ESCUELA POLITÉCNICA NACIONAL
FACULTAD DE INGENIERÍA DE SISTEMAS
INGENIERÍA EN CIENCIAS DE LA COMPUTACIÓN

Proyecto de 1er Bimestre
Sistema de Recuperación de Información

Nombre: Alexis Vera

1. Descripción del Corpus:

El corpus utilizado corresponde a un conjunto de documentos en formato CSV, cargado desde Kaggle para efectos del desarrollo y pruebas del sistema de recuperación de información.

Este corpus corresponde a un conjunto de detalles y características sobre criticas y catas de vinos, enfocaremos nuestro sistema de recuperación en 2 columnas principales, la id del vino en cuestión y al critica que recibió.
Cada documento incluye principalmente texto libre, almacenado en una columna específica que posteriormente fue sometida a un proceso de preprocesamiento para permitir su uso eficiente en los modelos de recuperación.

Para el preprocesamiento se aplicó:

- Convertir texto en minúsculas
- Eliminar caracteres no alfanuméricos
- Tokenizacion
- Stopwords
- Eliminación de tokens cortos

Los tokens generados fueron agregados a una nueva columna dentro del dataset para construir representaciones basadas en Jaccard, TF-IDF y BM25, como para la ejecución de consultas de texto libre y la evaluación del desempeño del sistema completo.

El corpus utilizado puede ser encontrado en el siguiente enlace: https://www.kaggle.com/datasets/zynicide/wine-reviews

2. Explicación de las decisiones de diseño

Se decidió trabajar con una representación basada en tokens normalizados, ya que esto permite:

- Reducir la variabilidad del texto.
- Mejorar la coincidencia entre consulta y documentos.
- Facilitar el cálculo de métricas como Jaccard y TF-IDF.
- La tokenización uniforme también permitió la construcción de un índice invertido.

Construcción del índice invertido

El índice invertido se implementó como un diccionario donde cada término apunta a la lista de documentos en los que aparece. Esta estructura permite:

- Recuperación eficiente basada en presencia o ausencia de términos.
- Reducción del espacio de búsqueda.
- Uso de métricas basadas en conjuntos y frecuencias de términos.

Métodos de recuperación implementados

- **Jaccard**: Basado en la similitud entre conjuntos de tokens. Ventajoso por su simplicidad, pero limitado para textos largos.
- **TF-IDF + Coseno**: Permite capturar la importancia diferencial de los términos y la similitud semántica entre documentos.
- **BM25**: Modelo basado en probabilidad y saturación logarítmica del término, ampliamente utilizado en sistemas de búsqueda modernos.

Interfaz de línea de comandos

Se desarrolló una CLI para permitir con el objetivo de realizar consultas por parte del usuario, con la posibilidad de elegir el método de búsqueda y obtener resultados a modo de ranking de documentos relevantes según cada método. Esta interfaz facilita la experimentación y mejora la comprensión del funcionamiento del sistema.

Qrels para evaluación

Debido a la ausencia de etiquetas reales de relevancia, se definió un conjunto manual de documentos relevantes por consulta, lo que permitió medir métricas estándar como Precision, Recall y MAP.

3. Ejemplos de consulta y resultados

Consultas con Jaccard

```
#Funcionalidad
consulta = ["fruity", "aroma", "wine"]

resultados = recuperar_por_jaccard(df, consulta, top_k=5)

#Visualizar resultados
for doc_id, score in resultados:
    print(f"Doc: {doc_id}, Similitud: {score}")
    print(df.loc[doc_id, 'description'])
    print("-----")
```

```
Doc: 6486, Similitud: 0.2
This wine has a geranium aroma and a strangely perfumed character. It is light and fruity although not likely to develop much.
-----
Doc: 48045, Similitud: 0.2
This is a straightforward, clean, fruity wine, with a green apple flavor.
-----
Doc: 25921, Similitud: 0.1875
Very ripe in aroma, almost thick in texture and syrupy in flavor, this wine is full bodied, super fruity and smooth.
-----
Doc: 108778, Similitud: 0.1875
With an aroma like cherry cola, and fruity, almost sweet flavors, this medium-bodied wine is easy to drink.
-----
Doc: 56946, Similitud: 0.18181818181818182
Light and fruity, the wine is dilute and gently textured. It is ready to drink.
-----
```

Consultas TF-IDF

```python
#Funcionalidad
consulta = ["fruity", "aroma", "wine"]

resultados = recuperar_por_coseno(consulta, tfidf_matrix, vectorizador, top_k=5)

#Visualizacion
for doc, score in resultados:
    print(f"Doc {doc} → Score: {score:.4f}")
    print(df.loc[doc, 'description'])
    print("-----")
```

```
Doc 108778 → Score: 0.5110
With an aroma like cherry cola, and fruity, almost sweet flavors, this medium-bodied wine is easy to drink.
-----
Doc 25921 → Score: 0.4186
Very ripe in aroma, almost thick in texture and syrupy in flavor, this wine is full bodied, super fruity and smooth.
-----
Doc 14546 → Score: 0.4124
This is a light, soft and creamy wine, with an apricot aroma and flavor, fresh acidity and an attractive fruity aftertaste. It is
-----
Doc 120116 → Score: 0.3998
A vanilla aroma is followed by old oak flavors and sweetly textured fruits. Not a wine for aging, this is already soft and fruity.
-----
Doc 84146 → Score: 0.3992
In the house style of this producer this wine is soft and fruity. Its strawberry aroma is followed by a gentle, ripe wine with red
-----
```

Consultas BM25

```python
#Funcionalidad
consulta = ["fruity", "aroma", "wine"]

resultados = bm25.search(consulta, top_k=5)

#Visualizacion
for doc_id, score in resultados:
    print(f"Doc: {doc_id}  →  Score: {score:.4f}")
    print(df.loc[doc_id, 'description'])
    print("-----")
```

```
Doc: 6486  →  Score: 10.0350
This wine has a geranium aroma and a strangely perfumed character. It is light and fruity although not likely to develop much.
-----
Doc: 108778  →  Score: 10.0350
With an aroma like cherry cola, and fruity, almost sweet flavors, this medium-bodied wine is easy to drink.
-----
Doc: 25921  →  Score: 9.8838
Very ripe in aroma, almost thick in texture and syrupy in flavor, this wine is full bodied, super fruity and smooth.
-----
Doc: 52725  →  Score: 9.8838
A caramel aroma is followed by sweet strawberry fruit from the Touriga Nacional, lending a wine that is freshly fruity and soft.
-----
Doc: 52014  →  Score: 9.7371
This fruity wine has an aroma of tobacco and a spicy black-currant taste. It is full bodied, and it should be kept for another yea
```

Resultados con búsqueda de texto libre

```python
#Ingreso de la Query
query_usuario = input("Ingrese su consulta de texto libre: ")
```

```
Ingrese su consulta de texto libre: sweet grape wine
```

Jaccard

```
#Funcionalidad
resultados = buscar_jaccard_texto_libre(query_usuario, df, top_k=5)

#Visualizar resultados
for doc_id, score in resultados:
    print(f"Doc: {doc_id}, Similitud: {score}")
    print(df.loc[doc_id, 'description'])
    print("-----")
```

```
Doc: 13022, Similitud: 0.25
A flat tasting, slightly sweet, generic white wine.
-----
Doc: 123451, Similitud: 0.2222222222222222
As sweet as a dessert wine, with simple pineapple jam flavors.
-----
Doc: 14735, Similitud: 0.2
As sweet and sugary as a dessert wine, with watery berry flavors.
-----
Doc: 18349, Similitud: 0.2
Made with undisclosed grape varieties, this rustic, softly sweet wine has apricot, honey and herb flavors.
-----
Doc: 117009, Similitud: 0.2
Made with undisclosed grape varieties, this rustic, softly sweet wine has apricot, honey and herb flavors.
-----
```

4. Análisis de métricas de evaluación

Para evaluar el sistema se utilizaron las métricas estándar en recuperación de información:

- Precision
- Recall
- MAP (Mean Average Precision)

Los valores de relevancia (qrels) fueron asignados manualmente para cada consulta, permitiendo comparar objetivamente el rendimiento de cada modelo.

El sistema desarrollado demuestra cómo diferentes técnicas de recuperación de información afectan el rendimiento al resolver consultas ingresadas por el usuario. La comparación entre Jaccard, TF-IDF y BM25 permite evidenciar las fortalezas y limitaciones de cada enfoque.
Dando mejores resultados según el tamaño del corpus, presencia de términos y múltiples términos relevantes