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

TÉCNICAS DE WEB SCRAPING APLICADAS A LAS ESTADÍSTICAS DE INSERCIÓN LABORAL

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

Mª Rosa Cánovas Balboa
Instituto de Estadística y Cartografía de Andalucía
rosa.canovas@juntadeandalucia.es

Jose E. Molina Trapero
Instituto de Estadística y Cartografía de Andalucía
josee.molina@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. Cuando este producto se desarrolló ya existían otras aproximaciones a nivel nacional y autonómico, de las que destacamos las siguientes:

- Encuesta de Inserción Laboral de los titulados universitarios del INE, siendo el año 2019 la última vez que se realizó y sin periodicidad establecida. Igualmente lo ha realizado para egresados de FP, siendo el último año 2019.

- El Ministerio de Ciencia, Innovación y Universidades ha elaborado el informe Indicadores de afiliación a la Seguridad Social de los egresados universitarios siendo el último analizado el correspondiente a la promoción 2013-2014. Para este último curso se hace seguimiento a la inserción laboral de los egresados hasta el año 2018. El Ministerio de Educación y Formación Profesional ha realizado la correspondiente a FP hasta el curso 2015-2016.

- Las oficinas estadísticas de otras comunidades autónomas, como el Instituto Canario de Estadística y el Instituto de Estadística de las Islas Baleares han realizado sus propias estadísticas de inserción laboral de egresados basadas en la fusión de registros administrativos.

- También con base en la integración de información administrativa, el Observatorio ARGOS del Servicio Andaluz de Empleo elabora un informe anual denominado tanto para egresados de FP como para universitarios.

Es el modelo de los institutos de estadística insulares el más parecido al elegido para la publicación. Para más detalles, consultar la ponencia "Estadísticas de inserción laboral de los egresados en Andalucía. Universidad y FP" de estas JECAS.

Objetivos

El objetivo final de este trabajo es 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 titulados de FP (público o privados) y universidades públicas que han finalizado sus estudios en Andalucía.

Para conseguir este objetivo es necesario alcanzar dos objetivos intermedios:

- Clasificar a los individuos en función de su relación con el mercado laboral y académico.

- Determinar los vacíos de información que se derivan de las fuentes administrativas utilizadas.

- Determinar fuentes de información externa que puedan complementar la información existente.

- Evaluar problemas derivados de la incorporación de esta información.

Metodología

En primer lugar se ha realizado un análisis exhaustivo de las fuentes originales de procedencia de los datos para poner de manifiesto posibles déficit de información. En segundo lugar se han buscado fuentes externas que permitan recuperar información total o parcial de la que no se disponga en las fuentes originales. Finalmente, se han utilizado técnicas de web scraping para recuperar información de terceros.

- Base de Datos de Población de Andalucía

Este sistema de información tiene como objetivo integrar, reutilizar y facilitar el seguimiento de la información recopilada en diferentes registros estadísticos y administrativos de base poblacional. El núcleo de la BDLPA está formado por los efectivos de la población y sus características, que se obtienen vinculando e integrando la información individual procedente de censos de población, padrones municipales de habitantes, y las variaciones residenciales derivadas de los mismos, así como los flujos de defunciones y nacimientos procedentes del Movimiento Natural de la Población.

La base tiene una doble función. La primera es proporcionar variables nucleares como lugar de residencia, sexo o año de nacimiento. La segunda es permitir la comunicación con otras fuentes de información sectoriales y permitir intercomunicar estas fuentes auxiliares. Uno de los subproductos de la BDLPA es la posibilidad de obtener la población residente para cualquier día del año.

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

Se reciben trimestralmente los afiliados a la Seguridad Social dados de alta en alguno de los regímenes del Sistema de la Seguridad Social, tanto por cuenta ajena como por cuenta propia. La información temática aportada comprende el tipo de contrato, el régimen y grupo de cotización, la actividad económica de la empresa y la provincia de la cuenta de cotización.

Es necesario remarcar que el fichero de afiliaciones contiene las personas afiliadas que trabajan en Andalucía.

- Fichero de mutualistas titulares de MUFACE

La información individual integrada en la BDLPA se refiere a mutualistas titulares residentes en Andalucía. También se recibe trimestralmente.

- Ficheros de contratos de trabajo y personas demandantes de empleo

El IECA recibe mensualmente, con fecha de referencia el último día laborable del mes, un fichero que contiene todas las demandas registradas en alguna de las oficinas del Servicio Andaluz de Empleo, así como los contratos de trabajo registrados.

- Sistema Integrado de Información Universitaria

Las universidades suministran los ficheros periodicidad anual: rendimientos académicos y de avance de las matriculaciones en títulos universitarios oficiales impartidos en universidades públicas de Andalucía conforme a las especificaciones técnicas del Sistema Integrado de Información Universitaria (SIIU). El fichero de rendimientos académicos está referido a 30 de septiembre (fecha de finalización del curso académico) y de él se extraen los registros correspondientes a los egresos producidos en el curso de referencia.

- Sistema Séneca de Educación no universitaria

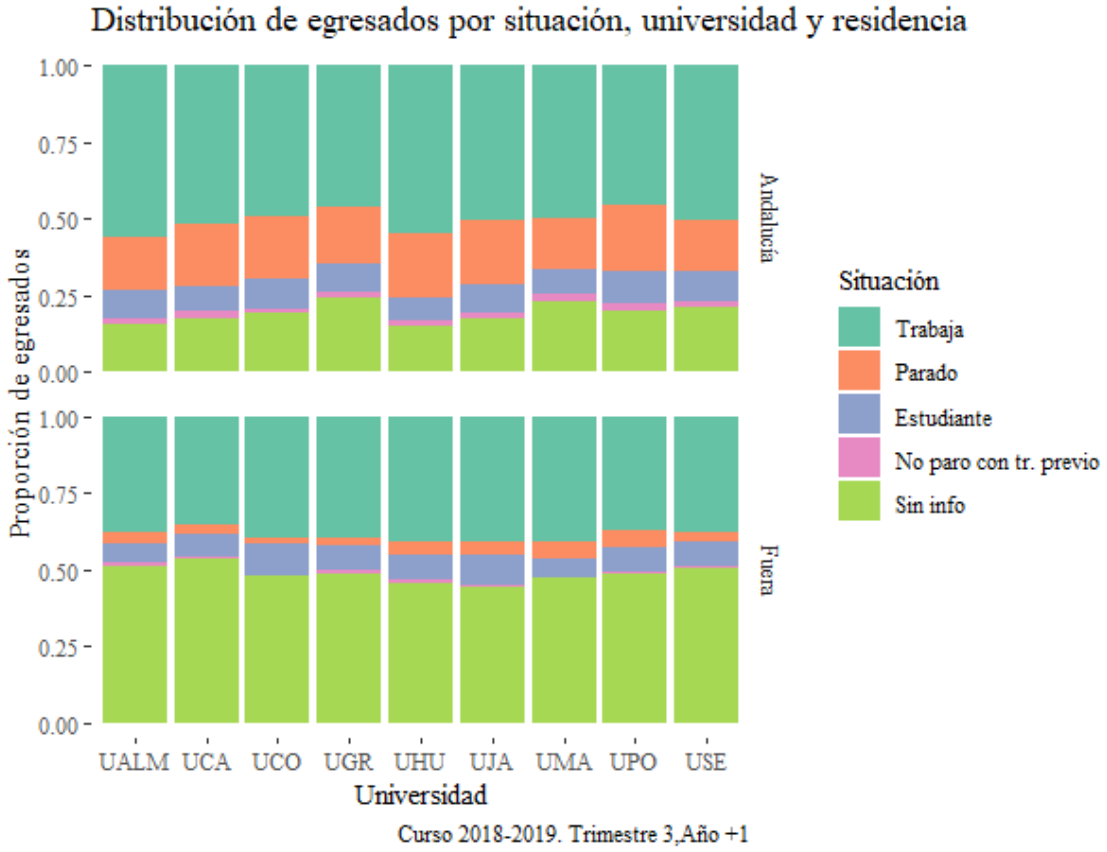
Este fichero proporciona matriculaciones y resultados académicos (curso o titulación superados) y permite extraer los registros que corresponden a los egresos producidos en cada curso de referencia en las diferentes especialidades profesionales ofertadas por el sistema educativo andaluz.

Clasificación de los individuos y sesgos asociados

Con toda esta información, se toman como referencia el último día de cada trimestre del año posterior al egreso y el último día del tercer trimestre entre el segundo y el cuarto año. En cada punto temporal de referencia, se asigna a cada individuo una situación de relación con el mercado de trabajo, según la siguiente clasificación. Si una persona está en más de una categoría, se determina el principal de manera jerárquica. La clasificación de las personas queda de la siguiente forma:

- Trabajando
- Demandante de empleo que había trabajado anteriormente (en cuenta de cotización de Andalucía).
- Demandante de empleo que no había trabajado anteriormente (en cuenta de cotización de Andalucía).
- Continuando su formación en el sistema público andaluz (otro título universitario en caso de egresados universitarios).
- No trabaja ni está inscrito en Andalucía como demandante, pero ha trabajado durante el trimestre.
- Otras situaciones especiales: cobrando pensión de jubilación, fallecido.
- Resto.

Un primer gráfico es muy elocuente en cuanto al déficit de las fuentes.



Si queremos dar un indicador de empleabilidad, a priori podemos considerar que es el porcentaje de trabajadores respecto al total. Si quisiéramos aproximar una tasa de paro, el problema sería mucho más complejo.

Si volvemos al gráfico, vemos que la distribución de los que residían (empadronados) en Andalucía durante el último curso de estudios y los que no, nos ofrece un panorama muy diferente. La distribución de aquellos sin información, que a priori no estudian, no trabajan ni buscan empleo a través del servicio público es totalmente distinta. Podríamos tener la tentación de decir que la inserción o la búsqueda de empleo es mayor entre los residentes, pero tenemos que leer la letra pequeña de las categorías que hemos establecido y de las fuentes que hemos utilizado.

Es precisamente el conjunto de personas sin información la que nos impide dar un aproximador de una tasa de paro o de actividad. Este colectivo que hemos denominado resto, teóricamente inactivos, que podríamos denominar "ninis" que no estudian, ni trabajan ni buscan empleo (al menos esto último desde el punto de vista de inscripción en una oficina de empleo) no es tal. El problema es que la falta de exhaustividad derivada de las fuentes disponibles hacen que no estemos seguros de la composición de este colectivo, ya que entre otras situaciones, pueden haber sucedido las siguientes:

- La persona está inscrita en una oficina de empleo de fuera de Andalucía.
- La persona es mutualista de una mutua diferente de MUFACE, que es la única de la que disponemos de información.
- La persona se encuentra trabajando o estudiando en el extranjero.
- La persona ha continuado su formación en una universidad privada o de fuera de Andalucía.

Estos matices son muy importantes, ya que la composición del alumnado de las 9 universidades públicas andaluzas es muy diversa. Por ejemplo, el porcentaje de egresados en 2018 que no se encontraban empadronados en Andalucía al comienzo del curso en el que finalizaron sus estudios, va desde el 7,5% de la Universidad de Málaga hasta el 25,3% de la Universidad de Granada, que además tiene la particularidad de tener campus en Ceuta y Melilla, fuera de la Comunidad Autónoma.

En cuanto a egresados extranjeros, la cifra va desde el 3,7% de Jaén hasta el 7,7% de Granada. Por lo tanto, realizar una aproximación a una tasa de paro con estos perfiles distantes de población podría introducir sesgos relevantes. Puestos a hacerlo, sería mucho más fiable para los residentes.

Una de las cosas que se ha conseguido con esta publicación es distinguir si la afiliación es en Andalucía o fuera, lo cual permite ver a cuántos egresados absorbe el mercado laboral andaluz y cuántos buscan su futuro fuera. Para ilustrar con cifras esta mejora, para los estudiantes españoles no empadronados en Andalucía que egresaron en el curso 2018/2019, el porcentaje de egresados empleados que no tenían su alta en una cuenta de cotización de Andalucía iba desde el 60% de la Universidad de Almería hasta el 77% de la Universidad de Granada. En cualquier caso, más de de la mitad de las altas no se producían en la comunidad, es decir, los estudiantes volvían a su lugar de procedencia o bien buscaban una tercera vía fuera de Andalucía.

Para los españoles empadronados en Andalucía, las cifras están muy por debajo, pero no por ello resultan despreciables. Respecto al total de trabajadores detectados en las fuentes, el porcentaje de los que trabajaban en cuentas de fuera de Andalucía iba desde el 12,7% de Almería hasta el 23,6% de Jaén, personas no asimiladas por el mercado laboral regional.

Tras esto, hay un dato que realmente llama la atención. Mientras que para ciudadanos españoles el porcentaje de los localizados como trabajadores un año después de la finalización de los estudios es ligeramente superior al 50%, para extranjeros es de tan solo el 15%. ¿A qué se debe esto? Como parece evidente, a que no disponemos información administrativa de fuera de España (salvo algún contrato aislado gestionado por las oficinas de empleo, por ejemplo con Gibraltar).

Pero no todos los problemas vienen por la falta de información de los que se marchan al extranjero. Dentro de España existen otras mutualidades distintas de MUFACE, como MUGEJU (Justicia) o ISFAS (Fuerzas Armadas). Incluso determinados colectivos de profesionales colegiados (arquitectos técnicos, gestores administrativos, abogados, arquitectos superiores, químicos, procuradores, peritos e ingenieros técnicos industriales, médicos) tienen la opción de ser mutualistas de mutualidades de previsión social alternativas, en cuyo caso se encuentran exentos de la obligación de alta en el Régimen Especial de los Trabajadores por Cuenta Propia o Autónomos de la Seguridad Social. Esto hace que algunas titulaciones puedan estar viéndose afectadas por esta falta de información.

Al margen de estos problemas, hay otros relacionados con la estacionalidad. El primero de ellos es el de los egresados en Medicina, que pasado un año desde el egreso tienen un porcentaje de alumnos trabajando muy elevado. Sin embargo, la incorporación del grueso de esta titulación se produce casi masivamente en un trimestre, que depende de cuando haya sido el examen de MIR y la posterior elección de plazas. Por lo tanto, un mal dato en un trimestre en este colectivo puede deberse simplemente a la preparación de las pruebas de acceso, que no constan en ningún registro, aunque es evidente que se asimilaría más a la categoría de "sigue estudiando".

El otro caso es el de los egresados del sector educativo, fundamentalmente de la educación obligatoria. En estos casos las pruebas de acceso para formar parte del profesorado se realizan cada dos años y las incorporaciones se realizan en el mes de septiembre del año en el que se realizan las pruebas. Esto hace que el indicador fluctúe cuando se compara en clave longitudinal con otras promociones.

Otro problema adicional que se produce con las categorías "Sigue estudiando" y "Resto", es debido a que, aunque disponemos de la información de egresados y matriculados en las universidades públicas andaluzas, la información se refiere a titulaciones oficiales, por lo que los títulos propios o cursos de experto quedan fuera de la información de referencia. Tampoco existe información sobre universidades privadas ni sobre universidades del resto de España y por supuesto tampoco del extranjero.

Resultados

Ante los problemas que se muestran, en el IECA nos planteamos explorar nuevas vías para estudiar si alguna fuente adicional, pública o privada, podía darnos al menos una estimación (aunque fuera a la baja) del grado de desviación de los datos administrativos respecto a la situación real.

La extracción de los datos se ha realizado con el software R y los paquetes empleados han sido rvest para la extracción de la información de la web, Rselenium para generar sesiones virtuales del navegador, hacer click en los botones y rellenar los formularios y tabulizer para extraer tablas de ficheros en pdf. Además se ha utilizado tesseract para la lectura de información en formato imagen.

La búsqueda se ha centrado en cuatro fuentes fundamentalmente:

- Boletín Oficial del Estado (BOE)
- LinkedIn
- Research Gate
- Teseo

A continuación detallaremos los principales objetivos, virtudes y defectos de cada una de estas fuentes.

Boletín Oficial del Estado

El objetivo de acudir al BOE se ha centrado en extraer las listas de aprobados de las oposiciones 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.

Por ejemplo, en los BOE más antiguos, las tablas con los aprobados aparecen como imagen y con nombre, apellidos y DNI. Luego aparecieron pdf legibles y ahora ya se puede consultar en html, pero los aprobados ya solo aparecen con el DNI.

En función de la tipología se emplearon los paquetes tesseract para el reconocimiento óptico de los que venían como imagen, tabulizer para los pdf en que se leía en formato tabla y rvest para los que venían en un html.

Si bien de cara a futuro (corto plazo) el html será el utilizado, el uso de tesseract podría ser de gran utilidad en otras estadísticas de carácter histórico.

Ventajas:

- Podría hacerse extensible a cualquier tipo de nombramiento de obligada publicación.
- La información es muy veraz por las implicaciones que tiene.

Inconvenientes:

- Habría que hacer un barrido de todos los BOE y se conoce quienes son los aprobados, pero no necesariamente el centro de destino ni la categoría profesional.

Para el curso 2015-2016 se localizaron algo más de 100 personas por este método.

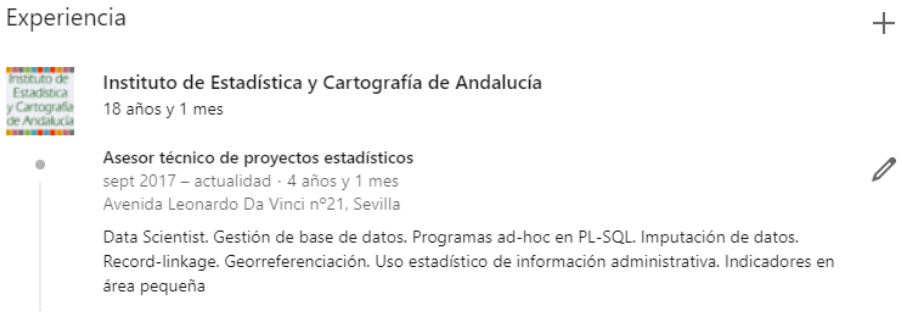
LinkedIn

LinkedIn es probablemente la red profesional más conocida. En ella se pueden consultar desde ofertas de empleo a post, 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.

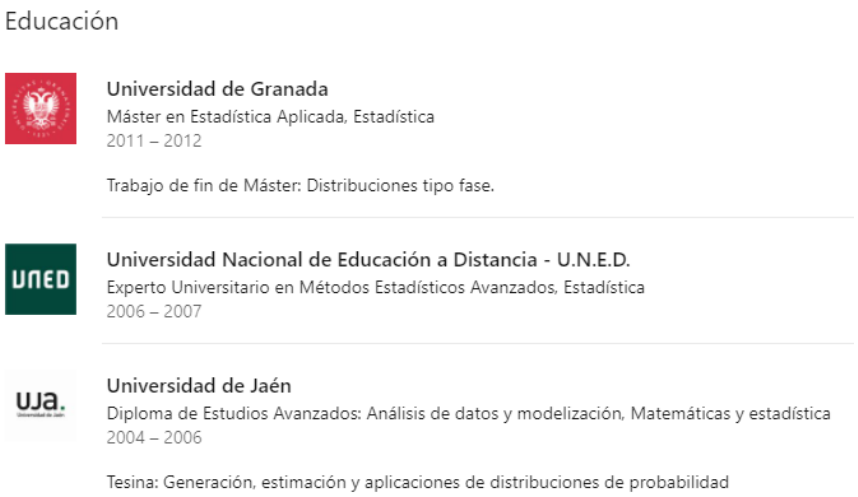
Lo habitual al entrar en el perfil de una persona es encontrar fundamentalmente tres bloques de información: tarjeta resumen, experiencia laboral y estudios.

El primero de ellos es una especie de tarjeta de presentación con nombre y apellidos, con una descripción del puesto de trabajo y la empresa y unos iconos de la empresa y el lugar de estudios. 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.

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.



Por último, en el tercer apartado aparece la trayectoria académica reportada, con estructura similar al apartado de experiencia.



Ventajas:

- En muchos casos se dispone del periodo de permanencia en una titulación o un empleo.
- Permite observar si ha estudiado en otra universidad al finalizar los estudios en una universidad andaluza.
- Permite observar si ha tenido empleos fuera de España, aunque necesitaría un tratamiento posterior.
- La información está relativamente estructurada

Inconvenientes:

- No es una fuente exhaustiva y la presencia de egresados varía con el tipo de titulación.
- La información extraída necesita de varios procesos de limpieza y no siempre tiene la misma estructura
- Se debe presuponer la veracidad de la información.
- Es posible que no se declaren todos los empleos porque a la persona no le interese mostrar ciertos contratos en su currículum.
- Al hacer consultas masivas, LinkedIn bloquea las consultas temporalmente, al menos si no se es usuario “Premium”.

Como curiosidad, indicaremos que el primer caso que buscamos declaraba un empleo en Andalucía en la fecha de referencia que no estaba registrado ni como contrato en los datos del Servicio Andaluz de Empleo ni como alta en Seguridad Social.

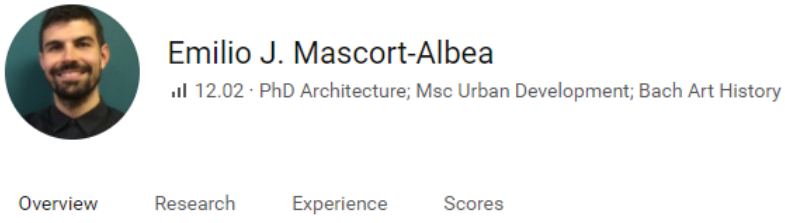
Por citar algunos resultados, se tomó una pequeña muestra de un 3.5% de individuos clasificados como otras situaciones. De ellos, cerca de la tercera parte fueron localizados con un perfil en LinkedIn. Para más del 40% se localizó que en la fecha de referencia

(aproximada) se encontraban fuera (a través de la ubicación declarada), 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”.

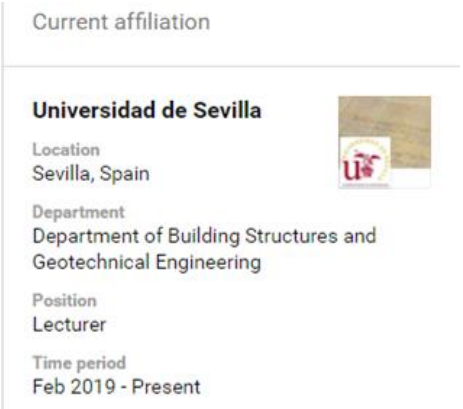
Research Gate

Se trata de un portal destinado a investigadores. De hecho, para darse de alta como usuario se exige alguna prueba documental de que se es 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.

Al igual que LinkedIn, tiene una pantalla resumen inicial en la que hay una cabecera con foto, nombre y titulación.



Dispone de un apartado denominado “Current affiliation” que señala los datos de su trabajo actual. En este caso, no disponemos de trabajos anteriores (salvo que aparezcan en la pestaña Experience) por lo que podríamos seguir con la incertidumbre en nuestra fecha de referencia. En el caso de que esté vigente, nos da la fecha de inicio y la ubicación.



ORIGINAL PAPER

Check for updates

Analysing the life index of diamond cutting tools for marble building stones based on laboratory and field investigations

Maziyar Bahri[1] · Ebrahim Ghasemi[2] · Mohammad Hossein Kadkhodaei[2] · Rocío Romero-Hernández[1] · Emilio J. Mascort-Albea[1]

Received: 22 January 2021 / Accepted: 15 July 2021

✉ Maziyar Bahri
mazbah@alum.us.es

[1] Escuela Técnica Superior de Arquitectura, Departamento de Estructuras de La Edificación E Ingeniería del Terreno, Instituto Universitario de Arquitectura Y Ciencias de La Construcción, Universidad de Sevilla, Avenida de Reina Mercedes 2, 41012 Sevilla, Spain

[2] Department of Mining Engineering, Isfahan University of Technology, 8415683111 Isfahan, Iran

Published online: 28 July 2021

El apartado Research es otro de los que nos puede 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).

Por el perfil de usuario, esta fuente es de especial utilidad para personas con un doctorado. Entre este colectivo la red tienen un gran calado, mucho más que LinkedIn.

Ventajas:

- Muy extendida entre investigadores
- Permite observar traslados fuera de España

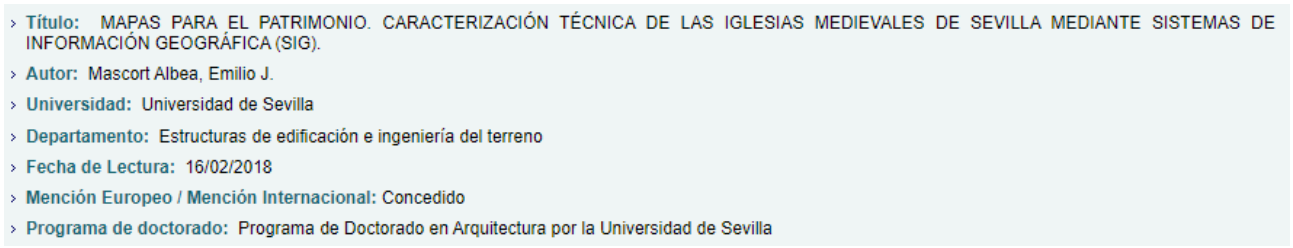
Inconvenientes:

- La información del centro de trabajo es muy compleja de sacar de los artículos.

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. Permite la búsqueda tanto por número de documento como por nombre y apellidos.

En el resultado se devuelve toda la información relevante de la tesis, incluido además un resumen de la tesis, el tribunal, dirección de tesis...aunque para lo que andábamos buscando lo útil era el departamento y el programa de doctorado.



El acceso se hizo de la siguiente forma. Se accedía a la página de TESEO con una sesión virtual de Google Chrome utilizando el paquete Rselenium, se tomaba un DNI con información faltante mediante consulta a nuestra base de datos interna y se pegaba esa información de manera automática en el buscador. A partir de ahí se continuaba de forma automatizada la navegación hasta la página de navegación y se extraía con las órdenes del paquete rvest y finalmente se guardaba en un objeto dataframe.

Ventajas:

- Permite buscar por DNI, nombre y apellidos o incluso por universidad, curso y departamento.
- Es (casi) exhaustiva

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.

La fiabilidad de la información reportada, a nuestro juicio, no sería tan problemática. Por ejemplo, cuando una persona responde a una encuesta de un organismo estadístico no nos planteamos que pueda estar proporcionando información falsa y, además, en los casos que estamos analizando poseemos información que se puede contrastar, como la

universidad en la que realizó los estudios, el año de finalización o la titulación obtenida, o incluso trabajos previos en empresas de Andalucía.

Otros problemas son que cualquier cambio en la estructura de la web y la distribución de la información lleva a tener que reprogramar por completo las búsquedas, o que las páginas protejan la información de sus usuarios limitando el número de consultas, lo que dificulta la búsqueda masiva de datos.

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 y ante un replanteamiento de cómo hacer las estadísticas.

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