Investigación en Ciencias de la Computación CI-0134

Avance 2 - Identificación de elementos web faltantes usando técnicas de procesamiento de lenguaje natural

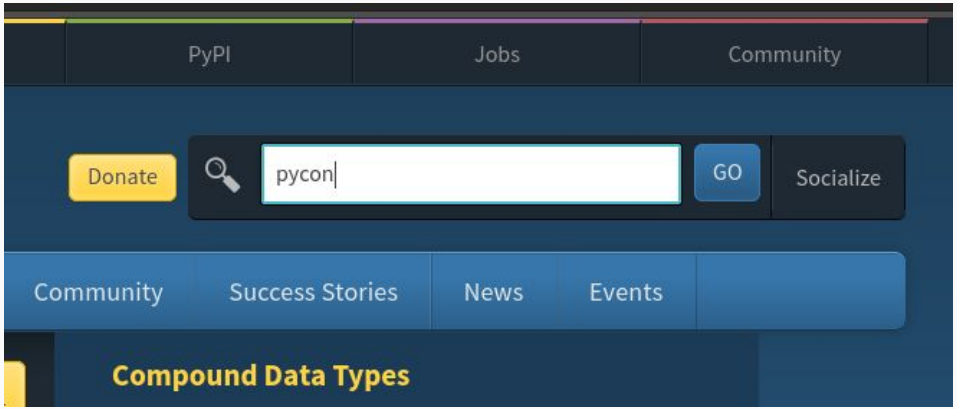
Gilbert Márquez Aldana - B94560
Daniel Artavia Cordero - B70771

El Problema

Problema

```python
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.common.by import By

driver = webdriver.Firefox()
driver.get("http://www.python.org")
# Get the HTML element with the name 'q'
elem = driver.find_element(By.NAME, "q")  # This is what usually fails
elem.clear()
# Type 'pycon' in the search box and hit Enter
elem.send_keys("pycon")
elem.send_keys(Keys.RETURN)
# Check that it found results
assert "No results found." not in driver.page_source
driver.close()
```




```
whitespace
<label class="screen-reader-text" for="id-search-field">Search This Site</label>
<input id="id-search-field" class="search-field" name="q" type="search" role="textbox" placeholder="Search" v
whitespace
<button id="submit" class="search-button" type="submit" name="submit" title="Submit this Search" tabindex="3"
<!--[if IE]><input type="text" style="display: none;" disabled="disabled" size="1" tabindex="4"><![endif]-->
```

Problema

```
driver = webdriver.Firefox()
driver.get("http://www.python.org")
# Get the HTML element with the name 'q'
elem = driver.find_element(By.NAME, "search")
elem.clear()
# Type 'pycon' in the search box and hit Enter
elem.send_keys("pycon")
```

```
(Selenium) fignewton@pop-os:~/Documents/Selenium/src$ /home/fignewton/Documents/Selenium/bin/python /home/fignewton/Documents/Selenium/src/test.py
Traceback (most recent call last):
  File "/home/fignewton/Documents/Selenium/src/test.py", line 8, in <module>
    elem = driver.find_element(By.NAME, "search")  # This is what usually fails
  File "/home/fignewton/Documents/Selenium/lib/python3.10/site-packages/selenium/webdriver/remote/webdriver.py", line 831, in find_element
    return self.execute(Command.FIND_ELEMENT, {"using": by, "value": value})["value"]
  File "/home/fignewton/Documents/Selenium/lib/python3.10/site-packages/selenium/webdriver/remote/webdriver.py", line 440, in execute
    self.error_handler.check_response(response)
  File "/home/fignewton/Documents/Selenium/lib/python3.10/site-packages/selenium/webdriver/remote/errorhandler.py", line 245, in check_response
    raise exception_class(message, screen, stacktrace)
selenium.common.exceptions.NoSuchElementException: Message: Unable to locate element: [name="search"]
Stacktrace:
RemoteError@chrome://remote/content/shared/RemoteError.sys.mjs:8:8
WebDriverError@chrome://remote/content/shared/webdriver/Errors.sys.mjs:180:5
NoSuchElementError@chrome://remote/content/shared/webdriver/Errors.sys.mjs:392:5
element.find/</<@chrome://remote/content/marionette/element.sys.mjs:134:16
```

¿Qué se propone?

- Utilizar técnicas de NLP
- Analizar el contexto en el DOM
- Analizar etiquetas y atributos
- Identificar los cambios en localizadores

Metodología

Tecnologías

Beautiful Soup

- Parsing del HTML

Long Short-Term Memory (LSTM) Network - TensorFlow

- Predicción de secuencias
- Predice la próxima palabra o elemento
- Mejor desempeño ante palabras fuera del vocabulario

Tokenización

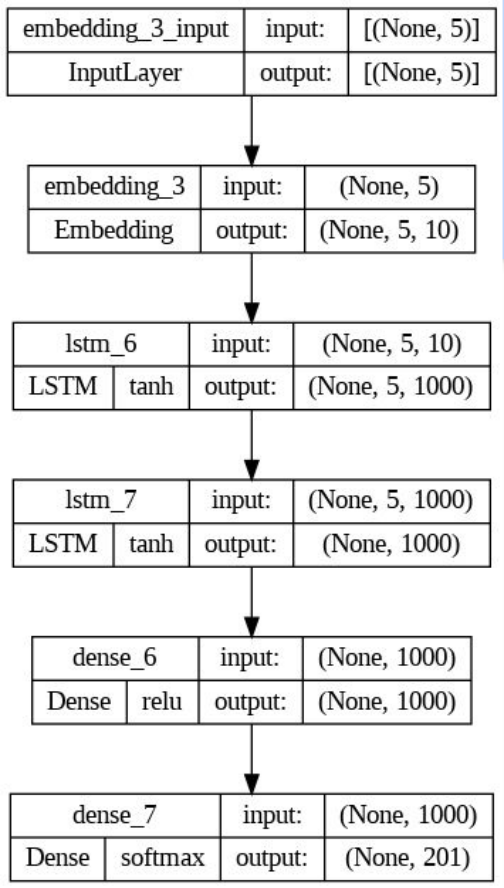
```
<!DOCTYPE html>
<html lang="es" dir="ltr" prefix="content: http://purl.org/rss/1.0/modules/content/ dc: http://
2000/01/rdf-schema# sioc: http://rdfs.org/sioc/ns# sioct: http://rdfs.org/sioc/types# skos: htt
<head>
  <link rel="profile" href="http://www.w3.org/1999/xhtml/vocab" />
  <meta charset="utf-8">
  <meta name="viewport" content="width=device-width, initial-scale=1.0">
  <link rel="apple-touch-icon" href="/sites/all/themes/ecci_bootstrap/img/apple-touch-icon.png"
  <link href="https://fonts.googleapis.com/css?family=Open+Sans:300,400,600,700,800%7CShadows+I
  <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<meta name="Generator" content="Drupal 7 (http://drupal.org)" />
<link rel="canonical" href="/inicio" />
<link rel="shortlink" href="/node/147" />
<link rel="shortcut icon" href="https://www.ecci.ucr.ac.cr/sites/default/files/icon.ico" type="
  <title>Escuela de Ciencias de la Computación e Informática | Escuela de Ciencias de la Comput
  <link type="text/css" rel="stylesheet" href="https://www.ecci.ucr.ac.cr/sites/default/files/c
<link type="text/css" rel="stylesheet" href="https://www.ecci.ucr.ac.cr/sites/default/files/css
<link type="text/css" rel="stylesheet" href="https://www.ecci.ucr.ac.cr/sites/default/files/css
<link type="text/css" rel="stylesheet" href="https://www.ecci.ucr.ac.cr/sites/default/files/css
<style>background-color:white;
</style>
<link type="text/css" rel="stylesheet" href="/sites/all/themes/ecci_bootstrap/vendor/bootstrap/
<link type="text/css" rel="stylesheet" href="https://www.ecci.ucr.ac.cr/sites/default/files/css
  <script src="/sites/all/themes/ecci_bootstrap/vendor/modernizr/modernizr.min.js"></script>
</head>
<body class="html front not-logged-in no-sidebars page-node page-node- page-node-147 node-type-
  <div id="skip-link">
    <a href="#main-content" class="element-invisible element-focusable">Pasar al contenido prin
  </div>
    <div class="body front">
```

```
html head link meta meta link link meta meta link link link title titleend link link link link
bodyfronthtmli18nesnosidebarsnodetypepagenotloggedinpagenodepagenodepagenode147 divskiplink ael
divcontainer divcollg8colmd8colsm12colxs12 a imgimgresponsive aend divend divcollg4colmd4colsm1
ifafauser iend aend liend li a ifafaat iend aend liend li a ifafaphone iend aend liend li a aen
divcontainerheadercontainer divheaderrow divheadercolumn divheaderrow divheadersearchhiddenxs d
formcontentsearchformsearchsearchblockform div div h2elementinvisible h2end divinputgroup input
spanglyphiconglyphiconsearchicon spanend buttonend spanend divend divformactionsformgroupformwra
input divend divend formend sectionend divend divend navheadernavtop ulnavnavpills lihiddenxs a
iend aend liend li spanwsnowrap a ifafaphone iend aend spanend liend ulend navend divend divhea
divcollapseheadernavmainheadernavmaineffect1headernavmainsubeffect1 nav ulmenunavnavbarnav liact
aend liend li a aend liend lidropdownsubmenu adropdowntoggle aend uldropdownmenu li a aend lien
uldropdownmenu li a aend liend li a aend liend ulend liend lidropdownsubmenu adropdowntoggle ae
aend uldropdownmenu li a aend liend li a aend liend lidropdownsubmenu adropdowntoggle aend uldro
uldropdownmenu li a aend liend li a aend liend li a aend liend ulend liend lidropdownsubmenu ad
aend uldropdownmenu li a aend liend ulend liend ulend liend liend lidropdown adropdowntog
uldropdownmenu li a aend liend li a aend liend li a aend liend li a aend liend li a aend liend
adropdowntoggle aend uldropdownmenu li a aend liend ulend liend li a aend liend li a aend liend
uldropdownmenu li a aend liend li a aend liend li a aend liend ulend liend li a
aend liend li a aend liend li a aend liend ulend liend li a aend liend li a aend liend ulend li
lidropdown adropdowntoggle aend uldropdownmenu li a aend liend ulend liend ulend navend divend
divrevsliderwrapperslidercontainervisiblelgvisiblemd divregionregionslider sectionblockblockrev
divbannerfullwidthbannerrsbanner ul li img divbtlgbtnbtnquaternarysfttpcaptiontpresizeme a br a
divskewfromrighttpcaptiontpresizeme img divend divlfltpcaptiontpresizeme img divend divlargebol
divfademediumlargelightwhitetpcaptiontpresizeme divend divfadetpcaptiontpresizeme divend divske
img divbtnbtnlgbtnquaternarysfttextlefttpcaptiontpresizeme br br br divend liend li img divsftt
divsfttpcaptiontpresizeme img divend liend ulend divtpbannertimer divend divend divend divend s
sectionblockblockblockblockblockclearfixblockblock102 divcontainer divrow divcolmd8 p span span
sectionend divend divend divcontainer divcenterrow divcolmd12spaceblocks divregionregioncontent
articleclearfixnodenodepagenodepromotednode147 header spanelementhiddenrdfmeta spanend spaneleme
divfielditems divevenfielditem divcontainer divcenterrow divcolmd12spaceblocks plead pend diven
sectiongrayblock divcontainer divcolsm12spaceblocks divcolmd9 divregionregionhighlighted sectio
```

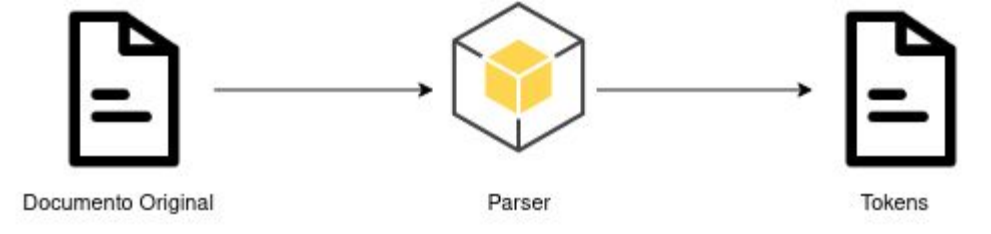
Entrenamiento

- Recorrer documento por “ventanas” o secuencias
- El tamaño de la ventana es un parámetro configurable

Modelo

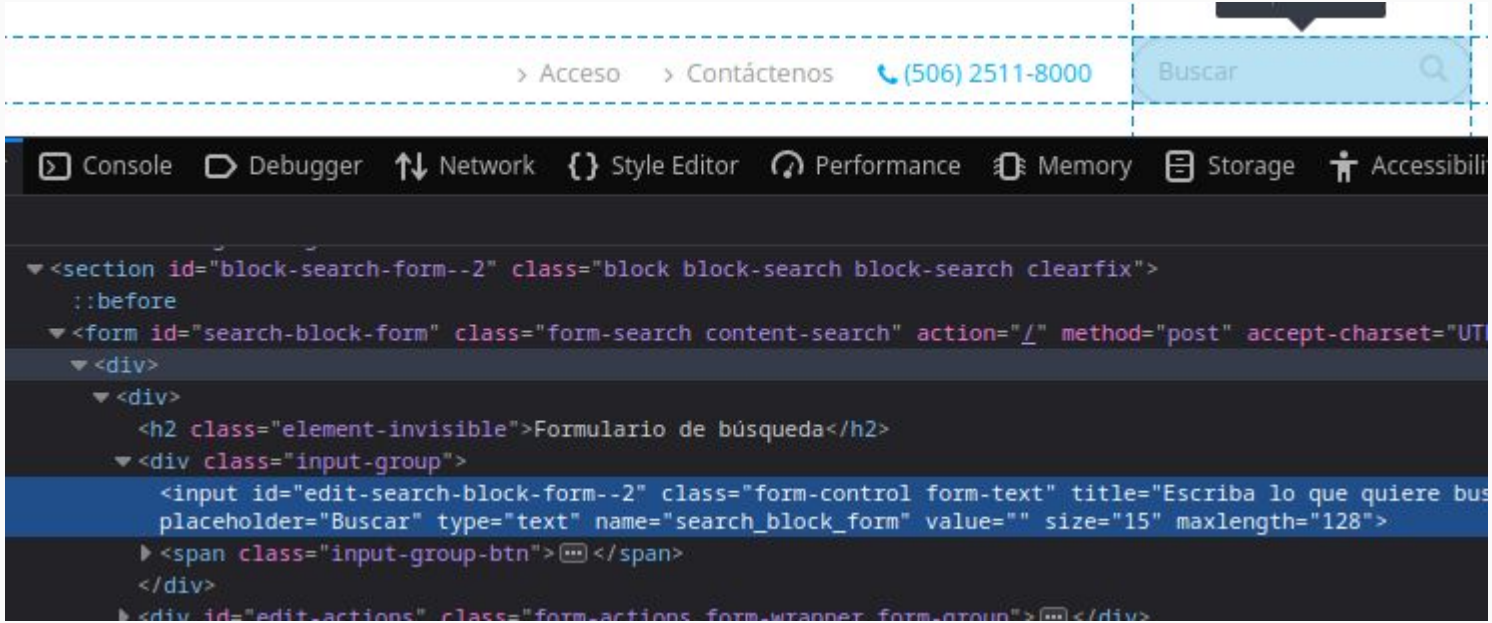

embedding_3_input | input: | [(None, 5)]
InputLayer | output: | [(None, 5)]
embedding_3 | input: | (None, 5)
Embedding | output: | (None, 5, 10)
lstm_6 | input: | (None, 5, 10)
LSTM | tanh | output: | (None, 5, 1000)
lstm_7 | input: | (None, 5, 1000)
LSTM | tanh | output: | (None, 1000)
dense_6 | input: | (None, 1000)
Dense | relu | output: | (None, 1000)
dense_7 | input: | (None, 1000)
Dense | softmax | output: | (None, 201)

Nuestro Producto (Paso 1)



Nuestro Producto (Paso 2)

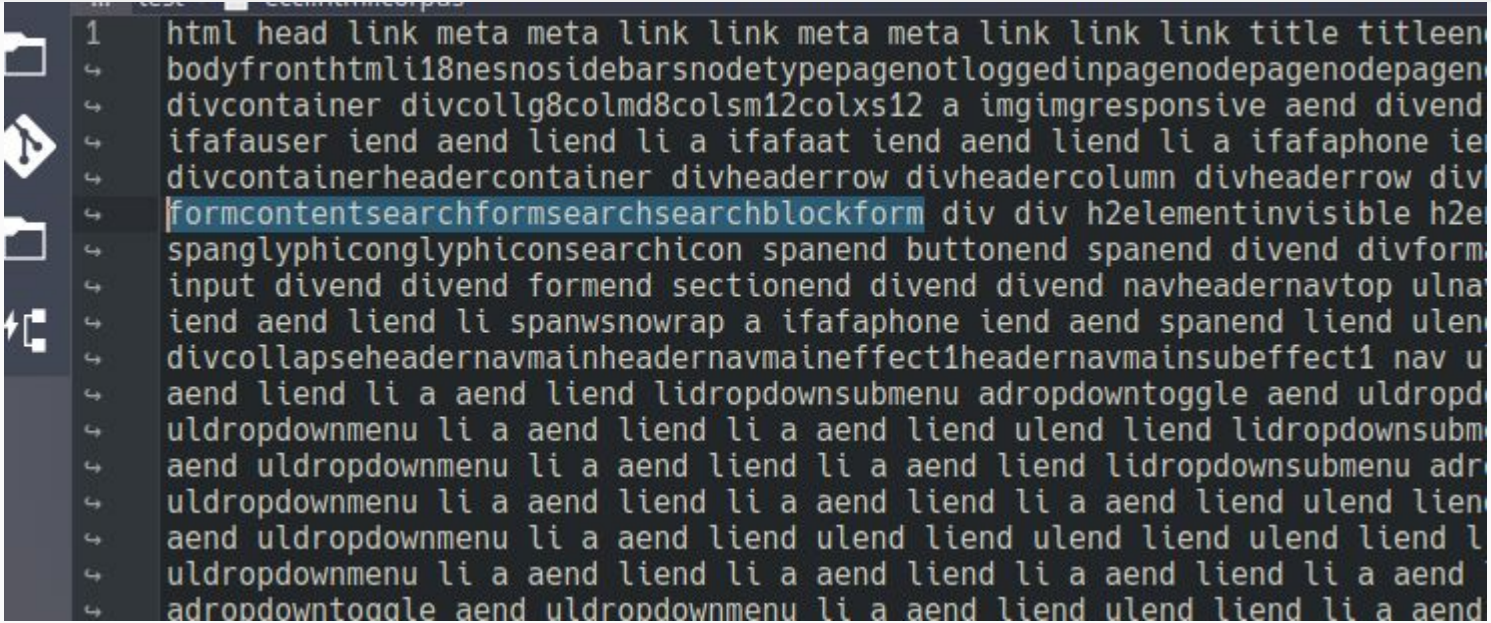
Ejemplo:


› Acceso
› Contáctenos
(506) 2511-8000
Buscar
Console
Debugger
Network
Style Editor
Performance
Memory
Storage
Accessibili
<section id="block-search-form--2" class="block block-search block-search clearfix">
::before
<form id="search-block-form" class="form-search content-search" action="/" method="post" accept-charset="UT

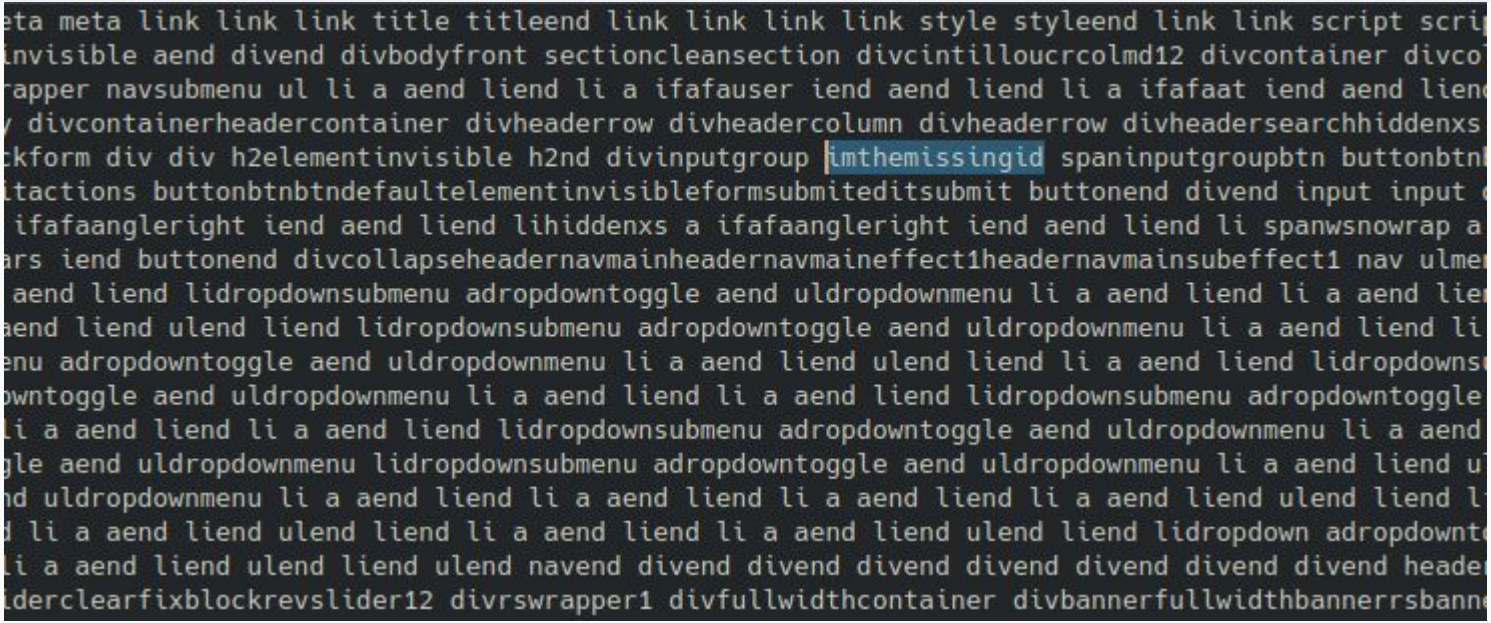
<h2 class="element-invisible">Formulario de búsqueda</h2>

<input id="edit-search-block-form--2" class="form-control form-text" title="Escriba lo que quiere bus
placeholder="Buscar" type="text" name="search_block_form" value="" size="15" maxlength="128">
…

Ejemplo:


1
html head link meta meta link link meta meta link link link title titleen
bodyfronthtmli18nesnosidebarsnodetypepagenotloggedinpagenodepagenodepagen
divcontainer divcollg8colmd8colsm12colxs12 a imgimgresponsive aend divend
ifafauser iend aend liend li a ifafaat iend aend liend li a ifafaphone ie
divcontainerheadercontainer divheaderrow divheadercolumn divheaderrow div
formcontentsearchformsearchsearchblockform div div h2elementinvisible h2e
spanglyphiconglyphiconsearchicon spanend buttonend spanend divend divform
input divend divend formend sectionend divend divend navheadernavtop ulna
iend aend liend li spanwsnowrap a ifafaphone iend aend spanend liend ulen
divcollapseheadernavmainheadernavmaineffect1headernavmainsubeffect1 nav u
aend liend li a aend liend lidropdownsubmenu adropdowntoggle aend uldropd
uldropdownmenu li a aend liend li a aend liend ulend liend lidropdownsubm
aend uldropdownmenu li a aend liend li a aend liend lidropdownsubmenu adr
uldropdownmenu li a aend liend li a aend liend li a aend liend ulend lien
aend uldropdownmenu li a aend liend ulend liend ulend liend ulend liend l
uldropdownmenu li a aend liend li a aend liend li a aend liend li a aend
adropdowntoggle aend uldropdownmenu li a aend liend ulend liend li a aend

Ejemplo:


eta meta link link link title titleend link link link link style styleend link link script scri
invisible aend divend divbodyfront sectioncleansection divcintilloucrcolmd12 divcontainer divco
rapper navsubmenu ul li a aend liend li a ifafauser iend aend liend li a ifafaat iend aend lien
divcontainerheadercontainer divheaderrow divheadercolumn divheaderrow divheadersearchhiddenxs
kform div div h2elementinvisible h2nd divinputgroup imthemissingid spaninputgroupbtn buttonbtn
tactions buttonbtnbtndefaultelementinvisibleformsubmiteditsubmit buttonend divend input input
ifafaangleright iend aend liend lihiddenxs a ifafaangleright iend aend liend li spanwsnowrap a
rs iend buttonend divcollapseheadernavmainheadernavmaineffect1headernavmainsubeffect1 nav ulme
aend liend lidropdownsubmenu adropdowntoggle aend uldropdownmenu li a aend liend li a aend lie
end liend ulend liend lidropdownsubmenu adropdowntoggle aend uldropdownmenu li a aend liend li
nu adropdowntoggle aend uldropdownmenu li a aend liend ulend liend li a aend liend lidropdowns
wntoggle aend uldropdownmenu li a aend liend li a aend liend lidropdownsubmenu adropdowntoggle
i a aend liend li a aend liend lidropdownsubmenu adropdowntoggle aend uldropdownmenu li a aend
le aend uldropdownmenu lidropdownsubmenu adropdowntoggle aend uldropdownmenu li a aend liend u
d uldropdownmenu li a aend liend li a aend liend li a aend liend li a aend liend ulend liend l
li a aend liend ulend liend li a aend liend li a aend liend ulend liend lidropdown adropdownt
i a aend liend ulend liend ulend navend divend divend divend divend divend divend divend heade
derclearfixblockrevslider12 divrswrapper1 divfullwidthcontainer divbannerfullwidthbannerrsbanne

Ejemplo:

Original Element ID: inputformcontrolformtexteditsearchblockform2

Expected New Element ID: imthemissingid

```
# Travel the new document
new_data = open('/content/drive/MyDrive/ecci.new.html.corpus', 'r', encoding='utf-8').read().split()
original_id = 'inputformcontrolformtexteditsearchblockform2'

new_id = None

for i in range(0, len(new_data) - window):
    words = ' '.join(new_data[i:i+window])
    if predict(words) == original_id:
        new_id = new_data[i+window]
        break
```

```
[ ] print('Could not find the new element' if new_id == None else 'The new ID is: ' + new_id)
```

```
The new ID is: imthemissingid
```

Gracias