Sistemas de Recuperación de Información

Proyecto Final: Implementación de un motor de búsqueda

Autores:

Laura Victoria Riera Pérez

Leandro Rodríguez Llosa

Marcos Manuel Tirador del Riego

Grupo: C-311

Facultad de Matemática y Computación

Universidad de La Habana, Cuba

Diciembre, 2022

Índice general

Diseño del sistema

Modelos implementados

Modelo booleano

Descripción

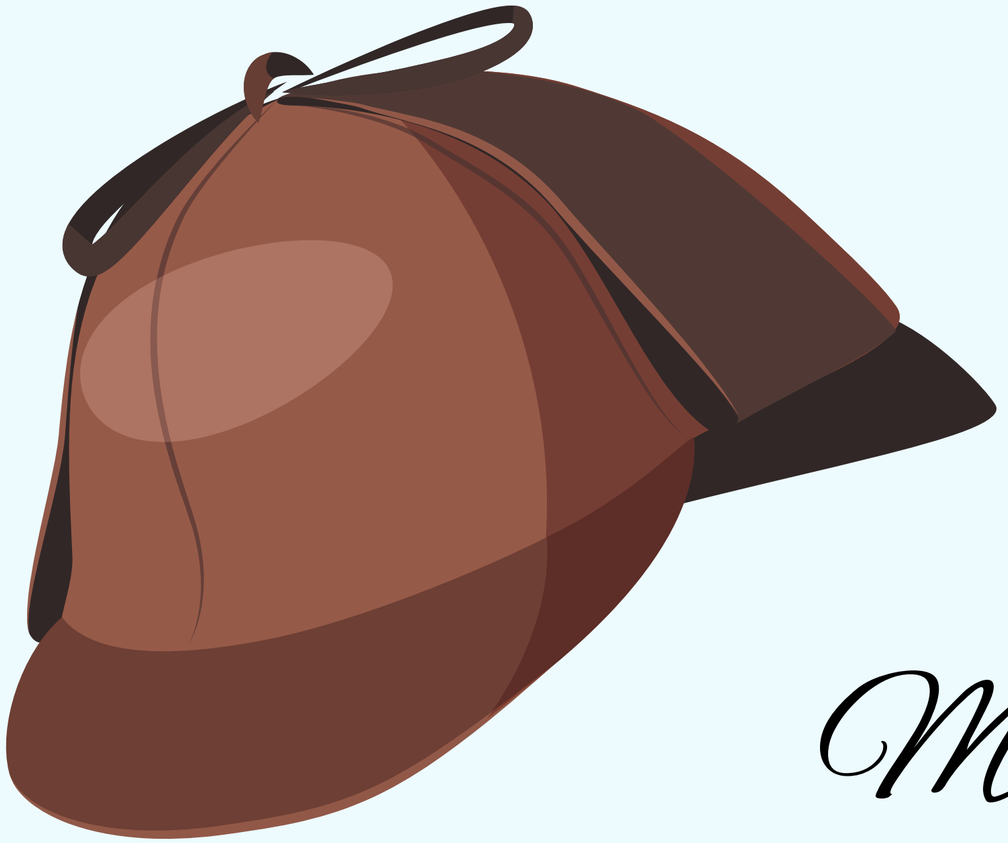
- El modelo booleano clásico expuesto es un modelo simple de Recuperación de Información basado en teoría de conjuntos y álgebra booleana.
- Las consultas son expresiones booleanas que utilizan los operadores lógicos AND, OR y NOT, y sólo se recuperan los documentos que tengan coincidencias exactas a las mismas.
- Es un modelo eficiente y formal, de fácil comprensión e implementación, recomendado para el trabajo con expertos sobre un tema específico.
- Un documento es relevante a una consulta si, luego de convertir esta última a Forma Normal Disyuntiva, todos los términos de alguna componente conjuntiva están contenidos en el documento

Implementación

- Se utilizan los símbolos &, | y ~ para representar las operaciones AND, OR y NOT respectivamente.
- En caso de que no aparezca ningún operador se toman con AND entre ellas.
- Al procesar la consulta sólo quedan en ella letras, los símbolos de los operadores y paréntesis (para poder agrupar términos).
- Para convertir la consulta de string a expresión y hallar entonces su forma normal disyuntiva se utilizan los m´etodos sympify y to_dfn de la biblioteca sympy.

Implementación

- Luego de terminar el procesamiento de la consulta, se tendrá una lista en donde en cada posición se tiene otra lista con todos los términos de una componente conjuntiva.
- Para hallar las coincidencias a documentos simplemente se recorre cada componente conjuntiva y se añaden a la respuesta los documentos que contengan a todos los términos de la misma con un score igual a 1.



Modelo vectorial

Descripción

- En el modelo vectorial cada documento se representa como un vector de términos indexados, donde lo único que interesa saber es qué términos aparecen en el documento y qué pesos tiene cada uno de esos términos en el documento.
- Luego dada una consulta, que también se representa como un documento, hallar el conjunto de documentos relevantes se reduce a encontrar el conjunto de vectores documento más similares a la consulta.

Implementación

Modelo fuzzy

Descripción

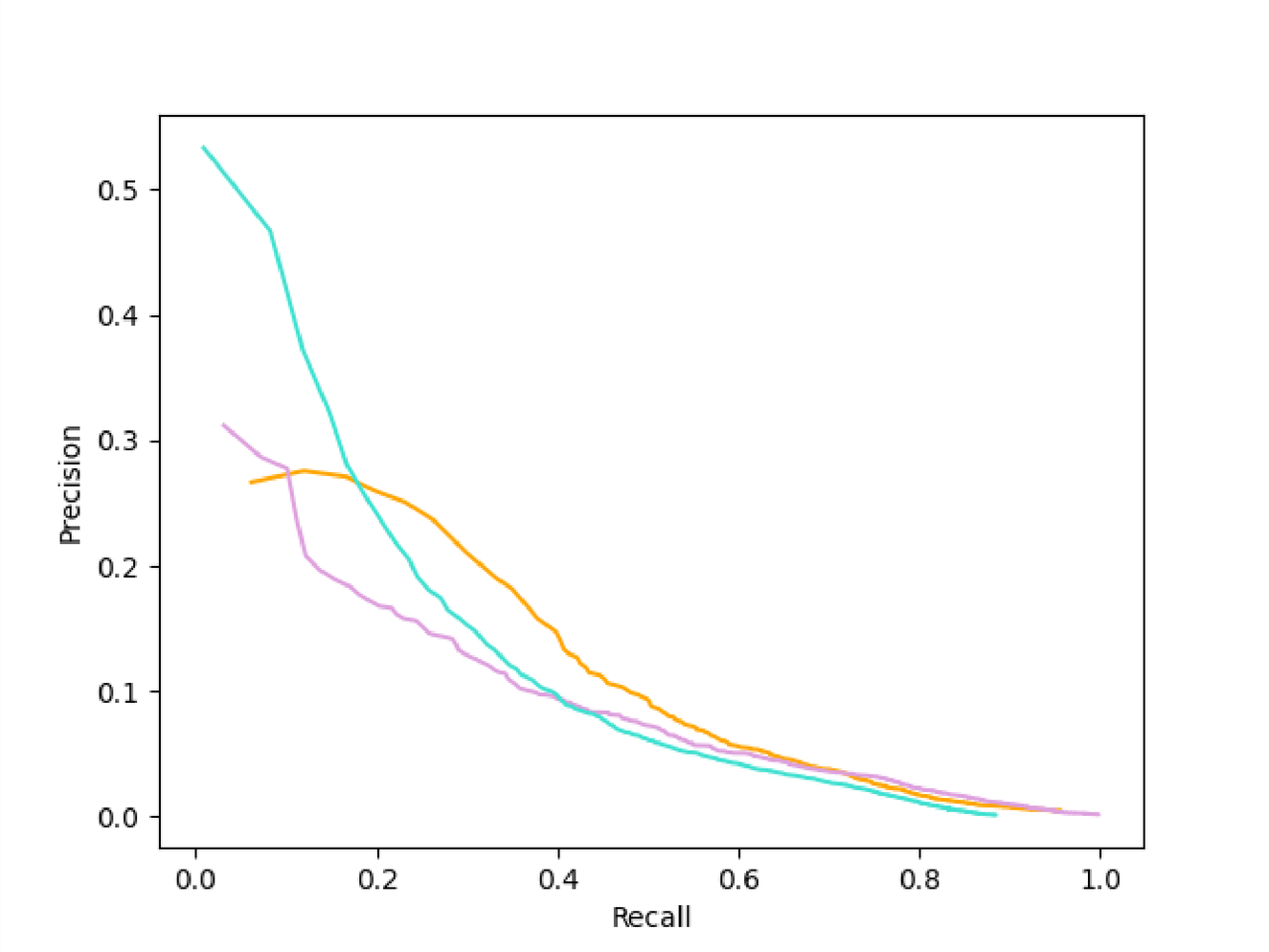
- Es una alternativa al modelo booleano en donde se define un conjunto difuso por cada palabra de la consulta y se determina un grado de pertenencia de cada documento a este conjunto, para luego, basado en esto, poder asignar un grado de relevancia de cada documento a la consulta dada.
- Se basan en el uso de la relación entre términos para expandir los términos de las consultas, de forma que se encuentren más documentos que sean relevantes a las necesidades del usuario
- Esta correlación se basa en la idea de que si dos palabras están relacionadas aparecerán, con frecuencia, juntas en un mismo documento.

Implementación

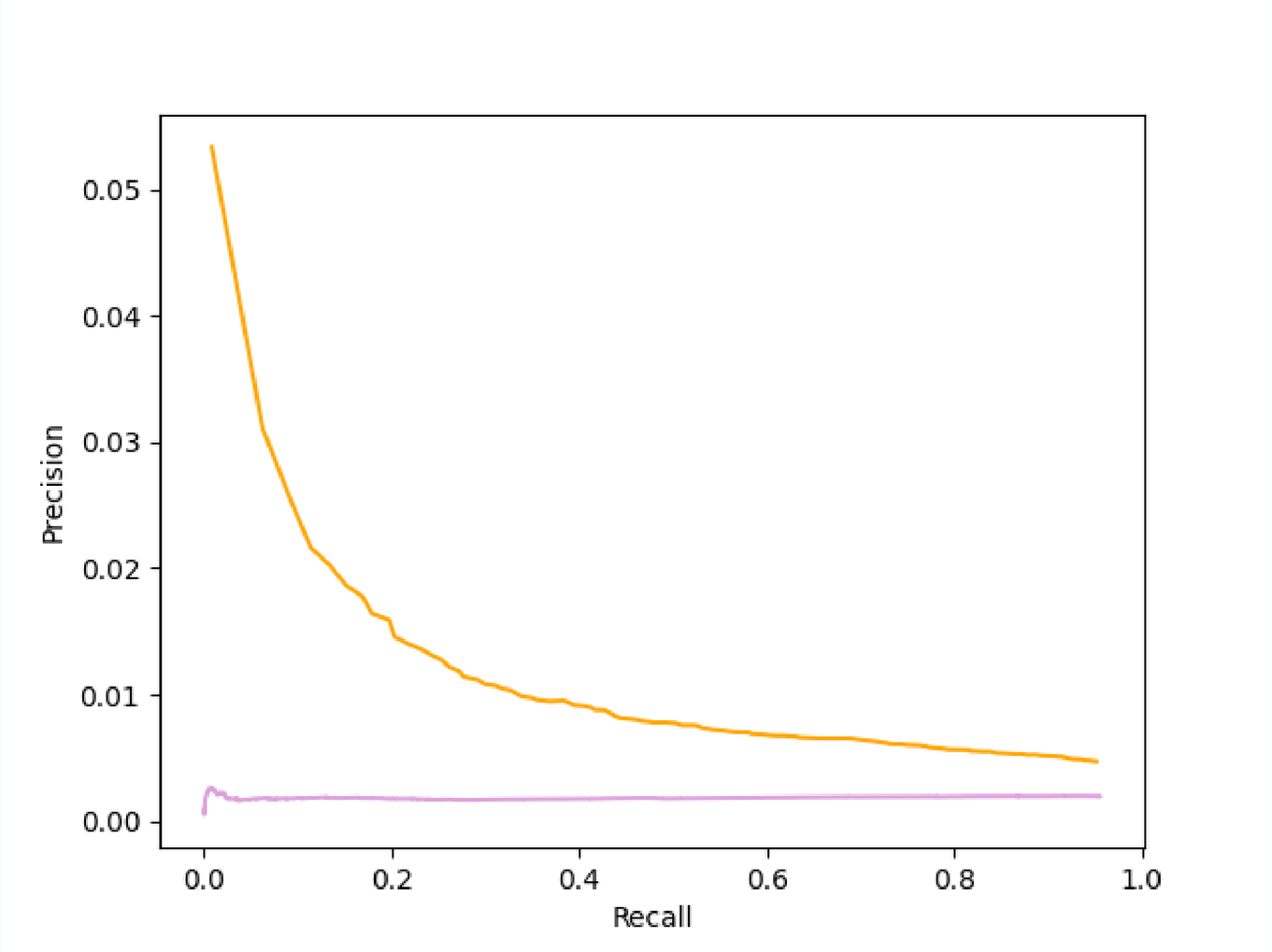
- Basado en el procesamiento de los documentos y consultas que se expusieron en el modelo booleano.
- Se modifica entonces el método que computa el resultado de una consulta, asignando en esta ocasión una puntuación a cada documento respecto a la consulta.
- Los documentos serán recuperados estableciendo un orden según este valor.
- La correlación entre términos se calcula una sola vez para el total de pares de términos que aparecen en todos los documentos, y se guarda en el almacenamiento físico de la computadora.

Evaluación de los modelos

Precisión vs. Recobrado utilizando el modelo **Vectorial**



Precisión vs. Recobrado utilizando el modelo **Fuzzy**



Retroalimentación

Agrupamiento

Descripción

Los algoritmos de agrupamiento reúnen un con junto de documentos en subconjuntos o cl´usteres.

- Los grupos formados deben tener un alto grado de asociaci´on entre los documentos de un mismo grupo y un bajo grado entre miembros de diferentes grupos.
- Es la forma m´as com´un de aprendizaje no supervisado.
- Generalmente se utiliza la distancia euclidiana para medir la similitud entre documentos.

Hipótesis de agrupamiento

Los documentos en el mismo grupo se com portan de manera similar con respecto a la relevancia para las necesidades de información.

K-means

Es uno de los algoritmos de agrupamiento más importantes.

- Este algoritmo se ejecuta sobre un conjunto espacio de documentos representados como vectores dimensionales.
- Fija inicialmente de forma aleatoria los centroides de los cl´usteres y en cada iteración los va moviendo de forma que se disminuya en cada paso la suma de los cuadrados de las distancias de cada documento a su clúster más cercano (RSS).

 Objetivo

Mejorar la recuperación de documentos en el modelo vectorial.

Implementación

- Se implementó un nuevo modelo de recuperación de la información que se sustenta en el modelo vectorial antes explicado.
- Se modificó el método de recuperación de documentos a una consulta para añadir el criterio por clústeres.
- Se utilizó la biblioteca de python sk-learn.
- Para la elección de un número k de clústeres adecuados. se optó por una decisión en funci´on de minimizar los RSS, penalizando el número de clústeres usados.

Gracias