Una herramienta de Machine Learning para la actualización y el desarrollo del Directorio de Empresas y Establecimientos con actividad en Andalucía

Nuria Gómez-Vargas
Instituto de Matemáticas de la Universidad de Sevilla
ngvargas@us.es

Jasone Ramírez-Ayerbe
Instituto de Matemáticas de la Universidad de Sevilla
mrayerbe@us.es

Emilio Carrizosa Priego
Instituto de Matemáticas de la Universidad de Sevilla
ecarrizosa@us.es

Rafael Blanquero Bravo
Instituto de Matemáticas de la Universidad de Sevilla
rblanquero@us.es

Marina Enguídanos Weyler
Instituto de Estadística y Cartografía de Andalucía
marina.enguidanos@juntadeandalucia.es

Elisa Isabel Caballero Ruiz
Instituto de Estadística y Cartografía de Andalucía
elisai.caballero@juntadeandalucia.es

Ana Gema Galera Pozo
Instituto de Estadística y Cartografía de Andalucía
gema.galera@juntadeandalucia.es

PALABRAS CLAVE
Python, Inteligencia artificial, Innovación, Directorio, Empresas

RESUMEN

En el marco de la colaboración conjunta entre el Instituto de Estadística y Cartografía de Andalucía (IECA) y el Instituto de Matemáticas de la Universidad de Sevilla (IMUS) se está desarrollando un proyecto con dos objetivos: la actualización automática del Directorio de empresas y establecimientos con actividad en Andalucía, y el desarrollo de un modelo de inteligencia artificial que permita la identificación del carácter innovador de estas empresas. Para estas dos finalidades se han desarrollados herramientas en código abierto utilizando el lenguaje de programación Python haciendo uso intensivo de librerías existentes de aprendizaje automático, análisis de textos web, *web scraping* y manejo de bases de datos.

La actualización del Directorio de empresas y establecimientos en Andalucía implica recurrir a fuentes privadas y el acceso a registros administrativos cada año, por lo que es muy necesario el estudio de fuentes alternativas y el desarrollo de procedimientos de recuperación automática de información que eviten la dependencia de fuentes con tratamiento complejo de la información y que no tienen asegurada su continuidad. Los trabajos realizados se han centrado en la identificación de fuentes web de información complementarias, el desarrollo de herramientas en Python para la recuperación automática de información de las fuentes seleccionadas y el análisis de la calidad de sus datos, para contrastarlos mediante la definición de diferentes algoritmos con la información disponible en el Directorio.

Respecto a la determinación del carácter innovador de las empresas, actualmente el Instituto Nacional de Estadística, determina mediante encuestas el carácter innovador de las empresas de más de diez empleados, un enfoque que, por definición, excluye a muchas empresas andaluzas. Con el objetivo de definir una metodología para caracterizar la innovación que no dependa del uso de encuestas y que sea así, extrapolable a pequeñas empresas, se ha desarrollado una herramienta de aprendizaje automático implementada en Python que, usando bosques aleatorios, permite clasificar a las empresas según su innovación a partir del análisis de su página web, concretamente del texto y de distintas variables que la caracterizan.