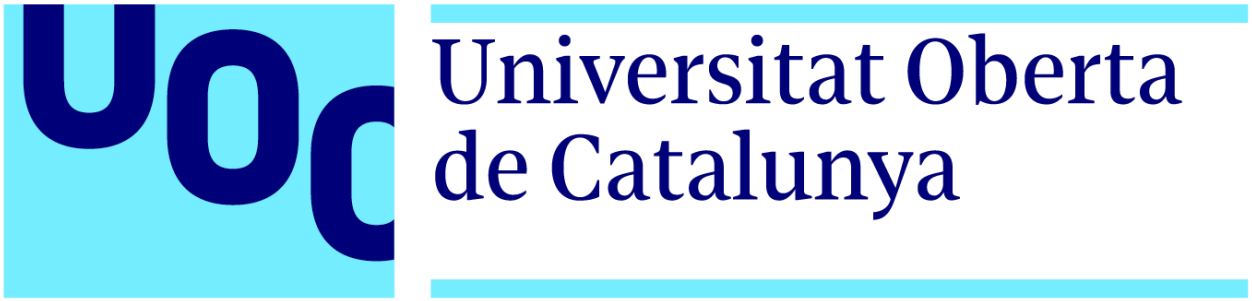


Práctica:
Web scraping

Alumnos: María Sonia Rodriguez Cepedano
Carlos E. Jiménez
José Luis Esteban González

Repositorio: https://github.com/jestebango/uoc/

Asignatura: Tipología y ciclo de vida de los datos
Estudios: Máster en Ciencia de Datos
Fecha: 13/11/2017

Índice de contenido

1. Contexto: diseño y estructura del proyecto

El proyecto está integrado por tres personas, y la infraestructura se ha coordinado desde la plataforma Github, en concreto en el repositorio ***jestebango/uoc.***

En una primera fase se realizaron varias propuestas con el fin de partir de un problema inicial global, así como posibles divisiones de tareas. Las distintas propuestas quedaron registradas como *issues* en el repositorio, que se cerraron tras existir un consenso pero sin una definición completa debido a ligeras discrepancias.

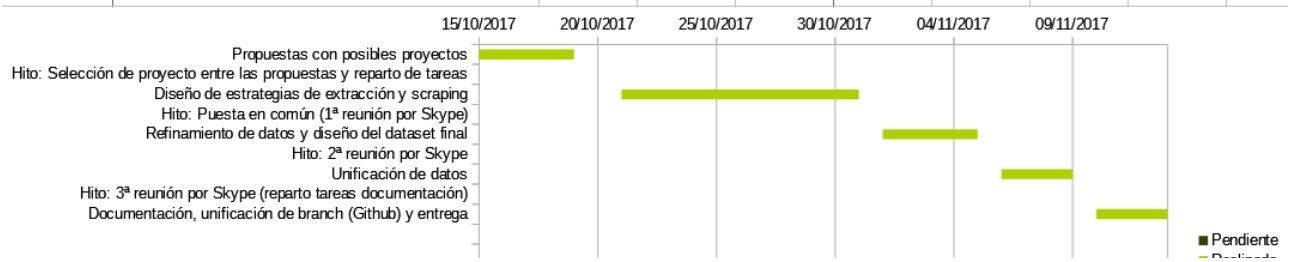
En general, existían dos puntos de partida o formas de dividir las tareas para que cada integrante implementara la técnica de web scraping de forma individual:

1. Establecer un tema común y que cada uno se ocupara de distintas fuentes (división según fuentes de los datos).
2. Establecer varios temas y una posible aplicación en la que se pudieran relacionar (división por temáticas con una relación).

Se optó por la segunda opción, con lo que cada integrante del proyecto se ha responsabilizado de seleccionar las fuentes y extraer los datos del tema asignado. La división de tareas, quedó establecida en la primera reunión establecida por Skype:

También se fijó una fase para exponer resultados, así como de posibles datasets para unificarlos. Las distintas fechas de reunión se han dejado registradas en el repositorio, y el calendario de planificación se muestra en los siguientes apartados. El calendario fijado que se acordó se muestra a continuación:

Fase	Inicio	Fin	Duración	Completado	Pendiente	Realizado
Propuestas con posibles proyectos	10/10/17	19/10/17	9	100,00 %	0	9
Hito: Selección de proyecto entre las propuestas y reparto d	20/10/17	20/10/17	0	100,00 %	0	0
Diseño de estrategias de extracción y scraping	21/10/17	31/10/17	10	100,00 %	0	10
Hito: Puesta en común (1ª reunión por Skype)	01/11/17	01/11/17	0	100,00 %	0	0
Refinamiento de datos y diseño del dataset final	01/11/17	05/11/17	4	100,00 %	0	4
Hito: 2ª reunión por Skype	05/11/17	05/11/17	0	100,00 %	0	0
Unificación de datos	06/11/17	09/11/17	3	100,00 %	0	3
Hito: 3ª reunión por Skype (reparto tareas documentación)	09/11/17	09/11/17	0	100,00 %	0	0
Documentación, unificación de branch (Github) y entrega	10/11/17	13/11/17	3	100,00 %	0	3



2. Definición de objetivos y ámbito

En la última década se han creado en España un gran número de empresas cuya actividad principal es el juego en sus diferentes modalidades: apuestas deportivas, loterías, etc. Algunas de estas actividades están teniendo bastante acogida a nivel nacional, y ya eran tradicionales en algunos países europeos como Inglaterra.

En esta línea se creó la Dirección General de Ordenación del Juego, dependiente de la Secretaría de Estado de Hacienda, para promover el "juego responsable" entre la población. Es este organismo el que también encargó un estudio para analizar los factores de riesgo en los trastornos derivados del juego, y en el que las adicciones son la manifestación más clara.

El estudio publicado en [1], identifica los factores de riesgo asociados al Juego Patológico o Trastorno del Juego de una muestra de afectados de diferentes regiones geográficas a nivel nacional, procedentes de 28 centros de salud y con diferentes niveles socioeconómicos, con un 92.4% de hombres de 43 años de edad media.

A modo de resumen, algunos de los factores predictores de una mayor gravedad en el trastorno son:

- En hombres: estar soltero, activo laboralmente, no disponer de ayudas sociales, padecer problemas de salud, mayor impulsividad cognitiva, etc

- En mujeres: estar soltera o separada/divorciada, edades más avanzadas que los hombres, situación laboral inactiva, recibir ayudas sociales, mayor impulsividad cognitiva que los hombres,etc.

Respecto a las preferencias de juego con mayor prevalencia y en orden decreciente han sido:

- En hombres: ***máquinas tragaperras***, loterías, quinielas, salas de juego, loterías instantáneas y apuestas por Internet.

- En mujeres: ***máquinas tragaperras***, loterías y bingos.

A partir de las conclusiones generales de este estudio, se puede deducir que, si la actividad de las empresas prestadoras de estos servicios tienen un crecimiento económico, las posibilidades de que las prevalencias de estos trastornos aumenten en un futuro y puedan visualizarse en los próximos estudios.
La manifestación de este crecimiento económico se puede comprobar si se consultan los datos publicados en el propio organismo de Ordenación del Juego, en su última publicación

de datos de juego online del segundo trimestre de 2017 (referencia [2]). En este informe se puede ver el crecimiento depósitos y retiradas por meses y hasta el año 2012. A efectos de comparación, se muestra en la siguiente tabla las fechas límite del informe:

(€)	***Junio 2012***	***Junio 2017***
Depósitos	29.002.347	113.280.810,12
Retiradas	17.406.487	82.587.027,39
Diferencia	11.595.860	30.693.782,73

Con los datos económicos oficiales y las conclusiones del estudio se ven varios proyectos de análisis de datos que podrían llevarse a cabo una vez extraídos los datos relevantes en esta área:

- Búsqueda de relaciones entre localizaciones de empresas (físicas) con actividades de juego e indicadores judiciales.
- Búsqueda de relaciones a nivel de censo entre este tipo de empresas e indicadores demográficos, como:
  - ¿Áreas de juego más “densas” a nivel de juego tienen peores servicios públicos?
  - ¿Este tipo de zonas tendrían una tasa de prevalencia más alta en otro tipo de trastornos psicológicos?,…

Una vez establecido el ámbito de actuación, hay que mencionar el principal objetivo que se busca con esta tarea. Como no hay una fuente de datos pública que muestre un directorio con las ubicaciones de las empresas relacionadas con el juego, se trata de usar la técnica web scraping para extraer estos datos a partir de distintas webs.

El resultado de esta extracción web será un datasets de empresas que desarrollan la actividad del juego en España, con un identificador, dirección física asociada, dirección online en caso de disponer de ella y modalidad de juego que realiza. Se resume en la siguiente tabla:

Atributo / Variable	***Descripción***
Identificador (nombre)	Denominación de la empresa
Ubicación	Ubicación de la empresa y delegaciones
URL	Web (obligatorio en empresas de apuestas online)
Modalidad	Tipo de juego según clasificación sectorial

3. Estrategia de diseño y aspectos éticos y legales

Dado que se trata de buscar información de empresas, y por tanto, que están ya creadas y tienen una actividad, las fuentes de búsqueda que se han consultado en este caso proceden de sitios webs públicos. Se podría haber optado por ampliar este espectro accediendo a otras fuentes de datos, por ejemplo, sitios de la Deep Web de apuestas clandestinas, pero en este caso no sería posible asociar una ubicación fija ni de dominio, aunque podrían servir para otro tipo de análisis.

Una vez fijado el tipo de datos a extraer, se han seleccionado las fuentes públicas siguientes como origen de los datos:

1. Listado de URLs de operadores con licencia de la Dirección General de Ordenación del Juego (DGOJ).

2. Censo Nacional de Empresas, que unifica los datos de las diferentes Cámaras regionales distribuidas por toda España.

El proceso completo puede resumirse en el siguiente diagrama:

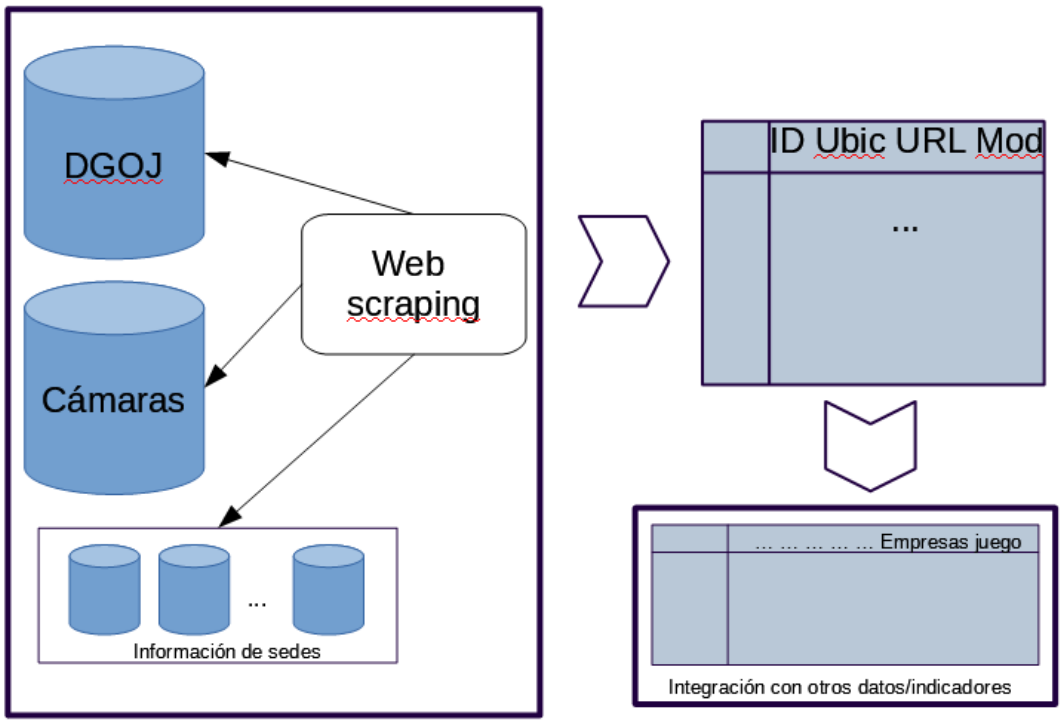


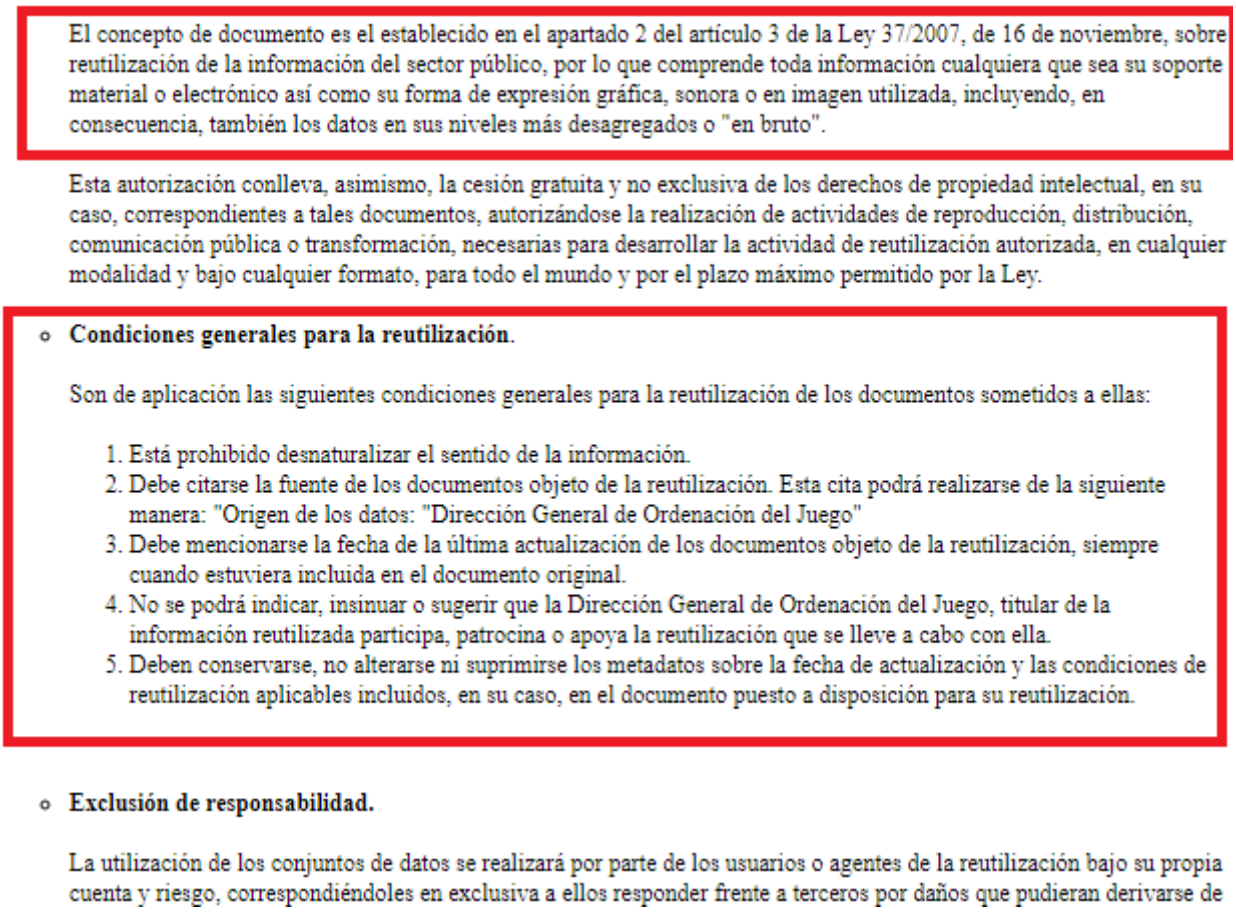
Ilustración 2: Diagrama del proceso

3.2 Cuestiones éticas y legales (fuentes de datos)

Las fuentes principales son registros públicos (DGOJ y Cámaras) que no disponen de alternativas para recuperar los datos objetivo. Carecen de API y/o datasets publicados para su uso. En este sentido no hay apartados ni licencias específicas a los que estarían sujetos estos supuestos APIs o datasets.

La primera verificación ha sido el análisis el apartado “Aviso Legal” en el portal DGOJ, con las siguientes conclusiones:

- No es aplicable el apartado de propiedad intelectual, ya que no se van a recopilar partes del diseño, logos u otras imágenes.
- No se hace mención de forma explícita a la recuperación de información automática.



El concepto de documento es el establecido en el apartado 2 del artículo 3 de la Ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público, por lo que comprende toda información cualquiera que sea su soporte material o electrónico así como su forma de expresión gráfica, sonora o en imagen utilizada, incluyendo, en consecuencia, también los datos en sus niveles más desagregados o "en bruto".

Esta autorización conlleva, asimismo, la cesión gratuita y no exclusiva de los derechos de propiedad intelectual, en su caso, correspondientes a tales documentos, autorizándose la realización de actividades de reproducción, distribución, comunicación pública o transformación, necesarias para desarrollar la actividad de reutilización autorizada, en cualquier modalidad y bajo cualquier formato, para todo el mundo y por el plazo máximo permitido por la Ley.

- **Condiciones generales para la reutilización**.

Son de aplicación las siguientes condiciones generales para la reutilización de los documentos sometidos a ellas:

1. Está prohibido desnaturalizar el sentido de la información.
2. Debe citarse la fuente de los documentos objeto de la reutilización. Esta cita podrá realizarse de la siguiente manera: "Origen de los datos: "Dirección General de Ordenación del Juego"
3. Debe mencionarse la fecha de la última actualización de los documentos objeto de la reutilización, siempre cuando estuviera incluida en el documento original.
4. No se podrá indicar, insinuar o sugerir que la Dirección General de Ordenación del Juego, titular de la información reutilizada participa, patrocina o apoya la reutilización que se lleve a cabo con ella.
5. Deben conservarse, no alterarse ni suprimirse los metadatos sobre la fecha de actualización y las condiciones de reutilización aplicables incluidos, en su caso, en el documento puesto a disposición para su reutilización.

- **Exclusión de responsabilidad.**

La utilización de los conjuntos de datos se realizará por parte de los usuarios o agentes de la reutilización bajo su propia cuenta y riesgo, correspondiéndoles en exclusiva a ellos responder frente a terceros por daños que pudieran derivarse de

Ilustración 3: Apartados a destacar del sitio web DGOJ

- Aplicaría el apartado de condiciones generales de reutilización, que se muestra en la siguiente captura. El apartado 3 no aplica en este caso, porque no se van a recuperar documentos y los datos usados no incorporan fechas (se aplicaría si se recuperaran los datos con las fechas de cada licencia singular obtenida).

Del mismo modo se actúa en el sitio web de Cámaras, donde se va a recuperar información de direcciones físicas. Hay que tener en cuenta que las personas jurídicas no están sujetas a la LOPD (para personas físicas), y que el censo de consulta es público.

Para consultar las restricciones existe el apartado “Legal y privacidad”. En este caso las condiciones de acceso son más restrictivas que en el caso anterior, se muestra un fragmento resaltando las posibles limitaciones a realizar con web scraping.

2. Condiciones de acceso y utilización

El sitio web y sus servicios son de acceso libre y gratuito, no obstante, la CCE condiciona la utilización de algunos de los Servicios a la previa cumplimentación del correspondiente formulario. El Usuario garantiza la autenticidad y actualidad de todos aquellos datos que comunique a la CCE y será el único responsable de las manifestaciones falsas o inexactas que realice.El Usuario se compromete expresamente a hacer un uso adecuado de los contenidos y servicios de camara.es y a no emplearlos para:

a) Difundir contenidos delictivos, violentos, pornográficos, racistas, xenófobo, ofensivos, de apología del terrorismo o, en general, contrarios a la ley o al orden público.

b) Introducir en la red virus informáticos o realizar actuaciones susceptibles de alterar, estropear, interrumpir o generar errores o daños en los documentos electrónicos, datos o sistemas físicos y lógicos de la CCE o de terceras personas; así como obstaculizar el acceso de otros Usuarios al sitio web y a sus servicios mediante el consumo masivo de los recursos informáticos a través de los cuales la CCE presta sus servicios.

c) Intentar acceder a las cuentas de correo electrónico de otros usuarios o a áreas restringidas de los sistemas informáticos de la CCE o de terceros y, en su caso, extraer información.

d) Vulnerar los derechos de propiedad intelectual o industrial, así como violar la confidencialidad de la información de la CCE o de terceros.

e) Suplantar la identidad de otro Usuario, de las administraciones públicas o de un tercero.

f) Reproducir, copiar, distribuir, poner a disposición o de cualquier otra forma comunicar públicamente, transformar o modificar los Contenidos, a menos que se cuente con la autorización del titular de los correspondientes derechos o ello resulte legalmente permitido.

g) Recabar datos con finalidad publicitaria y de remitir publicidad de cualquier clase y comunicaciones con fines de venta u otras de naturaleza comercial sin que medie su previa solicitud o consentimiento.

Todos los contenidos del sitio web, como textos, fotografías, gráficos, imágenes, iconos, tecnología, software, así como su diseño gráfico y códigos fuente, constituyen una obra cuya propiedad pertenece a la CCE, sin que puedan entenderse cedidos al Usuario ninguno de los derechos de explotación sobre los mismos, más allá de lo estrictamente necesario para el correcto uso de la web.

Ilustración 4: Condiciones de uso del sitio camaras.es

El apartado f) no se vulnera ya que los datos de empresas en censos públicos están legalmente permitidos y no hay que pedir autorización. El apartado g) tampoco aplica ya que el fin de los datos recopilados es para un fin didáctico, y en ningún caso se va a comerciar con los datos.

Para finalizar las comprobaciones habría que comprobar que las ubicaciones de acceso a los datos no están limitadas mediante un fichero robots.txt en cada dominio. Para ello, se ha hecho uso de la aplicación online [4]. El sitio web DGOJ no dispone de restricciones (no recupera robots.txt), y para el sitio Camaras.es se comprueba que la ubicación de la

aplicación PHP (formulario web) no está restringida; en la siguiente captura de imagen se puede comprobar.

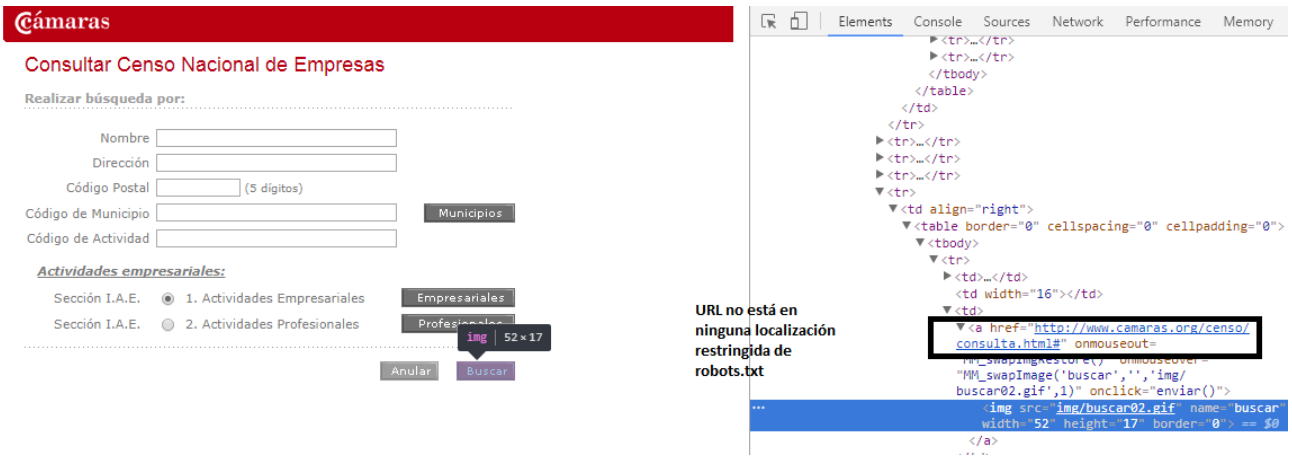


Ilustración 5: Ubicación accedida con el formulario /censo/...

```
#
# robots.txt
#
# This file is to prevent the crawling and indexing of certain parts
# of your site by web crawlers and spiders run by sites like Yahoo!
# and Google. By telling these "robots" where not to go on your site,
# you save bandwidth and server resources.
#
# 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: *
Crawl-delay: 10
# Directories
Disallow: /includes/
Disallow: /misc/
Disallow: /modules/
Disallow: /profiles/
Disallow: /scripts/
Disallow: /themes/
# Files
Disallow: /CHANGELOG.txt
Disallow: /cron.php
Disallow: /INSTALL.mysql.txt
Disallow: /INSTALL.pgsql.txt
Disallow: /INSTALL.sqlite.txt
Disallow: /install.php
Disallow: /INSTALL.txt
Disallow: /LICENSE.txt
Disallow: /MAINTAINERS.txt
Disallow: /update.php
Disallow: /UPGRADE.txt
Disallow: /xmlrpc.php
# 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: /?q=admin/
Disallow: /?q=comment/reply/
Disallow: /?q=filter/tips/
Disallow: /?q=node/add/
Disallow: /?q=search/
Disallow: /?q=user/password/
Disallow: /?q=user/register/
Disallow: /?q=user/login/
Disallow: /?q=user/logout/
# Modificado DIC 2016
Allow: /misc/*.js
```

Ilustración 6: Archivo robots.txt

3.2 DGOJ

Este sitio web almacena de forma estructurada el listado de operadores de juego que disponen de licencias. Este será el punto de partida para localizar las empresas.

Según se especifica en [3], cada juego requiere de una licencia singular (previo requisito de haber obtenido la licencia general), y esta será la condición con la que realizar las consultas.

El formulario de consulta es accesible mediante la opción del menú “Listado de URLs de operadores con licencia”. Se excluyen las rifas, juego ocasional, juegos reservados de loterías y entidades de certificación que figuran en otros apartados.

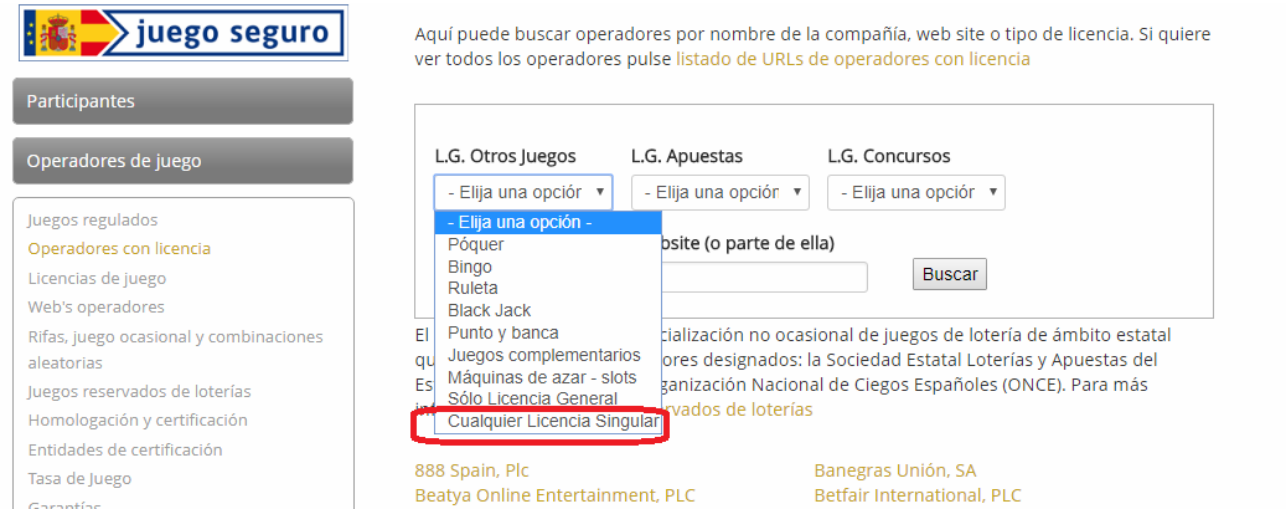


Ilustración 7: Criterios de las consultas

En este formulario son necesarias tres consultas diferentes, no un única consulta que se interpretaría como un AND (con licencias siingulares en las tres categorías):

L.G. Otros Juegos → Cualquier Licencia Singular
L.G. Apuestas → Cualquier Licencia Singular
L.G. Concursos → Cualquier Licencia Singular

Como se devuelven tres conjuntos de resultados, y puede haber resultados que tengan licencias de dos o tres categorías, se eliminan duplicados y se unifican en una única tabla.

El número de variables a recuperar es el nombre, dominio (si es un web operador) y las licencias singulares (modalidades de juego). Estas variables son fundamentales para la búsqueda de la información final en el resto de fuentes.

3.3 Cámaras

La consulta a la base de datos del censo de empresas se realiza con el formulario mostrado en la siguiente captura de pantalla. Los datos sobre los que se realiza cada consulta son las empresas registradas en las diferentes localidades con su dirección física y código de actividad económica.

Ilustración 8: Formulario de consulta

Aquí el uso de web scraping está justificado porque los datos son accesibles mediante un sistema de consultas con los parámetros:

- Nombre
- Dirección
- Código Postal
- Código de Municipio
- Código por IAE (actividad económica con la que se ha registrado)

La limitación se realiza en el número de resultados devueltos, y que corresponde a un máximo de 50 empresas por consulta. Sin este límite el problema tendría una resolución relativamente sencilla: realizar consultas globales por IAE y de cada una realizar consultas a cada enlace de resultados

consulta IAE=9692 + consulta IAE=9693 + … + consulta IAE=9697 (se excluye 9691)

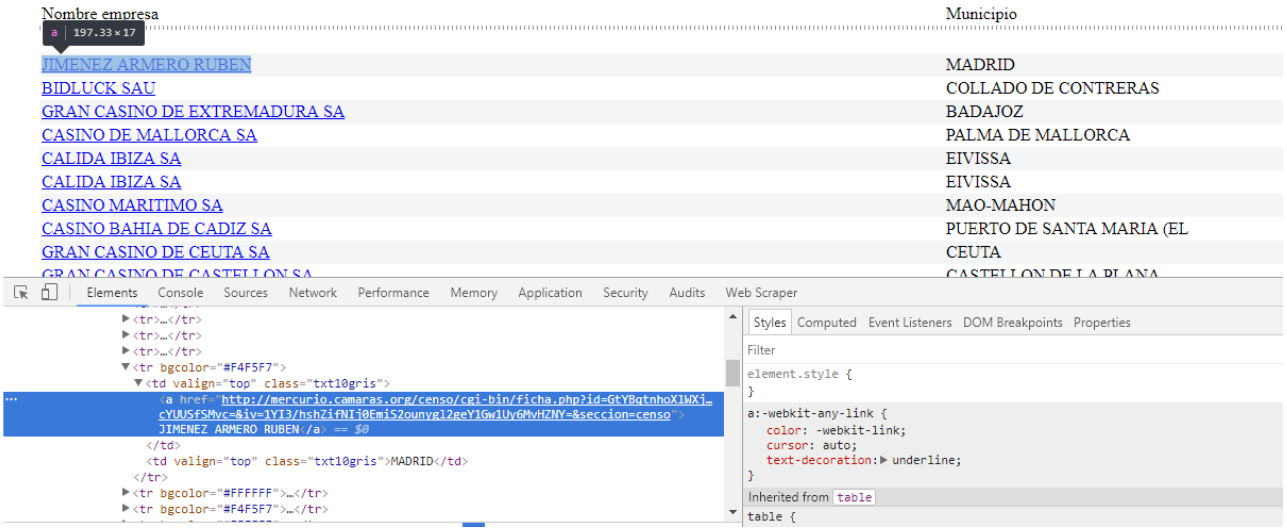


Ilustración 9: Scraping primario de URLs, sobre las que realizar scraping secundario de los datos

Para acceder a los datos finales, se realiza un scraping primario para recuperar URLs con los criterios de búsqueda, y scraping secundario (tantos como URLs de resultados) para acceder a los datos objeto de interés.

Una búsqueda por localidades no sería adecuada para aquellas con muchas empresas (como las capitales de provincia), ya que también van a exceder del límite máximo de 50.

También hay que tener en cuenta que una empresa con varias sedes se va a devolver en forma de diferentes registros con el mismo nombre, como se muestra en la siguiente captura (IAE=9692).

La limitación de la búsqueda estará en aquellos códigos postales que tengan más de 50 entradas (delegaciones/sedes). La estrategia de consultas

implementada mediante el formulario de la Cámara se puede comprender en el siguiente diagrama:

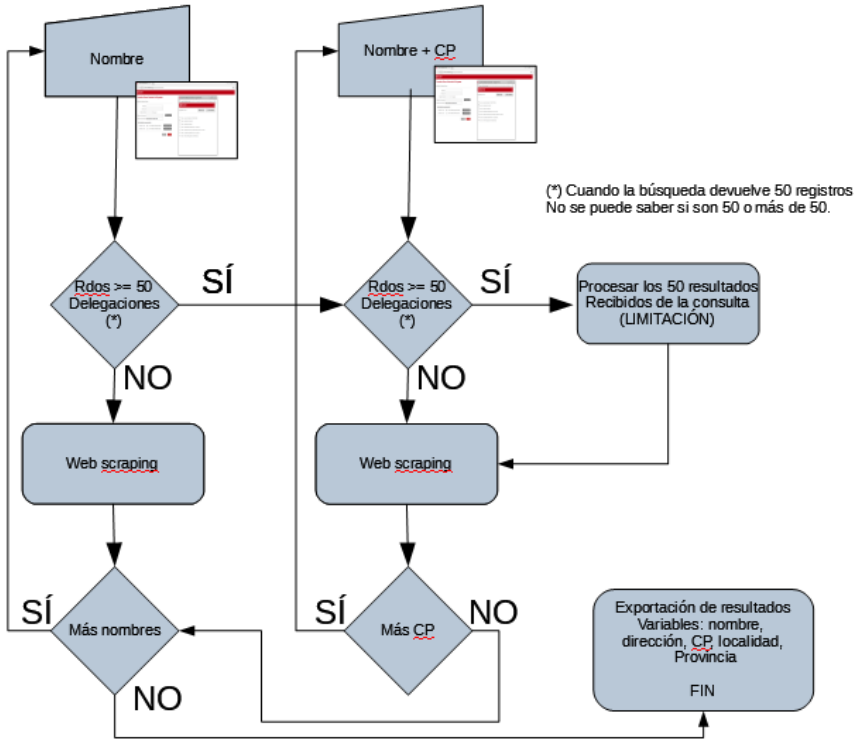


Ilustración 11: Estrategia de consultas para obtener la máxima cobertura

Finalmente, no se ha podido implementar de forma completa la estrategia del diagrama anterior por los siguientes motivos:

- Bloqueo temporal en la primera ejecución del script que no tenía retardo entre peticiones.
- Se aumenta el número de peticiones POST al formulario de forma considerable.
- Incapacidad de saber cuáles son las localidades con más resultados (sólo muestra las 50 primeras ordenadas alfabéticamente).

De todas las causas, la tercera es la que ha llevado finalmente al descarte. No obstante, en un proyecto real en el que se necesiten todos los datos, se habría implementado.

3.4 Extracción de Twitter

El ámbito de este apartado serían los operadores de juego con licencia registrados en DGOJ que disponen de algún dominio. Antes de proceder al uso de la API se ha realizado una búsqueda manual de usuarios Twitter.

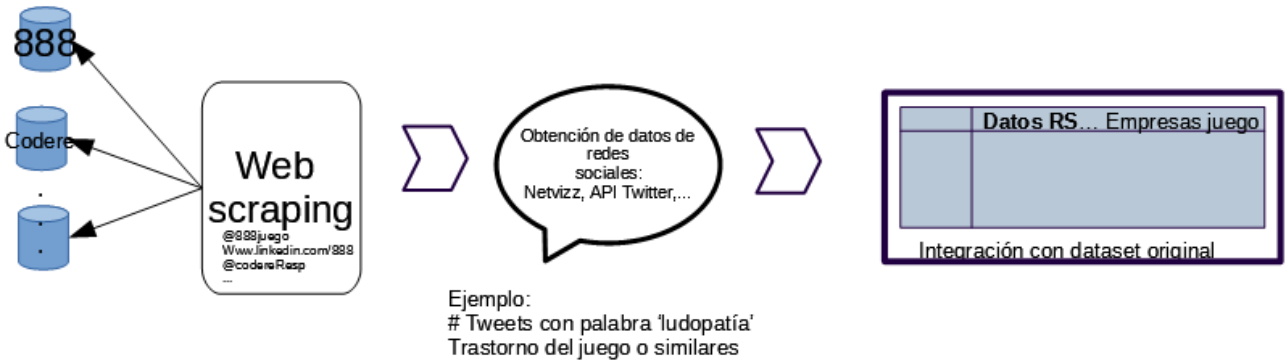


Ilustración 12: Idea de ampliación mediante scraping a webs individuales

En el caso de la búsqueda en Twitter, no se impone limitación de tiempo, aunque se recoge la fecha de inicio o creación del usuario, ya que lo que nos interesa es el volumen acumulado de dicha actividad a fecha de hoy.

Para la recogida de los datos hemos creado en primer lugar una API (Juego) a través del sitio para desarrolladores de twitter, como instrumento a utilizar en la exploración y extracción dentro de Twitter. Creamos en R un script que comunica con la API y que ejecuta la exploración y recogida de los resultados de la exploración en un conjunto de datos. Utilizamos para ello las librerías de Twitter:

- library(twitteR)
- library(ROAuth)

Generamos las credenciales dadas por la API y conectamos con ella a través de la función setup_twitter_oauth.

A partir del webscraping buscando la lista completa de las empresas de juego que operan en España, buscamos en la red social a partir del usuario de twitter de las mismas, para extraer la información relevante para nuestro objetivo. Hemos utiizado la función lookupUsers, para no sobrecargar la API y hacer una petición única con todos los usuarios al mismo tiempo.

Las variables o atributos extraidos han sido 17:

1. description, describe el usuario
2. statusesCount, recoge la actividad del usuario, incluyendo retweets
3. followersCount, recoge el número de seguidores del usuario
4. favoritesCount, recoge los tweets o retweets que han sido marcados como favoritos por el usuario
5. friendsCount, recoge el número de usuarios a los que sigue este usuario
6. url, incorpora la dirección web que ha incluido el usuario
7. name, recoge el nombre del usuario en su cuenta
8. created, indica en que fecha fue creado el usuario
9. protected, indica si los tweets e información de la cuenta está o no protegida
10. verified, indica si el usuario ha verificado la cuenta
11. screenName, indica el nombre o alias del usuario
12. location, aparece recogido si existe una localización específica de usuario
13. lang, indica el idioma autoidentificado por el usuario
14. id, es el número de identificación único del usuario para twitter
15. listedcount, indica el número de listas públicas del que el usuario es parte
16. followRequestSent, si se ha enviado solicitud de amistad
17. profileImageUrl, sitio donde se encuentra la imagen del perfil del usuario

Añadimos una variable más que vinculará los datos extraidos con el dataset principal del proyecto. Es la siguiente variable:

18. UsuarioTwitter, variable-conexión propia que vincula al usuario con otras tablas de este proyecto, en función de su perfil en twitter. Coincide con la variable name dada por twitter.

Recogemos los resultados en un conjunto de datos y, a partir del mismo, trabajamos filtrando las variables y seleccionando seis de ellas que son las que a nuestro entender nos aportan más información para la respuesta a las preguntas y los propósitos buscados en este proyecto. Estas variables son:

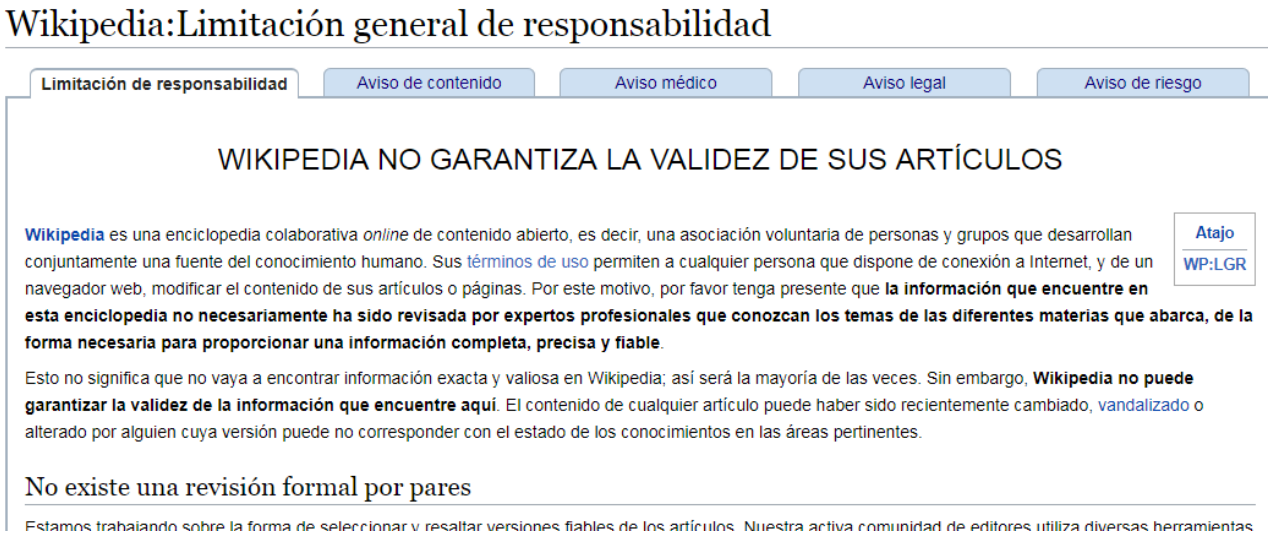
1. statusesCount, recoge la actividad del usuario, incluyendo retweets
2. followersCount, recoge el número de seguidores que tiene el usuario
3. friendsCount, recoge el número de usuarios a los que sigue este usuario
4. created, indica en que fecha fue creado el usuario
5. listedcount, indica el número de listas públicas de otros usuarios que han incluido a este usuario como parte de ella
6. UsuarioTwitter, variable-conexión propia que vincula al usuario con otras tablas de este proyecto, en función de su perfil en twitter

3.5 Extracción de Wikipedia y Web

En un primer momento se considera tener en cuenta la información de las provincias de España. En cuanto a población y Comunidad autónoma.

Para ello planteamos realizar web Scraping sobre la página web https://es.wikipedia.org/wiki/Anexo:Provincias_y_ciudades_aut%C3%B3nomas_de_Espa%C3%B1a

Primeramente, comprobamos desde “Limitación de responsabilidad” la posibilidad de la utilización de los datos.



Concretamente en la pestaña “Limitación de responsabilidad” aparece un link hacía “esquema de licenciamiento de Wikipedia”

Wikipedia:Derechos de autor

Esta página describe una política oficial de Wikipedia en español.
Ha sido elaborada por la comunidad y su cumplimiento es obligatorio para todos los editores. Puedes editarla para mejorar su redacción y formato, pero si deseas cambiar alguna cuestión de fondo, busca el consenso comunitario primero.

Los derechos de autor de los textos que figuran en Wikipedia corresponden a los editores que han colaborado en ellos y tales derechos se encuentran protegidos automáticamente en el marco del Convenio de Berna, que establece los estándares mínimos de protección de derechos de autor que deben acatar y adaptarse los países ratificantes cuando estos los regulen libremente en sus leyes nacionales.[1][2][3][4][5] La legislación aplicable a todos los proyectos Wikimedia es la de los Estados Unidos de América,[6] que en el caso de los derechos de autor corresponde al Título 17 del Código de los Estados Unidos establecido en el *Copyright Act of 1976*.[7]

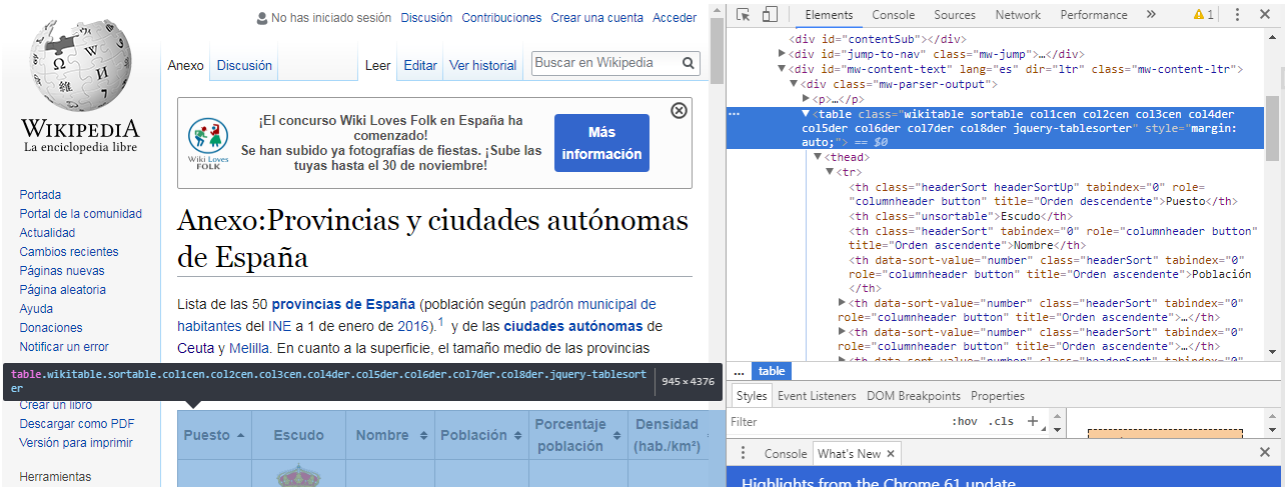
La mayoría de los textos de Wikipedia son licenciados bajo la Creative Commons Attribution/Share-Alike License 3.0 (Unported) así como bajo la GNU Free Documentation License (sin versionar, sin secciones, textos de tapa o contratapa invariantes). Los reutilizadores pueden elegir la licencia que deseen. Estas licencias sí permiten el uso comercial de los contenidos reutilizados, siempre y cuando que los usos que se les den respeten las condiciones establecidas en la licencia respectiva.

Atajos
WP:DR
WP:DA
WP:CR
WP:©

Políticas de Wikipedia
Los cinco pilares
Lo que Wikipedia no es
Estándares de artículos
Punto de vista neutral
Verificabilidad
Fuentes fiables
Derechos de autor
Principios globales
Ignora las normas
Usa el sentido común
Conflicto de interés

Al leer el contenido del texto nos aseguramos que la información puede ser utilizada para nuestro propósito.
Comenzamos con el scraping para obtener directamente los datos página anteriormente indicada.

Analizamos la web



Esta nos proporciona la información que necesitamos para comenzar el scraping con R.
En primer lugar, procedemos a instalar el paquete xml2 y las librerías xml2 y rvest.
Seguidamente obtenemos una imagen interna de la web. Teniendo en cuenta que esta no es más una composición de nodos, procedemos a extraer los nodos de tipo “tables” y la convertimos en una tabla con la que podemos trabajar y obtener información.

El script utilizado en R es el siguiente:

```r
install.packages("xml2")
library("xml2")
library("rvest")
url.provincias<- "https://es.wikipedia.org/wiki/Anexo:Provincias_y_ciudades_aut%C3%B3nomas_de_Espa%C3%B1a"
tabla_temp<- read_html(url.provincias)
tabla_temp <- html_nodes(tabla_temp, "table")
tabla_temp
length(tabla_temp)
sapply(tabla_temp,class)
html_table(tabla_temp,fill=TRUE)
#sapply(tabla_temp, function(x) dim(html_table(x, header=false, fill = TRUE)))
provincias<- html_table(tabla_temp,fill=TRUE)
provincias
write.csv(provincias, file="provincias.csv") #guardamos en un archivo CSV.
```

También, con R procedemos a realizar una selección de los datos que nos interesa y que acabamos de obtener.

```r
#Limpieza del DataSet
getwd()
datos_aux<-read.csv2("provincias.csv", sep=",",na.strings = "NA")
datos_aux
datos=data.frame(datos_aux$Nombre, datos_aux$Población,
                 datos_aux$Porcentaje.población,datos_aux$Densidad..hab..km².,
                 datos_aux$Comunidad.autónoma)
write.csv(datos, file="datos.csv")
```

Se piensa en obtener datos económicos de las empresas que hemos extraído de DGOJ.

Para ello nos dirigimos a la web http://ranking-empresas.eleconomista.es/sector-9200.html
Teniendo en cuenta la información

4. LIMITACIONES DE USO Y PROTECCIÓN DE LA INFORMACIÓN Y CONTENIDOS DE RANKING-EMPRESAS.ELECONOMISTA.ES

El contenido y la información a la que pueda acceder el usuario a través de la web de Ranking-empresas.eleconomista.es puede provenir de terceros, o de la Base de Datos de Información de Empresas y Empresarios propiedad de la EMPRESA o haber sido obtenida a través de Internet.

Valoramos la calidad de la información ofrecida, por ello, si usted detecta alguna información o contenido que no sea correcta, infrinja alguna ley, normativa o incumpla los términos de este contrato, debe comunicarlo lo antes posible a través del enlace al formulario Contacte con Nosotros.

Con el fin de mejorar la calidad de la información ofrecida, Usted reconoce y acuerda que la EMPRESA tiene derecho de modificación sobre el contenido de Ranking-empresas.eleconomista.es y ésta podrá en cualquier momento, sin necesidad de notificación, modificar o eliminar todo o parte del contenido de Ranking-empresas.eleconomista.es.

El acceso a la información de Ranking-empresas.eleconomista.es sólo puede realizarse a través de la dirección web www.ranking-empresas.eleconomista.es y aquellos otros dominios o subdominios que se especifiquen por contrato por la EMPRESA. Está prohibido acceder a Ranking-empresas.eleconomista.es de cualquier manera automática y reiterada mediante robots, scripts o arañas que de forma automática realicen múltiples peticiones al sitio web. Está prohibido almacenar la información contenida en Ranking-empresas.eleconomista.es, sin autorización expresa de la EMPRESA. Está autorizado el acceso automatizado a Ranking-empresas.eleconomista.es por motores de búsqueda, como Google y Bing.

La EMPRESA realiza cambios puntuales en la información, insertando marcas de agua con tecnología de cifrado, que no alteran el contenido ni significado de la misma en grado alguno, pero que permiten la detección de comercializaciones ilícitas de su base de datos y la procedencia de dichas comercializaciones, así como la fecha e IP de extracción. La EMPRESA podrá probar la captación ilícita de información y, por tanto, exigirá las penas tipificadas en el código penal al respecto de la apropiación indebida de bases de datos.

No procedemos a utilizar esta información dentro del dataset.

4. Tecnologías de implementación

Para la implementación de las distintas extracciones se han usado las diferentes técnicas de extracción:

Web scraping en Python y R mediante librería BeautifulSoup
Consultas a la API de Twitter

En el caso de la exploración de twitter se aprovecha la potencia de las funciones ya existentes en las librerías:

- library(twitteR)
- library(ROAuth)

También se consultaron otras fuentes, y se hizo uso de open data (fichero INE para ver la correspondencia de los códigos de municipio).

5. Resultados

5.1 Datasets parciales

De los resultados de los diferentes scraping realizados, se han unificado los datos en el esquema del apartado 3, que corresponde a las empresas operadoras de juego con cualquier licencia singular publicadas.

No obstante, hay que mencionar algunos aspectos de los resultados obtenidos en el dataset:

- Los datos se limitan a los datos publicados, con licencia de operador. No figuran por tanto empresas que no tengan la licencia hasta la fecha de realización de este trabajo.

- Las direcciones de las sedes pueden estar incompletas, ya que los datos están limitados a los contenidos en el censo de las distintas cámaras. Este punto lo he comprobado en la ciudad donde resido (Guadalajara), ya que acaban de inaugurar sede de la empresa ‘CODERE’ que no figuran en la base de datos.

- Los datos pueden servir para distintos proyectos, con relaciones geográficas y que quieran buscar relaciones por modalidades de juego entre otros.

- Los datos resultantes pueden subdividirse en distintos datasets. Esto es importante porque al recuperar los datos de DGOJ también se muestran operadores online, muchos de los cuales no están ubicados físicamente en España (ni por tanto a nivel fiscal). Las alternativas que se presentaron fueron las siguientes:

    1. Descartar todos los operadores online: si se hubiera realizado, no figurarían empresas sin sede física pero que tienen sede social/fiscal.

2. Descartar operadores online que no disponen de sede física: un ejemplo sería el caso de '888 Spain, PLC' que tiene sede en Reino Unido pero tiene licencia de operador online con dominio en DGOJ (sin sedes registradas).

3. No descartar ningún operador con licencia singular (opción elegida). Se dejaría el dataset con el máximo de información, y dependiendo del uso posterior se filtrarían los datos en fases de preparación. Este es el motivo de que muchos de los campos no están completos (NULL/NA).

El primer dataset parcial de los 'operadoresDGOJ.csv' tiene las siguientes variables:

VARIABLE	***CARACTERÍSTICA***	***EJEMPLO***
Nombre	Nombre de la empresa (puede tener múltiples delegaciones / franquicias) con las que comparte las licencias	Bluesblock, SA
Licencias	Licencias singulares obtenidas para cada tipología. Puede estar vacío ('Sin licencia vigente'), lo que indica que está extinguida	Concursos
Dominios	Sitios web donde opera la empresa ofreciendo el servicio para el que tiene licencia. Puede no tener ningún dominio o varios (0,n)	www.bluesblock.es

Se muestra una captura de pantalla con algunas empresas:

	Nombre	Licencias	Dominios
2	Bluesblock, SA	Concursos	www.bluesblock.es
3	Codere Online, SAU	Ruleta,Black Jack,Máquinas de azar - slots,Deportivas de contrapartida,Hípicas de contrapartida,Otras de contra	www.codere.es,apuestas.codere.es,www.codereapuestas.es
4	Concursos Multiplataformas, SAU	Concursos	S,i,n, ,d,o,m,i,n,i,o,
5	Esgaming, SAU	Ruleta,Black Jack,Punto y banca,Máquinas de azar - slots	www.casinogranmadridtv.es,www.casinogranmadridonline.es
6	Eurojuego Star, SA	Concursos	www.eurojuego.es,www.eurojuegostar.es,
7	JS2015 Games, SAU		
8	Sociedad Estatal Loterías y Apuestas del E	Deportivas mutuas,Hípicas mutuas	www.loteriasyapuestas.es
9	Suertia Interactiva, SA	Ruleta,Black Jack,Máquinas de azar - slots,Deportivas mutuas,Deportivas de contrapartida	www.suertia.es
10	Unidad Editorial Juegos, SA	Ruleta,Black Jack,Punto y banca,Máquinas de azar - slots,Deportivas de contrapartida	apuestas.marca.es,marcaapuestas.es,
11	888 Spain, Plc	Póquer,Ruleta,Black Jack,Máquinas de azar - slots,Deportivas de contrapartida	www.888.es,www.888casino.es,www.888poker.es,www.888
12	Banegras Unión, SA	Ruleta,Black Jack,Punto y banca,Máquinas de azar - slots,Deportivas de contrapartida	www.juegging.es
13	Beatva Online Entertainment, PLC	Ruleta,Black Jack,Punto y banca,Máquinas de azar - slots,Deportivas de contrapartida	www.starvegas.es
14	Betfair International, PLC	Póquer,Ruleta,Black Jack,Máquinas de azar - slots,Deportivas de contrapartida,Deportivas cruzadas,Hípicas de	www.betfair.es
15	Betway Spain, PLC	Ruleta,Black Jack,Punto y banca,Máquinas de azar - slots	www.betway.es
16	Casino Barcelona Interactivo, SA	Póquer,Ruleta,Black Jack,Máquinas de azar - slots	www.casinobarcelona.es
17	Cirsa Digital, SAU	Póquer,Bingo,Ruleta,Black Jack,Punto y banca,Máquinas de azar - slots,Deportivas de contrapartida,Hípicas de	www.sportium.es
18	Comar Inversiones, SA	Ruleta,Black Jack,Punto y banca,Máquinas de azar - slots,Deportivas de contrapartida	www.ijuego.es
19	Ebingo Online España, SA	Bingo,Ruleta,Black Jack,Máquinas de azar - slots	ebingo.es
20	Ekasa Apuestas Online, S.A.	Deportivas de contrapartida,Otras de contrapartida	apuestas.retabet.es
21	Electraworks España, PLC	Póquer,Ruleta,Black Jack,Máquinas de azar - slots,Deportivas de contrapartida	www.bwin.es,www.partypoker.es,
22	Euroapuestas Online, SAU	Ruleta,Black Jack,Máquinas de azar - slots,Deportivas de contrapartida,Otras de contrapartida	www.paston.es
23	Eurobet International, SPA		
24	Eurobox, S.A.	Ruleta,Black Jack,Máquinas de azar - slots,Deportivas de contrapartida	www.circus.es
25	Golden Park Games, SA	Ruleta,Black Jack,Máquinas de azar - slots,Deportivas de contrapartida,Hípicas de contrapartida	http://www.goldenpark.es,www.todoslots.es,
26	Gtech Spain, SA	Ruleta,Black Jack,Punto y banca,Máquinas de azar - slots	
27	Hillside Spain New Media, PLC	Deportivas de contrapartida,Hípicas de contrapartida,Otras de contrapartida	www.bet365.es
28	Interwetten España, PLC	Ruleta,Black Jack,Máquinas de azar - slots,Deportivas de contrapartida	www.interwetten.es
29	Juego Online, EAD	Ruleta,Black Jack,Punto y banca,Máquinas de azar - slots,Deportivas de contrapartida	www.titanbet.es,poker.titanbet.es,casino.titanbet.es,www.dl

Ilustración 13: operadoresDGOJ.csv

El segundo dataset parcial sobre el que se implementó scraping fue la consulta de las distintas delegaciones o franquicias en España (físicas), registradas en las Cámaras de Comercio. Habría que tener en cuenta que habrá empresas internacionales que tengan licencias pero sin sede física.

Para obtener las direcciones, hubo que realizar un scraping adicional al servidor, para conseguir las URLs de acceso al backend donde se almacenan los datos finales. Para ello, se ha tenido que generar una tabla parcial con las siguientes variables:

VARIABLE	***CARACTERÍSTICA***	***EJEMPLO***
Nombre	Nombre de la empresa según aparece en la tabla operadoresDGOJ.csv	CODERE ONLINE, SAU
NombreDel	Nombre de la franquicia/delegación que aparece en www.camaras.es/censo	CODERE APUESTAS VALENCIA, S.A.
Enlace	Link a los datos finales, de donde extraer las direcciones	http://mercurio.camaras.org/censo/cgi-bin/ficha.php?id=1k77...-
Localidad	Localidad recuperada de las listas de resultados	PICANYA

La tabla se genera a partir sucesivas consultas en el formulario del censo de empresas, y tras obtener distintos listados que se van fusionando. El cuello de botella en este servidor está en la restricción de tiempo entre consulta y consulta, para evitar ataques de denegación de servicio. De hecho, se produjeron bloqueos de la MAC en los intentos iniciales debido a la inexperiencia, por lo que hubo que programar retardos de 10 segundos entre consulta y consulta para cumplir con los requisitos del archivo robots.txt.

Ilustración 14: Bloqueo temporal de usuario en el servidor y alternativa tras desbloqueo días después

Se muestra captura con los resultados:

	Nombre	NombreDel	Enlace	Localidad
2	Bluesblock, SA	BLUESBLOCK SA	http://mercurio.camaras.org/censo/cgi-bin/ficha.php?id=2oZJnIlp2l	BILBAO
3	Codere Online, SAU	CODERE APUESTAS ANDALUCIA SA	http://mercurio.camaras.org/censo/cgi-bin/ficha.php?id=XIcFvBgN	ANTEQUERA
4	Codere Online, SAU	CODERE APUESTAS ARAGON SLU	http://mercurio.camaras.org/censo/cgi-bin/ficha.php?id=LIESb/2LP	ZARAGOZA
5	Codere Online, SAU	CODERE APUESTAS ASTURIAS, S.A.	http://mercurio.camaras.org/censo/cgi-bin/ficha.php?id=dWxSuA3t	GIJON
6	Codere Online, SAU	CODERE APUESTAS BALEARES, S.A.	http://mercurio.camaras.org/censo/cgi-bin/ficha.php?id=KUsXJlcZ	CONSELL
7	Codere Online, SAU	CODERE APUESTAS CANTABRIA, S.A.	http://mercurio.camaras.org/censo/cgi-bin/ficha.php?id=UkjV0lNdC	SANTANDER
8	Codere Online, SAU	CODERE APUESTAS CASTILLA LA MANCHA,	http://mercurio.camaras.org/censo/cgi-bin/ficha.php?id=jtUa0BPec	TOLEDO
9	Codere Online, SAU	CODERE APUESTAS CASTILLA Y LEON, S.A.	http://mercurio.camaras.org/censo/cgi-bin/ficha.php?id=Nq5jbP7Tr	VALLADOLID
10	Codere Online, SAU	CODERE APUESTAS CATALUÑA, S.A.	http://mercurio.camaras.org/censo/cgi-bin/ficha.php?id=pgTMaLv9	PALAU DE PLEGAMANS
11	Codere Online, SAU	CODERE APUESTAS CEUTA SLU	http://mercurio.camaras.org/censo/cgi-bin/ficha.php?id=kyQ0lDs3	CEUTA
12	Codere Online, SAU	CODERE APUESTAS DEPORTIVAS S A U	http://mercurio.camaras.org/censo/cgi-bin/ficha.php?id=Okhleo1cf	NOAIN
13	Codere Online, SAU	CODERE APUESTAS EXTREMADURA S.A.U.	http://mercurio.camaras.org/censo/cgi-bin/ficha.php?id=qjz01uk0/l	CACERES
14	Codere Online, SAU	CODERE APUESTAS GALICIA SL	http://mercurio.camaras.org/censo/cgi-bin/ficha.php?id=sQ/cK9u9l	CORUÑA (A)
15	Codere Online, SAU	CODERE APUESTAS LA RIOJA SA	http://mercurio.camaras.org/censo/cgi-bin/ficha.php?id=us3q3t03Z	ARRUBAL
16	Codere Online, SAU	CODERE APUESTAS MELILLA, S.A.	http://mercurio.camaras.org/censo/cgi-bin/ficha.php?id=ABkXhbbM	MELILLA
17	Codere Online, SAU	CODERE APUESTAS MURCIA SL	http://mercurio.camaras.org/censo/cgi-bin/ficha.php?id=Z+IZWvY/	MURCIA
18	Codere Online, SAU	CODERE APUESTAS NAVARRA SA	http://mercurio.camaras.org/censo/cgi-bin/ficha.php?id=KyuWinwT	PAMPLONA
19	Codere Online, SAU	CODERE APUESTAS SA UNIPERSONAL	http://mercurio.camaras.org/censo/cgi-bin/ficha.php?id=Abbc2cFz	ALCOBENDAS
20	Codere Online, SAU	CODERE APUESTAS VALENCIA SA	http://mercurio.camaras.org/censo/cgi-bin/ficha.php?id=1k77tarzIo	PICANYA
21	Codere Online, SAU	CODERE INTERACTIVA SL	http://mercurio.camaras.org/censo/cgi-bin/ficha.php?id=vKEx4cDs	ALCOBENDAS

Ilustración 15: ficheroURL.CSV

Finalmente, con los datos anteriores se realiza una extracción de cada URL con la dirección final. El script añade el campo “dirección” a la tabla anterior y elimina el enlace, ya que no será relevante para el dataset final. Las variables se vuelven a mostrar en la siguiente tabla:

VARIABLE	***CARACTERÍSTICA***	***EJEMPLO***
Nombre	Nombre de la empresa según aparece en la tabla operadoresDGOJ.csv	CODERE ONLINE, SAU

NombreDel	Nombre de la franquicia/delegación que aparece en www.camaras.es/censo	CODERE APUESTAS VALENCIA, S.A.
DirCompleta	Se fusiona la dirección en una línea separando por comas la dirección, y el código y la localidad	http://mercurio.camaras.org/censo/cgi-bin/ficha.php?id=1k77...-
Localidad	Localidad recuperada de las listas de resultados	AV ALQUERIA MORET 00019 , 46210 PICANYA - VALENCIA

En la siguiente captura se puede comprender mejor la estrategia de extracción usada:

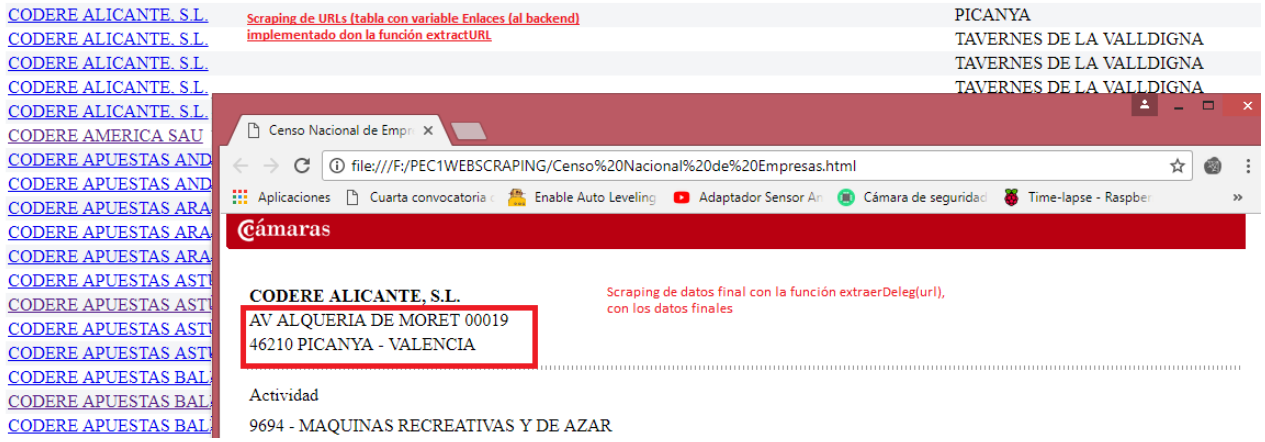


Ilustración 16: Dos niveles de extracción en el mismo servidor

Otro aspecto a tener en cuenta es que debido a incidencias con el sraping se simplificó en gran medida la extracción de sedes, y al final no se pudo implementar completamente el esquema planteado en el diseño, por tanto, faltarían delegaciones.

La idea inicial es que debido a la restricción del máximo de 50 resultados en los listados, se iban a fragmentar las consultas en el caso de búsquedas con más de 50 delegaciones. Sería el caso de Madrid por ejemplo, con empresas conocidas como Codere.

Otra aspecto determinante ha sido que una empresa puede haberse registrado con una actividad (código I.A.E.) y haber obtenido una licencia diferente posteriormente. Además, el

formulario sólo permite un máximo de 5 códigos, cuando las posibles licencias quedarían cubiertas por 9 códigos: "9692,9693,9694,9695,9696,9697,9821,9822,9825"

La estrategia de extracción final ha sido el lanzamiento del script por duplicado

- 1ª ejecución: con los códigos 9692,9693,9694,9695,9696
- 2ª ejecución: con códigos 9697,9821,9822,9825
- 3º unificación de datos y control de integridad

Los datos obtenidos con las direcciones al final no se han integrado totalmente en el dataset final, y servirían para estudios relacionados (para relacionar datos geográficos). Se han dejado bien relacionados mediante el campo Nombre.

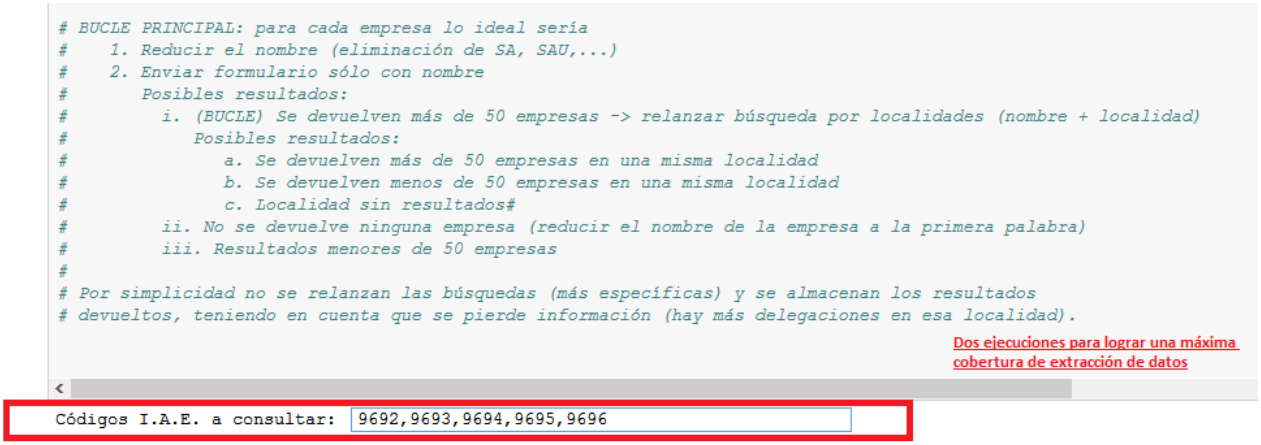


Ilustración 17: Input implementado para discriminar por código IAE

Otro aspecto a destacar del script sería la posibilidad de realizar consultas con un código de municipio. Aunque no se implementó en el bucle principal, sí se ha dejado preparado mediante descarga de dataset oficial del INE con la carga de los códigos de municipio y el municipio en sí.

Sobre esta dataset oficial sólo se han realizado consultas y se ha eliminado el último digito de control del código para obtener correspondencias directas.

	Nombre	NombreDel	Localidad	DirCompleta
2	Bluesblock, SA	BLUESBLOCK SA	BILBAO	CL URIBITARTE 00018 , 48001 BILBAO - VIZCAYA
3	Codere Online, SAU	CODERE APUESTAS ANDALUCIA SA	ANTEQUERA	CL TORRE DE HACHO 00003 , 29200 ANTEQUERA - MALAGA
4	Codere Online, SAU	CODERE APUESTAS ARAGON SLU	ZARAGOZA	CA JOSE PELLICER 00000 , 50007 ZARAGOZA - ZARAGOZA
5	Codere Online, SAU	CODERE APUESTAS ASTURIAS, S.A	GIJON	CA POLA DE SIERO 00008 , 33207 GIJON - ASTURIAS
6	Codere Online, SAU	CODERE APUESTAS BALEARES, S.L	CONSELL	CT PALMA A ALCUDIA 19400 , 07330 CONSELL - MALLORCA
7	Codere Online, SAU	CODERE APUESTAS CANTABRIA, S.L	SANTANDER	CA COLUMNA SAGARDIA 00003 , 39009 SANTANDER - SANTANDER
8	Codere Online, SAU	CODERE APUESTAS CASTILLA LA M	TOLEDO	CA JARAMA 00050 , 45007 TOLEDO - TOLEDO
9	Codere Online, SAU	CODERE APUESTAS CASTILLA Y LE	VALLADOLID	CA RECONDO 00011 , 47007 VALLADOLID - VALLADOLID
10	Codere Online, SAU	CODERE APUESTAS CATALUÑA, S.A	PALAU DE PLEGAMANS	CA MERCADERS 00001 , 08184 PALAU DE PLEGAMANS - BARCELONA
11	Codere Online, SAU	CODERE APUESTAS CEUTA SLU	CEUTA	GT DEL TENIENTE REINOSO 00000 , 51001 CEUTA - CARTAGENA
12	Codere Online, SAU	CODERE APUESTAS DEPORTIVAS S	NOAIN	P TALLUNTXE, CALLE C 00048 , 31110 NOAIN -
13	Codere Online, SAU	CODERE APUESTAS EXTREMADUR	CACERES	PO CAPELLANIAS 00105 , 10005 CACERES - CACERES
14	Codere Online, SAU	CODERE APUESTAS GALICIA SL	CORUÑA (A)	GT AMERICA 00000 , 15004 CORUÑA (A) - A CORUÑA
15	Codere Online, SAU	CODERE APUESTAS LA RIOJA SA	ARRUBAL	CA RIO PIQUERAS 00133 , 26151 ARRUBAL - RIOJA
16	Codere Online, SAU	CODERE APUESTAS MELILLA, S.A.	MELILLA	CA PUERTO DEPORTIVO 00011 , 52001 MELILLA - GIJON
17	Codere Online, SAU	CODERE APUESTAS MURCIA SL	MURCIA	LOS MARTINEZ 00004 , 30012 MURCIA - MURCIA
18	Codere Online, SAU	CODERE APUESTAS NAVARRA SA	PAMPLONA	PL DEL CASTILLO 00002 , 31110 PAMPLONA -
19	Codere Online, SAU	CODERE APUESTAS SA UNIPERSON	ALCOBENDAS	AV BRUSELAS 00026 , 28108 ALCOBENDAS - MADRID
20	Codere Online, SAU	CODERE APUESTAS VALENCIA SA	PICANYA	AV ALQUERIA MORET 00019 , 46210 PICANYA - VALENCIA

Ilustración 18: ficheroSedes.CSV

Otra técnica que se ha usado es la extracción de datos de redes sociales mediante API, en concreto datos de Twitter. Pero antes de realizar esta extracción, se ha realizado una **revisión manual** de las empresas para obtener los distintos usuarios de Twitter.

De esta API se extrayeron las siguientes variables:

- description, describe el usuario
- statusesCount, recoge la actividad del usuario, incluyendo retweets
- followersCount, recoge el número de seguidores del usuario
- favoritesCount, recoge los tweets o retweets que han sido marcados como favoritos por el usuario
- friendsCount, recoge el número de usuarios a los que sigue este usuario
- url, incorpora la dirección web que ha incluido el usuario
- name, recoge el nombre del usuario en su cuenta
- created, indica en que fecha fue creado el usuario
- protected, indica si los tweets e información de la cuenta está o no protegida
- verified, indica si el usuario ha verificado la cuenta
- screenName, indica el nombre o alias del usuario
- location, aparece recogido si existe una localización específica de usuario
- lang, indica el idioma autoidentificado por el usuario
- id, es el número de identificación único del usuario para twitter
- listedcount, indica el número de listas públicas del que el usuario es parte
- followRequestSent, si se ha enviado solicitud de amistad
- profileImageUrl, sitio donde se encuentra la imagen del perfil del usuario
- UsuarioTwitter, variable-conexión propia que vincula al usuario con otras tablas de este proyecto, en función de su perfil en twitter. Coincide con la variable name dada por twitter.

Sobre la lista anterior se realizó la selección de variables final, con las variables:

VARIABLE	*CARACTERÍSTICA*	*EJEMPLO*
Usuario Twitter	El usuario @...	CodereApuestas
statusesCount	Actividad del usuario, incluyendo retweets	14213
FollowersCount	Seguidores del usuario	22267
friendsCount	Usuarios a los que sigue	0
Created	Fecha de creación del usuario	2012-11-19 13:00:41
ListedCount	Listas públicas del que el usuario es parte	69

A continuación se muestra una captura con algunas de las entradas:

statusesCou	followersCo	friendsCoun	created	listedCount	UsuarioTwitter
14213	22267	0	2012-11-19 13:00:41	69	CodereApuestas
17331	4610	782	2011-01-07 12:29:41	77	CasinoGranMad
0	0	11	2011-08-12 08:05:20	0	Eurojuego
16612	22664	392	2008-12-18 09:37:56	209	loterias_es
16099	1177	184	2009-10-12 18:06:02	58	Suertia
20560	5521	323	2010-04-13 20:41:31	77	MARCAapuestas
6307	7983	932	2013-10-28 10:15:14	42	888pokerSpain
1239	503	98	2017-05-29 15:53:22	3	JueggingES
25570	14569	3632	2009-04-20 13:48:18	312	Betfair_ES
198	52	58	2017-03-27 15:33:33	0	BetwayES
13645	2500	268	2012-07-13 07:04:16	33	CasinoBcnES
21232	40572	564	2009-10-22 08:47:55	178	sportium
4796	759	624	2014-06-02 16:24:52	17	ebingo_es
11802	4153	376	2010-01-07 14:40:38	56	apuestasRETA
40609	187402	2617	2011-10-04 11:37:51	391	bwin_es
583	251	310	2016-11-21 16:46:42	4	Paston_es
270	159	234	2014-02-25 08:59:40	1	circus_es
22053	10950	4290	2011-07-28 13:27:40	52	GoldenPark_Esp
56267	112496	366	2011-10-26 16:26:11	258	bet365_es

Ilustración 19: usodetwitter_filtrado.CSV

Además, de la información de redes sociales se realizaron distintas extracciones para posibles relaciones en futuros análisis del dataset, aunque no se incluyeron en el dataset final. Se adjuntan también en el anexo y son scripts en R que extraen datos sociales de las empresas (se quería relacionar ubicaciones con censo poblacional de las distintas provincias, así como resultados obtenidos en un periodo concreto).

5.2 Dataset final

El número de variables que compone el dataset final es de 9, con un total de 53 observaciones (empresas).

La tabla de variables es la siguiente:

VARIABLE	***CARACTERÍSTICA***	***EJEMPLO***
Nombre	Nombre de la empresa (puede tener múltiples delegaciones / franquicias) con las que comparte las licencias	Codere Online, SAU
Licencias	Licencias singulares obtenidas para cada tipología. Puede estar vacío (‘Sin licencia vigente’), lo que indica que está extinguida	Ruleta, Black Jack,...
Dominios	Sitios web donde opera la empresa ofreciendo el servicio para el que tiene licencia. Puede no tener ningún dominio o varios (0,n)	Www.codere.es,...
Usuario Twitter	El usuario @...	CodereApuestas
statusesCount	Actividad del usuario, incluyendo retweets	14213
FollowersCount	Seguidores del usuario	22267
friendsCount	Usuarios a los que sigue	0
Created	Fecha de creación del usuario	2012-11-19 13:00:41
ListedCount	Listas públicas del que el usuario es parte	69

También se muestra una captura con el dataset unificado:

	A	B	C	D	E	F	G	H	I
	Nombre	**Licencias**	**Dominios**	**statusesCou**	**followersCo**	**friendsCoun**	**created**	**listedCount**	**UsuarioTwitter**
	Bluesblock, SA	Concursos	www.bluesblock.es						
	Codere Online, SAU	Ruleta,Black Jack,Máquinas de azar - s	www.codere.es,apues	14213	22267	0	2012-11-19 13:00:41	69	CodereApuestas
	Concursos Multiplataformas, SAU	Concursos							
	Esgaming, SAU	Ruleta,Black Jack,Punto y banca,Máqu	www.casinogranmadri	17331	4610	782	2011-01-07 12:29:41	77	CasinoGranMad
	Eurojuego Star, SA	Concursos	www.eurojuego.es,ww	0	0	11	2011-08-12 08:05:20	0	Eurojuego
	JS2015 Games, SAU								
	Sociedad Estatal Loterías y Apuestas d	Deportivas mutuas,Hípicas mutuas	www.loteriasyapuesta	16612	22664	392	2008-12-18 09:37:56	209	loterias_es
	Suertia Interactiva, SA	Ruleta,Black Jack,Máquinas de azar - s	www.suertia.es	16099	1177	184	2009-10-12 18:06:02	58	Suertia
	Unidad Editorial Juegos, SA	Ruleta,Black Jack,Punto y banca,Máqu	apuestas.marca.es,m	20560	5521	323	2010-04-13 20:41:31	77	MARCAapuestas
	888 Spain, Plc	Póquer,Ruleta,Black Jack,Máquinas de	www.888.es,www.888	6307	7983	932	2013-10-28 10:15:14	42	888pokerSpain
	Banegras Unión, SA	Ruleta,Black Jack,Punto y banca,Máqu	www.juegging.es	1239	503	98	2017-05-29 15:53:22	3	JueggingES
	Beatya Online Entertainment, PLC	Ruleta,Black Jack,Punto y banca,Máqu	www.starvegas.es		*Hay empresas que no tienen datos en Twitter, y algunas incluso sin dominio*				
	Betfair International, PLC	Póquer,Ruleta,Black Jack,Máquinas de	www.betfair.es	25570	14569	3632	2009-04-20 13:48:18	312	Betfair_ES
	Betway Spain, PLC	Ruleta,Black Jack,Punto y banca,Máqu	www.betway.es	198	52	58	2017-03-27 15:33:33	0	BetwayES
	Casino Barcelona Interactivo, SA	Póquer,Ruleta,Black Jack,Máquinas de	www.casinobarcelona	13645	2500	268	2012-07-13 07:04:16	33	CasinoBcnES
	Cirsa Digital, SAU	Póquer,Bingo,Ruleta,Black Jack,Punto	www.sportium.es	21232	40572	564	2009-10-22 08:47:55	178	sportium
	Comar Inversiones, SA	Ruleta,Black Jack,Punto y banca,Máqu	www.ijuego.es						
	Ebingo Online España, SA	Bingo,Ruleta,Black Jack,Máquinas de a	ebingo.es	4796	759	624	2014-06-02 16:24:52	17	ebingo_es
	Ekasa Apuestas Online, S.A.	Deportivas de contrapartida,Otras de co	apuestas.retabet.es	11802	4153	376	2010-01-07 14:40:38	56	apuestasRETA
	Electraworks España, PLC	Póquer,Ruleta,Black Jack,Máquinas de	www.bwin.es,www.pa	40609	187402	2617	2011-10-04 11:37:51	391	bwin_es
	Euroapuestas Online, SAU	Ruleta,Black Jack,Máquinas de azar - s	www.paston.es	583	251	310	2016-11-21 16:46:42	4	Paston_es

Ilustración 20: operadoresDGOJTwitter.CSV

5.3 Entregables (disponibles en Github)

Respecto a los entregables se adjuntan los siguientes como anexo, y están disponibles en el repositorio:

SCRIPTS

extractDGOJ.py:	extrae operadores con licencia
scrapCamaras.py:	extrae delegaciones
scrapWiki.R:	extrae datos censales
twitterEmpresasJuego.R:	extracción de datos de Twitter

A estos procesos hay que añadir procesos manuales de búsqueda de usuarios de Twitter, así como para la unificación final.

DATASET

10codcomun.xls:	extraído del INE, para consultar códigos de municipios y su equivalencia
operadoresDGOJ.csv:	datos de operadores, licencias y dominios (si los tiene)
ficheroURL.csv:	datos de las URLs para extraer datos de cada delegación
ficheroSedes.csv:	datos de las delegaciones/franquicias de cada empresa con licencia
provincias.csv:	datos socioeconómicos de provincias de España
usodetwitter_filtrado.csv:	dataset con el listado reducido de variables extraído del API

operadoresDGOJTwitter.csv: dataset final de estudio

Habría que tener en cuenta que hay 'missing values' en el dataset y habría que establecer un criterio para realizar algunos análisis. Para análisis descriptivos, como frecuencia de cada licencia en la población este criterio no es necesario.

En cambio, en otros análisis que usarán los datos obtenidos de Twitter se decidiría descartar las observaciones que no disponen en usuario Twitter, con lo que se reduce la población de estudio a 38 empresas, a las que habría que eliminar las correspondientes delegaciones si se deciden usar.

6. Licencia

La licencia considederada es **Released Under CC BY-NC-SA 4.0 License**.

Detallamos el motivo por el cual ha sido elegido esta licencia para ello debemos tener en cuenta un pequeño análisis de las licencias indicadas que luego nos servirá como base a la elección del tipo de licencia. Hay que tener en cuenta que todas las licencias CC(Creative Commons) permiten derechos de autor sin restricciones pero no puede existir ningún tipo de intencionalidad comercial ni realización de modificaciones.

- BY permite la utilización de la información tanto para copiar, transmitir, utilizar tanto de manera visual como origen de nuevos trabajos solamente si se da constancia del autor de la información.
- SA Permite que sea origen de nuevos trabajos, pero estos deben tener la misma licencia que el trabajo original.
- NC Permite copiar, transmitir, utilizar y realizar trabajos que deriven de este, pero con una finalidad no comercial.
- ND Permite copiar, transmitir y utilizar, pero no realizar trabajos que deriven de este.

4.0 Corresponde a la última versión de CC, estas son utilizadas en gran medida por las jurisdicciones. Teniendo en cuenta las licencias de cada una de las páginas web que hemos accedido se considera como las más apropiada la licencia Released Under CC BY-NC-SA 4.0 License.

En la página de la Dirección General de Ordenación del Juego se establece que las condiciones para la reutilización de la información que ofrece deben ser la siguiente:

1. No se debe modificar el sentido de la información
2. Debe citarse la fuente de los documentos objetos de la reutilización.
3. Se debe indicar la fecha en la que se han actualizado por última vez los datos.

4. No se debe de ninguna manera indicar que la Dirección General de la Ordenación del Juego tiene algo que ver como beneficio propio la reutilización de la información.
5. Hay que indicar en todo momento las condiciones de reutilización de los datos.

En el caso de Twitter no ha existido restricción, ya que las cuentas no están protegidas y la actividad de los usuarios en este sentido es pública. Pero debemos tener en cuenta que la información extraída por esta página no debe tener ningún fin abusivo, de engaño o incluso que causa un mal físico o moral a ninguna persona.

ANEXO: Scripts

```
----------------------
TwitterEmpresasJuego.R
----------------------

#Script para la extracción de Twitter a través de API  (Usuarios empresas juego)

  #App previamente creada para esta práctica
  #Sitio de la App "Juego" https://apps.twitter.com/app/14410452/show (con login twitter)

  #Carga de librería twitteR con las funciones de extracción
    library(twitteR)
  #Carga de la librería ROAuth para autenticación en la conexión
    library(ROAuth)

  #Credenciales para la conexión a través de dev.twitter.com
    clave_API <- "hash_clave_API"
    clave_API_Secreta <- "hash_clave_API_secreta"
    token <- "hash_token"
    token_secreto <- "hash_token_secreto"

  # Establecimiento de la conexión con la API de Twitter
    #Función setup_twitter_oauth gestona el conjunto de credenciales de autenticación
    #para una sesión de la librería TwitterR del paquete httr paquete que, a su vez,
    #permite configurar funciones para trabajar con protocolos de autenticación
    setup_twitter_oauth (clave_API, clave_API_Secreta, token, token_secreto)

  #Creamos lista completa para toda las empresas (tengan o no usuario de twitter)
    listadeusuarios <- c('Bluesblock, SA','CodereApuestas','Concursos_multiplataformas',
'CasinoGranMad','Eurojuego','JS2015Games', 'loterias_es', 'Suertia', 'MARCAapuestas', '888pokerSpain',
'JueggingES', 'Starvegas SA', 'Betfair_ES', 'BetwayES', 'CasinoBcnES', 'sportium', 'ijuego',
'ebingo_es','apuestasRETA','bwin_es','Paston_es','EurobetInternational','circus_es','GoldenPark_Esp','GtechS
pain','bet365_es','Interwetten_Esp','Titanbet_Espana','KambiSpain','luckia_es','Marathonbet_es','merkurmagic
','Paf
consulting','carcaj','StarCasino_ES','PokerStarsSpain','wanabet_es','planetwin365es','Kirolbet_es','WilliamH
illES','Yobingo','Casino777es','botemania','Cigagameonline','HillsideEspa_aLeisure','NetEntGamming','PrimaNe
tworks','PtEntretenimientoonline','enracha','BingoTombola','ventura24','vivelasuerte')

  #Lanzamos búsqueda datos sobre usuarios específicos que si tienen cuenta de Twitter
  #Función lookupUsers permite extraer la información simultaneamente de un conjunto
  #de usuarios sin cargar la API con cola de solicitudes
    usodetwitter <- lookupUsers(c('CodereApuestas','CasinoGranMad','Eurojuego','loterias_es', 'Suertia',
'MARCAapuestas', '888pokerSpain', 'JueggingES', 'Betfair_ES', 'BetwayES', 'CasinoBcnES',
'sportium','ebingo_es','apuestasRETA','bwin_es','Paston_es','circus_es','GoldenPark_Esp','bet365_es','Interw
etten_Esp','Titanbet_Espana','luckia_es','Marathonbet_es','merkurmagic','StarCasino_ES','PokerStarsSpain','w
```

```
anabet_es','planetwin365es','Kirolbet_es','WilliamHillES','Yobingo','Casino777es','botemania','enracha','Bin
goTombola','ventura24','vivelasuerte'), includeNA=TRUE)

  #Devuelve variables para cada usuario:
    #description, describe el usuario
    #statusesCount, recoge la actividad del usuario, incluyendo retweets
    #followersCount, recoge el número de seguidores del usuario
    #favoritesCount, recoge los tweets o retweets que han sido marcados como favoritos por el usuario
    #friendsCount, recoge el número de usuarios a los que sigue este usuario
    #url, incorpora la dirección web que ha incluido el usuario
    #name, recoge el nombre del usuario en su cuenta
    #created, indica en que fecha fue creado el usuario
    #protected, indica si los tweets e información de la cuenta está o no protegida
    #verified, indica si el usuario ha verificado la cuenta
    #screenName, indica el nombre o alias del usuario
    #location, aparece recogido si existe una localización específica de usuario
    #lang, indica el idioma autoidentificado por el usuario
    #id, es el número de identificación único del usuario para twitter
    #listedcount, indica el número de listas públicas del que el usuario es parte
    #followRequestSent, si se ha enviado solicitud de amistad
    #profileImageUrl, sitio donde se encuentra la imagen del perfil del usuario
  #almacenamos los resultados para las variables extraidas en matriz "usodetwitter"
```

```
----------------------
ExtractDGOJ.py
----------------------



import urllib
import ssl
import pandas as pd
from bs4 import BeautifulSoup

# Función para devolver una cadena única separando los elementos por comas a partir de los elementos de una lista
def lista_cadena(lista):
  enCadena=""

  if (len(lista)==1):
    enCadena=enCadena+lista[0]

  else:
    for elto in lista:

      enCadena=enCadena+elto
      if (elto!=lic[len(lic)-1]):
        enCadena=enCadena+','

  return enCadena


# Con archivos locales guardados con el navegador
file1 = "file:///F:\PEC1WEBSCRAPING\Buscador de Operadores _ Dirección General de Ordenación del Juego.html"
file2 = "file:///F:\PEC1WEBSCRAPING\Buscador de Operadores _ Dirección General de Ordenación del Juego2.html"
file3 = "file:///F:\PEC1WEBSCRAPING\Buscador de Operadores _ Dirección General de Ordenación del Juego3.html"

# Creación de contexto SSL
ctx = ssl.create_default_context()
ctx.check_hostname = False
ctx.verify_mode = ssl.CERT_NONE
```

```
# URL remotas
# CONCURSOS
url1 = "https://www.ordenacionjuego.es/es/operadores/buscar?field_got_tid=All&field_gat_tid=All&field_gct_tid=999993&field_dominio="
# APUESTAS
url2 = "https://www.ordenacionjuego.es/es/operadores/buscar?field_got_tid=All&field_gat_tid=999992&field_gct_tid=All&field_dominio="
# OTROS JUEGOS
url3 = "https://www.ordenacionjuego.es/es/operadores/buscar?field_got_tid=999991&field_gat_tid=All&field_gct_tid=All&field_dominio="

urls = (url1,url2,url3)
#urls = {file1,file2,file3}


# CON URL REMOTA

# Dos alternativas:
#   - CASO 1: No almacenar el tipo de licencia singular (concursos, apuestas u otros)
#   - CASO 2: Almacenar el tipo de licencia
# NOTA: Se ha implementado sólo el CASO 1, para el caso 2 habría que crear un campo adicional para almacenar el tipo de
# licencia; en cada iteración un tipo diferente por lo que habría empresas repetidas con mismo nombre, url pero distintas
# licencias.
# Almacenar el tipo de licencia podría ser útil por ejemplo si se desea realizar un estudio más específico por categorías

# CASO 1 (sólo es necesario un diccionario para las tres iteraciones, el diccionario no admite duplicados en el índice)
# ------------------------------------------------------------------------------------------------------------------
# datURL almacena los nombres y urls

datURL = dict()

for url in urls:

   pagina = urllib.request.urlopen(url,context=ctx)
   soup = BeautifulSoup(pagina,"lxml")
   contenidos = soup.find("div",class_="view-content").find_all("a")

   for empresa in contenidos:
      nombre = empresa.text.strip()
      link = "https://www.ordenacionjuego.es" + empresa.get("href").strip()
      datURL[nombre]=link
      print(nombre,link)

#print(datURL)

# Web scraping de empresas individuales
# Con los resultados almacenados en el diccionario datURL, se realiza scraping sobre cada página individual de cada empresa

datasetFinal = []

#url1 = "file:///F:\PEC1WEBSCRAPING\Bluesblock, SA _ Dirección General de Ordenación del Juego.html"

#url1="https://www.ordenacionjuego.es/es/op-antena3juegos"


#if url1 != '':
for url in datURL.values():

   #pagina = urllib.request.urlopen(url)
   pagina = urllib.request.urlopen(url, context=ctx)
   soup = BeautifulSoup(pagina,"lxml")
   contenidos = soup.find(attrs={'id':'operatorContent'})
```

```
    # Campo NOMBRE
    nombre = contenidos.find(attrs={'id':'operatorTitle'}).text.strip() # Strip para eliminar posibles espacios

    # Campo LICENCIAS(SINGULARES)
    licencias = contenidos.find(attrs={'id':'operatorSingular'}).find_all('li')
    lic=[]

    for licencia in licencias:
        lic.append(licencia.text)

    # Se procesa para que se almacenen con la forma Concursos,Lotería,... en una cadena de texto única
    # Se podrían haber almacenado como variables binarias (Concursos=Sí/No, Loterías=Sí/No,...)

    campoLicencias = lista_cadena(lic)

    # Campo DOMINIOS
    dominios = contenidos.find(attrs={'id':'opetatorBody'}).find_all('li')
    #print(dominios)
    dom=[]
    #if not dominios:
    #    dom = 'Sin_dominio'

    #else:
    #    for dominio in dominios:
    #        if dominio.find('a'):
            #            dom.append(dominio.find('a').text)

    if dominios:
        for dominio in dominios:
            if dominio.find('a'):
                dom.append(dominio.find('a').text)

    # Se procesan los dominios recuperados para dejarlos en una única cadena

    campoDominios = lista_cadena(dom)

    # Se añaden al total de observaciones

    datasetFinal.append((nombre,campoLicencias,campoDominios))

# Peculiaridades de los datos almacenados
#   1. Licencias: puede haber operadores que al buscar en el formulario aparezcan como poseedores de licencias, pero que
#      al acceder a su URL específica no muestren ninguna licencia (ej: antena3juegos). Si se consulta la tabla se puede ver
#      que tiene la licencia extinguida. Se especifica como 'Sin licencia vigente'
#   2. Dominios: en el caso de operadores que no tengan dominio, se especifica como 'Sin dominio'

# ALMACENAMIENTO EN DATAFRAME
#df['date'] = pd.to_datetime(df['date']) # yyyy-mm-dd

df = pd.DataFrame(datasetFinal, columns=['Nombre','Licencias','Dominios'])
#df.head()
#df.tail()
print(df)
df.to_csv('F:\\operadoresDGOJ.csv',index=False, encoding='utf-8')
#df = pd.read_csv('operadores.csv',encoding='utf-8')

#for valores in zip(datURL.keys(),datURL.values()):
#    print(valores)
```

```
---------------
ScrapCamaras.py
---------------

import urllib
import time as ti
import re
import pandas as pd
from bs4 import BeautifulSoup
import unicodedata

pd.set_option('display.max_colwidth', -1)

#-----------------------------------------------------------------------------------------------------------------
-
# FUNCIONES
#-----------------------------------------------------------------------------------------------------------------
-

# Devuelve una única separando los elementos por comas a partir de los elementos de una lista
def lista_cadena(lista):

    enCadena=""
    if (len(lista)==1):
        enCadena=enCadena+lista[0]
    else:
        for elto in lista:
            enCadena=enCadena+elto
            if (elto!=lic[len(lic)-1]):
                enCadena=enCadena+','
    return enCadena

# Comprueba acentos
def chkTil(s):
    if (unicodedata.category(c) != 'Mn'):
        return True
    else:
        return False

# Elimina acentos (en el censo aparecen búsquedas con "loterias" pero no con "loterías")
def sinTil(s):
   return ''.join((c for c in unicodedata.normalize('NFD', s) if unicodedata.category(c) != 'Mn'))

# Función que realiza una petición POST con los datos del formulario y devuelve una respuesta
# Sólo se realiza la petición cuando hay como mucho 5 epígrafes (LÍMITE IMPUESTO POR EL FORMULARIO)
# La clave es el nombre completo que sirve para hacer referencia a la "empresa" de las deleg.

def peticionPOST(nom='',direccion='',cpostal='',municipio='',epigrafes=''):

    urlSUBMIT = "http://www.camaras.org/censo/cgi-bin/listado.php"
    valoresForm      =      {'empresa'      :      nom,'direccion'      :      direccion,'cpostal'      :
cpostal,'municipios':municipios,'epigrafes' : epigrafes,'seccion':seccion}
    param = urllib.parse.urlencode(valoresForm)
    param = param.encode('ascii')
    req = urllib.request.Request(url=urlSUBMIT,data=param,method='POST')
    res = urllib.request.urlopen(req)
    return res
```

```
# Función que devuelve en forma de dataframe nombre, url y localidad de cada delegación
# de la lista de resultados

def extractURL(url,clave):

    soup = BeautifulSoup(url,"lxml")

    contenidos = soup.find("form").find_all(attrs={'class':'txt10gris'})

    nombre = list()
    enlace = list()
    localidad = list()

    for entrada in contenidos[::2]:

        nombre.append(entrada.find("a").text)
        enlace.append("http://mercurio.camaras.org/censo/" + entrada.find("a").get("href"))
        localidad.append(entrada.findNextSibling().text.strip())

    claves = clave * len(nombre)

    sedes                                                                                                    =
pd.DataFrame({'Nombre':claves,'NombreDel':nombre,'Enlace':enlace,'Localidad':localidad},columns=['Nombre','N
ombreDel','Enlace','Localidad'])
    return sedes


# Recibe el dataframe con nombre, enlace y delegación y lo completa accediendo a cada página con la dirección
def extraerDeleg(df):

    direcciones = list()

    for direc in df.Enlace:

        print('Extrayendo dirección de: ' + direc)
        res = urllib.request.urlopen(direc)
        soup = BeautifulSoup(res,"lxml")
        contenidos = soup.find_all(attrs={'class':'txt11gris'})
        #nomFinal = contenidos[0].text
        linea1 = contenidos[1].text
        linea2 = contenidos[2].text
        dirFinal = linea1 + ' , ' + linea2
        print(dirFinal)
        direcciones.append(dirFinal)
        # Para adapatrse al crawl-delay de robots.txt
        #ti.sleep(10)

    # Las direcciones se añaden al dataframe como una variable adicional
    df['DirCompleta'] = direcciones

    return(df)

#*********************************************************************************************

#codigos = "9692,9693,9694,9695,9696,9697,9821,9822,9825" # Problema: el formulario sólo admite 5 códigos como
máximo
#incluir 9825 ORGANIZ. APUESTAS DEPORTIVAS, LOTERIA
# Estrategia utilizada: se lanza el script 2 veces
# 1ª ejecución: códigos 9692,9693,9694,9695,9696 se guarda en ficheroSedes1.csv
#    tiempo de ejecución aproximado: 10s * 50
# 2ª ejecución: códigos 9697,9821,9822,9825 se guarda en ficheroSedes2.csv

#codigos = "9692,9693,9694,9695,9696"
```

```
    #codigos = "9697,9821,9822,9825"
codigos = input('Códigos IAE a consultar: ')
nombreCSV = input('Nombre de la tabla csv (sin la extensión): ')
direccion=''
cpostal=''
municipios=''
seccion='censo'

# DATASETS cargados en memoria (operadores y código de localidades obtenido del INE)

# El código de municipio se compone de cinco dígitos: los dos primeros corresponden
# al código de la provincia y los tres restantes al del municipio dentro de ésta.
# Asimismo se publica un sexto dígito de control que, asignado mediante una regla
# de cálculo, permite la detección de errores de grabación y codificación.


dfOperadores= pd.read_csv('operadoresDGOJ.csv',encoding='utf-8')

try:
    dfCM = pd.read_excel('10codmun.xls',header=1,dtype={'CPRO':str,'CMUN':str})
    dfCM['COD'] = dfCM['CPRO'] + dfCM['CMUN'] # Se unifican columnas y se elimina dígito de control
    dfCM['COD'] = dfCM['COD'].apply(lambda x: x[:-1])
    dfCM = dfCM.drop(['CPRO','CMUN'],axis=1) # Se eliminan columnas redundantes para liberar memoria
except:
    print("Fichero de códigos de municipios INE no está en la ruta del script")

# Almacena resultados de todas las consultas realizadas (en cada iteración una empresa)
dfURL = pd.DataFrame(columns=['Nombre','NombreDel','Enlace','Localidad'])

# PRIMER PASO: Scracping de URLs de las empresas operadoras de juego (resultados de las tablas)

for datos in dfOperadores.Nombre:

    nombreReducido = datos.split(',')[0]
    print('Procesando: ' + datos + ' #### Keyword en formulario: ' + nombreReducido)
    try:
        respuesta = peticionPOST(nom=nombreReducido,epigrafes=codigos)
        respuesta['Nombre'] = datos
        dfSedes = extractURL(respuesta,clave=datos)

        #Si no hay resultados, se vuelve a buscar pero sólo con la primera palabra de la empresa (Ej: CODERE
ONLINE -> CODERE)
        if dfSedes.empty:

            nombreReducido2 = datos.split(' ')[0]
            print(nombreReducido + " no muestra resultados. Probando con: " + nombreReducido2)
            respuesta = peticionPOST(nom=nombreReducido2,epigrafes=codigos)
            dfSedes = extractURL(respuesta,clave=datos)

        dfURL = dfURL.append(dfSedes)

    # Para adaptarse al crawl-delay de robots.txt
    #ti.sleep(10)

    except:
        print('Error al consultar los datos de ' + datos)

#dfURL.to_csv('ficheroURL.csv',index=False,encoding='utf-8')
#eliminar duplicados antes de consultas

# SEGUNDO PASO: Scraping de las direcciones de cada URL (delegación)

dfSedesFinal = extraerDeleg(dfURL)
```

```
dfSedesFinal = dfSedesFinal.drop('Enlace',axis=1)
dfSedesFinal.to_csv(nombreCSV + '.csv',index=False, encoding='utf-8')


# BUCLE PRINCIPAL: para cada empresa lo ideal sería
#    1. Reducir el nombre (eliminación de SA, SAU,...)
#    2. Enviar formulario sólo con nombre
#       Posibles resultados:
#        i. (BUCLE) Se devuelven más de 50 empresas -> relanzar búsqueda por localidades (nombre + localidad)
#            Posibles resultados:
#               a. Se devuelven más de 50 empresas en una misma localidad
#               b. Se devuelven menos de 50 empresas en una misma localidad
#               c. Localidad sin resultados#
#         ii. No se devuelve ninguna empresa (reducir el nombre de la empresa a la primera palabra)
#         iii. Resultados menores de 50 empresas
#
# Por simplicidad no se relanzan las búsquedas (más específicas) y se almacenan los resultados
# devueltos, teniendo en cuenta que se pierde información (hay más delegaciones en esa localidad).
```

```
-----------
scrapWiki.R
-----------

require("xml2")
require("rvest")
url.provincias<- "https://es.wikipedia.org/wiki/Anexo:Provincias_y_ciudades_aut%C3%B3nomas_de_Espa%C3%B1a"
tabla_temp<- read_html(url.provincias)
tabla_temp <- html_nodes(tabla_temp, "table")
tabla_temp
length(tabla_temp)
sapply(tabla_temp,class)
html_table(tabla_temp,fill=TRUE)
#sapply(tabla_temp, function(x) dim(html_table(x, header=false, fill = TRUE)))
provincias<- html_table(tabla_temp,fill=TRUE)
provincias
write.csv(provincias, file="provincias.csv") #guardamos en un archivo CSV.
```