Informe: Análisis TF-IDF y similitud coseno

Informe: Análisis TF-IDF y similitud coseno

Fecha: 2025-11-04

Resumen

Este informe resume el análisis realizado con la aplicación CLI `tfidf` (C++) incluida en este repositorio. El software procesa documentos `.txt`, aplica filtrado por stopwords y lematización, y calcula TF, IDF, TF-IDF y la matriz de similitud coseno entre documentos.

Alcance y limitaciones

- El repositorio solicitado en https://github.com/ull-cs/gestion-conocimiento/tree/main/recommeder-systems/examples-documents no pudo ser descargado automáticamente desde este entorno (restricciones de red). Para reproducir exactamente esos documentos, por favor provee los ficheros o descarga y colócalos en el workspace (por ejemplo en `ejemplos/` o `samples/`).
- En su lugar, se ha utilizado el conjunto de ejemplo local (`samples/doc1.txt`, `samples/doc2.txt`) y un conjunto nuevo de 10 documentos creados en `samples/new_docs/` (ver más abajo).

Metodología

- Tokenización: se normalizan los caracteres a minúsculas y se separan tokens por caracteres no alfabéticos/digitales.
- Stopwords: se eliminan tokens presentes en `samples/stopwords.txt`.
- Lematización: se sustituye cada token por su lema según `samples/lemmas.txt`.
- TF: recuento bruto de ocurrencias por documento.
- IDF: log(N / df) con N = número de documentos, df = número de documentos con el término.
- TF-IDF: TF * IDF.
- Similitud coseno: producto escalar de vectores TF-IDF normalizado por normas.

Datos analizados

- Documentos locales de ejemplo:
  - `samples/doc1.txt`
  - `samples/doc2.txt`

- Conjunto propuesto (10 nuevos documentos creados aquí):
  - `samples/new_docs/doc01.txt` — sobre un gato en la alfombra
  - `samples/new_docs/doc02.txt` — economía local y negocios
  - `samples/new_docs/doc03.txt` — lluvia y flores
  - `samples/new_docs/doc04.txt` — partido de fútbol y gol
  - `samples/new_docs/doc05.txt` — receta de sopa de verduras
  - `samples/new_docs/doc06.txt` — innovaciones tecnológicas, IA
  - `samples/new_docs/doc07.txt` — viaje por carretera y paisaje
  - `samples/new_docs/doc08.txt` — concierto de música
  - `samples/new_docs/doc09.txt` — salud pública y programas
  - `samples/new_docs/doc10.txt` — mercado de arte local

Resultados — documentos de ejemplo (`samples/doc1.txt`, `samples/doc2.txt`)

Los resultados para los documentos de ejemplo locales se almacenan en `outputs/doc_examples.txt` y se muestran a continuación (extracto):

```
$(cat outputs/doc_examples.txt)
```

Resultados — 10 documentos propuestos

Los resultados completos (TF/IDF/TF-IDF y matriz de similitud) para los 10 documentos nuevos están en `outputs/new_docs.txt`. Se incluye a continuación el contenido:

```
$(cat outputs/new_docs.txt)
```

Análisis y conclusiones

- Observaciones sobre TF/IDF:
  - En documentos pequeños y con vocabulario distinto por tema (arte, cocina, tecnología), muchos términos aparecen en un único documento -> IDF alto, TF-IDF alto para esos términos.
  - Términos muy comunes (artículos y conjunciones) reciben IDF cercano a 0 o son filtrados por stopwords, por lo que su TF-IDF es bajo o nulo.
- Observaciones sobre similitud coseno:
  - En el conjunto `new_docs` se observan similitudes bajas en general (valores ~0.00x–0.15), salvo casos donde hay vocabulario compartido (por ejemplo, textos sobre ocio/turismo pueden mostrar similitud más alta entre sí).
  - Valores mayores indican documentos con términos únicos pero relacionados (por ejemplo, `doc04` partido de fútbol muestra una similitud mayor con `doc08` concierto? — revisar matriz para confirmar temas relacionados por términos como "el", "la" y palabras no descartadas).
- Limitaciones:
  - Tokenización y lematización son simples y pueden fragmentar palabras (p. ej. truncados si hay acentos o caracteres especiales); mejorar normalización Unicode y lematización aumentaría calidad.
  - El IDF usado es log(N/df) sin suavizado; en corpus muy pequeño produce valores extremos.

Propuesta de 10 documentos adicionales (descripciones)

Listado (ya creados en `samples/new_docs/`):
1. doc01.txt — El gato negro duerme en la alfombra...
2. doc02.txt — Economía local, negocios pequeños...
3. doc03.txt — Lluvia y flores en el jardín...
4. doc04.txt — Partido de fútbol y gol de último minuto...
5. doc05.txt — Receta de sopa de verduras...
6. doc06.txt — Innovaciones tecnológicas e IA...
7. doc07.txt — Viaje por carretera y paisajes...
8. doc08.txt — Concierto de música clásica y contemporánea...
9. doc09.txt — Salud pública y programas comunitarios...
10. doc10.txt — Mercado de arte local; pinturas y esculturas...

Cómo reproducir (comandos)

Compilar con Makefile (genera `bin/tfidf`):

```bash
make
```

Ejecutar sobre los 10 documentos nuevos:

```bash
./bin/tfidf -d samples/new_docs -s samples/stopwords.txt -l samples/lemmas.txt > outputs/new_docs.txt
```

Exportar los resultados en PDF: este repositorio incluye `report.md` y `report.pdf` si la conversión fue exitosa.

Siguientes pasos recomendados

- Normalizar Unicode y acentos (por ejemplo con ICU o una librería de normalización)
- Añadir opción para normalizar TF (TF normalizado o log-scaling)
- Añadir salida CSV/TSV/JSON para facilitar análisis posterior
- Integrar scripts para crear automáticamente un resumen de top-k términos por documento

---

Apéndice: archivos de salida generados

- `outputs/doc_examples.txt` — salida para ejemplos locales
- `outputs/new_docs.txt` — salida para 10 documentos propuestos

Fin del informe.

Apéndice: Salida de la aplicación (ejemplos locales)

```
Documento: /workspaces/modelos_basados_contenido/samples/doc1.txt
Idx     Termino TF      IDF     TF-IDF

Documento: /workspaces/modelos_basados_contenido/samples/doc2.txt
Idx     Termino TF      IDF     TF-IDF

Matriz de similitud coseno:
        0       1
0       0       0
1       0       0
```

Apéndice: Salida de la aplicación (10 documentos nuevos)

```
Documento: /workspaces/modelos_basados_contenido/samples/new_docs/doc01.txt
Idx     Termino TF      IDF     TF-IDF
5       al      1       2.30259 2.30259
6       alfombra        1       2.30259 2.30259
38      duerme  1       2.30259 2.30259
40      el      1       0.356675        0.356675
41      en      2       1.60944 3.21888
49      gata    1       2.30259 2.30259
50      gato    2       2.30259 4.60517
58      la      3       0.356675        1.07002
66      maulla  1       2.30259 2.30259
70      mira    1       2.30259 2.30259
78      negro   1       2.30259 2.30259
79      noche   1       2.30259 2.30259
121     y       1       0.105361        0.105361

Documento: /workspaces/modelos_basados_contenido/samples/new_docs/doc02.txt
Idx     Termino TF      IDF     TF-IDF
0       a       1       2.30259 2.30259
15      buscan  1       2.30259 2.30259
22      clientes        1       2.30259 2.30259
30      crece   1       2.30259 2.30259
39      econom  1       2.30259 2.30259
58      la      1       0.356675        0.356675
60      lentamente      1       2.30259 2.30259
62      local   1       1.20397 1.20397
63      los     1       0.916291        0.916291
67      mejoran 1       2.30259 2.30259
77      negocios        1       2.30259 2.30259
82      os      1       1.60944 1.60944
92      peque   1       2.30259 2.30259
105     servicio        1       2.30259 2.30259
110     su      1       2.30259 2.30259
121     y       1       0.105361        0.105361

Documento: /workspaces/modelos_basados_contenido/samples/new_docs/doc03.txt
Idx     Termino TF      IDF     TF-IDF
4       agua    1       2.30259 2.30259
16      ca      1       2.30259 2.30259
31      crecer  1       2.30259 2.30259
32      da      1       2.30259 2.30259
34      del     1       1.60944 1.60944
47      flores  2       2.30259 4.60517
57      jard    1       2.30259 2.30259
58      la      1       0.356675        0.356675
59      las     2       1.60944 3.21888
61      lluvia  1       2.30259 2.30259
74      n       1       1.20397 1.20397
76      necesitan       1       2.30259 2.30259
88      para    1       2.30259 2.30259
101     reg     1       2.30259 2.30259
108     sol     1       2.30259 2.30259
```

```
121     y       1       0.105361        0.105361

Documento: /workspaces/modelos_basados_contenido/samples/new_docs/doc04.txt
Idx     Termino TF      IDF     TF-IDF
3       aficionados     1       2.30259 2.30259
20      celebraron      1       2.30259 2.30259
27      con     1       1.60944 1.60944
33      de      1       1.20397 1.20397
40      el      2       0.356675        0.71335
41      en      2       1.60944 3.21888
45      f       1       1.60944 1.60944
52      gol     1       2.30259 2.30259
58      la      1       0.356675        0.356675
63      los     1       0.916291        0.916291
64      ltimo   1       2.30259 2.30259
69      minuto  1       2.30259 2.30259
89      partido 1       2.30259 2.30259
94      plaza   1       2.30259 2.30259
112     tbol    1       2.30259 2.30259
114     termin  1       2.30259 2.30259
117     un      1       2.30259 2.30259

Documento: /workspaces/modelos_basados_contenido/samples/new_docs/doc05.txt
Idx     Termino TF      IDF     TF-IDF
7       apio    1       2.30259 2.30259
19      cebolla 1       2.30259 2.30259
24      cocinar 1       2.30259 2.30259
33      de      1       1.20397 1.20397
53      incluye 1       2.30259 2.30259
58      la      1       0.356675        0.356675
85      paciencia       1       2.30259 2.30259
91      patata  1       2.30259 2.30259
100     receta  1       2.30259 2.30259
102     requiere        1       2.30259 2.30259
109     sopa    1       2.30259 2.30259
118     una     1       2.30259 2.30259
119     verduras        1       2.30259 2.30259
121     y       1       0.105361        0.105361
122     zanahoria       1       2.30259 2.30259

Documento: /workspaces/modelos_basados_contenido/samples/new_docs/doc06.txt
Idx     Termino TF      IDF     TF-IDF
1       aceleran        1       2.30259 2.30259
10      artificial      1       2.30259 2.30259
17      cambio  1       2.30259 2.30259
40      el      1       0.356675        0.356675
51      gicas   1       2.30259 2.30259
54      industrias      1       2.30259 2.30259
55      innovaciones    1       2.30259 2.30259
56      inteligencia    1       2.30259 2.30259
59      las     1       1.60944 1.60944
103     rob     1       2.30259 2.30259
113     tecnol  1       2.30259 2.30259
115     tica    1       2.30259 2.30259
116     transforman     1       2.30259 2.30259
121     y       1       0.105361        0.105361

Documento: /workspaces/modelos_basados_contenido/samples/new_docs/doc07.txt
Idx     Termino TF      IDF     TF-IDF
18      carretera       1       2.30259 2.30259
25      comida  1       2.30259 2.30259
34      del     1       1.60944 1.60944
37      disfrut 1       2.30259 2.30259
40      el      1       0.356675        0.356675
46      familia 1       2.30259 2.30259
58      la      2       0.356675        0.71335
62      local   1       1.20397 1.20397
71      monta   1       2.30259 2.30259
72      mostr   1       2.30259 2.30259
82      os      1       1.60944 1.60944
83      osos    1       2.30259 2.30259
86      paisaje 1       2.30259 2.30259
87      paisajes        1       2.30259 2.30259
```

```
95      por     1       2.30259 2.30259
99      r       1       2.30259 2.30259
120     viaje   1       2.30259 2.30259
121     y       2       0.105361        0.210721

Documento: /workspaces/modelos_basados_contenido/samples/new_docs/doc08.txt
Idx     Termino TF      IDF     TF-IDF
8       aplaudieron     1       2.30259 2.30259
14      blico   1       2.30259 2.30259
21      cl      1       2.30259 2.30259
27      con     1       1.60944 1.60944
28      concierto       1       2.30259 2.30259
29      contempor       1       2.30259 2.30259
35      director        1       2.30259 2.30259
36      dirigi  1       2.30259 2.30259
40      el      3       0.356675        1.07002
63      los     1       0.916291        0.916291
65      m       2       2.30259 4.60517
74      n       1       1.20397 1.20397
75      nea     1       2.30259 2.30259
81      ofreci  1       2.30259 2.30259
84      p       1       1.60944 1.60944
90      pasi    1       2.30259 2.30259
106     sica    2       1.60944 3.21888
107     sicos   1       2.30259 2.30259
121     y       2       0.105361        0.210721

Documento: /workspaces/modelos_basados_contenido/samples/new_docs/doc09.txt
Idx     Termino TF      IDF     TF-IDF
2       actividad       1       2.30259 2.30259
12      bienestar       1       2.30259 2.30259
13      blica   1       2.30259 2.30259
26      comunitarios    1       2.30259 2.30259
40      el      1       0.356675        0.356675
45      f       1       1.60944 1.60944
48      fomentan        1       2.30259 2.30259
58      la      3       0.356675        1.07002
74      n       1       1.20397 1.20397
84      p       1       1.60944 1.60944
96      prevenci        1       2.30259 2.30259
97      programas       1       2.30259 2.30259
98      promueve        1       2.30259 2.30259
104     salud   1       2.30259 2.30259
106     sica    1       1.60944 1.60944
121     y       1       0.105361        0.105361

Documento: /workspaces/modelos_basados_contenido/samples/new_docs/doc10.txt
Idx     Termino TF      IDF     TF-IDF
9       arte    1       2.30259 2.30259
11      artistas        1       2.30259 2.30259
23      cnicas  1       2.30259 2.30259
33      de      1       1.20397 1.20397
40      el      1       0.356675        0.356675
42      esculturas      1       2.30259 2.30259
43      estilos 1       2.30259 2.30259
44      exhibe  1       2.30259 2.30259
62      local   1       1.20397 1.20397
63      los     1       0.916291        0.916291
68      mercado 1       2.30259 2.30259
73      muestran        1       2.30259 2.30259
80      nuevas  1       2.30259 2.30259
93      pinturas        1       2.30259 2.30259
111     t       1       2.30259 2.30259
121     y       2       0.105361        0.210721

Matriz de similitud coseno:
        0       1       2       3       4       5       6       7       8       9
0       1       0.00568828      0.00480499      0.149924        0.00560604      0.00204049
0.0124914       0.00468916      0.0193855       0.00218322
1       0.00568828      1       0.00184468      0.0143679       0.00215221      0.000178517
0.0643976       0.0109079       0.00646763      0.0368151
2       0.00480499      0.00184468      1       0.00159703      0.00181801      0.0705251
0.03613 0.0157357       0.0256271       0.000298716
```

3 0.149924 0.0143679 0.00159703 1 0.0230939 0.00385145
0.00714611 0.0499583 0.050011 0.0381355
4 0.00560604 0.00215221 0.00181801 0.0230939 1 0.000175936
0.00406746 0.00027695 0.00637412 0.0231032
5 0.00204049 0.000178517 0.0705251 0.00385145 0.000175936 1
0.00227059 0.00520663 0.00232005 0.00242415
6 0.0124914 0.0643976 0.03613 0.00714611 0.00406746 0.00227059 1
0.00509582 0.0142028 0.0244005
7 0.00468916 0.0109079 0.0157357 0.0499583 0.00027695 0.00520663
0.00509582 1 0.127058 0.0161615
8 0.0193855 0.00646763 0.0256271 0.050011 0.00637412 0.00232005
0.0142028 0.127058 1 0.00248234
9 0.00218322 0.0368151 0.000298716 0.0381355 0.0231032 0.00242415
0.0244005 0.0161615 0.00248234 1