ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA EN SISTEMAS
CARRERA DE INGENIERÍA EN COMPUTACIÓN

(ICCD753) - RECUPERACION DE INFORMACIÓN

GR1CC

Proyecto IB:

Informe de Evaluación

INTEGRANTES:

- Paola Aucapiña
- Kevin Maldonado
- Raquel Zumba

DOCENTE:
Iván Carrera, Ph.D.

FECHA DE ENTREGA:
19/06/2024

PERIODO:
2024-A

Contenido

Introducción

La Recuperación de la Información es una disciplina crucial en el ámbito de la ciencia de datos y la informática. Esta se enfoca en la búsqueda de documentos relevantes dentro de grandes colecciones de datos en respuesta a consultas específicas realizadas por los usuarios [1]. Con el auge de la digitalización y la generación masiva de datos, disponer de sistemas eficientes y precisos para la recuperación de información se ha vuelto indispensable en múltiples sectores, desde motores de búsqueda en la web hasta sistemas de recomendación en comercio electrónico.

Desarrollo

La evaluación de un sistema de recuperación de información es fundamental para garantizar que el sistema cumple con los requisitos de eficiencia y precisión necesarios para proporcionar resultados útiles y relevantes a los usuarios. A través de esta evaluación, se pueden identificar áreas de mejora, optimizar los algoritmos utilizados y asegurar que el sistema responde de manera efectiva a diferentes tipos de consultas.

Por lo tanto, durante la ejecución de este proyecto, se implementó la función del motor de búsqueda. Esto involucró la creación de la estructura para manejar las consultas de los usuarios, la implementación de la similitud el coseno, que se aplicó tanto en Bag of Words como en TF-IDF. Además, se implementó en el motor la lógica para presentar los mejores 10 resultados de manera ordenada, asi como la implementación de un umbral para filtrar resultados irrelevantes.

```python
def motor_busqueda_u(consulta, corpus_df, stopwords_path, umbral=0.0):
    """Realiza la búsqueda de la consulta utilizando las representaciones BoW y TF-IDF."""
    stopwords_set = set(leer_stopwords(stopwords_path))  # Convertir a conjunto para búsquedas rápidas

    # Crear representaciones BoW y TF-IDF
    bow_df = create_bow_representation(corpus_df)
    tfidf_df = create_tfidf_representation(corpus_df)

    # Crear vectorizadores
    bow_vectorizer = CountVectorizer(binary=True)
    bow_vectorizer.fit(corpus_df['Texto'])
    tfidf_vectorizer = TfidfVectorizer()
    tfidf_vectorizer.fit(corpus_df['Texto'])

    # Preprocesar la consulta
    consulta_procesada = preprocesar_consulta(consulta, stopwords_set)

    # Vectorizar la consulta
    query_vector_bow = vectorizar_consulta(consulta_procesada, bow_vectorizer)
    query_vector_tfidf = vectorizar_consulta(consulta_procesada, tfidf_vectorizer)

    # Matrices de documentos (excluyendo la columna 'Archivo')
    document_matrix_bow = bow_df.drop(columns=['Archivo'])
    document_matrix_tfidf = tfidf_df.drop(columns=['Archivo'])

    # Calcular similitudes
    similitudes_bow = calcular_similitud_coseno(query_vector_bow, document_matrix_bow)
    similitudes_tfidf = calcular_similitud_coseno(query_vector_tfidf, document_matrix_tfidf)

    # Crear resultados ordenados aplicando el umbral
    resultados_bow = [(archivo, similitud) for archivo, similitud in zip(bow_df['Archivo'], similitudes_bow) if similitud >= umbral]
    resultados_tfidf = [(archivo, similitud) for archivo, similitud in zip(tfidf_df['Archivo'], similitudes_tfidf) if similitud >= umbral]

    # Ordenar los resultados por similitud (en orden descendente)
    resultados_ordenados_bow = sorted(resultados_bow, key=lambda x: x[1], reverse=True)[:10]
    resultados_ordenados_tfidf = sorted(resultados_tfidf, key=lambda x: x[1], reverse=True)[:10]

    return resultados_ordenados_bow, resultados_ordenados_tfidf
```

Análisis de Resultados

Con la función `motor_busqueda_u` descrita anteriormente , se evaluó el desempeño del sistema bajo diferentes umbrales. Se utilizaron las métricas de precisión, recall y F1 para determinar el umbral óptimo que maximice la eficiencia y precisión del motor de búsqueda.

Motor de búsqueda sin ajuste de umbral

Para evaluar el desempeño del sistema de recuperación de información con un umbral de 0.0, se analizaron los resultados utilizando dos representaciones: Bag of Words (BoW) y TF-IDF, ambas con la métrica de similitud coseno. A continuación, se presentan los resultados detallados y el análisis de las métricas de evaluación: precisión, recall y F1.

- Evaluación del sistema con cada categoría como query de entrada:

Resultados para BoW:

	categoria	precision	recall	f1
0	nat-gas	0.2	1.0	0.333333
1	copper	1.0	1.0	1.000000
2	orange	0.7	1.0	0.823529
3	coconut-oil	0.1	1.0	0.181818
4	sorghum	0.7	1.0	0.823529
..	...	...	...	...
85	crude	1.0	1.0	1.000000
86	rubber	0.8	1.0	0.888889
87	castor-oil	0.1	1.0	0.181818
88	gas	0.0	0.0	0.000000
89	coconut	0.4	1.0	0.571429

[90 rows x 4 columns]

Resultados para TF-IDF:

	categoria	precision	recall	f1
0	nat-gas	0.6	1.0	0.750000
1	copper	1.0	1.0	1.000000
2	orange	0.6	1.0	0.750000
3	coconut-oil	0.2	1.0	0.333333
4	sorghum	0.8	1.0	0.888889
..	...	...	...	...
85	crude	1.0	1.0	1.000000
86	rubber	1.0	1.0	1.000000
87	castor-oil	0.1	1.0	0.181818
88	gas	0.0	0.0	0.000000
89	coconut	0.4	1.0	0.571429

[90 rows x 4 columns]

- Promedio de las métricas de evaluación del SRI:

df_resultados_umbral_0.csv

1 to 2 of 2 entries Filter

	Recall	Precision	F1
BoW	0.7222222222222222	0.38111111111111107	0.45668985774764925
TF-IDF	0.7666666666666667	0.4633333333333334	0.530604028849643

Con un umbral de 0.0, la representación TF-IDF demuestra un mejor rendimiento en comparación con BoW, especialmente en términos de precisión y F1 score. Sin embargo, el recall es alto en ambas representaciones, lo cual es un indicativo positivo de la capacidad del sistema para recuperar documentos relevantes.

Para determinar el umbral óptimo para las métricas de evaluación, es necesario experimentar con diferentes valores de umbral y analizar cómo afectan a las métricas de evaluación. Esto con la finalidad de identificar el umbral que proporcione el mejor equilibrio y rendimiento del sistema de recuperación de información.

Motor de búsqueda con distintos ajuste de umbral

Se fue modificando el parámetro `umbral` en la función `motor_busqueda_u` con los valores de 0.1, 0.2, 0.3 y 0.4.

Resultados con umbrales diferentes

Umbral de 0.1

- Evaluación del sistema con cada categoría como query de entrada:

```
Metricas de evaluación del SRI con un umbral de 0.1 por cada query/categoria
Resultados para BoW con similitud coseno:
    categoria  precision  recall        f1
0   groundnut        0.0     0.0  0.000000
1        ship        0.9     1.0  0.947368
2       cocoa        1.0     1.0  1.000000
3    l-cattle        0.0     0.0  0.000000
4    interest        0.3     1.0  0.461538
..        ...        ...     ...       ...
85      wheat        1.0     1.0  1.000000
86  meal-feed        0.8     1.0  0.888889
87    housing        0.1     1.0  0.181818
88        cpu        0.0     0.0  0.000000
89  palladium        0.5     1.0  0.666667

[90 rows x 4 columns]

Resultados para TF-IDF con similitud coseno:
    categoria  precision  recall        f1
0   groundnut        0.0     0.0  0.000000
1        ship        1.0     1.0  1.000000
2       cocoa        1.0     1.0  1.000000
3    l-cattle        0.4     1.0  0.571429
4    interest        0.6     1.0  0.750000
..        ...        ...     ...       ...
85      wheat        1.0     1.0  1.000000
86  meal-feed        0.8     1.0  0.888889
87    housing        0.3     1.0  0.461538
88        cpu        0.0     0.0  0.000000
89  palladium        0.5     1.0  0.666667

[90 rows x 4 columns]
```

- Promedio de las métricas de evaluación del SRI:

```
Promedio de las metricas de evaluación del SRI con un umbral de 0.1
              Recall  Precision        F1
0     BoW  0.700000   0.400926  0.473654
1  TF-IDF  0.744444   0.483951  0.548283
```

Al comparar los resultados con un umbral de 0.1, nuevamente se observa que la representación TF-IDF sigue siendo superior a BoW en términos de precisión y F1 score, mientras que ambas representaciones mantienen un alto nivel de recall. Esto sugiere que TF-IDF es más efectiva en filtrar documentos irrelevantes mientras mantiene una alta tasa de recuperación de documentos relevantes.

Umbral de 0.2

- Evaluación del sistema con cada categoría como query de entrada:

```
Metricas de evaluación del SRI con un umbral de 0.2 por cada query/categoria
Resultados para BoW con similitud coseno:
     categoria  precision  recall        f1
0    groundnut        0.0     0.0  0.000000
1         ship        0.9     1.0  0.947368
2        cocoa        1.0     1.0  1.000000
3     l-cattle        0.0     0.0  0.000000
4     interest        0.3     1.0  0.461538
..         ...        ...     ...       ...
85       wheat        1.0     1.0  1.000000
86   meal-feed        0.5     1.0  0.666667
87     housing        0.1     1.0  0.181818
88         cpu        0.0     0.0  0.000000
89   palladium        0.0     0.0  0.000000

[90 rows x 4 columns]

Resultados para TF-IDF con similitud coseno:
     categoria  precision  recall        f1
0    groundnut        0.0     0.0  0.000000
1         ship        1.0     1.0  1.000000
2        cocoa        1.0     1.0  1.000000
3     l-cattle        0.4     1.0  0.571429
4     interest        0.6     1.0  0.750000
..         ...        ...     ...       ...
85       wheat        1.0     1.0  1.000000
86   meal-feed        0.8     1.0  0.888889
87     housing        0.3     1.0  0.461538
88         cpu        0.0     0.0  0.000000
89   palladium        1.0     1.0  1.000000

[90 rows x 4 columns]
```

- Promedio de las métricas de evaluación del SRI:

```
Promedio de las metricas de evaluación del SRI con un umbral de 0.2
              Recall  Precision        F1
0     BoW  0.500000   0.318735  0.361363
1  TF-IDF  0.722222   0.505794  0.566358
```

Con este umbral, se observó que que TF-IDF muestra un recall significativamente más alto en comparación con BoW, además, una mayor precisión en comparación con BoW (0.505794 vs. 0.318735). TF-IDF tiene un F1-score más alto que BoW (0.566358 vs. 0.361363), lo que indica un mejor equilibrio entre precisión y recall. TF-IDF logra una mejor combinación de estas métricas en comparación con BoW bajo el umbral de 0.2.

Umbral de 0.3

- Evaluación del sistema con cada categoría como query de entrada:

```
Metricas de evaluación del SRI con un umbral de 0.3 por cada query/categoria
Resultados para BoW con similitud coseno:
     categoria  precision  recall        f1
0    groundnut        0.0     0.0  0.000000
1         ship        1.0     1.0  1.000000
2        cocoa        1.0     1.0  1.000000
3     l-cattle        0.0     0.0  0.000000
4     interest        0.3     1.0  0.461538
..         ...        ...     ...       ...
85       wheat        1.0     1.0  1.000000
86   meal-feed        0.0     0.0  0.000000
87     housing        0.2     1.0  0.333333
88         cpu        0.0     0.0  0.000000
89   palladium        0.0     0.0  0.000000

[90 rows x 4 columns]

Resultados para TF-IDF con similitud coseno:
     categoria  precision  recall        f1
0    groundnut   0.000000     0.0  0.000000
1         ship   1.000000     1.0  1.000000
2        cocoa   1.000000     1.0  1.000000
3     l-cattle   0.333333     1.0  0.500000
4     interest   1.000000     1.0  1.000000
..         ...        ...     ...       ...
85       wheat   1.000000     1.0  1.000000
86   meal-feed   0.666667     1.0  0.800000
87     housing   0.375000     1.0  0.545455
88         cpu   0.000000     0.0  0.000000
89   palladium   0.000000     0.0  0.000000

[90 rows x 4 columns]
```

- Promedio de las métricas de evaluación del SRI:

```
Promedio de las metricas de evaluación del SRI con un umbral de 0.3
             Recall  Precision        F1
0     BoW  0.233333   0.182500  0.195101
1  TF-IDF  0.600000   0.478889  0.512855
```

TF-IDF demuestra ser considerablemente más efectivo que BoW en la recuperación de información bajo un umbral de 0.3, según las métricas de recall, precisión y F1-score evaluadas. Además, la diferencia en las métricas entre TF-IDF y BoW destaca la capacidad de TF-IDF para ajustarse mejor a umbrales más altos de relevancia.

Umbral de 0.4

- Evaluación del sistema con cada categoría como query de entrada:

```
Metricas de evaluación del SRI con un umbral de 0.4 por cada query/categoria
Resultados para BoW con similitud coseno:
    categoria  precision  recall   f1
0   groundnut       0.00     0.0  0.0
1        ship       0.00     0.0  0.0
2       cocoa       0.00     0.0  0.0
3    l-cattle       0.00     0.0  0.0
4    interest       0.25     1.0  0.4
..        ...        ...     ...  ...
85      wheat       0.00     0.0  0.0
86  meal-feed       0.00     0.0  0.0
87    housing       0.00     0.0  0.0
88        cpu       0.00     0.0  0.0
89  palladium       0.00     0.0  0.0

[90 rows x 4 columns]

Resultados para TF-IDF con similitud coseno:
    categoria  precision  recall   f1
0   groundnut   0.000000     0.0  0.0
1        ship   1.000000     1.0  1.0
2       cocoa   1.000000     1.0  1.0
3    l-cattle   0.333333     1.0  0.5
4    interest   0.000000     0.0  0.0
..        ...        ...     ...  ...
85      wheat   1.000000     1.0  1.0
86  meal-feed   0.000000     0.0  0.0
87    housing   1.000000     1.0  1.0
88        cpu   0.000000     0.0  0.0
89  palladium   0.000000     0.0  0.0
```

- Promedio de las métricas de evaluación del SRI

```
Promedio de las metricas de evaluación del SRI con un umbral de 0.4
             Recall  Precision        F1
0     BoW  0.077778   0.069444  0.071111
1  TF-IDF  0.511111   0.433704  0.454587
```

TF-IDF sigue demostrando ser más efectivo que BoW en la recuperación de información bajo un umbral de 0.4, según las métricas de recall, precisión y F1-score evaluadas.

Conclusiones

En general, TF-IDF tiende a mostrar mejores métricas (recall, precisión y F1-score) en la mayoría de los umbrales evaluados. Esto sugiere que TF-IDF es más efectivo para identificar y recuperar documentos relevantes en comparación con BoW en este conjunto de datos y configuración específica.

A medida que se aumentó el umbral, se observó que BoW como TF-IDF mostraron una tendencia a reducir el recall y, en algunos casos, la precisión también. Esto fue algo esperado, ya que al aumentar el umbral, se vuelven más restrictivo el SRI en cuanto a qué documentos considera relevantes.

El umbral óptimo puede variar dependiendo de las necesidades específicas de la aplicación. Sin embargo, basándonos en estos resultados, podríamos considerar el umbral de 0.2 como un buen punto de equilibrio para TF-IDF, ya que muestra un buen balance entre recall (0.722222) y precisión (0.505794), con un F1-score (0.566358) relativamente alto. Para BoW, aunque los valores son más bajos en general, el umbral de 0.1 podría ser una opción si se valora más el recall sobre la precisión.

Referencias

[1] "¿Qué es la recuperación de información? | Una guía completa de la recuperación de información (IR) | Elastic." Accessed: Jun. 19, 2024. [Online]. Available: https://www.elastic.co/es/what-is/information-retrieval