Escuela Superior
de Ingeniería y Tecnología
Universidad de La Laguna

Informe del análisis

Sistema de recomendación

Modelos basados en el contenido

C/ Padre Herrera s/n
38207 La Laguna
Santa Cruz de Tenerife. España

T: 900 43 25 26

ull.es

Análisis de los casos

Tenemos varios ficheros de texto, en los cuales cada línea corresponde a un documento, analizaremos varios casos para posteriormente sacar conclusiones acerca de las pruebas y la utilidad de la práctica.

Para ello utilizaremos los 3 documentos proporcionados, todos ellos en inglés, y una parte del clásico “El Quijote”, para comprobar un documento con un tamaño bastante mayor.

Disponemos de unos resultados dados por 2 cálculos principales, los cuales son:

- TF(Term Frequency): (Número de veces que aparece un término en el documento)/ (Número total de términos en el documento). Por lo que el número debe ser siempre entre 0 y 1.
- IDF(Inverse Document Frequency): Log(Número total de documentos/Número de documentos que contienen el término).

Como pruebas utilizaremos los 3 documentos en inglés proporcionados en el Github de la actividad, y un extracto en texto plano del libro “El Quijote” para ver cómo funciona el programa con un documento más extenso, en este caso, necesitaremos pasar por argumentos los ficheros en español, tanto “stop_words” como el “corpus.”

Ejecución documents-01.txt :

```
Document 8:
     Index       Term      TF     IDF  TF_IDF
0        0     accent  0.0000  2.7047  0.0000
1        1    acidity  0.1750  1.3185  0.2308
2        2      apple  0.0000  2.2993  0.0000
3        3     aromas  0.0000  1.7885  0.0000
4        4  astringent  0.0000  2.7047  0.0000
..     ...        ...     ...     ...     ...
120    120   unripened  0.0000  2.7047  0.0000
121    121      whiff  0.0000  2.7047  0.0000
122    122      white  0.0000  2.7047  0.0000
123    123       wine  0.2132  1.6061  0.3425
124    124     winter  0.0000  2.7047  0.0000

[125 rows x 5 columns]

Document 9:
     Index       Term      TF     IDF  TF_IDF
0        0     accent  0.2778  2.7047  0.7514
1        1    acidity  0.0000  1.3185  0.0000
2        2      apple  0.0000  2.2993  0.0000
3        3     aromas  0.0000  1.7885  0.0000
4        4  astringent  0.0000  2.7047  0.0000
..     ...        ...     ...     ...     ...
120    120   unripened  0.0000  2.7047  0.0000
121    121      whiff  0.0000  2.7047  0.0000
122    122      white  0.0000  2.7047  0.0000
123    123       wine  0.1650  1.6061  0.2650
124    124     winter  0.0000  2.7047  0.0000

[125 rows x 5 columns]

Document 10:
     Index       Term      TF     IDF  TF_IDF
0        0     accent  0.0000  2.7047  0.0000
1        1    acidity  0.1433  1.3185  0.1890
2        2      apple  0.2500  2.2993  0.5748
3        3     aromas  0.0000  1.7885  0.0000
4        4  astringent  0.0000  2.7047  0.0000
..     ...        ...     ...     ...     ...
120    120   unripened  0.0000  2.7047  0.0000
121    121      whiff  0.0000  2.7047  0.0000
122    122      white  0.0000  2.7047  0.0000
123    123       wine  0.0000  1.6061  0.0000
124    124     winter  0.0000  2.7047  0.0000

[125 rows x 5 columns]

           Document1 Document2 Document3 Document4 Document5 Document6 Document7 Document8 Document9 Document10
Document1         #    0.0168    0.0157    0.0589       0.0    0.0719    0.1282    0.0843    0.1472      0.069
Document2    0.0168         #    0.0462       0.0    0.0255    0.0145    0.1873    0.1365    0.0865     0.0818
Document3    0.0157    0.0462         #    0.0951    0.0238    0.1176    0.0173    0.0573    0.1095     0.0763
Document4    0.0589       0.0    0.0951         #       0.0    0.0594    0.0648       0.0    0.0489        0.0
Document5       0.0    0.0255    0.0238       0.0         #     0.038       0.0    0.0316    0.0244        0.0
Document6    0.0719    0.0145    0.1176    0.0594     0.038         #     0.063     0.018    0.0579     0.0418
Document7    0.1282    0.1873    0.0173    0.0648       0.0     0.063         #    0.0763    0.0866     0.0971
Document8    0.0843    0.1365    0.0573       0.0    0.0316     0.018    0.0763         #    0.0352     0.3124
Document9    0.1472    0.0865    0.1095    0.0489    0.0244    0.0579    0.0866    0.0352         #     0.0357
Document10     0.069    0.0818    0.0763       0.0       0.0    0.0418    0.0971    0.3124    0.0357          #

Se han registrado los resultados completos en el fichero: documents-01-result.xlsx
```

Ejecución documents-02.txt

```
Document 8:
     Index       Term      TF     IDF  TF_IDF
0        0     abound  0.0000  2.7047  0.0000
1        1     abrupt  0.0000  2.7047  0.0000
2        2   accented  0.0000  2.7047  0.0000
3        3    accents  0.0000  2.7047  0.0000
4        4    acidity  0.0000  2.7047  0.0000
..     ...        ...     ...     ...     ...
170    170     winery  0.0000  2.7047  0.0000
171    171       wood  0.1970  2.7047  0.5329
172    172  woodspice  0.1970  2.7047  0.5329
173    173      years  0.0000  2.7047  0.0000
174    174      zesty  0.0000  2.7047  0.0000

[175 rows x 5 columns]

Document 9:
     Index       Term      TF     IDF  TF_IDF
0        0     abound  0.0000  2.7047  0.0000
1        1     abrupt  0.0000  2.7047  0.0000
2        2   accented  0.0000  2.7047  0.0000
3        3    accents  0.0000  2.7047  0.0000
4        4    acidity  0.0000  2.7047  0.0000
..     ...        ...     ...     ...     ...
170    170     winery  0.0000  2.7047  0.0000
171    171       wood  0.0000  2.7047  0.0000
172    172  woodspice  0.0000  2.7047  0.0000
173    173      years  0.0000  2.7047  0.0000
174    174      zesty  0.0000  2.7047  0.0000

[175 rows x 5 columns]

Document 10:
     Index       Term      TF     IDF  TF_IDF
0        0     abound  0.0000  2.7047  0.0000
1        1     abrupt  0.0000  2.7047  0.0000
2        2   accented  0.0000  2.7047  0.0000
3        3    accents  0.0000  2.7047  0.0000
4        4    acidity  0.0000  2.7047  0.0000
..     ...        ...     ...     ...     ...
170    170     winery  0.0000  2.7047  0.0000
171    171       wood  0.0000  2.7047  0.0000
172    172  woodspice  0.0000  2.7047  0.0000
173    173      years  0.0000  2.7047  0.0000
174    174      zesty  0.0000  2.7047  0.0000

[175 rows x 5 columns]

           Document1 Document2 Document3 Document4 Document5 Document6 Document7 Document8 Document9 Document10
Document1         #    0.0764       0.0    0.0541    0.0473       0.0    0.0421       0.0    0.0265     0.0321
Document2    0.0764         #    0.0292    0.0285       0.0    0.0234    0.0978    0.0384    0.1453     0.0883
Document3       0.0    0.0292         #       0.0    0.0223     0.129    0.0902    0.1202    0.0553     0.0767
Document4    0.0541    0.0285       0.0         #    0.0352    0.0483    0.0313    0.0208    0.0257     0.0625
Document5    0.0473       0.0    0.0223    0.0352         #    0.0178       0.0       0.0    0.0695     0.0844
Document6       0.0    0.0234     0.129    0.0483    0.0178         #    0.1046    0.0612    0.0646     0.0613
Document7    0.0421    0.0978    0.0902    0.0313       0.0    0.1046         #    0.0454    0.0122     0.0148
Document8       0.0    0.0384    0.1202    0.0208       0.0    0.0612    0.0454         #    0.0743     0.0257
Document9    0.0265    0.1453    0.0553    0.0257    0.0695    0.0646    0.0122    0.0743         #     0.1746
Document10   0.0321    0.0883    0.0767    0.0625    0.0844    0.0613    0.0148    0.0257    0.1746          #

Se han registrado los resultados completos en el fichero: documents-02-result.xlsx
```

Ejecución documents-03.txt

```
Document 18:
     Index     Term      TF     IDF  TF_IDF
0        0  acidity  0.2900  2.6582  0.7709
1        1      add  0.0000  3.3514  0.0000
2        2   almond  0.0000  3.3514  0.0000
3        3    apple  0.0000  3.3514  0.0000
4        4    aroma  0.0000  3.3514  0.0000
..     ...      ...     ...     ...     ...
237    237    white  0.2900  2.6582  0.7709
238    238   widely  0.0000  3.3514  0.0000
239    239     wild  0.0000  3.3514  0.0000
240    240     wine  0.1701  1.5596  0.2654
241    241  zealand  0.0000  2.9459  0.0000

[242 rows x 5 columns]

Document 19:
     Index     Term      TF     IDF  TF_IDF
0        0  acidity  0.0000  2.6582  0.0000
1        1      add  0.1977  3.3514  0.6625
2        2   almond  0.0000  3.3514  0.0000
3        3    apple  0.0000  3.3514  0.0000
4        4    aroma  0.0000  3.3514  0.0000
..     ...      ...     ...     ...     ...
237    237    white  0.0000  2.6582  0.0000
238    238   widely  0.0000  3.3514  0.0000
239    239     wild  0.0000  3.3514  0.0000
240    240     wine  0.0920  1.5596  0.1435
241    241  zealand  0.1738  2.9459  0.5119

[242 rows x 5 columns]

Document 20:
     Index     Term      TF     IDF  TF_IDF
0        0  acidity  0.0000  2.6582  0.0000
1        1      add  0.0000  3.3514  0.0000
2        2   almond  0.0000  3.3514  0.0000
3        3    apple  0.0000  3.3514  0.0000
4        4    aroma  0.0000  3.3514  0.0000
..     ...      ...     ...     ...     ...
237    237    white  0.0000  2.6582  0.0000
238    238   widely  0.0000  3.3514  0.0000
239    239     wild  0.0000  3.3514  0.0000
240    240     wine  0.2716  1.5596  0.4237
241    241  zealand  0.0000  2.9459  0.0000

[242 rows x 5 columns]

           Document1 Document2 Document3 Document4 Document5 Document6 Document7 Document8  ... Document13 Document14 Document15 Document16 Document17 Document18 Document19 Document20
Document1          #    0.162       0.0       0.0    0.0215    0.0214       0.0    0.1001  ...     0.1755     0.0182        0.0     0.0612        0.0     0.0282     0.0322     0.3759
Document2      0.162        #       0.0       0.0     0.016     0.016       0.0    0.1376  ...      0.131     0.0788        0.0     0.0133     0.0818      0.021     0.0517     0.0336
Document3        0.0      0.0         #     0.023    0.0781    0.0326    0.1027    0.0485  ...     0.0594      0.016     0.0229      0.036        0.0     0.0275     0.0221     0.0382
Document4        0.0      0.0     0.023         #    0.0404       0.0    0.0178    0.1463  ...        0.0     0.1041     0.0282        0.0        0.0      0.034     0.0319        0.0
Document5     0.0215    0.016    0.0781    0.0404         #    0.0647    0.0165    0.0156  ...     0.1503     0.1899     0.0261     0.0199        0.0      0.186     0.1681        0.0
Document6     0.0214    0.016    0.0326       0.0    0.0647         #    0.0634    0.0156  ...     0.0785     0.0128     0.1187     0.0459        0.0     0.0481     0.0342     0.0384
Document7        0.0      0.0    0.1027    0.0178    0.0165    0.0634         #    0.0376  ...        0.0     0.0124     0.0904     0.0633        0.0     0.1026     0.0636     0.0296
Document8     0.1001   0.1376    0.0485    0.1463    0.0156    0.0156    0.0376         #  ...     0.0538     0.0299     0.0239      0.013     0.0532     0.0492     0.0465     0.0802
Document9     0.0508   0.0379    0.0757       0.0    0.0188    0.0188    0.1004    0.0734  ...     0.2285      0.016     0.0428     0.0156        0.0     0.0247     0.0839     0.1447
Document10    0.1678   0.1345    0.0467       0.0    0.0424    0.0697    0.0424    0.0426  ...     0.1832     0.0125      0.106     0.0499        0.0     0.0386     0.0761     0.1065
Document11       0.0      0.0    0.1739       0.0    0.0668    0.0596    0.0512       0.0  ...        0.0     0.0366     0.0286     0.0218     0.1023     0.0631        0.0        0.0
Document12    0.1419   0.0167    0.0599     0.027     0.025    0.1191    0.0361    0.0391  ...     0.0982     0.0187     0.1037     0.0731        0.0     0.0553     0.0124     0.1348
Document13    0.1755    0.131    0.0594       0.0    0.1503    0.0785       0.0    0.0538  ...          #     0.0233     0.0727     0.0227        0.0     0.1404     0.0816     0.2107
Document14    0.0182   0.0788     0.016    0.1041    0.1899    0.0128    0.0124    0.0299  ...     0.0233          #     0.0196        0.0     0.1009     0.0724     0.1132        0.0
Document15       0.0      0.0    0.0229    0.0282    0.0261    0.1187    0.0904    0.0239  ...     0.0727     0.0196          #     0.0558        0.0     0.0337     0.0271        0.0
Document16    0.0612   0.0133     0.036       0.0    0.0199    0.0459    0.0633     0.013  ...     0.0227        0.0     0.0558          #     0.0476     0.0836     0.0099     0.0648
Document17       0.0   0.0818       0.0       0.0       0.0       0.0       0.0    0.0532  ...        0.0     0.1009        0.0     0.0476          #        0.0     0.0278        0.0
Document18    0.0282    0.021    0.0275     0.034     0.186    0.0481    0.1026    0.0492  ...     0.1404     0.0724     0.0337     0.0836        0.0          #     0.0157     0.1026
Document19    0.0322   0.0517    0.0221    0.0319    0.1681    0.0342    0.0636    0.0465  ...     0.0816     0.1132     0.0271     0.0099     0.0278     0.0157          #      0.025
Document20    0.3759   0.0336    0.0382       0.0       0.0    0.0384    0.0296    0.0802  ...     0.2107        0.0        0.0     0.0648        0.0     0.1026      0.025          #

[20 rows x 20 columns]

Se han registrado los resultados completos en el fichero: documents-03-result.xlsx
```

Ejecución del Quijote: (El documento tenía un total de 500 documentos y más de 2 millones de líneas, así que no pudimos subirlo a Github por cuestión de peso. Se puede acceder a su output a través de este enlace: [Quijote-Output](#)).

Matriz de Similaridad de Coseno:

	Document1	Document2	Document3	Document4	Document5	Document6	Document7	Document8	Document9	Document10	Document11	Document12	Document13	Document14	Document15	Document16
Document1	-	0.2875	0.3534	0.0	0.3253	0.1505	0.4484	0.0772	0.3664	0.1682	0.2877	0.3379	0.1893	0.2585	0.2457	0.3096
Document2	0.2875	-	0.1682	0.0	0.2204	0.1084	0.3069	0.0	0.2157	0.0961	0.1992	0.2052	0.1129	0.1242	0.1524	0.2069
Document3	0.3534	0.1682	-	0.0	0.213	0.1428	0.2873	0.0	0.238	0.1114	0.1936	0.2233	0.1554	0.1467	0.1636	0.2152
Document4	0.0	0.0	0.0	-	0.0676	0.1424	0.048	0.0	0.0	0.0	0.0403	0.0712	0.0	0.0	0.0	0.0
Document5	0.3253	0.2204	0.213	0.0676	-	0.1086	0.3294	0.0	0.253	0.1475	0.2024	0.2397	0.1223	0.144	0.2003	0.2048
Document6	0.1505	0.1084	0.1428	0.1424	0.1086	-	0.2372	0.0	0.1824	0.0571	0.2047	0.1685	0.063	0.0822	0.0986	0.1276
Document7	0.4484	0.3069	0.2873	0.048	0.3294	0.2372	-	0.0248	0.4305	0.1902	0.3503	0.3494	0.2144	0.2475	0.2508	0.332
Document8	0.0772	0.0	0.0	0.0	0.0	0.0	0.0248	-	0.0	0.0	0.0251	0.0	0.0	0.0	0.0	0.0
Document9	0.3664	0.2157	0.238	0.0	0.253	0.1824	0.4305	0.0	-	0.1121	0.2401	0.2751	0.1252	0.1798	0.1803	0.243
Document10	0.1682	0.0961	0.1114	0.0	0.1475	0.0571	0.1902	0.0	0.1121	-	0.1109	0.1913	0.0626	0.1654	0.0953	0.124
Document11	0.2877	0.1992	0.1936	0.0403	0.2024	0.2047	0.3503	0.0251	0.2401	0.1109	-	0.2569	0.104	0.1592	0.2027	0.2519
Document12	0.3379	0.2052	0.2233	0.0712	0.2397	0.1685	0.3494	0.0	0.2751	0.1913	0.2569	-	0.1726	0.1833	0.2084	0.2166
Document13	0.1893	0.1129	0.1554	0.0	0.1223	0.063	0.2144	0.0	0.1252	0.0626	0.104	0.1726	-	0.1035	0.1221	0.1403
Document14	0.2585	0.1242	0.1467	0.0	0.144	0.0822	0.2475	0.0	0.1798	0.1654	0.1592	0.1833	0.1035	-	0.1362	0.1663
Document15	0.2457	0.1524	0.1636	0.0	0.2003	0.0986	0.2508	0.0	0.1803	0.0953	0.2027	0.2084	0.1221	0.1362	-	0.314
Document16	0.3096	0.2069	0.2152	0.0	0.2048	0.1276	0.332	0.0	0.243	0.124	0.2519	0.2166	0.1403	0.1663	0.314	-
Document17	0.2283	0.17	0.1332	0.0	0.1576	0.0957	0.2148	0.0	0.187	0.0675	0.2088	0.1598	0.1025	0.0828	0.166	0.2295
Document18	0.2557	0.1632	0.2105	0.0	0.182	0.1156	0.2697	0.0	0.2058	0.0975	0.1552	0.2001	0.1131	0.1233	0.1617	0.1837
Document19	0.1568	0.1181	0.1284	0.0501	0.1111	0.0813	0.2326	0.0	0.1511	0.0593	0.1862	0.1052	0.0581	0.0977	0.1597	0.2347
Document20	0.1201	0.105	0.0912	0.0	0.1066	0.0582	0.1366	0.0	0.1082	0.0584	0.1062	0.1077	0.1054	0.0375	0.0445	0.1882
Document21	0.2713	0.1488	0.1696	0.0	0.1716	0.1265	0.3202	0.0	0.2452	0.1126	0.1854	0.2346	0.1194	0.1653	0.2066	0.2478
Document22	0.2952	0.1869	0.1794	0.0	0.1854	0.1284	0.3296	0.0	0.2397	0.1086	0.2693	0.2166	0.13	0.1428	0.1967	0.2543
Document23	0.2162	0.1635	0.1335	0.0	0.1688	0.0864	0.2387	0.0	0.1577	0.1647	0.1769	0.1791	0.0813	0.1358	0.2189	0.2663
Document24	0.0693	0.1227	0.0676	0.0	0.0477	0.0445	0.1113	0.0	0.0761	0.048	0.076	0.0718	0.0209	0.0434	0.0398	0.0934
Document25	0.3534	0.2092	0.2053	0.0418	0.2206	0.1837	0.333	0.0	0.287	0.1204	0.2674	0.2523	0.1199	0.1835	0.2168	0.2477
Document26	0.3829	0.2199	0.2659	0.0594	0.2824	0.1892	0.3641	0.0	0.2717	0.1386	0.2666	0.2613	0.1456	0.1914	0.2647	0.2896
Document27	0.2962	0.182	0.2043	0.0	0.1752	0.1254	0.2993	0.0	0.2141	0.1175	0.1963	0.2186	0.138	0.201	0.1919	0.231
Document28	0.3202	0.1849	0.225	0.0	0.1801	0.1423	0.3092	0.0	0.2436	0.1179	0.2509	0.2244	0.1453	0.1812	0.195	0.2625
Document29	0.3107	0.1943	0.1987	0.0	0.2535	0.1199	0.3045	0.0	0.2362	0.1101	0.2595	0.24	0.1414	0.136	0.223	0.2377
Document30	0.3523	0.2122	0.2231	0.0225	0.236	0.2021	0.3748	0.0	0.2907	0.1651	0.3368	0.2873	0.1513	0.1937	0.2541	0.3173
Document31	0.2349	0.1373	0.1614	0.0	0.1787	0.0926	0.235	0.0	0.1827	0.0922	0.1979	0.1743	0.1175	0.1567	0.1449	0.1718

Document464	0.187	0.1435	0.1162	0.0	0.1162	0.0592	0.1739	0.0	0.1253	0.0808	0.1295	0.1231	0.0636	0.0957	0.1063	0.174
Document465	0.0	0.0268	0.0	0.0	0.0	0.0242	0.0301	0.0	0.0	0.0	0.0	0.0178	0.0	0.0	0.0213	0.011
Document466	0.0682	0.0298	0.0336	0.0	0.0319	0.0261	0.0597	0.0	0.0621	0.0313	0.0517	0.0483	0.0189	0.1153	0.0424	0.0692
Document467	0.2087	0.1392	0.0992	0.0	0.1622	0.1172	0.1915	0.0	0.1519	0.0599	0.1396	0.1176	0.0635	0.1117	0.0822	0.1081
Document468	0.2811	0.1549	0.17	0.0292	0.1783	0.1143	0.2659	0.0	0.1978	0.0969	0.1961	0.1942	0.1165	0.1679	0.1692	0.1966
Document469	0.1753	0.1118	0.1167	0.0	0.1377	0.06	0.1956	0.0	0.1264	0.0799	0.1498	0.1386	0.0608	0.1061	0.1125	0.1281
Document470	0.1748	0.0873	0.1295	0.0794	0.0898	0.072	0.1541	0.0	0.1435	0.0547	0.159	0.1309	0.0495	0.0857	0.1025	0.109
Document471	0.2456	0.1371	0.167	0.0	0.1779	0.0996	0.2381	0.0	0.193	0.0834	0.1397	0.1546	0.0907	0.1034	0.142	0.1611
Document472	0.053	0.045	0.0312	0.0	0.0287	0.0194	0.0917	0.0	0.0415	0.0518	0.0466	0.0514	0.0141	0.0373	0.0666	0.0752
Document473	0.203	0.1249	0.1373	0.0	0.1248	0.0631	0.1949	0.0	0.1761	0.0956	0.1289	0.1477	0.1088	0.1116	0.1122	0.1408
Document474	0.1189	0.0914	0.0783	0.0	0.0904	0.0618	0.1166	0.0	0.0887	0.0506	0.0678	0.1034	0.0696	0.058	0.094	0.1069
Document475	0.2676	0.1898	0.1912	0.0	0.1816	0.1052	0.3083	0.0	0.2162	0.1137	0.188	0.2292	0.1555	0.1749	0.1972	0.2503
Document476	0.0561	0.0385	0.0267	0.0	0.0246	0.0167	0.0521	0.0	0.0277	0.02	0.0399	0.0328	0.0121	0.0736	0.035	0.0403
Document477	0.1837	0.1167	0.1097	0.0	0.1149	0.0853	0.2598	0.0	0.1843	0.0684	0.1199	0.1561	0.0659	0.1066	0.1252	0.1504
Document478	0.0579	0.027	0.0528	0.0	0.0232	0.0039	0.0434	0.0	0.0457	0.0	0.0398	0.0319	0.0038	0.0043	0.0218	0.0333
Document479	0.1602	0.1164	0.1227	0.0	0.1173	0.0624	0.1831	0.0	0.1338	0.0649	0.143	0.1183	0.0665	0.0983	0.115	0.1254
Document480	0.0935	0.0904	0.0589	0.0	0.0642	0.0276	0.0881	0.0	0.074	0.0378	0.0603	0.0703	0.0303	0.0476	0.0537	0.0829
Document481	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0129	0.0	0.0	0.0485	0.0279
Document482	0.1137	0.094	0.0769	0.0	0.0777	0.0685	0.1399	0.0	0.0964	0.0509	0.1042	0.1084	0.0637	0.0745	0.0908	0.1148
Document483	0.1221	0.0457	0.0448	0.2174	0.0673	0.031	0.0894	0.0	0.0412	0.0372	0.0678	0.0816	0.0224	0.1586	0.0577	0.06
Document484	0.0104	0.0169	0.016	0.0	0.0065	0.0093	0.0159	0.0	0.0247	0.0	0.021	0.0068	0.009	0.0102	0.0181	0.0139
Document485	0.0081	0.0263	0.0	0.0	0.05	0.0	0.0756	0.0	0.0776	0.0764	0.0405	0.038	0.0	0.0	0.0	0.0572
Document486	0.205	0.126	0.127	0.0	0.1504	0.0731	0.2271	0.0	0.1497	0.1559	0.1558	0.135	0.1152	0.1912	0.13	0.1699
Document487	0.1522	0.1028	0.0853	0.0	0.0813	0.0305	0.139	0.0	0.118	0.0842	0.0521	0.0905	0.0444	0.0419	0.0497	0.1115
Document488	0.1105	0.038	0.0373	0.0	0.0348	0.0258	0.0908	0.0	0.0805	0.0309	0.0688	0.0854	0.0187	0.1319	0.0604	0.0642
Document489	0.007	0.0228	0.0	0.0	0.0	0.0	0.0301	0.0	0.0123	0.0663	0.0	0.0102	0.0	0.0	0.0	0.0497
Document490	0.2756	0.169	0.1729	0.0	0.1779	0.0732	0.2236	0.0	0.2146	0.0908	0.1368	0.1943	0.0764	0.1468	0.1284	0.1348
Document491	0.0046	0.0149	0.0	0.0	0.0	0.0	0.0196	0.0	0.008	0.0432	0.0	0.0066	0.0	0.0	0.0	0.0323
Document492	0.0	0.0	0.0	0.0	0.016	0.0	0.0131	0.0	0.0204	0.0	0.013	0.0084	0.0	0.0	0.0	0.0
Document493	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Document494	0.1811	0.1204	0.1506	0.0694	0.1216	0.1236	0.2094	0.0	0.1351	0.0804	0.1792	0.1573	0.0645	0.1544	0.181	0.1942
Document495	0.11	0.0913	0.0796	0.0	0.092	0.0634	0.1288	0.0	0.1102	0.0454	0.1033	0.0754	0.0401	0.0553	0.0719	0.1095

503707	4732	retorno	0.0000	6.5134	0.0000
503708	4733	retraer	0.0000	6.5134	0.0000
503709	4734	retratar	0.0000	6.5134	0.0000
503710	4735	reuma	0.0000	6.5134	0.0000
503711	4736	revalidar	0.0000	6.5134	0.0000
503712	4737	reventaba	0.0000	6.5134	0.0000
503713	4738	reventar	0.0000	6.1080	0.0000
503714	4739	reverencia	0.0000	6.5134	0.0000
503715	4740	reveses	0.0000	6.5134	0.0000
503716	4741	reviente	0.0000	6.5134	0.0000
503717	4742	revueltas	0.0000	6.5134	0.0000
503718	4743	revueltos	0.0000	6.5134	0.0000
503719	4744	rey	0.0000	4.2621	0.0000
503720	4745	reyes	0.0000	5.8203	0.0000
503721	4746	rezaba	0.0000	6.5134	0.0000
503722	4747	rezando	0.0000	6.5134	0.0000
503723	4748	riberas	0.0000	6.5134	0.0000
503724	4749	rica	0.0000	6.1080	0.0000
503725	4750	ricamente	0.0000	6.5134	0.0000
503726	4751	ricamonte	0.0000	6.5134	0.0000
503727	4752	ricas	0.0000	6.5134	0.0000
503728	4753	rico	0.0000	5.0094	0.0000
503729	4754	ricos	0.0000	5.5971	0.0000
503730	4755	rienda	0.0000	5.5971	0.0000
503731	4756	riendas	0.0418	5.0094	0.2095
503732	4757	riendo	0.0000	6.1080	0.0000
503733	4758	riera	0.0000	6.5134	0.0000
503734	4759	riesgo	0.0000	6.5134	0.0000
503735	4760	rige	0.0000	6.5134	0.0000
503736	4761	rigor	0.0000	5.2607	0.0000
503737	4762	rigurosa	0.0000	5.5971	0.0000
503738	4763	rigurosamente	0.0000	6.5134	0.0000
503739	4764	riguroso	0.0000	6.1080	0.0000
503740	4765	rijoso	0.0000	6.5134	0.0000

Lo que podemos apreciar en esta captura es que los valores de TF son casi todos 0, y es lógico, hay tantísimos documentos que la mayoría de palabras no aparecerán en ese documento concreto, mientras que IDF son valores más altos puesto que hay tantas palabras diferentes y documentos que es normal que el ratio de aparición de las palabras sea muy bajo, puesto que cuanto menos se acerca IDF al 1 menos común es el término.

Conclusiones

En primer lugar, se ha notado mucho la diferencia entre procesar los documentos del Github y “El Quijote” debido principalmente a la cantidad de documentos de cada uno de ellos, mientras que los documentos apenas tardaron unos pocos segundos en procesar todo, con el libro pasaron unos 20 o 30 minutos. Por otro lado, evidentemente los tamaños de las matrices son muy diferentes.

Esta práctica nos ha servido para entender la utilidad que tienen estos sistemas para detectar y visualizar claramente, en este caso, el número de veces que se repetía un término en cada documento de manera proporcional, entendiendo el peso que cada uno tenía, tanto en un mismo documento como en el texto entero.

Aplicándolo en otro ámbito, es útil para entender que esto podría aplicarse en otros aspectos, por ejemplo, a la hora de recomendar cosas en diferentes áreas, si sabemos cómo de frecuente es un término y el peso que tiene, podemos extraer mucha información raíz de estos datos.

Integrantes:

JONATHAN MARTÍNEZ PÉREZ - alu0101254098@ull.edu.es

EDUARDO GONZÁLEZ PÉREZ - alu0101319001@ull.edu.es