Módulo de tecnologías cognitivas

El desarrollo del Módulo de Tecnologías Cognitivas es un componente transversal que debe utilizar diferentes tecnologías y que en gran parte es usado en el módulo de automatización tanto para la toma de decisiones como para la resolución de distintas partes de la automatización. Este módulo también podría ser invocado a través de API REST para que cualquiera puede utilizarlo (solamente haría falta su despliegue y ejecución como API REST).

A continuación, se define los siguientes procesos desarrollados y donde encajarían dentro de los puntos mencionados por el pliego en el módulo de tecnologías cognitivas.

Módulo 1 : Reconocimiento de caracteres (OCR)

Esta herramienta podrá obtener texto de cualquier tipo de imagen, aparte de utilizar ficheros PDFs como entrada, será capaz de reconocer los caracteres que se encuentren en una imagen cualquiera, así como de documentos escaneados, impresos o escritos a mano.

Tecnologías Cognitivas utilizadas:

- OCR.

Repositorio: https://github.com/hercules-rpa/module_cognitive_lib/blob/main/module_cognitive_treelogic/PDF2Table.py

Módulo 2 : Extracción de tablas en archivos PDF

Librería capaz de extraer tablas de archivos PDFs, siendo capaz de detectar cambios de formatos en dicha tabla en lo que respecta a la orientación (texto en vertical), es decir, la librería es capaz de extraer cualquier tabla en cualquier orientación, esto incluye detección de cambios de formato en los documentos de concesiones, ya que si las tablas cambiaran de orientación está funcionalidad seguiría detectándolas.

Esta librería servirá de apoyo para la extracción de tablas Proceso 3 : Automatización de procesos de gestiones documentales parametrizables, en la extracción de concesiones, ya que su extracción significará la búsqueda de información y automatización de procesos. También se apoya en el módulo de Reconocimiento de caracteres (OCR), para aquellas tablas que se extraigan de imágenes y se necesite reconocer el texto de las celdas.

Tecnologías Cognitivas utilizadas:

- OCR.
- Detección de cambios formato.
- Analítica de Textos.

Repositorio: https://github.com/hercules-rpa/module_cognitive_lib/blob/main/module_cognitive_treelogic/PDF2Table.py

Módulo 3 : Minería de datos

Este módulo tiene como objetivo hacer minería de datos explotando los datos de los distintos subsistemas Hércules, en concreto sobre los datos de grupos de investigación y su producción con objeto de realizar clasificación y categorización que permitan identificar agrupaciones y similitudes, para ello se hace uso de tecnologías de procesamiento de lenguaje natural (NLP) y de aprendizaje (Machine Learning).

El funcionamiento de esta librería se divide en:

- Extracción de datos de los subsistemas HÉRCULES.
- Tratamiento de datos.
- Vectorización y creación de vocabulario (NLP) utilizando las etiquetas relacionadas con los trabajos de los investigadores.
- Reducción de la dimensión usando UMAP (Machine Learning).
- Agrupación utilizando técnicas de clustering (Machine Learning).

Módulo de Tecnologías Cognitivas:

- Procesamiento del lenguaje natural (NLP).
- Categorización y clasificación.
- Machine Learning.

Repositorio: https://github.com/hercules-rpa/module_cognitive_lib/blob/main/module_cognitive_treelogic/DataMining.py

Módulo 4 : Extracción de información utilizando recorrido de documentos XML

El proceso de recorrido de documentos XML nos permite obtener una abstracción de la estructura XML obteniendo solo los nodos necesarios para los procesos RPA.

Un ejemplo de la aplicación que puede tener este módulo es el recorrido del documento XML del BOE, utilizado enProceso 3 : Automatización de procesos de gestiones documentales parametrizablespara la obtención de las bases reguladoras de convocatorias.

Para el ejemplo mencionado anteriormente, el proceso recorrerá la estructura XML, dada una lista de nodos padre y de nodos hijos, en este caso, el nodo padre será el nodo con la etiqueta "item" y los nodos hijos de dicho nodo padre serán los nodos "titulo" y "urlPdf", y creará una lista donde se relaciona cada nodo "item" con sus nodos hijos. Gracias a esto el proceso de extracción de bases reguladoras solo tratará los nodos que necesita y obtendrá la información de ellos.

Tecnologías Cognitivas utilizadas:

- Categorización y clasificación.
- Analítica de Textos.

Repositorio:

https://github.com/hercules-rpa/module_cognitive_lib/blob/main/module_cognitive_treelogic/ExtractXML.py

Módulo 5 : Web-scraping

El módulo cognitivo estará dotado de una librería que será capaz de recoger, agrupar y categorizar las distintas convocatorias distribuidas en las páginas que se usarán como bases de datos. Estas páginas no tienen API y por lo tanto, la única forma de atacarlas es usando web-scraping y recuperar los resultados como si de un humano se tratase.

Con esta librería se podrá visualizar, de una manera general, todas las convocatorias de las distintas fuentes de una manera rápida y sencilla otorgando una mayor accesibilidad a ellas y pudiendo aplicar un filtro dentro de las mismas.

Esta funcionalidad se divide en:

- https://confluence.um.es/confluence/pages/viewpage.action?pageId=397534869
- https://confluence.um.es/confluence/pages/viewpage.action?pageId=397534871

Tecnologías Cognitivas utilizadas:

- Diseño de un sistema de web-scraping para identificar anuncios de convocatorias de interés para los perfiles de investigadores desde diferentes fuentes suministradas al sistema.

Repositorio: https://github.com/hercules-rpa/module_cognitive_lib/blob/main/module_cognitive_treelogic/WebScrapping.py