Universidad Técnica Particular de Loja
Ingeniería en Sistemas Informáticos y Computación

Sistemas Basados en Conocimientos

Proyecto Bimestral - Modelo semi-formal del vocabulario/ontología

- Fuertes Alejandro
- Román Andrés

QA over text

Para trabajar un sistema QA a partir de las páginas de Wikipedia obtenidas debemos extraer la información en texto plano que hay en cada link obtenido. Para ello se utilizarán dichos links en la API de wikipedia y su función extracts que permite obtener texto plano desde las páginas dadas, esto se convertirá en nuestro dataset para poder trabajar con Haystack. Este framework open source nos permitirá diseñar nuestro sistema QA a partir de la base de conocimiento extraída desde Wikipedia.

Haystack busca sus respuestas mediante el almacenamiento de documentos y para este apartado utiliza Elasticsearch ya que tiene características precargadas que ayudan a la construcción del sistema QA. Para el procesado de los documentos, Haystack permite convertir archivos en texto, limpiar el texto y separarlo para después escribirlos en el almacenado de documentos de Elasticsearch. Es aquí donde ingresará nuestro dataset sobre la información que existe en los links de las Wikipages obtenidas en la recolección de datos realizada con anterioridad.

El siguiente paso realizado es inicializar los módulos de recepción, lectura y búsqueda de Haystack. El primero ayudará a delimitar un alcance para que el segundo módulo de lectura reciba un texto más pequeño con el fin de responder a la pregunta. En este paso se utilizará nuevamente un algoritmo de Elasticsearch. Posterior a esto, el módulo de lectura escanea el texto recibido y extrae un número posible de respuestas, esto se hace mediante el uso de modelos de deep learning. El último paso es la ejecución del pipeline de Haystack los cuales se pueden configurar según el uso que se vaya a hacer.

Para la codificación se utilizará como base el siguiente código de Haystack compartido en Google Colab, cambiando el dataset con el que fue obtenido en la recolección de Wikipages y su posterior paso a texto plano mediante la API de Wikipedia.

Modelo semiformal

El siguiente modelo es la representación del conjunto de datos obtenidos durante la recolección, donde la tabla utilizada para realizar el sistema QA será “owl:Class” en sus diferentes instanciaciones según la clase obtenida a partir de el concepto raíz. El atributo utilizado como link de la página de Wikipedia es “foaf:isPrimaryTopicOf”, el cual nos redirige a la url de Wikipedia para dicha instancia.