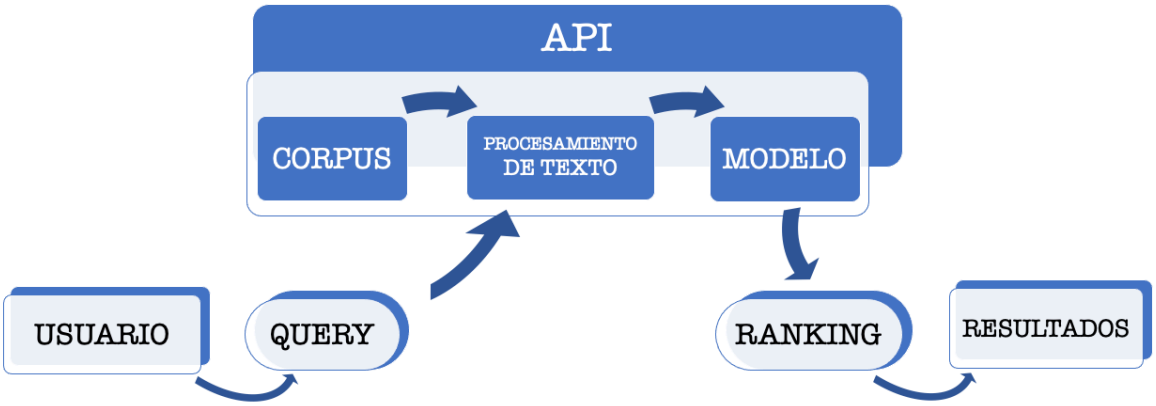
PROYECTO FINAL DE SRI

Preentrega Final

Ciencias de la Computación
Universidad de La Habana

El siguiente esquema presenta la arquitectura general que refleja el motor de búsqueda. El flujo del funcionamiento se desarrolla de la siguiente forma:



El usuario realiza una consulta en lenguaje natural, que luego es preprocesada. Entonces, teniendo ya el corpus de documentos también preprocesado de forma *offline*, se aplica un modelo de recuperación de información (en este caso el modelo vectorial) para devolver un conjunto de documentos.

1. Preprocesamiento del corpus de documentos: Consta de dos partes fundamentales:

   - ***Parser***: Dado un corpus de documentos permite obtener una lista que representa el conjunto de documentos. Dado que este proceso depende en gran medida del formato del corpus, hemos decido tener una clase por cada uno de los tipos de corpus que se trabajen (por el momento solo "*cranfield*").

   - ***Indexer***: Permite obtener los términos indexados que se tendrán en cuenta en el modelo de recuperación de información. En este caso aplicamos algunas técnicas de procesamiento del lenguaje natural, como:
     - Eliminación de *stopwords* o palabras de paso: Se refiere a la eliminación de palabras que puede que no tengan utilidad semántica dentro de un uso o contexto determinado. La elección de estas palabras depende del idioma (en este caso inglés).
     - Lematización: Consiste en determinar el lema de las palabras.
     - Etiquetado gramatical: Asignar a cada palabra su categoría gramatical.

   A partir del empleo de estas técnicas, solo consideramos como términos indexados a aquellos que no representen palabras de paso y cuya categoría gramatical sea sustantivo, verbo o adjetivo.

   Para aportar eficiencia al motor de búsqueda se calcula la frecuencia de los términos indexados en esta fase de preprocesamiento.

2. Procesamiento de la consulta: Se realiza de manera análoga al procesamiento de los documentos.

3- Aplicación del modelo vectorial: Aplicando las formulas correspondientes al modelo vectorial se calculan:
    - Los pesos de los términos en las consultas y en los documentos. En el caso de las consultas se usa 0.5 como amortiguado.
    - La similitud de cada uno de los documentos con la consulta.

4- Establecimiento de Ranking: Usando un *heap*, mantenemos ordenados los documentos en orden creciente atendiendo a su similitud respecto a la consulta. Solo se recuperan los documentos con similitud mayor o igual a 0.21.