

Proyecto WebScraping Integración Salas Corte Suprema

Javier Wilenmann
Universidad Adolfo Ibáñez

Introducción

- Corte Suprema funciona organizada en cuatros salas especializadas
- Función: estabilizar “jurisprudencia”
- Pero función afectada por alta variación en integración
- Cada sala está integrada por 5 ministros titulares, pero usualmente faltan por diversos motivos y deben “integrar” abogados externos
- Proyecto de investigación UAI: causas y efectos en dispersión de criterios



Sitio Web que contiene la información

<https://www.pjud.cl/tribunales/corte-suprema>

Datos Básicos

Localización


Datos Básicos

Dirección
📍 COMPAÑIA N° 1140 - 2° PISO

RUT
60301000-0

Teléfono
(02) 873-5000

📄



🔒 Permisos

📅 Programación de Salas

📅 Estados Diarios

🚩 Estado de Causa

🔄 Causas en Acuerdo

🏹 Causas Falladas

🔒 Visitadores 13-14

👤 Integraciones

🖥 Monitor de Salas

Integraciones

2019	Septiembre	17	Filtrar
Archivo	Fecha Integración	Descripción	
Sala 1.pdf 📄	17-09-2019	Sala 1	
Sala 2.pdf 📄	17-09-2019	Sala 2	
Sala 3.pdf 📄	17-09-2019	Sala 3	
Sala 4.pdf 📄	17-09-2019	Sala 4	

Cerrar

- ❖ Poder Judicial pone a disposición de usuarios información variada sobre Corte Suprema, incluyendo integraciones
- ❖ Pestaña integraciones se despliega en menú dinámico que filtra por año, mes y día
- ❖ Al presionar “filtrar” un año, mes y día, se despliega un PDF con el acta de integración de cada sala

Selección tecnología, estructura scrapping y almacenamiento de datos

- ❖ Extracción de información requiere paso por un menú dinámico de filtro y descarga de PDF
- ❖ Tecnología Scrapping: carácter dinámico justifica seleccionar Selenium
- ❖ Idea: Script debe llevar a bajar PDF con acta de integración de cada sala cada día relevante. Luego, proceso de extracción de datos
- ❖ Escalamiento, pseudo-código:
 - ❖ For año in años (2015:2021):
 - ❖ For mes in meses:
 - ❖ For día de funcionamiento en días del mes:
 - ❖ For sala in salas (1, 4):
 - ❖ Click descargar PDF, dejar en PDF-Files
- For PDF in PDF Files:
 - Buscar información relevante
 - Append in dataframe con
- Columnas df:
 - Fecha (datetime)
 - Sala (chr)
 - Integrantes (chr)
 - Justificaciones ausentes (chr)

Desarrollo loop

❖ Loop años y meses por selección Xpath a partir de lista:

```
for ano in anos:
    # Estos indices se obtienen de las opciones de la pagina, cada uno es un mes, cambiar para obtener mas meses
    #for index in range(1, 2):
        # Se envia la info a la pagina
    select_option(driver, '/html/body/div[11]/div/div/div[2]/form/div[2]/div[1]/div/select', str(ano))
for mes in meses:
    select_option(driver, '/html/body/div[11]/div/div/div[2]/form/div[2]/div[2]/div/select', mes)
    select = Select(driver.find_element(By.XPATH, "/html/body/div[11]/div/div/div[2]/form/div[2]/div[3]/div/select")) #get all the options into a list
    optionsList = []
```

Problema en funcionamiento del loop actual: tiende a descargar varias veces los archivos

❖ Loop días, al no ser siempre los mismos, a partir de lista obtenida por Python con valor que se despliegan tras seleccionar mes:

```
optionsList = []
for item in select.options:
    optionsList.append(item.get_attribute("value"))
for optionValue in optionsList:
    select_option(driver, '/html/body/div[11]/div/div/div[2]/form/div[2]/div[3]/div/select', optionValue
```

```
wait = WebDriverWait(driver, 10)
por salas cambia último tr: 1, 2, 3, 4 - no siempre está disponible, el mismo día, las mismas salas
try:
    boton_descarga = wait.until(EC.visibility_of_element_located((By.XPATH, '/html/body/div[11]/div/div/div[2]/table/tbody/tr[1]/td[1]/a')))
    boton_descarga.click()
except:
    pass
try:
    boton_descarga = wait.until(EC.visibility_of_element_located((By.XPATH, '/html/body/div[11]/div/div/div[2]/table/tbody/tr[2]/td[1]/a')))
    boton_descarga.click()
except:
    pass
try:
    boton_descarga = wait.until(EC.visibility_of_element_located((By.XPATH, '/html/body/div[11]/div/div/div[2]/table/tbody/tr[3]/td[1]/a')))
    boton_descarga.click()
except:
    pass
try:
    boton_descarga = wait.until(EC.visibility_of_element_located((By.XPATH, '/html/body/div[11]/div/div/div[2]/table/tbody/tr[4]/td[1]/a')))
    boton_descarga.click()
except:
    pass
```

❖ Salas no fue posible hacer con loop, sino con estrategia de intentar pulsar sobre cada elemento desplegado, por Xpath:

Estrategia extracción información PDFs

- ❖ PDFs no escaneados: PDF read
- ❖ Patrones:
 - ❖ “INSTALACI”: permite reconocer si es tipo de acta búsqueda
 - ❖ Lo que antecede a SALA: número de sala
 - ❖ Lo que sigue a ÑOR: nombre integrantes (problema: mujeres con A antes)
 - ❖ Lo que sigue a firma digital: fecha en palabras
 - ❖ Causas ausencia: extracción a partir de lista
- ❖ Trabajo de limpieza



CORTE SUPREMA ACTA DE INSTALACIÓN

PRIMERA SALA

Santiago, veintiuno de julio de dos mil veintiuno

En cumplimiento a lo dispuesto en el N°1 del artículo 105 del Código Orgánico de Tribunales, se procedió a instalar la **Primera Sala** de la Corte Suprema, quedando constituida como sigue:

PRESIDENTE:	SEÑORA MAGGI
MINISTRO:	SEÑORA MUÑOZ SEÑOR SILVA C. SEÑORA REPETTO
AB. INTEGRANTE:	SEÑOR MUNITA

No asisten los Ministros señora **Egnem** y señor **Fuentes** por estar en comisión de servicio, para cumplir funciones en el Tribunal Calificador de Elecciones, con motivo de las elecciones Primarias para Presidente de la República, realizadas el día 18 de julio de 2021.

Para debido testimonio se levanta la presente acta que autoriza el señor Secretario.

Estrategia extracción información PDFs escaneados

- ❖ PDFs no escaneados: OCR
- ❖ Patrones:
 - ❖ “INSTALACI”: permite reconocer si es tipo de acta búsqueda
 - ❖ Presidente, Ministros, Abogados: lo que sigue a esos conceptos
 - ❖ Fecha: lo que sigue a “En Santiago”
 - ❖ Causas: extracción por lista

```
palabras = ['licencia', 'comisión', 'permiso', 'feriado', 'inhabilidad', 'subroga', 'vacancia']
```

- ❖ Trabajo de limpieza



CORTE SUPREMA ACTA DE INSTALACION

PRIMERA SALA
En Santiago, ~~catorce~~ de abril de dos mil quince, en cumplimiento a lo dispuesto en el N°1 del artículo 105 del Código Orgánico de Tribunales, se procedió a instalar la PRIMERA SALA de la Corte Suprema, quedando constituida como sigue:

PRESIDENTE: SEÑOR SEGURA.

MINISTROS: SEÑORES VALDÉS, SILVA SEÑORA MAGGI y EL

ABOGADO INTEGRANTE: SEÑOR FIGUEROA.

No asiste el Ministro señor Fuentes por hacer uso de permiso.

Para debido testimonio se levanta la presente acta que autoriza señora Secretaria.

SECRETARIA

PRESIDENTE

	Sala	Presidente	Integrantes	Fecha2	Integrante1	Integrante2	Integrante3	Integrante4	Causa1	Causa2	Causa3
0	PRIMERA	SEGURA	VALDES SILVA FUENTES LECAROS	2015-01-12 00:00:00	VALDES	SILVA	FUENTES	LECAROS	feriado	NaN	NaN
1	PRIMERA	SEGURA	VALDES SILVA MAGGI LECAROS	2015-01-15 00:00:00	VALDES	SILVA	MAGGI	LECAROS	permiso	NaN	NaN
2	PRIMERA	VALDES	SILVA ARANGUIZ BARA ONA LECAROS	2015-01-29 00:00:00	SILVA	ARANGUIZ	BARAONA	LECAROS	licencia	permiso	feriado
3	PRIMERA	SEGURA	VALDES SILVA SEEÑORA MAGGI LECAROS	2015-01-14 00:00:00	VALDES	SILVA	SEEÑORA	MAGGI	permiso	NaN	NaN
4	PRIMERA	SEGURA	VALDES SILVA LAGOS PRIETO	2015-01-19 00:00:00	VALDES	SILVA	NaN	LAGOS	NaN	NaN	NaN
...
2310	CUARTA	DOMESTCH	— JUICA MUÑOZ CARREÑO	2017-11-03 00:00:00	—	JUICA	MUÑOZ	CARREÑO	comisión	feriado	NaN
2311	CUARTA	DOLMESTCH	CARREÑO KUNSEMULLER SILVA	2017-12-15 00:00:00	CARREÑO	KUNSEMULLER	SILVA	NaN	comisión	permiso	NaN
2312	CUARTA	DOLMESTCH	JUICA MUÑOZ VALDES CARREÑO	2017-12-22 00:00:00	JUICA	MUÑOZ	VALDES	CARREÑO	NaN	NaN	NaN
2313	CUARTA	DOLMESTCH	JUICA MUÑOZ CARREÑO	2017-12-29 00:00:00	JUICA	MUÑOZ	CARREÑO	NaN	licencia	comisión	feriado
2314	CUARTA	DOLMES TCH	JUICA CARREÑO KUNSEMULLER	2017-12-07 00:00:00	JUICA	CARREÑO	KUNSEMULLER	NaN	licencia	comisión	NaN

2315 rows × 11 columns