JUNTA DE ANDALUCIA

CDAU -GEOCODER.IA
Descripción

Versión: 1.0.0

Fecha: 21/03/24

1.0.0

HOJA DE CONTROL

Organismo	Instituto de Estadística y Cartografía de Andalucía		
Proyecto	Callejero Digital de Andalucía Unificado - GEOCODER.IA		
Entregable	Documento de presentación		
Autor	IECA		
Versión/Edición	1.0.0	**Fecha Versión**	27/03/24
Aprobado por		**Fecha Aprobación**	
		Nº Total de Páginas	10

REGISTRO DE CAMBIOS

Versión	Causa del Cambio	Responsable del Cambio	Fecha del Cambio
1000	Versión inicial	jamoreno	27/03/24

CONTROL DE DISTRIBUCIÓN

Nombre y Apellidos

ÍNDICE

1 DESCRIPCIÓN DEL PROYECTO GEOCODER.IA 4
2 REPERCUSIÓN PARA EL CIUDADANO Y ADMINISTRACIONES 7
3 EQUIPO DE DESARROLLO Y PROVEEDORES 8
4 VALORACIÓN ECONÓMICA 9
5 PLAZOS DE CUMPLIMIENTO 10

1 DESCRIPCIÓN DEL PROYECTO GEOCODER.IA

El proyecto Callejero Digital de Andalucía Unificado – CDAU tiene entre otras muchas funciones servir como marco para la georreferenciación de información que cuenta entre sus atributos con una dirección postal u otra información a través de la que se puede ubicar un elemento en el territorio.

La Administración autonómica andaluza cuenta con grandes conjuntos de datos de carácter administrativo susceptibles de ser georreferenciados para múltiples fines, entre ellos, la realización de analítica avanzada para determinar y predecir patrones en los que el territorio tiene un peso determinante.

Y uno de los problemas a los que tradicionalmente nos enfrentamos cuando se pretende posicionar sobre el territorio cualquier dato es que la fuente de procedencia no suele recoger de forma ordenada y normalizada la información postal, lo que en muchas ocasiones dificulta e imposibilita su correcta ubicación sobre el territorio.

Para solventar dicho problema, el Instituto de Estadística y Cartografía de Andalucía – IECA, ha desarrollado un proceso cuyo objetivo es ofrecer un proceso orientado a la georreferenciación de conjuntos de datos basado en la aplicación de inteligencia artificial, en el marco del proyecto CDAU.

A continuación se describen los pasos que se han seguido en la implementación de dicho proceso para la búsqueda de una vía en CDAU a partir de un nombre de vía que haya sido escrito en cualquier sistema de información o registro sin seguir una normalización. Este proceso de búsqueda está desarrollado mediante un tratamiento inteligente de los datos para mejorar los resultados respecto a los métodos de búsqueda que tradicionalmente se han venido usando en CDAU.

Descripción del proceso implementado. Definición del conjunto de datos (datos de entrada)

En primer lugar se define el conjunto de datos que va a ser utilizado para obtener el grado de similitud mediante los nombres de las vías en el modelo de IA.

Se toma una muestra de aquellos registros para los que no se ha obtenido el identificador de la vía de CDAU (id vial), mediante los procesos de normalización previos que se realizan tradicionalmente (búsqueda en varianteros del IECA, búsquedas en CDAU, normalizador de la herramienta aLink).

Estos datos se cruzan con los de CDAU obteniendo para cada registro aquellos nombres de vía e id vial de CDAU en los que coincidan con el tipo de vía y el código postal.

Estos datos son exportados a un csv que serán los datos de entrada para el modelo de IA.

A continuación se muestra un subconjunto de los datos:

id	id_vial_cdau	nv_espand	nv_cdau
522248	748000325	CARTEROS SEVEROS OCHOA	DON JUAN TENORIO
522248	748000254	CARTEROS SEVEROS OCHOA	DOCTOR FLEMING
522248	748000382	CARTEROS SEVEROS OCHOA	GIBRALTAR
522248	748000259	CARTEROS SEVEROS OCHOA	FRANCISCO ORELLANA
522248	748000006	CARTEROS SEVEROS OCHOA	SAN FERNANDO
522248	748000110	CARTEROS SEVEROS OCHOA	TOCINA
522248	748000290	CARTEROS SEVEROS OCHOA	MURCIA
522248	748000396	CARTEROS SEVEROS OCHOA	SEVERO OCHOA
522248	748000256	CARTEROS SEVEROS OCHOA	EMILIO CASTELAR
522248	748000013	CARTEROS SEVEROS OCHOA	TEODORA OCAÑA
522248	748000331	CARTEROS SEVEROS OCHOA	ALCALA DE GUADAIRA
522248	748000015	CARTEROS SEVEROS OCHOA	VALDES LEAL
522248	748000025	CARTEROS SEVEROS OCHOA	GOLONDRINAS (LAS)
522248	748000045	CARTEROS SEVEROS OCHOA	BLAS INFANTE

Descripción del proceso implementado. Modelo IA

En el siguiente paso del proceso se ha implementado un script de Python que usando las librerías Scikit-Learn y NLTK obtiene el grado de similitud entre el nombre de la vía con sus correspondientes en CDAU.

La librería Scikit-Learn es una biblioteca de Python para aprendizaje automático. Mientras que la librería NLTK es una biblioteca para el procesamiento de lenguaje natural en Python que proporciona herramientas para el análisis de texto y lingüística computacional.

El grado de similitud resultante del proceso, que tomará valores comprendido entre 0 y 1, se genera siguiendo los siguientes pasos:

- Limpieza de texto: En primer lugar se han procesado los textos eliminando las stopwords y lematizando los términos.
- Aplicación del método Tf-idf (del inglés Term frequency – Inverse document frequency). Se aplica este método a los textos limpios y proporciona para cada término la frecuencia en la que aparece en el texto.
- Se obtiene el grado de similitud entre los textos aplicando el modelo inteligente a los elementos obtenidos en el método Tf-idf.
- Por último, mediante un modelo de inteligencia artificial explicable, se establecen criterios de clasificación de los elementos en base al resultado del grado de similitud.

Se hacen uso de procesos de inteligencia artificial para obtener el rango del grado de similitud apropiado para obtener unos resultados de calidad del proceso completo. Al utilizar modelos explicables se pueden interpretar los resultados obtenidos. De hecho, del modelo de clasificación utilizado se extrajeron las siguientes conclusiones:

- Los resultados cuyo grado de similitud estén entre 0 a 0.7 se consideran que no son resultados válidos.
- Si el proceso encuentra algún nombre de vía cuyo coeficiente de similitud supera el umbral establecido por el proceso inteligente, en este caso igual o superior a 0.8, se considerará el resultado como válido.

En este mismo proceso se crea la tabla parametrizada TablaIA en base de datos, volcando solo aquellos pares en los que el grado de similitud es mayor a 0.8. Este grado de similitud se puede ajustar (manualmente o mediante modelos inteligentes) en el script en la línea donde se lanza el proceso.

En la tabla resultante se agregan además los campos con los nombres de las vías procesados y el número de palabras que componen dichos campos.

Por último se selecciona el par nombre vía e id vial de CDAU según estos criterios:

- aquellos casos en los que el grado de similitud es mayor o igual a 0.8.
- aquellos con grado de similitud mayor que 0.6 y el nombre de las calles (con el tratamiento) tienen una diferencia de número de palabras menor que dos.

De esta forma, se complementan los resultados de un modelo de inteligencia artificial con la experiencia humana enriqueciendo así los resultados obtenidos.

Caso de uso

Uno de los casos de uso en los que se ha aplicado este proceso es en la búsqueda de vías escritas por los ciudadanos en alguno de los formularios disponibles en los sistemas de información de la Junta de Andalucía. Ejemplos de uso son los que se derivan de los procesos que diariamente realiza la Consejería de Salud para normalizar y estandarizar todas aquellas direcciones de los andaluces que a lo largo del día facilitan datos de carácter postal, ya sea para solicitar médico, pedir la tarjeta sanitaria, etc.

Y a través de los servicios que provee CDAU a través de la plataforma de interoperabilidad de la Junta de

Andalucía, NEXO, se está procediendo a la modernización de los sistemas de información corporativos en lo que tiene que ver con la recogida de la información postal, garantizando con la integración de dichos servicios a través de los formularios de entrada, la recogida de información postal normalizada, estandarizada y georreferenciada conforme al dato que CDAU provee.

Para ello, en dichos formularios de entrada se proporcionan campos autocompletables que hacen uso de los servicios REST de CDAU para cargar la información. Pero en muchos casos, es posible que el ciudadano opte por la opción de introducir la dirección postal de forma manual.

Y es en esos casos en los que se provee información postal directamente introducida por el ciudadano, cuando se hace uso del método geodocoder actualizado con tratamiento de IA, geocoder.IA. Una vez registrada la dirección postal en el sistema, se llama al método geocoder.IA que devuelve un listado de las vías de CDAU cuyo grado de similitud con la dirección postal de origen sea mayor de 0.7 aplicando la algoritmia descrita en el punto anterior, de forma que se podrá optar por la selección de la vía que se considere más adecuada. Y en el caso de que geocoder.IA no devuelva resultados, se manda la dirección postal a CDAU, para que el técnico de ayuntamiento decida si se incluye o no, minimizando los riesgos de rechazo de inclusión en el sistema así como la sobrecarga de trabajo a los verificadores de CDAU.

2 REPERCUSIÓN PARA EL CIUDADANO Y ADMINISTRACIONES

Este proceso tiene como finalidad que las direcciones que entran en cualquiera de los sistemas que pivotan sobre CDAU tengan una mayor probabilidad de ser encontradas y por tanto, puedan normalizarse y estandarizarse conforme a lo establecido en la fuente oficial, que es la que está consolidada en CDAU. Y eso no solo permite su correcta ubicación sobre el territorio sino que limita el número de falsas direcciones que entran en el flujo de mantenimiento del Callejero sobrecargando el trabajo de los verificadores del sistema y de los técnicos de los Ayuntamientos, que son quien en definitiva tienen que rechazarla o consolidarla en el sistema según proceda.

Desde el punto de vista del ciudadano, se facilita que cuando se acceda a un formulario en el que se estén consumiendo los servicios que provee CDAU se aumenten la posibilidades de enlazar la dirección que introduce con una ya existente en el sistema, lo que garantiza que esa información postal que se almacena en el sistema de información sobre el que está interactuando, se recoja conforme a la denominación oficial y georreferenciada. Y este hecho supone un salto cualitativo a los gestores de dicho sistema, que podrán explotar sus datos desde una óptica espacial casi de forma inmediata, combinando los datos introducidos con otras variables diferentes, cuestión que permite la construcción de modelos matemáticos orientados a la analítica de datos así como otras tareas entre las que destaca la generación de nuevos productos estadísticos a desarrollar en el marco de las estadísticas experimentales, recogidas en el Plan Estadístico y Cartográfico de Andalucía 2023-2029, la evaluación de políticas públicas, etc.

Y ya hay casos de uso, como el descrito en el apartado anterior pero mucho más consolidados, que se benefician del desarrollo de este proceso pues se maximizan los resultados relacionados con la georeferenciación de los conjuntos de datos, previa y necesaria para su desarrollo y puesta en producción. Es el caso de la Malla de población, donde se toma como referencia una malla regular que permite relacionar el territorio con la población que lo habita a través de la georreferenciación de los individuos y determinadas variables procedentes de distintos registros administrativos y a través de la que se ofrece información sociodemográfica que permite conocer la población residente según grandes grupos de edad, nacionalidad, lugar de nacimiento en relación al lugar de residencia, tiempo de residencia, estado de afiliación, percepción de pensiones contributivas de la Seguridad Social, ingresos medianos de cada tipología de pensión y demandantes de empleo del Servicio Andaluz de Empleo. O el proyecto Espacios Productivos de Andalucía (ESPAND), que ofrece información tanto de espacios productivos y sus infraestructuras, como de las empresas instaladas en ellos en todo el territorio de Andalucía a partir de la integración de diferentes Infraestructuras de datos, como son el Directorio de Empresas y Establecimientos con Actividad Económica en Andalucía, el Callejero Digital de Andalucía Unificado, la base de datos cartográfica y alfanumérica de la Dirección General de Catastro, la capa de espacios productivos de los Datos Espaciales de Referencia de Andalucía, así como información proporcionada por las empresas de suministros como son la infraestructura eléctrica (líneas y subestaciones), gasoductos y cobertura de redes ultrarrápidas. Para este tipo de proyectos es fundamental contar con la correcta geolocalización de las personas y las empresas en el territorio. En cuanto a las personas, una vez conocida su ubicación y dado que tenemos información procedente de diversas fuentes administrativas asociadas a ellas, podemos generalizar los datos para cada una de las celdillas de 250 por 250 metros que conforman el territorio andaluz, lo que permite disponer de una foto muy precisa con mucha información socio económica y a una escala de resolución jamás lograda. Y en cuanto a establecimientos con actividad económica y empresas, su ubicación precisa sobre cada uno de los Espacios Productivos permite caracterizar cada uno de dichos espacios en función de las actividades económicas predominantes, buscar sinergias entre empresas del sector además de enclaves para la ubicación de nuevas empresas con la consiguiente generación de valor en el territorio andaluz.

3 EQUIPO DE DESARROLLO Y PROVEEDORES

Por parte del IECA, se ha contado con un equipo multidisplinar compuesto por técnicos de los siguientes Servicios:

- Servicio de Gestión de la Información: técnicos que prestan sus servicios en el proyecto Callejero Digital de Andalucía Unificado – CDAU.
- Servicio de Planificación: técnicos que prestan sus servicios en el proyecto Alink, herramienta orientada a la normalización, estandarización y geocodificación de direcciones postales.
- Servicio de Estadísticas Económicas: técnicos que prestan sus servicios en el proyecto Directorio de Establecimientos con actividad económica.
- Servicio de Producción Cartográfica: técnicos que prestan sus servicios en el proyecto SIPOB y DERA.
- Servicio de Infraestructura Geográfica: técnicos que se encargan de la publicación de los servicios y mantenimiento del visor.
- Gabinete de Analítica de la Agencia Digital de Andalucía – ADA: técnicos que dan soporte a trabajos de desarrollo realizados con empresas externas.

Y en cuanto a proveedores, se ha contado con la contratación de una empresa externa, Guadaltel, que ha sido la adjudicataria para llevar a cabo el desarrollo del proceso presentado en el marco del proyecto CDAU.

4 VALORACIÓN ECONÓMICA

47.726,08, IVA incluido

5 PLAZOS DE CUMPLIMIENTO

El proyecto se desarrolló entre los meses de enero de 2022 y junio de 2023.