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

PALABRAS CLAVE
Webscraping, egresados universitarios, linkedIN, Researchgate, TESEO

RESUMEN

La estadística de "Inserción laboral de los egresados en universidades públicas de Andalucía" ofrece, a partir de la explotación estadística de los ficheros de matrícula y rendimiento académico de los estudiantes de las universidades públicas de Andalucía, y tras el enlace con sus datos de empleo y desempleo disponibles en la Base de Datos Longitudinal de Población de Andalucía, información relativa al proceso de inserción laboral de los egresados universitarios, proporcionando un indicador de empleabilidad y tasas de inserción laboral.

Las fuentes administrativas disponibles para este análisis son insuficientes para cubrir exhaustivamente todas las posibles situaciones de relación con la actividad, quedando un colectivo aparente de "ninis" para los que no hay información sobre si se incorporan al mercado laboral o sobre continuidad en estudios de este colectivo. En este colectivo pueden aparecer individuos que trabajan en el extranjero, mutualistas pertenecientes a MUGEJU o ISFAS y otros sistemas de mutualidad, estudiantes en universidades de fuera de Andalucía o que realizan estudios no oficiales. Estos déficits pueden dar una medición sesgada del grado de inserción laboral de determinadas titulaciones.

Para estudiar en qué grado podría mejorarse la información se hizo una primera prueba de uso de técnicas de web scraping para suplir este déficit, apoyadas en las librerías rvest, tabulizer y Rselenium de R. Entre las acciones acometidas, se usó la información publicada en BOE de nombramientos de miembros de MUGEJU e ISFAS y también de las webs de perfiles profesionales linkedIN y Researchgate (fundamentalmente para doctores). Así mismo, se usó la información de la base de datos de Tesis Doctorales (TESEO) para completar la información de los doctorados incluidos en el fichero de explotación, enlazando cada doctorando con su tesis para recuperar las variables rama de enseñanza y ámbito/campo de estudio a partir del título de la tesis y el departamento al que estaba adscrita.

Esta nueva técnica abre una vía de posibilidades para aprovechar nuevas fuentes de información que completen lo reportado en los registros administrativos y que no se habían planteado debido a la dificultad para obtenerlas, ya que sin una forma de sistematizar y automatizar las búsquedas, sería necesario un trabajo absolutamente manual.