Sistemas de Recuperación de Información: Proyecto Final

Nadia González Fernández
José Alejandro Labourdette-Lartigue Soto
C-512

5to Año - Ciencias de la Computación - Curso 2022
Facultad de Matemática y Computación
Universidad de La Habana, La Habana, Cuba

Abstract. *El presente informe describe los sistemas de recuperación de información basado en el modelo clásico Vectorial y Booleano. Se explica el diseño del sistema, las ventajas y desventajas de los modelos escogidos y las herramientas utilizadas.*

Keywords: Model Vectorial, Modelo Booleano, spaCy, Python, Consultas, Motor de Búsqueda

1 Introducción

La recuperación de información es una disciplina que, con el incremento de la informatización y los volúmenes de datos en la red, cada vez se hace más necesaria la utilización de sistemas de recuperación de información.

Un sistema de información es un conjunto de componentes interrelacionados que permiten capturar, procesar almacenar y distribuir la información para apoyar la toma de decisiones y el control en una organización.

En el presente proyecto se realiza un sistema de recuperación de información con la utilización de dos modelos: el vectorial y el booleano. El sistema se especializa en la recuperación de documentos de texto dada su relevancia respecto a una consulta. Nos proponemos comparar los dos modelos con distintas colecciones de datos.

El proyecto se encuentra en github: `https://github.com/nala7/information_retrieval_models`

2 Diseño del Sistema

Etapas de la Recuperación de Información:

2.1 Procesamiento de la consulta hecha por un usuario

Las consultas son recibidas como string. Luego son tokenizadas y lematizadas con la biblioteca de Python spaCy.

2.2 Representación de los documentos y la consulta

Los documentos y las consultas son expresados en las clases `Document` y `Query` respectivamente. Un `Document`, como estructura de datos, contiene el nombre del mismo y las palabras que fueron devueltas por el algoritmo de procesamiento de texto. Una `Query` contiene los términos que fueron devueltos por el algoritmo de procesamiento de texto para eliminar stopwords y demás elementos del lenguaje sin carga semántica.

La clase `DocumentCollection` es la que representa la colección entera de documentos, en ella tiene un diccionario que permiten saber la frecuencia de ocurrencia de los términos en cada documento.

Como parte de la estrategia para ahorrar memoria se asignan a los nombres de los documentos y a los términos valores numéricos únicos. 4 diccionarios se usan para mapear esta representación numérica.

2.3 Funcionamiento del motor de búsqueda

Los motores de búsqueda es son las clases `VectorFramework` y `BooleanFramework` correspondientes a cada modelo. Estos que al ser instanciados se les específica la colección de documentos sobre la cual se desea trabajar. En el caso del modelos vectorial el propio constructor da la instrucción de computar los pesos de los documentos. El proceso de computar los pesos sigue los pasos establecidos por el motor de búsqueda para calcular pesos. Dichos valores son almacenados en la propia instancia de `DocumentCollection`.

Para realizar búsquedas existe la función find(), definida en el propio framework que recibe la query sobre la que se desea buscar. La propia función computa los pesos de la query. En el modelo vectorial también se calcula la similitud con cada documento y se devuelven los documentos ordenados.

2.4 Obtención de los resultados

Modelo Vectorial:

Para la obtención de resultados se define un límite de similitud mínima necesaria para considerar un documento relevante.Los nombres de los documentos con similitud mayor que ese mínimo son devueltos, ordenados de mayor a menor similitud.

Modelo Booleano:

Los resultados se obtienen comprobando la existencia de cada uno de los términos relevantes de la query en los documentos. Se consideran términos relevantes aquellos devueltos por el proceso de depuración de texto, que elimina los stopword y lleva las palabras a su lexema.

3 Herramientas Utilizadas

- `spaCy`

Para el procesamiento de los documentos y las consultas se utilizó `spaCy`. Esta es una biblioteca de software para el procesamiento de lenguajes naturales desarrollado por Matt Honnibal y programado en lenguaje Python. Es software libre con Licencia MIT su repositorio se encuentra disponible en Github. Esta es utilizada en la clase **read_content.py**. (`https://spacy.io/`)

```
nlp = spacy.load('en_core_web_sm')

nlp.max_length = 5030000  # or higher
doc = nlp(text)

# Tokenization and lemmatization
lemma_list = []
for token in doc:
    lemma_list.append(token.lemma_)

# Filter the stopword
filtered_sentence: list[Any] = []
for word in lemma_list:
    lexeme = nlp.vocab[word]
    if not lexeme.is_stop:
        filtered_sentence.append(word)

# Remove punctuation
punctuations = "?:!.,;"
for word in filtered_sentence:
    if word in punctuations:
        filtered_sentence.remove(word)
    if word == '\n':
        filtered_sentence.remove(word)
return filtered_sentence
```

- `ir_datasets`

  Se utilizó la biblioteca `ir_datasets` para la obtención de colecciones de datos. Esta es una API de Python que puede ser utilizada para acceder a coleciones de datos o para crearlas. (`https://ir-datasets.com/`)
- `pickle`

  El modulo `pickle` implementa protocolos binarios para serializar y deserializar una estructura de objetos Python. En este caso se utilizan para serializar la lectura de las colecciones de documentos ya que el procesamiento de estos es lento.

Para mayor eficiecia

4 Evaluación del Sistema

Las métricas empleadas para la evaluación del sistema fueron la **Precision**, el **Recobrado** y la **Medida F1**. Para la evaluación se utilizaron dos colecciones de pruebas: "Cranfield" y "Vaswani".

En las siguientes gráficas se analiza cuál es un mejor umbral a utilizar para el modelo Vectorial. Para ello se halló la media de la precision, el recobrado y la medida f1 de todas las consultas con diferentes umbrales.

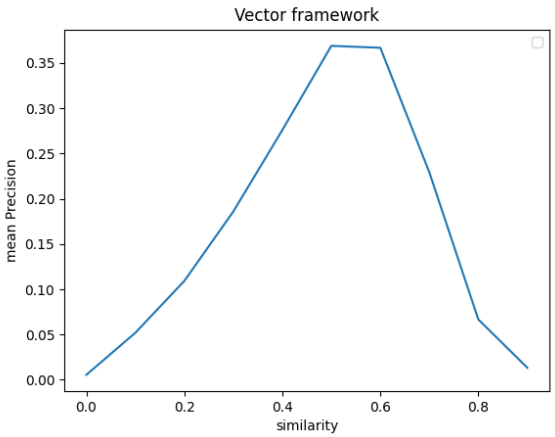

Fig. 1. Análisis del umbral con la colección "Cranfield"

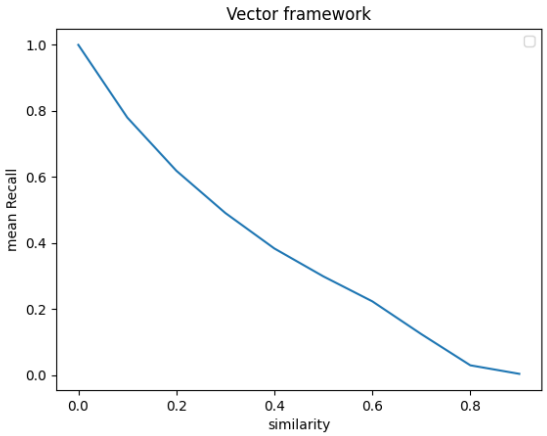

Fig. 2. Análisis del umbral con la colección "Cranfield"

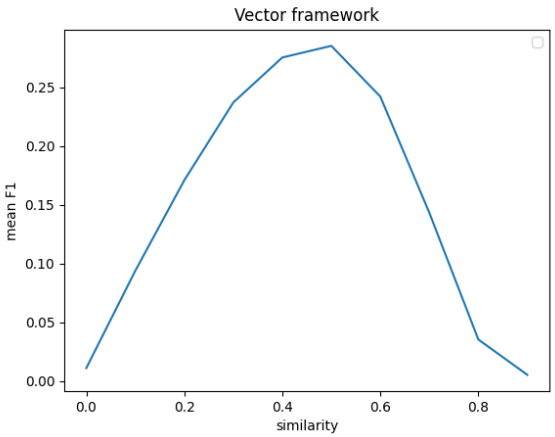

Fig. 3. Análisis del umbral con la colección "Cranfield"

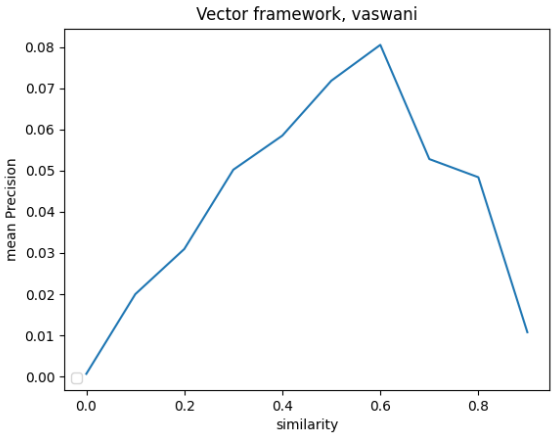

Fig. 4. Análisis del umbral con la colección "Vaswani"

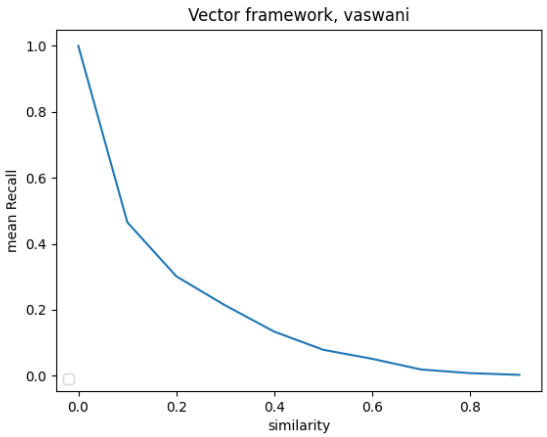

Fig. 5. Análisis del umbral con la colección "Vaswani"

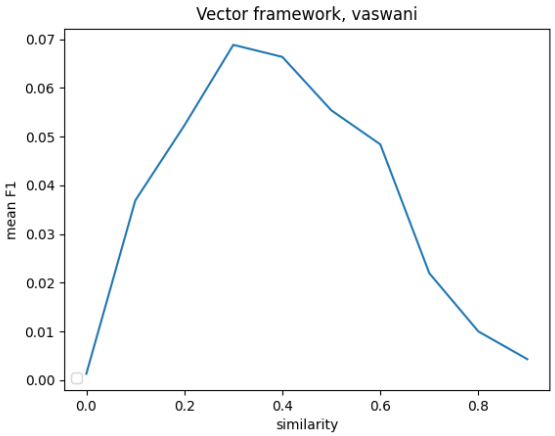

Fig. 6. Análisis del umbral con la colección "Vaswani"

En las siguientes gráficas se compara el rendimiento de los modelos con la colección "Crandfield":

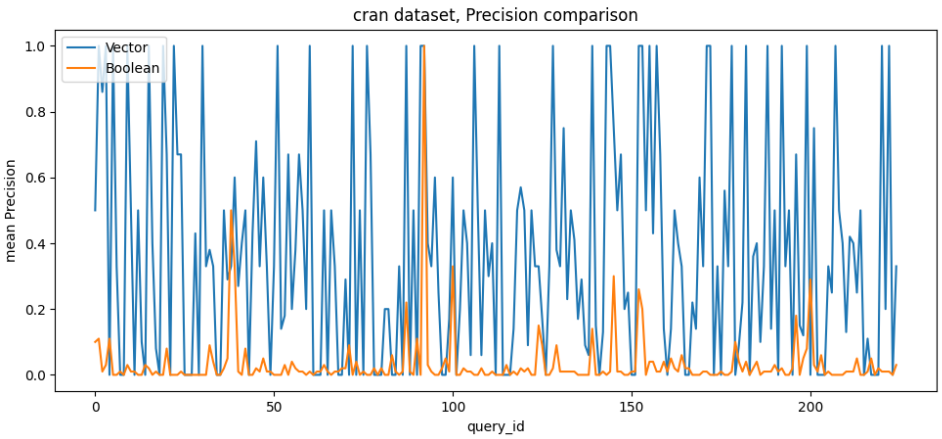

Fig. 7. Comparación de la precisión los modelos con la colección "Cranfield"

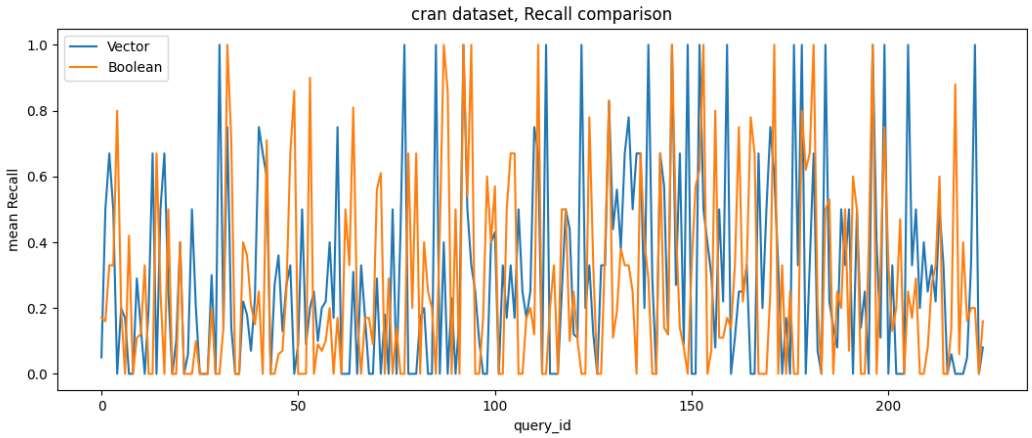

Fig. 8. Comparación del recobrado los modelos con la colección "Cranfield"

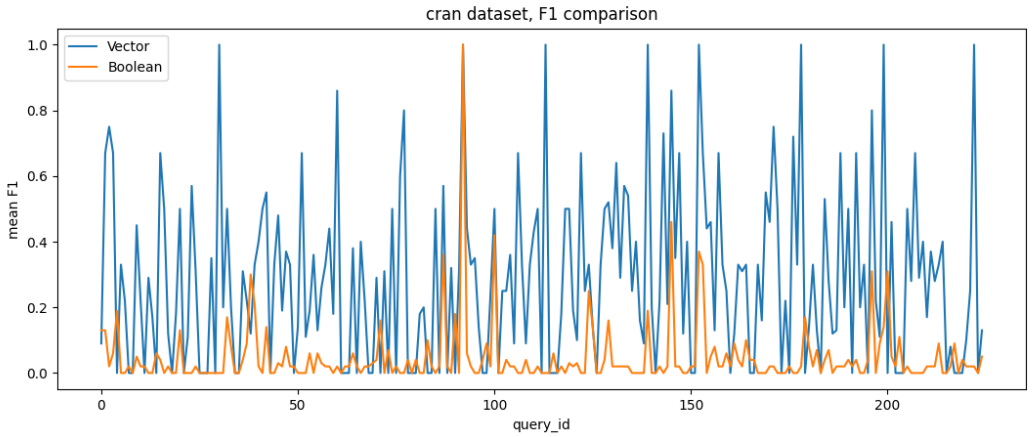

Fig. 9. Comparación de la medida F1 los modelos con la colección "Cranfield"

En las siguientes gráficas se compara el rendimiento de los modelos con la colección "Vaswani":

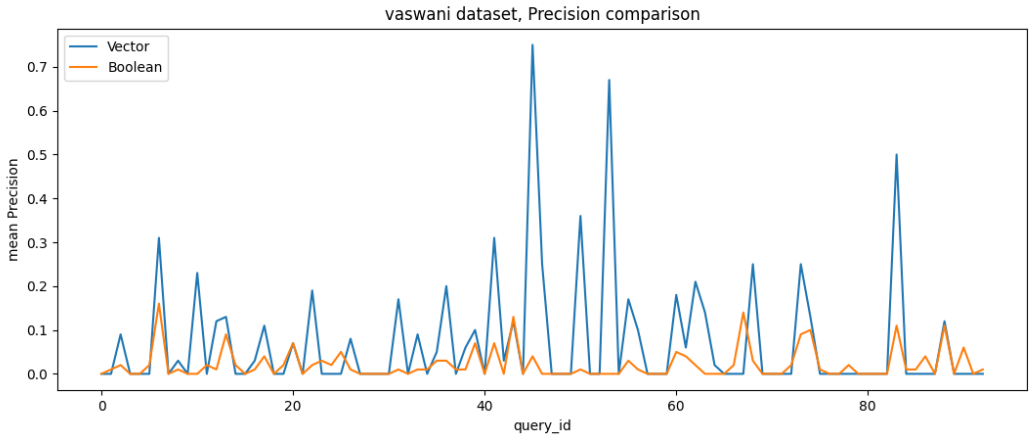

Fig. 10. Comparación de la precisión los modelos con la colección "Vaswani"

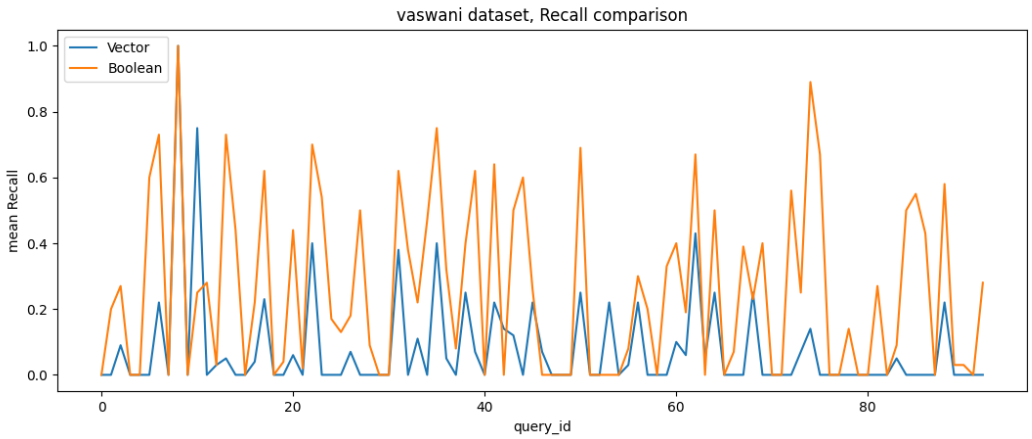

Fig. 11. Comparación del recobrado los modelos con la colección "Vaswani"

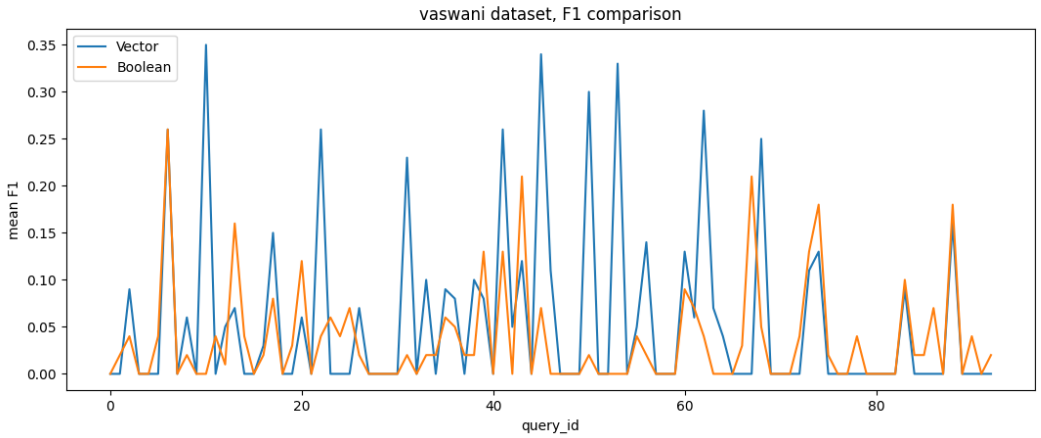

Fig. 12. Comparación de la medida F1 los modelos con la colección "Vaswani"

5 Ventajas y Desventajas

5.1 Modelo vectorial

En el modelo Vectorial el esquema de ponderación $tf - idf$ de los documentos resulta en un buen rendimiento de la recuperación. La estrategia de coincidencia parcial permite la recuperación de documentos que se aproximen a los requerimientos de la consulta. Además la fórmula del coseno ordena los documentos de acuerdo al grado de similitud con la consulta. Por otra parte, el modelo vectorial tiene como desventaja que asume que los términos indexados son mutuamente independientes, sin embargo, esto hace que su rendimiento sea mejor.

Con respecto al MRI Booleano, el Vectorial tiene como ventaja que permite hacer un ranking de los documentos y da una correspondencia parcial entre documentos y consultas. En comparación con el modelo Probabilístico se ha demostrado que el MRI Vectorial tiene un mejor desempeño.

El modelo Vectorial es simple, rápido y, en algunos casos, brinda mejores resultados en la recuperación de información que el resto de los MRI clásicos.

5.2 Modelo Booleano

El modelo Booleano es simple y fácil de comprender e implementar. También realiza consultas con precisión semántica. Por otro lado, no tiene noción de ranking y solo recupera documentos donde la coincidencia es exacta. Tampoco tiene da diferencia la importancia entre términos y recupera muchos o pocos documentos.

6 Recomendaciones

Recomendaciones para trabajos futuros que mejoren la propuesta.

References

1. Prof. Carlos Fleitasa Aparicio, Profe. Marcel E. Sánchez Aguilar, Departamento de Programación, Facultad MATCOM, Universidad de La Habana (2021)
2. Text normalization with spacy and nltk, `https://towardsdatascience.com/text-normalization-with-spacy-and-nltk-1302ff430119`
3. Documentación oficial de Spacy `https://spacy.io/`
4. ir datasets: Python API `https://ir-datasets.com/python.html/`