

# Proyecto WebScraping Integración Salas Corte Suprema

Javier Wilenmann  
Dany Rubiano  
Universidad Adolfo Ibáñez

# Introducción

- El proyecto sirve de insumo a un proyecto de investigación de un profesor de en torno a la pregunta: ¿Por qué se producen tantas diferencias entre fallos del tribunal supremo en Chile?
- Corte Suprema. Función teórica: estabilizar “jurisprudencia”. Hacer que casos similares difíciles sean resueltos del mismo modo. En Chile esta función no se cumple. La propia Corte resuelve de modo disímil. ¿Por qué?
- El trasfondo de la investigación se encuentra en una hipótesis organizacional. La dispersión está condicionada por ciertas decisiones de organización y funcionamiento.
- La Corte Suprema funciona organizada en cuatro salas especializadas (ver foto), cada una de las cuales está encargada de estabilizar jurisprudencia en un ámbito específico (civil, penal, público).
- Pero el cumplimiento de la función está 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. Rotación condiciona criterios.
- 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. Pero bajar PDFs con Selenium es costoso y poco confiable. Segunda tecnología: usar requests para bajar directamente los PDFs con información de URL obtenida por 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):
            - ❖ Selenium: obtener URL de cada PDF
            - ❖ Usar lista para descargar cada una con requests
- 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 a partir de lista de años a considerar. Meses y días por obtención de lista de opciones

```
for anio in list_anios:
    select_option(driver, '/html/body/div[11]/div/div/div[2]/form/div[2]/div[1]/div/select', str(anio))
    select_mes = Select(driver.find_element(By.XPATH, '/html/body/div[11]/div/div/div[2]/form/div[2]/div[2]/div/select')) #get all the options into a list
    options_meses_list = [item.get_attribute("value") for item in select_mes.options]
    options_meses_list.remove('null')
    list_meses = list(dict_mes_dias.keys()) # meses ya scrapeados
    # Obtencion de meses a considerar para la consulta
    option_meses = compare_lists(options_meses_list, list_meses)

# Loop de meses
for mes in option_meses:
    select_option(driver, '/html/body/div[11]/div/div/div[2]/form/div[2]/div[2]/div/select', mes)
    select_dias = 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
    options_dias_list = [item.get_attribute("value") for item in select_dias.options]
    options_dias_list.remove('null')
    list_dias = dict_mes_dias[mes] if mes in dict_mes_dias.keys() else [] # dias ya scrapeados

# Obtencion de dias a considera para la consulta
option_dias = compare_lists(options_dias_list, list_dias)
```

# Obtención de información y descarga PDFs

## ❖ Utilización de Selenium para obtener URL de PDFs

```
# Loop de dias a consultar
for optionValue in option_dias:
    select_option(driver, '/html/body/div[11]/div/div/div[2]/form/div[2]/div[3]/div/select', optionValue)
    driver.find_element(By.XPATH, '/html/body/div[11]/div/div/div[2]/form/div[2]/div[4]/button').click()
    wait = WebDriverWait(driver, 10)

# Loop que recorre las 4 salas
for i in range(1,5):
    try:
        boton_descarga = wait.until(EC.visibility_of_element_located((By.XPATH, f"/html/body/div[11]/div/div/div[2]/table/tbody/tr[{i}]/td[1]/a")))
        url_downloads.append(boton_descarga.get_attribute('href')) # Agrega link de descarga a lista
    except:
        pass
```

## ❖ Utilización de Requests, a partir de lista generada por Selenium, para descargar

```
def downloader_files(list_urls, dir_output):
    """Recibe la lista de urls para descarga de los archivos de pdf y los guarda en la carpeta dada"""

    for url_file in list_urls:
        urllib.request.urlretrieve(url_file, f"{dir_output}/acta_{''.join(filter(str.isdigit, url_file))}.pdf")
```

# 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

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

❖ 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

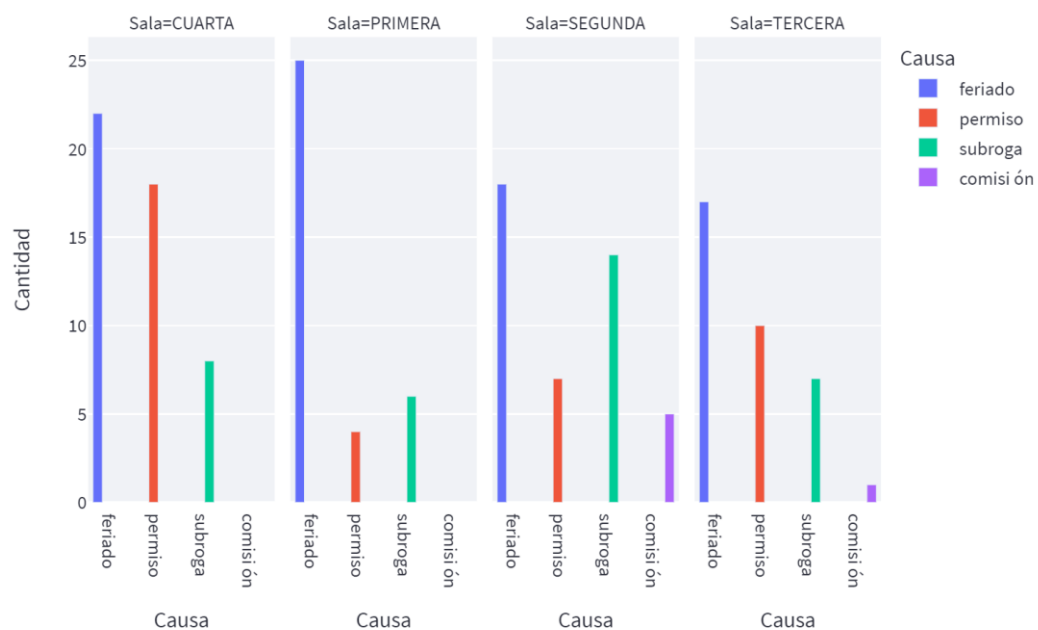


	Sala	Presidente	Integrantes	Fecha2	Integrante1	Integrante2	Integrante3	Integrante4	Causa1	Causa2	Causa3
0	PRIMERA	SEGURA	VALDES SILVA FUENTES LECAROS	2015-01-12 00:00:00	VALDES	SILVA	FUENTES	LECAROS	feriado	NaN	NaN
1	PRIMERA	SEGURA	VALDES SILVA MAGGI LECAROS	2015-01-15 00:00:00	VALDES	SILVA	MAGGI	LECAROS	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

# APP para visualización y obtención sencilla de información

- ❖ Disponible en: <https://scraping-cs-eoakmwhgya-uc.a.run.app/>
- ❖ Genera visualizaciones sencillas agregadas de causas de ausencia por período y por sala



- ❖ Permite descargar archivos con .csv



Descargar resultados