Islas Canarias
Del 15 al 19 de noviembre de 2021

Junta de Andalucía
Consejería de Transformación Económica, Industria, Conocimiento y Universidades
Instituto de Estadística y Cartografía de Andalucía

UN CASO DE USO DE R APLICADO AL WEB SCRAPING EN ESTADÍSTICAS DE INSERCIÓN LABORAL

Víctor Montañés Cobo
Instituto de Estadística y Cartografía de Andalucía
victor.montanes@juntadeandalucia.es

Juan De Dios Ramos Poyatos
Instituto de Estadística y Cartografía de Andalucía
juand.ramos@juntadeandalucia.es

Introducción

En los dos últimos años, el IECA ha desarrollado dos nuevas publicaciones estadísticas relacionadas con la incorporación al mercado laboral de los estudiantes. La primera de ellas se centra en los estudiantes universitarios que han realizado y finalizado estudios en universidades públicas andaluzas. La segunda, focalizada en los alumnos egresados de formación profesional. Estas publicaciones se han realizado con la cooperación del Servicio Andaluz de Empleo (SAE), la Dirección General de Universidades (DGU) y la Consejería de Educación y Deporte.

Su objetivo es estudiar la empleabilidad de los egresados de estas titulaciones en distintas ventanas temporales de forma que se muestre cómo evoluciona la incorporación al mercado laboral desde la finalización de los estudios hasta 4 años después. Este trabajo se centra en la inserción de los egresados universitarios tratando de paliar de forma experimental la información deficitaria de los registros administrativos en algunas situaciones.

Objetivos

Se establecen tres objetivos en el presente trabajo:

- Identificar las lagunas de información que tienen los registros administrativos de base por su propia concepción o por su ámbito competencial.
- Estudiar el potencial uso de las nuevas técnicas de recuperación masiva de datos a través de la web (web scraping) para mejorar las estadísticas de inserción laboral de los egresados universidades públicas que han finalizado sus estudios en Andalucía.
- Identificar situaciones en las que las técnicas de Web Scraping tampoco son capaces de ofrecer información.

Metodología

1. Análisis de fuentes

En primer lugar es necesario realizar un análisis exhaustivo de las fuentes originales de procedencia de los datos para poner de manifiesto posibles déficit de información.

- Fichero de mutualistas titulares de MUFACE

Información recibida: Mutualistas titulares residentes en Andalucía al finalizar el trimestre.

Déficit: Solo residentes en Andalucía

- Fichero de afiliaciones en alta laboral y cuentas de cotización a la Seguridad Social.

Información recibida: Se reciben trimestralmente los afiliados a la Seguridad Social dados de alta en alguno de los regímenes del Sistema de la Seguridad Social con fecha de referencia del último día del trimestre.

Déficit: No se recibe información completa de las altas fuera de Andalucía y no se recibe nada de trabajadores en el extranjero. No se recibe información de mutualidades, salvo la que disponemos del fichero de MUFACE.

- Ficheros de contratos de trabajo y personas demandantes de empleo

Información recibida: Demandas registradas en alguna de las oficinas del Servicio Andaluz de Empleo (SAE) registradas a último día laborable de cada mes. Contratos registrados durante el mes.

Déficit: Solo contratos gestionados desde oficinas SAE o bien que el trabajador o la empresa indiquen que residen (o tienen sede) en Andalucía.

- Sistema Integrado de Información Universitaria

Información recibida: Rendimiento académico, acceso y avance en títulos universitarios oficiales impartidos en universidades públicas de Andalucía conforme a las especificaciones técnicas del Sistema Integrado de Información Universitaria (SIIU).

Déficit: No hay información sobre universidades privadas ni de fuera de Andalucía. No hay información sobre títulos propios (algunos máster y cursos de experto).

Con la información disponible se puede clasificar a una parte de la población objetivo en las siguientes categorías de forma jerárquica:

- Trabajando
- Demandando empleo en el Servicio Andaluz de Empleo
- Continuando su formación al estar matriculado en otro título universitario en una universidad pública andaluza

- No trabaja ni está inscrito en Andalucía como demandante, pero ha trabajado durante el trimestre. (Siendo laxos, sería asimilable a un parado tipo EPA)
- Otras situaciones especiales: cobrando pensión de jubilación, fallecido.
- Resto.

Nuestra atención se centrará precisamente en esta última categoría de resto y el siguiente gráfico es un punto de partida que motiva nuestro foco en este colectivo.

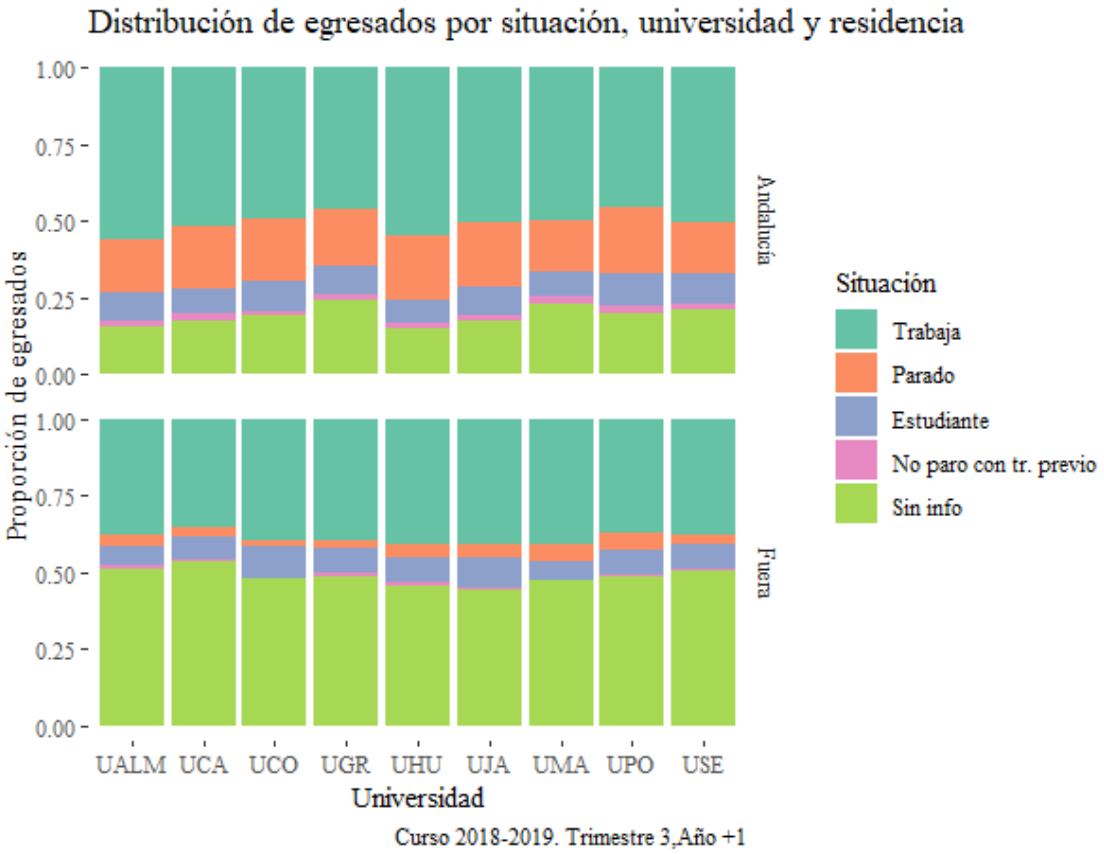


Figura 1: Distribución de egresados por relación con la actividad, universidad y lugar de residencia un año después del egreso. Curso 2018-2019.

Salta a la vista que este resto, o sin información es mucho más numeroso para estudiantes que no estaban empadronados en Andalucía (información que se obtiene de la Base de Datos Longitudinal de Población de Andalucía). Este gráfico no es más que la visualización de las carencias de información que tenemos con las fuentes disponibles.

Resumiendo, no tenemos información de trabajo o estudios en otras comunidades autónomas salvo alguna excepción. Quizás en un futuro sería viable consultar datos del Ministerio de Trabajo o de Educación para paliar este defecto, pero en un corto o medio plazo no parece una opción plausible. Pero sobre todo, no tenemos ojos fuera de España y no existe una fuente administrativa que nos permita obtener información de trabajo o estudios en el extranjero. En el caso de mutualistas, salvo con MUFACE, tampoco disponemos de información.

Además, para considerar que un egresado continúa estudiando, estamos restringiendo nuestra búsqueda a las propias universidades públicas andaluzas.

2. Búsqueda en Internet

Conocidos los ángulos muertos de nuestra información, el objetivo que se plantea es intentar recuperar información disponible en la web, tanto del ámbito público como del privado. Para este estudio hemos centrado la atención en 4 fuentes: Boletines Oficiales, TESEO, LinkedIn y ResearchGate.

Boletines oficiales

Los boletines oficiales (BOE, BOJA y análogos) ofrecen información personal que es de obligada publicación. El origen de pensar en su uso reside en que existen listados de nombramientos de funcionarios que se publican en ellos. En principio, el foco se ha centrado en extraer las listas de aprobados de teóricos miembros de la mutualidad ISFAS y la administración de Justicia en el BOE. A lo largo de los años el BOE ha ido modificando su formato, tanto el texto como la información mostrada para estos cuerpos que históricamente han tenido una especial sensibilidad con la protección de los datos personales de sus integrantes.

En los BOE más antiguos, de escasa utilidad para la publicación de egresados, las listas aparecen como formato imagen e incluyen el nombre, apellidos y DNI de los aprobados. Para extraer la información se ha utilizado el paquete tesseract que hace reconocimiento óptico de caracteres.

Los BOE posteriores ya se publicaban en pdf y se puede utilizar el paquete tabulizer para extraer tablas que vienen dentro de estos pdf.

Finalmente, los BOE actuales se distribuyen en formato html pero con una mayor protección de carácter personal. Podemos encontrar listados con nombre y apellidos y unas pocas posiciones del DNI o todas las posiciones del DNI pero sin literales identificativos. De cualquier forma, la información se puede obtener de una forma bastante sencilla con el paquete rvest identificando los nodos de información.

Si bien este es el trabajo que se ha llevado a la práctica, se podrían llevar otros trabajos adicionales en lo sucesivo, como por ejemplo:

- Automatizar la búsqueda de todos los boletines oficiales[1] , al menos el capítulo de "Autoridades y Personal". Nos permitiría conocer personas que trabajan en el sector público fuera de Andalucía o en empleos que no aparecen en las Altas de Seguridad Social.

- Descargar o scrapear (según el método de publicación) los listados de inscritos y aprobados en distintas fases de los procesos de oposiciones, fundamentalmente los de exámenes de oposiciones de primaria y secundaria. Nos permitiría conocer

[1]Carlos Gil Bellosta, DataAnalytics. "Aventuras de Web Scraping: Cómo bajarse todo el BOE" https://www.datanalytics.com/2014/04/24/aventuras-de-web-scraping-como-bajarse-todo-el-boe/

una parte de estos “ninis” que no trabajaban ni demandaban empleo[2] porque se encontraban opositando (al menos han pagado unas tasas de examen).

Podemos ilustrar con cifras un caso concreto, el de los graduados en Educación Primaria, cuyo volumen supera los 2.500 egresados residentes anuales. Si nos centramos en el porcentaje de egresados que continúa estudiando, se observa un patrón estacional cada dos años.

Los años en los que hay oposiciones baja el número de egresados que continúa estudiando, por lo que es de esperar que al menos la diferencia entre años consecutivos se deba a personas que están estudiando oposiciones. Se recopiló el listado de presentados y aprobados en estas oposiciones en 2019 del Boletín Oficial de la Junta de Andalucía (BOJA) para estudiar la presencia del colectivo sin información en estas pruebas selectivas.

El escrapeo de los boletines oficiales tiene la ventaja de que la información es muy fiable, por las implicaciones que conlleva, pero sin embargo le falta completitud, ya que dentro de los aprobados no tendríamos necesariamente ni el centro ni la fecha de incorporación, si bien en el caso de los maestros podríamos tenerla a través de la información que proporciona el sistema Séneca de la Junta de Andalucía.

TESEO

La información recibida por parte de las universidades tenía ciertas deficiencias para las promociones más antiguas en el caso de los egresados a través de un doctorado. Por ejemplo, había una gran cantidad de información faltante sobre el campo de estudio. Para poder realizar una imputación fiable se recurrió a TESEO, que es una base de datos del Ministerio de Educación con información sobre las tesis doctorales leídas en universidades españolas desde 1976. Por tanto, esta es una modificación utilizada para mejorar las variables de clasificación pero no modificaría de ninguna manera el dato de inserción.

TESEO permite la búsqueda tanto por número de documento como por nombre y apellidos, que es la opción que hemos utilizado por disponer la fuente original de ella y buscar solo la información faltante, pero también se puede hacer una búsqueda a ciegas por universidad y curso académico.

[2]Una de las causas de exclusión del paro registrado es: “1083. Estudiantes menores de 25 años o que demandan primer empleo si tienen más de 25 años”

Ilustración 1: Cuadro de búsqueda de TESEO

El acceso se ha hecho con el paquete Rselenium de R. Este paquete necesita disponer de una versión del navegador Google Chrome dentro de un listado válido que se puede consultar con la orden

binman::list_versions("chromedriver")

Una vez que disponemos de una versión compatible, se abre una sesión virtual a través de un puerto libre del PC (en principio en blanco), apareciendo un mensaje que indica que "Un software automatizado de pruebas está controlando Chrome". A partir de entonces la navegación es controlada en remoto por nuestro código.

Llegados a este punto, se abre la página de TESEO[3] y se recorre con un bucle el listado de DNI de doctorados con información faltante. Se identifica el nombre de la caja de búsqueda del DNI y del botón "Buscar".

La siguiente página contiene el título de las tesis compatibles con la búsqueda, en nuestra búsqueda por DNI casi siempre única, y es necesario marcar con una casilla de verificación la tesis que nos interesa y clicar en "Ver selección". Finalmente llegamos a una pantalla en la que se obtiene toda la información relativa a la tesis, incluyendo un resumen de la tesis, el tribunal, dirección de tesis, fecha de lectura...aunque para lo que andábamos buscando lo útil era el departamento y el programa de doctorado.

[3]https://www.educacion.gob.es/teseo/irGestionarConsulta.do

› **Título:** MAPAS PARA EL PATRIMONIO. CARACTERIZACIÓN TÉCNICA DE LAS IGLESIAS MEDIEVALES DE SEVILLA MEDIANTE INFORMACIÓN GEOGRÁFICA (SIG).
› **Autor:** Mascort Albea, Emilio J.
› **Universidad:** Universidad de Sevilla
› **Departamento:** Estructuras de edificación e ingeniería del terreno
› **Fecha de Lectura:** 16/02/2018
› **Mención Europeo / Mención Internacional:** Concedido
› **Programa de doctorado:** Programa de Doctorado en Arquitectura por la Universidad de Sevilla

Ilustración 2: Ejemplo de información devuelta por TESEO

Esta información se obtiene de una manera sencilla utilizando el paquete rvest. Finalmente se guarda la información en un data.frame (o similar) y es posible identificar el campo utilizando patrones y troceando con el paquete stringr.

LinkedIn

LinkedIn es probablemente la red profesional más conocida a nivel mundial. En ella se pueden consultar desde ofertas de empleo a publicaciones propias, intereses y currículos de las personas que forman parte de ella. A diferencia del BOE, y por obvio que parezca hay que indicarlo, cada persona elige libremente si quiere formar parte de ella e incluso dispone de diversos niveles de privacidad sobre qué información se quiere mostrar y a quien.

Hay una caja de búsqueda en la que se escribe el nombre de la persona buscada y aparecen generalmente varios candidatos, y no necesariamente el primero que aparece es el más parecido al texto buscado, sino el que LinkedIn considera más relevante (contactos comunes, coincidencia con lugar de trabajo, misma universidad de estudios…). Para esta prueba experimental se ha tomado el primer candidato únicamente.

Lo habitual al entrar en el perfil de una persona es encontrar fundamentalmente tres bloques de información: tarjeta resumen, experiencia laboral y estudios.

En la tarjeta resumen se muestran nombre y apellidos, una descripción del puesto de trabajo y la empresa y unos iconos de la empresa (si procede) y la institución en la que estudió. También suele aparecer el ámbito territorial en el que se mueve la persona, que no deja claro si es el de residencia, trabajo o búsqueda de empleo.

Ilustración 3: Ejemplo de cabecera de usuario de LinkedIn

El segundo bloque es el de experiencia autodeclarada y puede estar más o menos adornado. Por lo general se indica el nombre de la empresa, el cargo, periodo de tiempo

que se ha trabajado en la empresa y la localización de la empresa. Y decimos que "por lo general" porque hay mucha libertad a la hora de que el usuario deje apartados en blanco. Por último, en el tercer apartado aparece la trayectoria académica reportada, con estructura similar al apartado de experiencia.

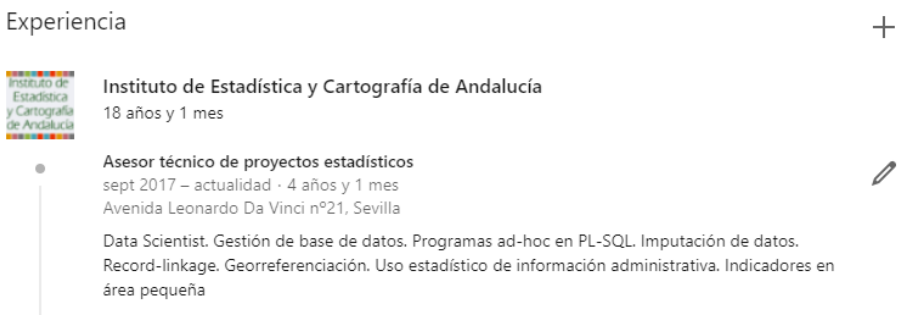


Ilustración 4: Ejemplo de datos de Experiencia de LinkedIn

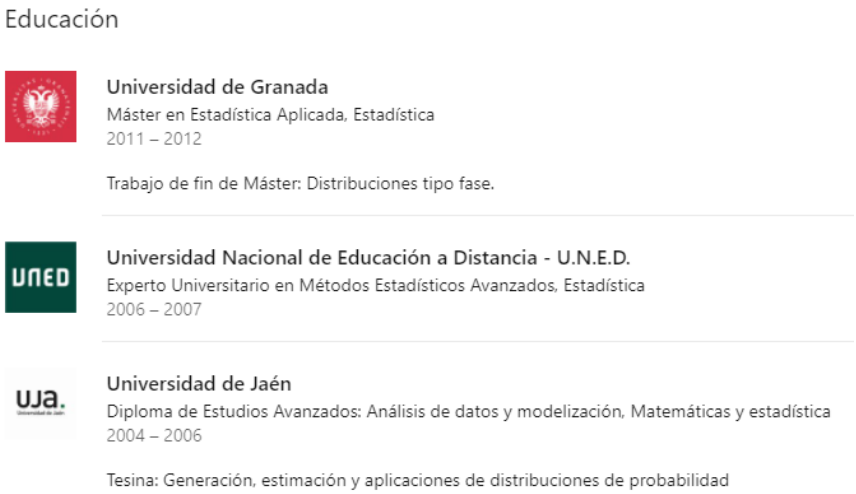


Ilustración 5: Ejemplo de datos de Educación de LinkedIn

El procedimiento utilizado pasa por el uso de los paquetes Rselenium y rvest de forma combinada. Al igual que hacíamos en TESEO, se abre una sesión de Chrome en forma remota y se va navegando hasta la página principal, rellenando las credenciales necesarias. Al ser la búsqueda de información más compleja, se crea una función de extracción en la que se pasa como primer parámetro la página de un currículum y como segundo parámetro qué parte nos interesa, ya que nos interesa diferenciarlas a la hora de almacenarlas.

Realizamos la búsqueda de una persona con el listado que tenemos (a través de la conexión a nuestra base de datos Oracle) y hacemos clic en el primer candidato. Una vez que aparece el currículum almacenamos en una tabla la clave de la persona junto con el enlace de su currículum para facilitar posteriores búsquedas. Aplicamos la función de extracción de la información del currículum y se guarda la salida, tras una limpieza del texto con stringr.

La parte final consiste en validar la información del candidato. Esta validación es doble, ya que por un lado se controla la proximidad entre las cadenas de caracteres de nombre y

apellidos en LinkedIn frente a las de nuestra Base de Datos Longitudinal de Población de Andalucía-BDLPA (con el paquete stringdist) y por otro validamos si la universidad en la que realizó los estudios y el año de finalización según LinkedIn es compatible con la información proporcionada por las universidades.

ResearchGate

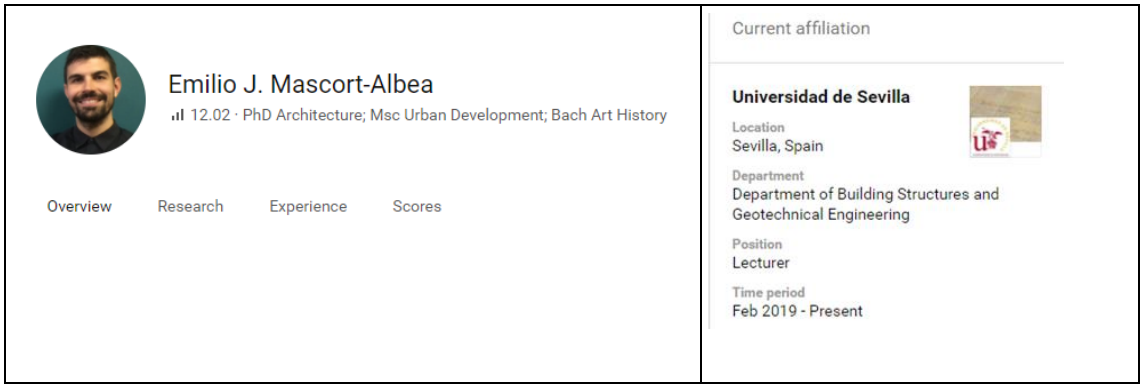
Se trata de un portal destinado a investigadores. De hecho, para darse de alta como usuario se exige alguna prueba documental de que ser investigador. El uso común de este portal es crear redes con otros investigadores y compartir publicaciones, independientemente de su índice de impacto. Es, nuevamente, una web cuya inscripción es totalmente voluntaria.

El potencial uso es determinar si un investigador está trabajando en una institución extranjera (generalmente universidad) y es útil tanto para doctorados como para doctorandos.

La búsqueda se hace de manera análoga a las anteriores. Disponemos de un listado de individuos para los que queremos recuperar información y, abriendo una sesión virtual con RSelenium utilizamos el cuadro de búsqueda por nombre y buscamos. Ahora hay una diferencia con LinkedIn porque se pueden dar 3 casos de forma habitual:

- Se encuentra a la persona dada de alta
- No se encuentra dada de alta, pero se encuentra un artículo en el que participa (o se le menciona)
- No se encuentra ningún tipo de información

Recuperamos únicamente en el caso de que la persona esté dada de alta, ya que al acceder al perfil del usuario hay una cabecera con foto, nombre, titulación y un apartado denominado “Current affiliation” que señala los datos de su trabajo actual. En el caso de que esté vigente, nos da la fecha de inicio y la ubicación.



El apartado Research es otro de los que nos podría proporcionar una gran información. En él aparecen las publicaciones en las que ha participado y una fecha de publicación. En algunos casos, no en todos, se puede acceder al texto del artículo, total o parcialmente. No obstante, no nos permite ver directamente en qué organismo trabajaba cuando realizó la publicación. Lo más habitual es que tras el nombre del autor aparezca un superíndice y que en otra parte de la primera página aparezca asociado a ese superíndice el lugar de

desarrollo del trabajo. Por lo tanto, para rastrear los artículos se necesita que sean accesibles, la descarga del pdf, la lectura OCR, la identificación del autor y la detección de la empresa en la que trabaja (generalmente una universidad). Extraer esta información conlleva una alta complejidad y se descartó en este ensayo.

Existen otras web de carácter similar, como por ejemplo Google Académico (Scholar)

Resultados

En el caso de búsquedas en BOE se centró el foco en teóricos miembros de las mutualidades ISFAS y MUGEJU. Con este proceso, en el que se buscaron los BOE se recuperaró información de algo más de 100 trabajadores.

Se realizó también una búsqueda de las personas presentadas a las oposiciones de Educación Primaria de 2019, para los que se obtuvieron su nombre, apellidos, cuatro posiciones del DNI y sus calificaciones. Para los egresados en el curso 2016-2017, se buscaron los resultados en la oposición de Junio de 2019 y se contrastó con su relación con el mercado laboral a 1 de Octubre de 2018. De aquellos empadronados de los que no teníamos información, un 25% pagó las tasas para la realización del examen y la mitad de ellos superó todas las pruebas. Asumiendo que estas personas que aprueban se están dedicando al estudio de las pruebas, tendríamos que 1 de cada 8 egresados de los que se desconoce su situación, estaban preparando (lo suficientemente en serio como para haber aprobado) las pruebas de acceso.

La cifra más baja en cuanto a aprobados se sitúa en el grupo de los que estaban trabajando en la fecha de referencia (17% presentados y 38% de ellos superaron las pruebas). Pasa lo mismo si nos vamos retrotrayendo a cursos previos.

En el caso de TESEO, se recuperó la información requerida de la práctica totalidad de los individuos.

En el caso de LinkedIn, por citar algunos resultados, se tomó una pequeña muestra de un 3.5% de individuos egresados del curso 2015-2016 y clasificados como “Resto”. De ellos, cerca de la tercera parte fueron localizados con un perfil en LinkedIn. Para más del 40% de ellos se localizó que en la fecha de referencia (aproximada) se encontraban en el extranjero, trabajando (otros regímenes, se entiende) o estudiando. El colectivo de estudiantes suponía casi un tercio de los perfiles buscados y llamaba la atención la continuidad en los estudios en organismos de tipo “Escuela de Negocio”, que quedan fuera del ámbito de nuestras fuentes de tipo administrativo.

La información de ejemplo, de un egresado en el curso 2015-2016, quedaría de la siguiente forma.

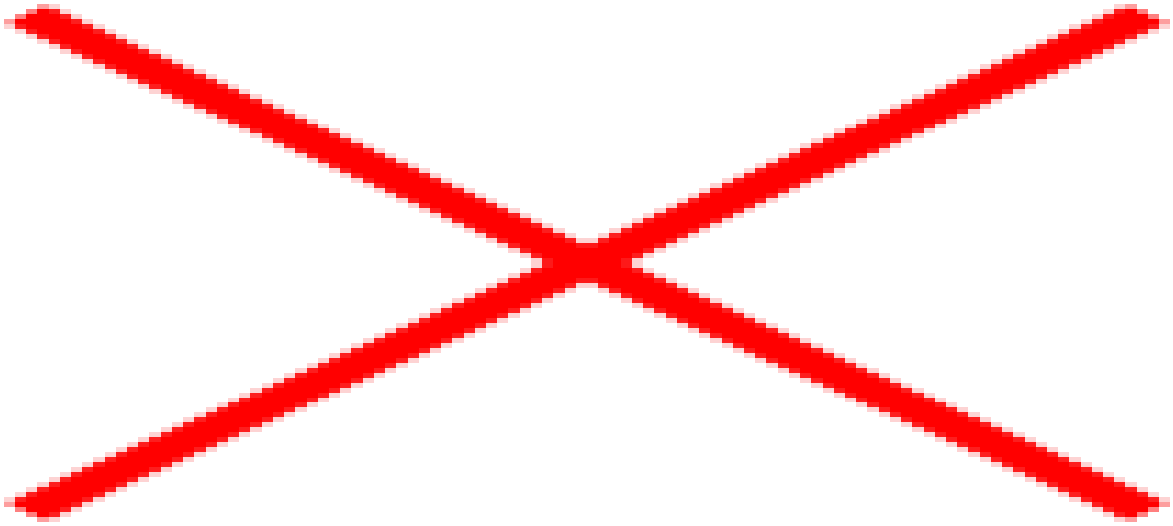


Ilustración 6: Información extraída de LinkedIn para un individuo clasificado como "Resto"

Como vemos, el motivo de no disponer de información en algunos trimestres es su permanencia en el extranjero.

La fuente se muestra útil, aunque insuficiente, para el propósito de corregir este grupo de personas sin información y además necesita de un proceso de limpieza y revisión posterior. No obstante, el hecho de que los estudios y trabajos vengan con un rango de fechas, es muy útil para comparar con nuestras fechas de referencia.

Un problema podría ser que debemos presuponer la veracidad de la información reportada, aunque si esto no supone un problema en una encuesta, no debería suponerlo aquí.

Por último, se hace necesario hacer el estudio de manera conjunta para todas las situaciones de relación con la actividad de la publicación para establecer si hay un sesgo en lo relativo a la inscripción en esta red social profesional en función de si se busca empleo o se está ocupado.

En el caso de Research Gate, el porcentaje de usuarios localizados está por debajo del de LinkedIn, al menos en cuanto a los clasificados en “Resto”. También es de esperar que si no trabajan en un organismo de investigación o universidad, no se hayan interesado por mantener un perfil en esta página, por lo que parece que hay un sesgo en favor de los que sí están activos.

Otros resultados, o cuanto menos aprendizajes de este proceso, han sido los siguientes:

- El código de escrapeo está totalmente sujeto a cambios en la estructura interna de la web. Evidentemente, el propietario de la web no avisa de los cambios que va a hacer y es necesario reprogramar el código casi desde 0.

- A las empresas no les gusta que se obtenga información de ellas y limitan el número de búsquedas que se pueden realizar en un día.

- Es necesario tomar precauciones para que en algunas webs se entienda que somos humanos y no robots. Una de ellas es establecer tiempos de espera aleatorios cada vez que realizamos una acción. Estos tiempos también nos ayudan si la página tarda en cargar.

- Es muy útil utilizar la función try. Esta función permite ejecutar líneas de código que no se detienen si hay un error. Es muy útil cuando después de una búsqueda no sabemos si vamos a obtener resultados.

Conclusión

El web scraping se muestra como una herramienta útil para complementar la información presente en los registros administrativos. No obstante, plantea nuevos problemas como la estimación de la cobertura de las fuentes no administrativas y los posibles sesgos debidos al perfil de usuario de cada individuo. Por ejemplo, es menos probable que alguien que trabaje en el sector público educativo (primaria y secundaria) esté dado de alta en una plataforma de este tipo que un autónomo con una titulación de ingeniería.

Un porcentaje relevante de individuos que teóricamente no tienen información están realizando estudios en ámbitos que no cubren las fuentes administrativas, como escuelas de negocio o academias de preparación de oposiciones.

También hace que nos replanteemos la forma de hacer estadísticas. Si tomamos la información de registros administrativos, podemos publicar una cifra sin (casi) ningún tipo de temor, porque tendría un carácter censal. Si introducimos elementos no exhaustivos como páginas de inscripción voluntaria, ya estaríamos en el terreno de la estimación pero con el problema de que el diseño está realizado por un tercero y la selección de unidades no es aleatoria.

Por otro lado, se hace necesaria que la relación empresa-administración esté equilibrada, ya que la información no fluye de igual manera en ambas direcciones y de esta forma, en el terreno de las estadísticas laborales, tanto los datos de un lado como los del otro tienen carencias.

Principales referencias bibliográficas

- Medición del número de viviendas turísticas en España y su capacidad. INE. (2020). Metodología

- El reto Big Data para la estadística pública. Alberto González Llanes. (2018). Trabajo de fin de máster.

- Estadísticas de inserción laboral de los egresados en Andalucía. Universidad y FP. (XXI Jecas, 2021). Ponencia