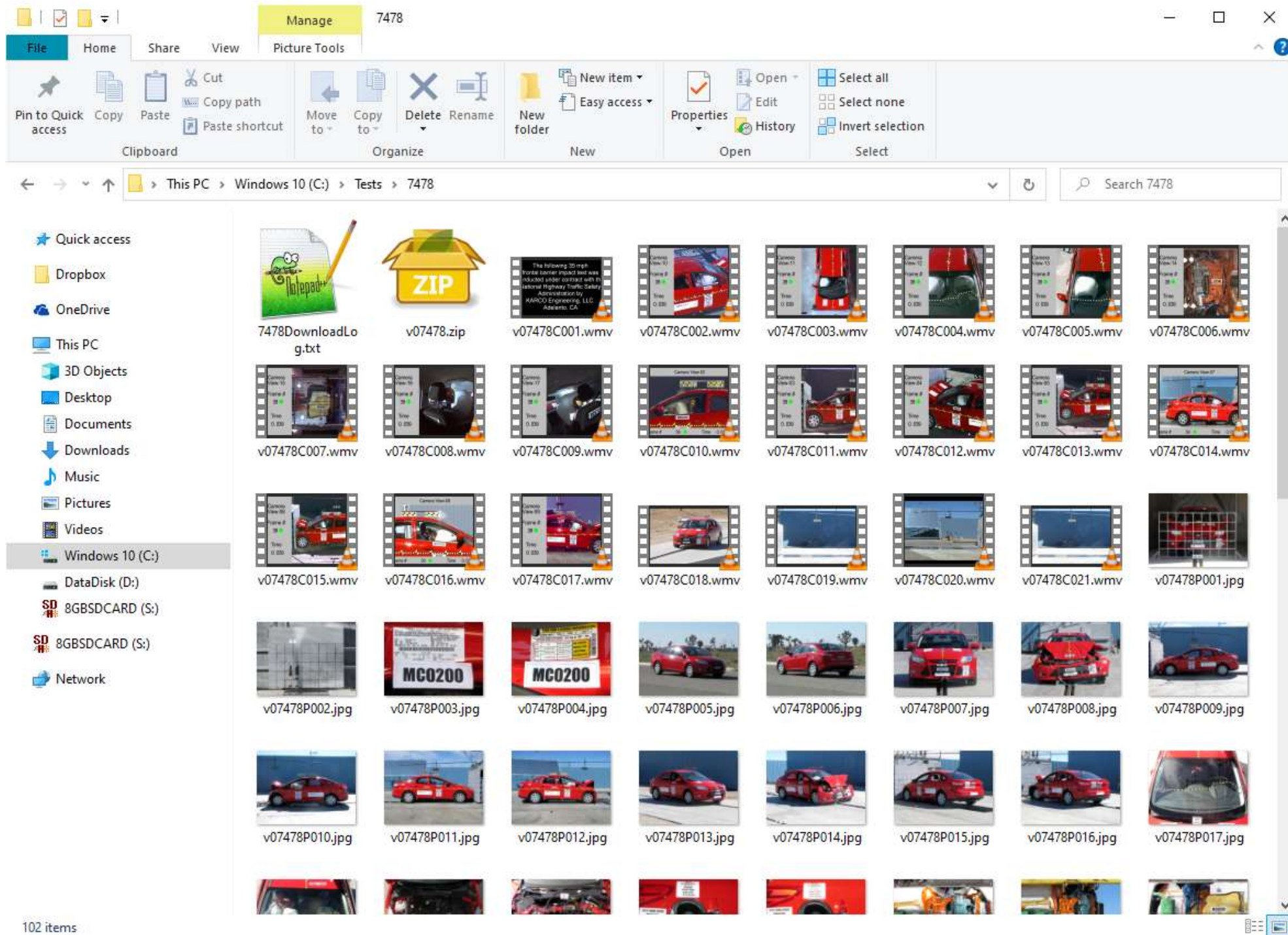
<https://nrd-static.nhtsa.dot.gov/compress/ev/vehdb/v00000/v07400/v07478.zip>

" target="_blank">

Hyperlink to file

NHTSA EV5 ASCII X-Y

File description



```
1 Crash Test No. 7478
2 =====
3
4 Reports
5 v07478R002.pdf    v7478_vin1FAHP3F26CL215691_ACM.PDF
6 v07478R001.pdf    MC0200 Complete Report Ford Focus Frontal.pdf
7
8 Data File
9 v07478.zip    NHTSA EV5 ASCII X-Y
10
11 Photographs
12 77 photographs available
13 v07478P001.jpg    No. - 001. Load Cell Location
14 v07478P002.jpg    No. - 002. Load Cell Wall
15 v07478P003.jpg    No. - 003. Manufacturer Label
16 v07478P004.jpg    No. - 004. Tire Placard
17 v07478P005.jpg    No. - 005. 2012 Ford Focus Frontal as Delivered
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58 v07478P076.jpg    2012 Ford Focus 4-Door Sedan AS DELIVERED FOR WEB
59 v07478P077.jpg    2012 Ford Focus 4-Door Sedan Frontal Impact FOR WEB
60 All 77 images downloaded
61
62 Videos
63 21 videos available
64 v07478C004.wmv    Camera No. 12 - Driver Windshield
65 v07478C021.wmv    2012 Ford Focus 4-Door Sedan Frontal Impact FOR WEB
66 v07478C017.wmv    Camera No. 9 - Right Front Half
67 v07478C013.wmv    Camera No. 5 - Steering Column - Top
68 v07478C014.wmv    Camera No. 7 - Right Overall
69 v07478C002.wmv    Camera No. 10 - Right Angle
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112 v07478C006.wmv    Camera No. 11 - Left Front
113 v07478C005.wmv    Camera No. 13 - Passenger Windshield
114 v07478C020.wmv    Camera No. 19 - Real Time Right View of Impact
115 All 21 videos downloaded
```



Anyone want to go into
competition with the big 