Universidad Técnica Particular de Loja
Ingeniería en Sistemas Informáticos y Computación

Sistemas Basados en Conocimientos

Proyecto Bimestral

- Fuertes Alejandro
- Román Andrés

1. **Definir el dominio de trabajo sobre el cual realizar su proyecto**

   Para desarrollar este trabajo bimestral, se trabajará sobre Wikipedia, fuente de información sobre la cual se realizarán las consultas; haciendo uso de la MediaWiki API extraeremos los documentos para su post-procesamiento en el sistema.

   Al hacer uso de la Wikipedia, se establece que el dominio de datos sobre el que trabajaremos puede ser de cualquier tipo, siempre y cuando la información se encuentre en Wikipedia, es decir, podremos hacer consultas sobre:

   - Corpus de conocimiento de diferentes temáticas, cuya información se encuentre en Wikipedia.

2. **Seleccionar fuentes de datos del dominio seleccionado**

   La fuente de información será Wikipedia, la cual es una fuente de datos semi-estructurada, a la cual accederemos mediante la API que se menciona anteriormente, para la cual haremos uso de un token para poder acceder a la API. Dicha API es de libre acceso y se puede realizar consultas desde cualquier lenguaje de programación, no existe un límite de consultas y se puede emplear un sistema “from scratch” para consumir la API. Está fuente de datos permitirá hacer búsquedas de cualquier tema que requiera el usuario, teniendo como fuente a esta gran enciclopedia en línea que recopila datos mediante redacciones de voluntarios, esto tal vez pueda generar desconfianza pero la veracidad de sus publicaciones está supervisada para garantizar su confiabilidad y el uso de voluntarios hace qué la actualidad de los datos sea casi inmediata. En la actualidad, Wikipedia cuenta con más de 178 millones de páginas con información varia, y más concretamente, en inglés existen más de 41 millones.

   Para este trabajo se usará la api en inglés, ya que cuenta con mayor cantidad de información y contenido que cualquier otro idioma, para esto se usará la dirección de API: https://en.wikipedia.org/w/api.php

3. **Seleccionar el tipo de aplicación que se construirá para explotar el grafo de conocimiento**

   La aplicación que construiremos será una aplicación de pregunta respuesta (NLP).

La aplicación será direccionada a Preguntas y Respuestas, que funciona bajo el siguiente esquema:

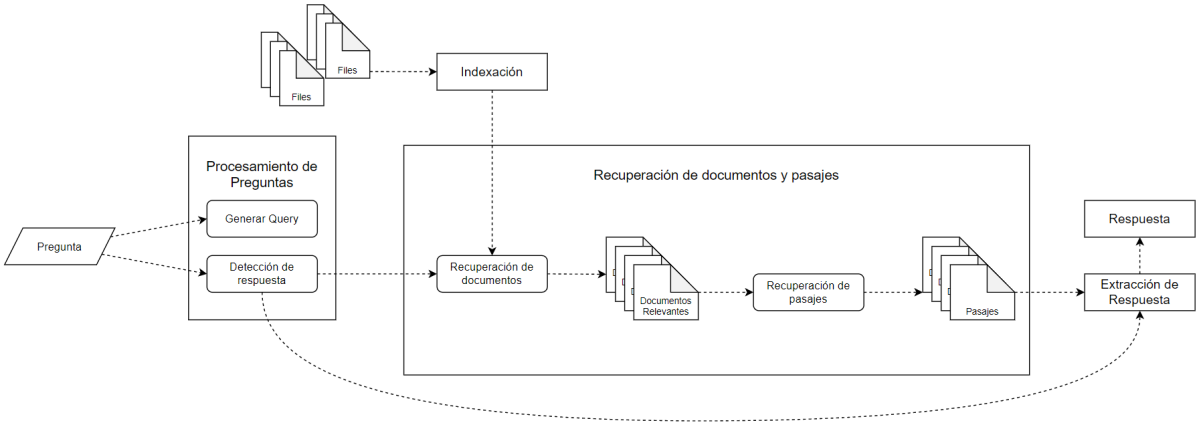

Fig. 1. Flujograma del sistema.

El sistema que se plantea diseñar se basa en 4 procesos fundamentales que se pueden apreciar en el flujograma.

1. Procesamiento de Query
2. Recuperación de documentos
3. Recuperación de pasajes
4. Extracción de respuesta

Procesamiento de Query

Este proceso transforma una pregunta en un Query, se eliminan palabras que no signifiquen nada en la oración, además, con el uso de Deep Learning es posible crear un vector que exprese el significado de la oración.

Esta vez, se generará una consulta de búsqueda usando spaCy. Primero, la pregunta se analiza y se etiqueta con etiquetas de parte del discurso. A continuación, eliminamos las palabras en función de sus etiquetas de parte del discurso. Específicamente, se eliminan las palabras que no sean nombres propios (PROPN), números (NUM), verbos (VERB), sustantivos (NOUN) y adjetivos (ADJ).

Recuperación de documentos

En este proceso recuperamos documentos relevantes usando la consulta generada en el proceso anterior. En el procesamiento siguiente, dado que la respuesta se extraerá de estos documentos, se requiere buscar los documentos que puedan contener la respuesta en la mayor medida posible.

Como hemos mencionado anteriormente, buscaremos Wikipedia. Wikipedia proporciona una API llamada MediaWiki API . Podemos utilizar la API para buscar los documentos relacionados con la consulta. El proceso aquí consta de dos pasos.

Primero, enviamos una consulta para obtener una lista de páginas relacionadas. Luego buscamos el contenido de las páginas individuales.

Recuperación de pasajes

En la recuperación de pasajes, los documentos se dividen en unidades más pequeñas (pasajes) como oraciones y párrafos, y se seleccionan pasajes que probablemente contengan una respuesta.

A continuación, seleccionaremos pasajes que son similares a la pregunta. Para calcular la similitud, usamos BM25 para crear vectores de preguntas y pasajes. Una vez que hemos creado los vectores, podemos usar el producto escalar y la similitud del coseno para calcular la similitud entre la pregunta y el pasaje.

Extracción de respuesta

La extracción de respuestas extrae respuestas de pasajes. Aquí, la pregunta y el pasaje se ingresan en el modelo de extracción de respuestas, y el modelo genera la respuesta acompañada de la puntuación. Luego, clasifica las respuestas según la puntuación y presenta al usuario las respuestas con las puntuaciones más altas como respuesta final.