

Asignatura: Tipología y Ciclo de Vida de los Datos
Período: 2021-I
Evaluación: Práctica 1
Alumnos: Rodolfo Méndez Marcano (RMM) y Concha Román Santos (CRS)
Fecha: 08 / 11 / 2021

Para la selección del ejemplo a tratar con Web Scraping en esta práctica, los miembros de este equipo consideramos varios ejemplos relacionados con nuestra experiencia profesional común con BBVA en las áreas de la estadística y la econometría (RMM) y la ciencia de datos (CRS). En particular, el ejemplo finalmente seleccionado, vinculado al cambio climático, se relaciona directamente con la participación de uno de nosotros ([RMM](#)) en proyectos en este campo junto al equipo de BBVA Research (específicamente con su Cluster de Economía del Cambio Climático) dirigidos a proveer al banco de información e indicadores sobre el cambio climático, sus consecuencias y las políticas públicas asociadas que sirvan (una vez incorporados a sus [cuadros de mando](#)) para guiar a sus directivos en la adaptación de las estrategias del banco a los mismos, además de cumplir con las [exigencias regulatorias](#) que le obligan, en tanto banco de dimensiones sistémicas, a adecuar su capital y provisiones antes los llamados riesgos de transición y físicos asociados al cambio climático.

Por otra parte, nuestra dinámica de trabajo en equipo se vio muy beneficiada al estar ambos habituados a desarrollar nuestro trabajo profesional, y en especial el desarrollo de modelos, códigos y aplicaciones, bajo los principios de la metodología [Agile](#) (alternando en el caso de la práctica el rol de desarrollador y usuario entre ambos).

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información.

1.1. Explicar en qué contexto se ha recolectado la información

La data recolectada mediante scraping procede de la página o sitio web del [UNFCCC](#) (United Nations Framework Convention for Climate Change), el grupo de trabajo de la UN (United Nations Organization) creado (por el convenio al que hace referencia su nombre) hace casi 30 años para coordinar en adelante los esfuerzos conjuntos de los países miembros de la organización (casi todos los del mundo) dirigidos a afrontar los retos planteados por el cambio climático o calentamiento global que está experimentando el planeta.

Más específicamente, la data procede de la subpágina del sitio correspondiente al [NDC Registry](#) el registro de los documentos contentivos de los [NDCs](#) (Nationally Determined Contributions) de los países suscriptores del Paris Agreement y **cuyos URLs (es decir, los identificadores de estos documentos en dicho registro) constituyen el contenido principal de nuestra base de datos.**

El Paris Agreement, como el Kyoto Protocol antes, forma parte de los acuerdos intergubernamentales sobre cambio climático cuya elaboración, firma y seguimiento **constituyen la razón de ser de la existencia del UNFCCC**, para lo cual este organiza periódicamente reuniones con los representantes de los gobiernos suscriptores (llamadas Conferences of Parties, COPs: la más reciente, el COP26, realizada en Glasgow entre el 31 de Octubre y el 12 de Noviembre de 2021).

En particular, el Paris Agreement se suscribió en la COP21 realizada en París en 2015 y ha sido firmado por casi todos los Gobiernos del mundo desde entonces, comprometiéndose así todos ellos con el objetivo central del Acuerdo: **reducir a cero las emisiones netas de gases de efecto invernadero hacia la atmósfera causadas por la actividad humana antes de finalizar el 2050**. De acuerdo al [IPCC](#) (Intergovernmental Panel on Climate Change), brazo científico de la UNFCCC, alcanzar este objetivo constituye una condición necesaria para detener el calentamiento global antes de que sus efectos adversos para la humanidad alcancen niveles catastróficos.

1.2. Explicar por qué el sitio web elegido proporciona dicha información.

Uno de los principales instrumentos del Paris Agreement para conseguir su objetivo (Artículo 4, Parágrafo 2) es el imponer a cada país firmante la obligación de elaborar y mantener actualizados **y a disposición pública** su NDC, que no es más que un documento en el que el país explicita (1) la magnitud de la reducción de emisiones desde su territorio nacional (a partir 2021) con las que se compromete (y que constituiría su contribución a la consecución del objetivo del Acuerdo) y (2) la estrategia de políticas públicas que llevara adelante para cumplir con dicho compromiso.

Cómo ya se ha explicado, el UNFCCC es la entidad oficial que coordina y da seguimiento al Paris Agreement, es en su site donde los Gobiernos deben hacer públicos, tal como exige el Acuerdo, sus NDCs, y más específicamente en el apartado denominado [NDC Registry](#), que es el repositorio oficial para estos documentos.

2. Título. Definir un título que sea descriptivo para el dataset.

El título general de la base de datos es:

- NationallyDeterminedContributions_URLs

Y el nombre del file que la contiene es:

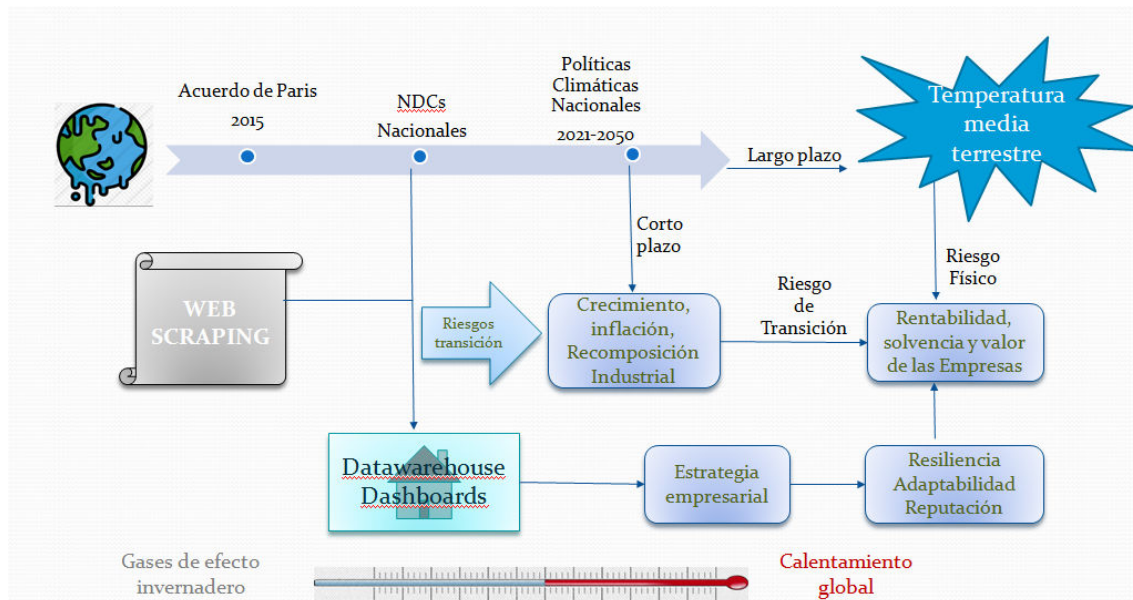
- NDC_URL_DiaMesAño.csv

3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido.

Se trata de una base de datos de los enlaces (URLs) a los documentos en formato PDF (en todas sus distintas versiones y actualizaciones a la fecha indicada en la base de datos) que contienen los objetivos de reducción de gases de efecto invernadero para los próximos años (Nationally Determined Contributions, NDCs) a los que se compromete cada uno de los países firmantes del **Paris Agreement** junto a un esbozo de la estrategia que seguirán para alcanzar dichos objetivos.

4. Representación gráfica. Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.

El siguiente diagrama ilustra el lugar e importancia de los NDCs (y de nuestra base de datos de los URLs de los mismos) desde la perspectiva empresarial que nos inspiró.



Interpretación (partiendo de la esquina superior inversa)

- El calentamiento global y sus efectos adversos para la humanidad han llevado a los países a suscribir el Paris Agreement
- El Acuerdo obliga a cada país a elaborar y mantener actualizado y a disposición pública su NDC: documentos en PDF donde cada país explicita su objetivo de reducción de las emisiones originadas en su territorio y la estrategia de políticas públicas que seguirá para conseguirlo.
- Estas metas y estrategias públicas impactarán la rentabilidad, solvencia y valor de las empresas tanto a corto, mediano y largo plazo: a corto plazo (lo que se denomina Riesgo de Transición), pues la decarbonización puede impactar negativamente al desempeño macroeconómico de las economías, penalizando más a las actividades económicas más dependientes (directa o indirectamente) de las emisiones, y a largo plazo (Riesgos Físicos), porque un aumento sustancial del calentamiento global alterará la habitabilidad y productividad relativa de las distintas regiones del planeta y de las distintas ramas de la actividad productiva.
- Por lo tanto la viabilidad de las empresas requerirá que adopten con suficiente anticipación estrategias de relocalización de sus activos, inversiones y operaciones entre regiones y ramas económicas para minimizar los impactos referidos

- Por todo lo anterior, debe ser claro que la información de los NDCs resulta clave para que los directivos logren este reajuste óptimo de sus estrategias, y nuestra práctica se dirige precisamente a facilitarles el uso de la misma.
- Concretamente, nuestra práctica utiliza el Web Scraping para construir una base de datos con los URLs a los NDCs de los distintos países, para facilitar incorporar selectivamente (sólo los países de interés) esta información en el almacén de datos y/o los cuadros de mando de los directivos de las empresas.

5. **Contenido.** Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

5.1. Campos y periodo de tiempo

Se trata de una base relacional muy simple (una única tabla) en la que cada registro (fila) se refiere a un documento PDF (alguna de las versiones del NDC de un país) y contiene los siguientes campos (columnas):

- **Country:** Nombre del País al que corresponde el NDC
 - Formato: Texto
 - Ejemplo valores: Afghanistan, Albania, Andorra, Angola, ...
- **Version:**
 - Formato: Texto
 - Ejemplo valores: First, First (Archived), First (Updated Submission), Second, etc
- **Language:**
 - Formato: Texto
 - Ejemplo valores: English, Spanish, etc
 -
- **Date:** Fecha de subida del documento al repositorio
 - Formato: DD/MM/YYYY
 - Ejemplo valores: 24/03/2016, 31/10/2021
 -
- **PDF:** URL a NDC en formato PDF
 - Formato: Texto
 - Ejemplo registro (European Union):
https://www4.unfccc.int/sites/ndcstaging/PublishedDocuments/European%20Union%20First/EU_NDC_Submission_December%202020.pdf

A la fecha de la última actualización de la base de datos los documentos disponibles abarcaban el período desde el 24 de Marzo de 2016 hasta el 04 de Noviembre de 2021.

5.2. Recogida de los datos

Como ya se ha explicado los datos (URLs a los NDCs) provienen de la página web del NDC_Registry (ver captura de pantalla siguiente) y, al no estar disponibles directamente ni en tabla ni en fichero para descargar, se ha debido recurrir a la técnica de Web Scraping, optando utilizar **Python** y algunas de sus librerías tanto para ello como para el ensamblaje de la base de datos final.

A. Fracaso del Static Web Scraping

Ahora bien, por razones que explicamos a continuación, una estrategia simple de Web Scraping Estático (directamente implementable con la librería **Beautiful Soup**) no resultó adecuada a nuestros propósitos (era imposible acceder los datos) y debió emplearse entonces una estrategia más compleja de Web Scraping Dinámico (para lo cual requerimos utilizar la librería **Selenium**).

La clave la encontramos en que (ver [referencia](#)) una estrategia de SWS (Static Web Scraping) requiere que todos los datos a descargar se muestren directamente en el web site, lo que se traduce en que también se muestren desplegados en el código fuente que puede descargarse manualmente de la página web (botón derecho del mouse). Sólo en ese caso, como nos tocó descubrir, Beautiful Soup es capaz de mostrar en el árbol de objetos Python que crea, los datos buscados.

Por el contrario, cuando, como en nuestro caso (ver capturas de pantallas siguientes), el acceso a distintas porciones de los datos (la localización de los documentos PDF) requiere hacer click en distintos sublinks en el código fuente directamente descargable los distintos segmentos de datos se encuentran plegados entre líneas de instrucciones de JavaScript que, al ejecutarse, son las encargadas de desplegarlos y mostrar los datos.

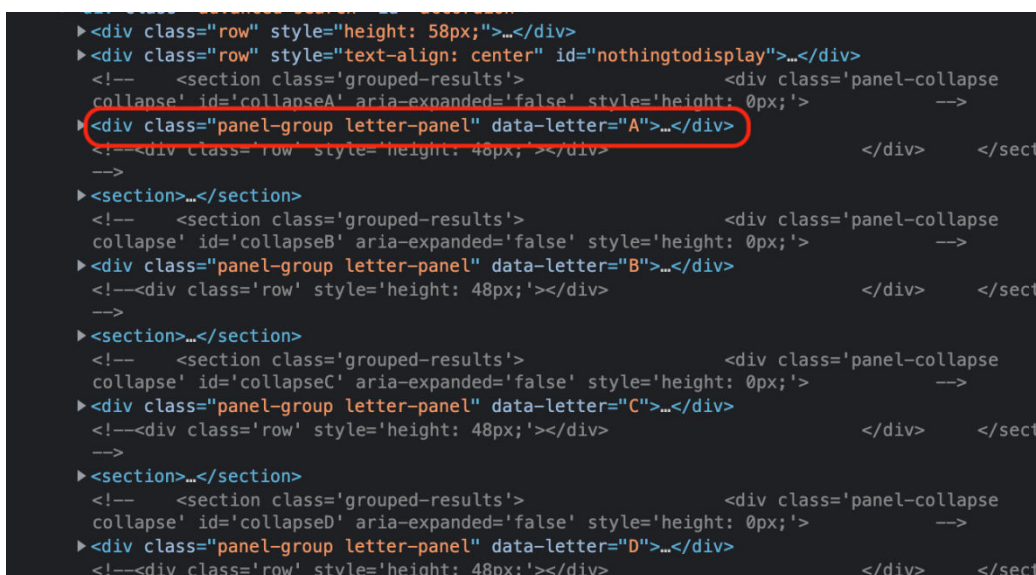
B. Nuestra estrategia de Dynamic Web Scraping

La solución al problema fue recurrir al DWS (Dynamic Web Scraping) con la ayuda de Selenium, la cual consiste en lo siguiente:

- Se utiliza Selenium para obtener una versión completamente desplegada del código fuente de la página web, para lo cual la librería simula la apertura de la página en el browser (Chrome en nuestro caso) y la realización de click con el mouse en cada uno de los sublinks de la página desplegándolos (rendering).
- Con código fuente así desplegado (y por tanto sin líneas de JavaScript insertas en él) ya puede procederse a utilizar BeautifulSoup de la manera usual para transformar el mismo en árbol de objetos de Python a través de cuyos nodos puedan localizarse y extraerse los datos de interés tal y como se describe en el resto de esta sección.

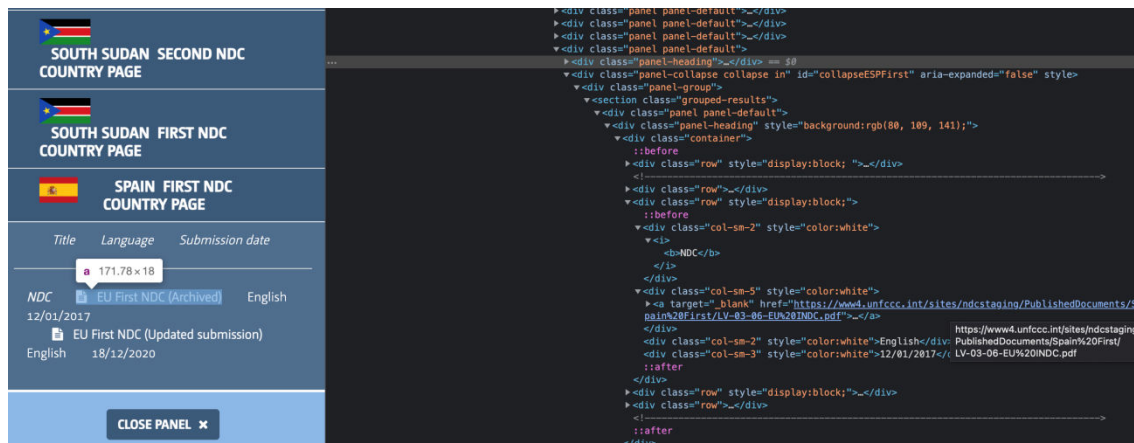
C. Extracción de los datos del árbol de objetos (soup)

Para extraer los datos era necesario entender con precisión la jerarquía de nodos que conducía a ellos, para lo cual realizamos una inspección del código mediante el navegador. Nos encontramos con que el código html de la página web se dispone en múltiples bloques “div” anidados. Los paneles de los países que tienen la misma inicial se incluyen dentro de un mismo div cuya clase es “panel-groupletter-panel”. Por tanto, hay una clase de este tipo para cada letra, identificados mediante el campo “data-letter” como podemos ver en la siguiente imagen.



```
<div class="row" style="height: 58px;">...</div>
<div class="row" style="text-align: center;" id="nothingtodisplay">...</div>
<!-- <section class='grouped-results'> <div class='panel-collapse
collapse' id='collapseA' aria-expanded='false' style='height: 0px;'> -->
<div class="panel-group letter-panel" data-letter="A">...</div>
<!--<div class='row' style='height: 48px;'></div> --> </div> </sect
-->
<section>...</section>
<!-- <section class='grouped-results'> <div class='panel-collapse
collapse' id='collapseB' aria-expanded='false' style='height: 0px;'> -->
<div class="panel-group letter-panel" data-letter="B">...</div>
<!--<div class='row' style='height: 48px;'></div> --> </div> </sect
-->
<section>...</section>
<!-- <section class='grouped-results'> <div class='panel-collapse
collapse' id='collapseC' aria-expanded='false' style='height: 0px;'> -->
<div class="panel-group letter-panel" data-letter="C">...</div>
<!--<div class='row' style='height: 48px;'></div> --> </div> </sect
-->
<section>...</section>
<!-- <section class='grouped-results'> <div class='panel-collapse
collapse' id='collapseD' aria-expanded='false' style='height: 0px;'> -->
<div class="panel-group letter-panel" data-letter="D">...</div>
<!--<div class='row' style='height: 48px;'></div> --> </div> </sect
```

Dentro de cada clase “panel-groupletter-panel” encontramos un div de clase “panel-panel-default” para cada panel de cada país. Y, a su vez, dentro de cada clase “panel-panel-default” encontramos un div de clase “panel-heading” que contiene los textos de los títulos y otro div de clase “panel-collapsecollapse in” que es la que contiene todos los datos que estamos buscando, el nombre del país, el nombre de versión del fichero, la fecha de subida del fichero y la url del fichero pdf. Podemos verlo en la siguiente imagen.



Como ya comentamos, intentamos inicialmente obtener este contenido mediante BeautifulSoup (la ya comentada librería de Python para "scrapear" documentos html), pero no logramos de ese modo obtener los datos requeridos de la página web (la soup que arroja BS no los contenía). Investigando comprendimos el origen del problema (codigo fuente plegado y necesidad de ejecutar JavaScript para desplegarla). La solución como ya se explico la encontramos en utilizar Selenium (y automatizar así Chrome).

De manera, que en adición a BeautifulSoup instalamos el paquete Selenium y el driver de Chrome y, de esta manera pudimos el contenido del html desplegado de la url indicada. A continuación, con el uso de la librería BeautifulSoup pudimos recorrer el html extrayendo la información de interés para nuestra práctica. El código implementado recorre los div anidados descritos anteriormente extrayendo el contenido.

El contenido obtenido se añade a un dataframe, empleando la librería Pandas, en el que se hacen una serie de transformaciones y procesamiento de los datos (documentado detalladamente en el notebook ipynb) para dividir los campos que queremos tener en la base de datos, limpiar algunos registros y, finalmente se exporta el dataset a un fichero Excel y un fichero csv.

D. Un último escollo

Vale la pena mencionar una dificultad adicional que surgió en esta última etapa de ensamblaje de la base de datos con Pandas, y que puede llevar fácilmente a error a quien se aproxime al problema de una manera puramente computacional y mecánica. Mientras que los registros para todos los países no europeos estaban correctos, los de los países pertenecientes a la Unión Europea al descargarlos aparecían con la etiqueta EU en lugar del nombre del país y adicionalmente aparecía un registro individual propio de la Unión Europea (con etiqueta "European Union") como tal. Optamos entonces por acceder a cada uno de los PDFs de todos estos países y de la UE y compararlos, para venir a descubrir (y así lo confirmamos con investigación adicional) que todos los países de la UE en realidad comparten un único NDC (en distintas versiones) y por lo tanto se podía eliminar sin pérdida de información todos los registros con etiqueta EU y dejar sólo el registro con la etiqueta European Union (con el cuidado eso sí, de incluir una nota al respecto en el README del repositorio de github).

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

6.1. Propietario del conjunto de datos

Los datos (NDCs) son documentos elaborados y guardados en el NDC_Registry (dentro del sitio web del UNFCCC) por los Gobiernos suscriptores del Paris Agreement.

6.2. Análisis similares

Respecto al uso de los datos por otros investigadores o personas, nuestra búsqueda (en internet y consultando a expertos) arroja los siguientes resultados:

- No encontramos casos de web scraping de los datos en cuestión (los URLs de los NDCs en el NDC_Registry) o de bases de datos que los contengan y faciliten incorporar estos documentos en dashboards adaptadas a las necesidades específicas de empresas como BBVA.
- Hay algunos portales que presentan información extraída de los documentos PDF en cuestión, incluyendo visualizaciones de alguna información numérica contenida en ellos:
 - <https://www.climatewatchdata.org/ndcs-explore>
 - <https://climatedata.imf.org/pages/re-indicators/#re1>
- En general, los reportes y análisis que evalúan el grado de cumplimiento del Paris Agreement hacen siempre referencia a la información de los NDCs sobre la magnitud de las reducciones de gases de efecto invernadero a las que se comprometen los países suscriptores del acuerdo, y especialmente a la suma de todas ellas. El más reciente de estos reportes de que tenemos conocimiento es el Emissions GAP Report de 2021 publicado por el UNEP (United Nations Environment Program), que demuestra la insuficiencia de la suma de las reducciones propuestas por los NECs actuales para evitar un incremento de la temperatura media terrestre inferior a 2,7 grados Celsius (muy superior al máximo de 2 al que aspira el Acuerdo), y que puede encontrarse en el siguiente link:
 - <https://www.unep.org/resources/emissions-gap-report-2021>

Para asegurarnos de que nuestra descarga y uso de los datos no violara ningún principio ético ni norma legal realizamos varias tareas:

(A) Analizamos el párrafo 12 del artículo 4 del Paris Agreement, confirmando que hace explícita la naturaleza pública de los NDCs cargados al registro oficial (el NDC_Registry)

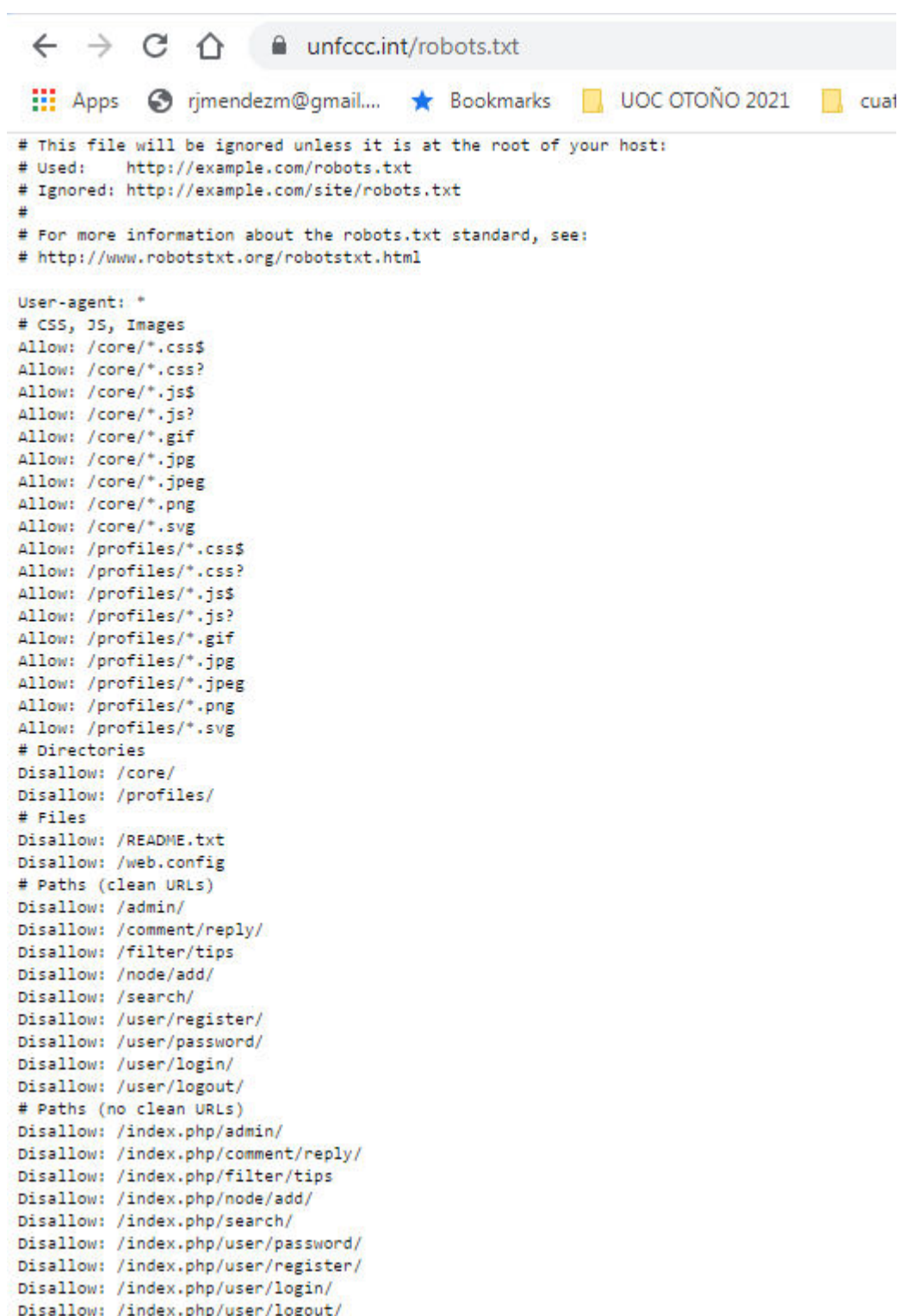
(B) Buscamos información confirmación en el propio NDC_Registry, hallándola en su sección [FAQ](#), como puede verse en la siguiente captura:

Who can access information published in the registry?

A public page of the registry is accessible to any interested users to browse and search information on submitted NDCs. Users can search information by Party or keywords using the search box.

(C) Consultamos el fichero robots.txt de la página del UNFCCC donde esta el NDC_Registy, mediante la siguiente consulta de la dirección <https://unfccc.int/robots.txt> (ver captura de pantalla siguiente). Se verifico que la localización de los NDCs (/sites/NDCStaging/Pages/All.aspx) no estaba entre los sub-sites o directorios con etiqueta DISALLOW, y por ende no hay restricciones a tareas de crawling o scraping sobre el mismo.

Ilustración 4: Fichero robots.txt de UNFCCC.INT



```
# This file will be ignored unless it is at the root of your host:
# Used: http://example.com/robots.txt
# Ignored: http://example.com/site/robots.txt
#
# For more information about the robots.txt standard, see:
# http://www.robotstxt.org/robotstxt.html

User-agent: *
# CSS, JS, Images
Allow: /core/*.css$
Allow: /core/*.css?
Allow: /core/*.js$
Allow: /core/*.js?
Allow: /core/*.gif
Allow: /core/*.jpg
Allow: /core/*.jpeg
Allow: /core/*.png
Allow: /core/*.svg
Allow: /profiles/*.css$
Allow: /profiles/*.css?
Allow: /profiles/*.js$
Allow: /profiles/*.js?
Allow: /profiles/*.gif
Allow: /profiles/*.jpg
Allow: /profiles/*.jpeg
Allow: /profiles/*.png
Allow: /profiles/*.svg
# Directories
Disallow: /core/
Disallow: /profiles/
# Files
Disallow: /README.txt
Disallow: /web.config
# Paths (clean URLs)
Disallow: /admin/
Disallow: /comment/reply/
Disallow: /filter/tips
Disallow: /node/add/
Disallow: /search/
Disallow: /user/register/
Disallow: /user/password/
Disallow: /user/login/
Disallow: /user/logout/
# Paths (no clean URLs)
Disallow: /index.php/admin/
Disallow: /index.php/comment/reply/
Disallow: /index.php/filter/tips
Disallow: /index.php/node/add/
Disallow: /index.php/search/
Disallow: /index.php/user/password/
Disallow: /index.php/user/register/
Disallow: /index.php/user/login/
Disallow: /index.php/user/logout/
```

7. Inspiración. Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

Cómo se ha explicado en el preámbulo de este documento, nuestra inspiración surge de nuestro trabajo con BBVA, y en particular de la participación de uno de nosotros (RMM) junto al Cluster de Cambio Climático de BBVA Research en la recolección, desarrollo e incorporación en cuadros de mando (consultar el siguiente [link](#)) de indicadores de cambio climático que ayuden a los gestores a adaptar las estrategias del banco al cambio climático y sus efectos.

Y como se explica e ilustra diagramáticamente en el punto 4, los NDC proveen información de gran utilidad para guiar a los directivos en el diseño de las estrategias de adaptación al cambio climático y nuestra base de datos con los URLs a los mismos facilitarán su incorporación a los cuadros de mando que los directivos utilizan para este diseño.

Entre las preguntas de interés podrá responder una empresa como BBVA a partir de la información de los NDC de los países de su interés (donde tiene inversiones y/o negocios actualmente o evalúa tenerlos en el futuro) se encuentran las siguientes:

- En primer lugar extraer de los documentos de los países de interés información que revele oportunidades y amenazas para el banco, creando con ellos indicadores particulares a incorporar en sus dashboards (como lo hacen los portales referidos en el punto 6).
- De manera similar los reportes de seguimiento del cumplimiento del París Agreement referidos en el punto 6 (por ejemplo, el Emission GAP Report), pero focalizándose en los países de interés, podrá evaluar el grado de cumplimiento del compromiso de decarbonización plasmado en el documento comparándolo con los indicadores de emisiones observadas (incorporados al mismo dashboard) y en función de ello y de criterio experto evaluar si ellos significara un debilitamiento del grado de compromiso o más bien apuntará a un ajuste al alza de la intensidad de la decarbonización en los años siguientes para compensar.
- Conocer anticipadamente las políticas públicas específicas de los Gobiernos para descarbonizar le permitirá a las empresas que operan en dicho país identificar oportunidades de inversión "verdes". Por ejemplo, un banco que opera localmente podrá conocer que infraestructuras verdes proyecta construir el Gobierno o tipo de empresas verdes planea subsidiar, y dirigir su capacidad crediticia hacia los mismos.

8. **Licencia.** Seleccionar una de estas licencias para el dataset resultante y justificar el motivo de su selección:

- Released Under CC0: Public Domain License.
- Released Under CC BY-NC-SA 4.0 License.
- Released Under CC BY-SA 4.0 License.
- Database released under Open Database License, individual contents under Database Contents License.
- Other (specified above).
- Unknown License.

Dado que deseamos que la base de datos y su contenido pueda ser utilizado con la mayor libertad por cualquier persona se ha seleccionado distribuir la base de datos como un todo bajo la licencia [Open Database License](#) (ODbL) y sus contenidos individuales bajo la licencia asociada [Database Contents Licence](#) (DbCL).

Las licencias ODbL/DbCL (ambas parte del [Open Data Commons](#)) otorgan libertad a cualquier persona de compartir, transformar y utilizar la base de datos libremente y de cualquier forma con la condición esencial de que cualquier distribución que la persona haga de la versión original o modificada de esta base de datos también se rija explícitamente por los términos de las licencias ODbL/DbCL.

9. **Código.** Adjuntar en el repositorio Git el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

Ver https://github.com/rjmendezm/NationallyDeterminedContributions_URLs

10. **Dataset.** Publicar el dataset obtenido(*) en formato CSV en Zenodo con una breve descripción. Obtener y adjuntar el enlace del DOI.

Ver <https://drive.google.com/file/d/1eJMoniUrB-iWIWpvsIF1HKPcFFAzHMOK/view>

Contribuciones	Firmas
Investigación previa	RMM, CRS
Redacción de las respuestas	RMM, CRS
Desarrollo del código	RMM, CRS