Sistema de Información
2019-2020

Leandro González Montesino
C-511

Facultad de Matemática y Computación
La Habana, Cuba

1- Resumen

En el siguiente reporte se muestra una solución al problema de cargar un directorio de archivos de texto y después realizar consulta sobre ello. Se utilizó el modelo clásico **boolean**. Se aplicaron medidas de evaluación como *Precisión, Recobrado, F-Medida y R-Precisión.*

2- Datos Técnicos

La solución brindada fue elaborada en python3 con gran peso en las siguientes bibliotecas:

1. NLTK
2. pdfminer
3. PyQt5

3- Contenido

El proyecto esta dividido en varias etapas:

1. Procesamiento del Texto
2. Modelos de recuperación
3. Creación de Indices
4. Evaluación del Sistema

3.1- Procesamiento del Texto

Para el procesamiento del texto se utilizó **NLTK**, primero dividiendo en tokens el texto, eliminando stopwords, aplicando stemmin y lemmatization.

3.2- Modelo de Recuperación boolean

El modelo clásico utilizado fue el boolean.

3.3- Creación de Indices

Los indices se crearon bajo la estructuras de diccionarios y a su vez se brinda una opción de guardar los daros en un fichero binario para futuros análisis.

3.4- Evaluación del Sistema

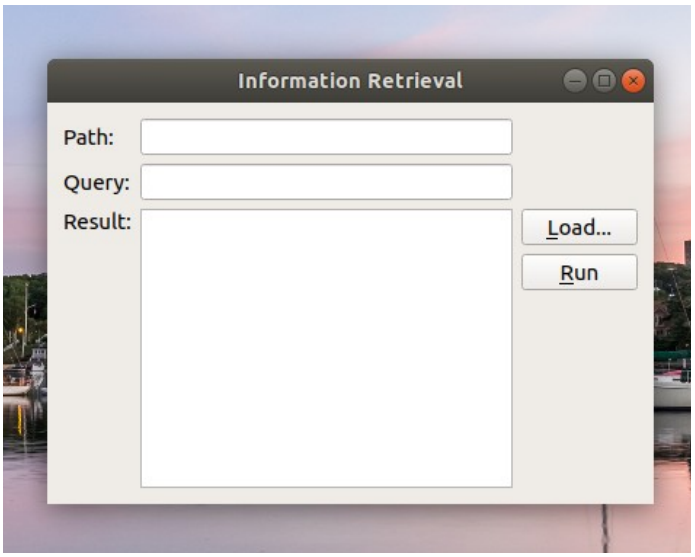
Se brinda la propuesta de 4 métodos de evaluación del sistema, Precision, Recall, F-Media y R-Precision.

4- Ejecución

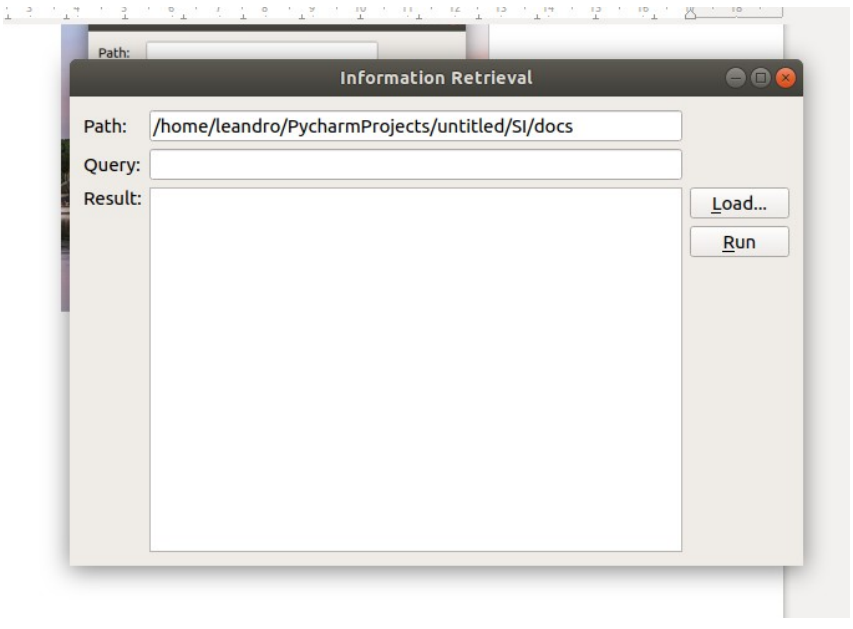
La ejecución del proyecto es muy sencilla, en cualquier interprete de paython3 con las dependencias requeridas corremos la linea siguiente:

>$ *python3 Visual.py*

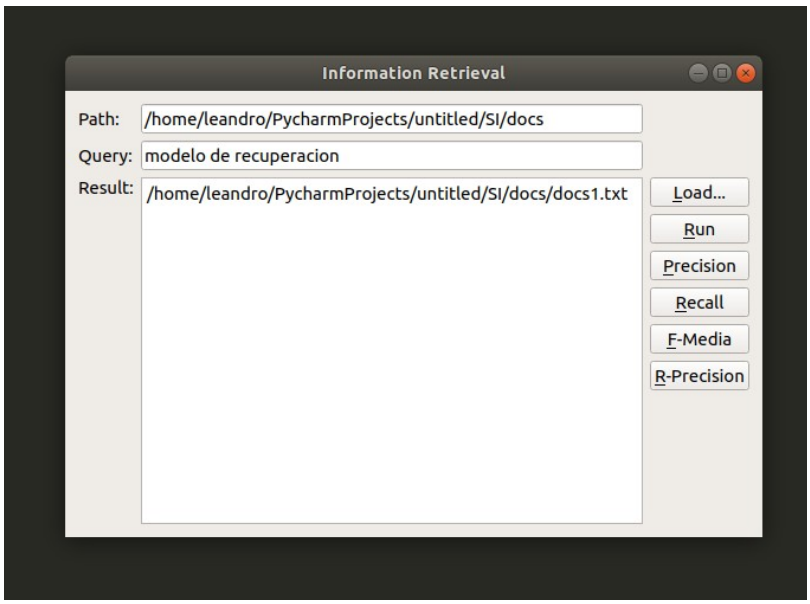
Instantáneamente ya podemos utilizar la aplicación mediante su interfaz, que se observa de la siguiente manera:



Mediante el botón **Load** añadimos el directorio donde están todos los documentos de texto que necesitamos procesar. Obteniendo la vista siguiente:



Se continua insertando una Query deseada y haciendo click en el botón **Run**



Como se muestra en la imagen anterior ya obtenemos los primeros resultados de la query requerida.

Por otra parte se muestran 4 nuevos botones que son para ejecutar las medidas de evaluación.
El caso particular al hacer click en **R-Precision** se despliega una nueva ventana donde debemos entrar el valor de *r.*