València, 22 de octubre de 2019

La Inteligencia Artificial desvela secretos de la historia de España

- **Investigadores de la Universitat Politècnica de València (UPV) y el Centro de Arqueología Subacuática del Instituto Andaluz del Patrimonio Histórico (IAPH) desarrollan en el proyecto Carabela una herramienta capaz de localizar con gran efectividad documentos manuscritos en fondos de Archivos Históricos.**
- **Entre otros hallazgos, el proyecto ha permitido sacar a luz información de principios del siglo XVIII sobre Australia.**
- **El proyecto ha recibido el apoyo del programa de Ayudas a Equipos de Investigación Científica de la Fundación BBVA, en el área de Humanidades Digitales. Los directores del proyecto presentarán mañana jueves sus resultados en una jornada organizada por el IAPH.**

Inteligencia Artificial y Aprendizaje Automático (Machine Learning) al servicio de los historiadores; algoritmos que localizan documentos de gran interés para la historia de España. Detrás de ello está Carabela, un proyecto desarrollado los últimos dos años por investigadores de la Universitat Politècnica de València (UPV) y el Centro de Arqueología Subacuática del Instituto Andaluz del Patrimonio Histórico. En él, han desarrollado y aplicado nuevas técnicas de IA/ML que permiten el acceso a los contenidos de más de 130.000 imágenes del Archivo General de Indias y el Archivo Histórico Provincial de Cádiz.

“Con estas técnicas podemos rastrear cualquier documento gráfico con la misma rapidez que un buscador web, identificando palabras concretas, combinaciones de palabras, frases, etc.... Todo ello gracias a modelos estadísticos que hemos entrenado a partir de ejemplos y que ahora son los grandes aliados para el estudio de estos fondos de la historia de España. Y los mismos métodos pueden aplicarse también a otros muchos documentos históricos”, destaca Enrique Vidal, investigador del centro Pattern Recognition and Human Language Technologies (PHRLT) de la Universitat Politècnica de València.

El proyecto ha recibido el apoyo del programa de Ayudas a Equipos de Investigación Científica de la Fundación BBVA, en el área de Humanidades Digitales.

Archivo General de Indias

Los fondos del Archivo General de Indias son de un interés excepcional para el estudio de la historia de España en América –desde el sur de Estados Unidos hasta Tierra de Fuego- y Filipinas durante los siglos XV al XIX. Se trata de manuscritos relacionados con viajes y comercio naval español, cuyo análisis no se puede hacer con las técnicas tradicionales de transcripción OCR -ya que están pensadas para texto impreso- ni

Área de Comunicación
Edificio Nexus (6G) - Camino de Vera, s/n - 46022 Valencia • Tel. 96 387 78 42 • Fax 96 387 78 49 • prensa@upv.es

tampoco con técnicas específicas para materiales manuscritos, pues los resultados que ofrecen cuando se aplican a estos textos históricos son demasiado imprecisos.

“Carabela ha permitido ir más allá, con técnicas de aprendizaje automático que permiten indexar imágenes de texto manuscrito en grandes colecciones de documentos históricos cuyo estado de conservación y enrevesados estilos de escritura hacen casi imposible la lectura de sus documentos por humanos,”, apunta Joan Andreu Sánchez, investigador también del PHRLT-UPV. Estas técnicas son capaces de identificar y discernir los distintos tipos de letras utilizados en cada una de las épocas en las que están datados los documentos e incluso analizar imágenes cuya calidad es muy baja.

La clave está en la capacidad de sus algoritmos para obtener modelos que se “aprenden” automáticamente a partir de ejemplos. “Dichos modelos necesitan una cantidad de datos de aprendizaje relativamente pequeña para obtener resultados muy satisfactorios. Estos métodos permiten responder satisfactoriamente a desafíos que los propios documentos plantean, como las diferencias de grafías, borrones, o calidad de la imagen.”, añade Enrique Vidal. En este caso, el aprendizaje se hizo con unas 500 páginas del Archivo de Indias, que fueron seleccionadas y transcritas por Carlos Alonso y su equipo de especialistas del Centro de Arqueología Subacuática del Instituto Andaluz del Patrimonio Histórico.

Pecios y Australia

Carabela ha sacado a la luz información de los manuscritos acerca de pecios que constituyen un patrimonio arqueológico de primera magnitud, debido a la gran riqueza histórica y cultural de su contenido. “Carabela contribuye así también a evitar el expolio del patrimonio sumergido”, explica Joan Andreu Sánchez.

Pero, sin duda, uno de los hallazgos más sorprendentes en estos fondos se produjo cuando, buscando términos relacionados con Australia -tales como “Tierra Austral Incógnita”- se encontró una carta de principios del siglo XVIII dirigida al rey Felipe V. “En esta misiva, escrita por el jesuita Andrés Serrano, hemos descubierto referencias muy precisas al continente austral datadas de 1705, mucho antes de que el capitán James Cook llegara hasta sus costas. Datos poco conocidos sobre la historia de Australia y que ahora descubrimos aplicando las técnicas de indexación y búsqueda probabilística desarrolladas en nuestro centro”, explica Enrique Vidal.

Presentación de resultados

Carlos Alonso y Enrique Vidal presentarán los resultados del proyecto mañana jueves en una jornada organizada por el Instituto Andaluz de Patrimonio Histórico. El evento tendrá lugar a partir de las 16.30 horas en la sede del IAPH.

READ, el siglo de Oro y Transkribus

En esta misma línea de trabajo, el equipo del PRHLT de la UPV ha participado en el proyecto europeo READ, que ha estudiado y analizado documentos del siglo de Oro de la literatura española, entre ellos manuscritos de

Área de Comunicación
Edificio Nexus (6G) - Camino de Vera, s/n - 46022 Valencia • Tel. 96 387 78 42 • Fax 96 387 78 49 • prensa@upv.es

Lope de Vega pertenecientes a la colección de la Biblioteca Nacional; correspondencia de los Hermanos Grimm, del Archivo Estatal de Marburgo. También del Archivo Nacional de Finlandia, del que se han indexado cerca de 150.000 páginas, y en futuros proyectos pretende llegar a indexar alrededor de 1 millón de páginas.

Además, en el marco del proyecto se ha desarrollado Transkribus, una plataforma software que permite anotar imágenes de documentos antiguos de gran valor historiográfico. Transkribus se utiliza fundamentalmente como herramienta de generación de datos de entrenamiento, ya que las técnicas de reconocimiento de texto manuscrito necesitan datos con los que aprender de manera automática. En un futuro próximo incorporará otras funcionalidades, como entrenamiento automático de modelos para otras lenguas.

READ ha concluido también con la creación de una cooperativa europea de la que la UPV es socia fundadora y que pone a disposición de todos los usuarios registrados el software Transkribus. Actualmente, la plataforma Transkribus cuenta con más de 30.000 usuarios de todo el mundo, lo que la convierte en una herramienta de referencia internacional para todos los historiadores.

Área de Comunicación
Edificio Nexus (6G) - Camino de Vera, s/n - 46022 Valencia • Tel. 96 387 78 42 • Fax 96 387 78 49 • prensa@upv.es