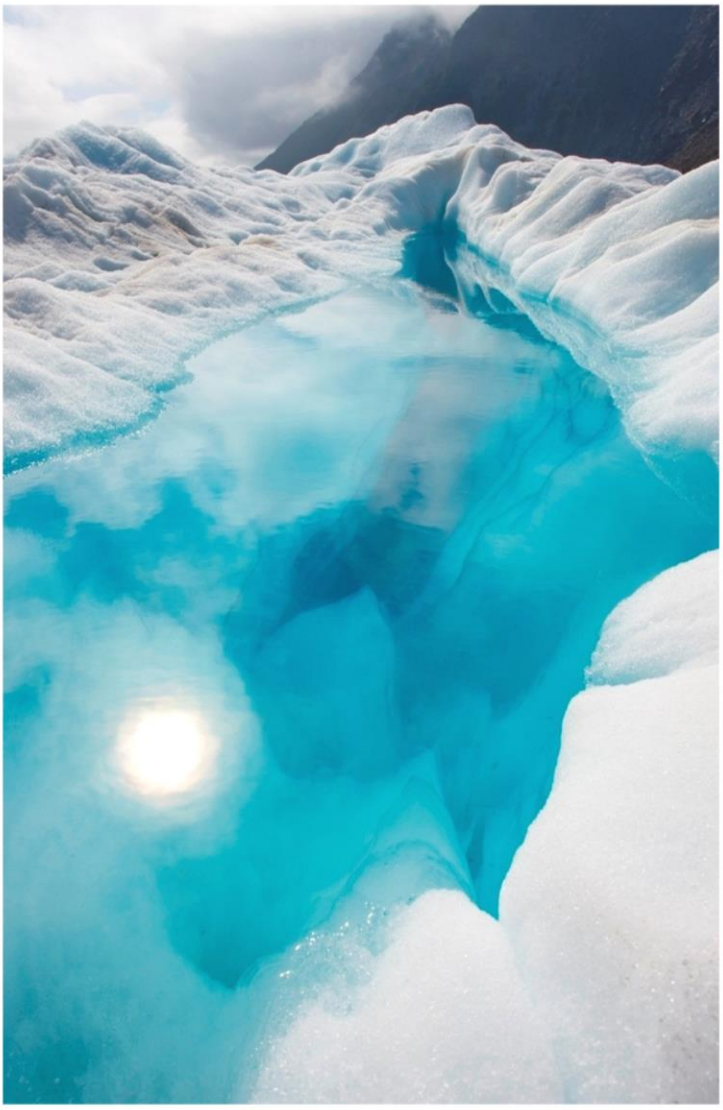

Vocabulario

LOS MISERABLES

Gerardo Gael García Camacho | Diplomado IA | 14 de marzo de 2025

Transformación a .csv

Al inicio no sabia como transformar un archivo .pdf a uno .csv, por lo que tuve que buscar diversas formas de lograrlo ya sea en Go o en la terminal, al final logre convertir primero de .pdf a .txt y luego de .txt a .csv, los código usados fueron los siguientes:

pdftotext -layout LM.pdf Lm.txt

sed 's/ \+/,/g' LM.txt > LM.csv

Con esto estábamos listos para iniciar el procesado de las palabras.

Retomando movielens

Ya una vez con el archivo en formato .csv era suficiente con recordar lo que se había estado usando en el código de movielens, tanto de apertura, lectura y guardado, prácticamente era reciclar código de la práctica y modificarlo para las necesidades actuales, como especificar al lector que él .csv no tiene la misma longitud en todos sus renglones o que en algunos renglones estaba vacío.

Igualmente usamos un mapa para guardar los datos que se iban a estar leyendo línea por línea (por ser más eficientes).

Limpieza del caracteres

Para esta parte usamos la paquetería de regular expresions de Go, para quitar caracteres no deseados en el archivo, como signos de interrogación, comas, guiones, comillas, puntos, espacios en blancos, acentos, etc. Después de esta limpieza ya lo agregábamos a nuestra base depurada.

Conteo

Mientras se hacia el proceso de limpieza as u vez llevábamos un contador de las palabras totales que pasaban la limpieza y un contador de cada palabra única en un mapa, para después pasarlas a un slice de duplas de un entero y una cadena, y así lograr ordenarlas de manera más fácil.

Resultados

Recordemos que las palabras se limpiaron para aparecer en minúsculas y sin acentos. Y realmente hay muchas más palabras más con solo una aparición en el texto.

I. Total de Palabras

```
Numero TOTAL de palabras en el texto: 151762
```

II. Palabras diferentes

```
Numero de palabras y numeros distintas en el texto: 13602
```

III. 100 palabras que más se repiten

1) de: 6587
2) la: 5903
3) que: 4990
4) y: 4644
5) el: 4633
6) a: 3982
7) en: 3598
8) se: 2862
9) no: 2167
10) un: 2093
11) su: 1706
12) los: 1705
13) una: 1668
14) lo: 1514
15) con: 1336
16) por: 1214
17) al: 1203
18) del: 1084
19) habia: 1038
20) las: 1036
21) es: 838
22) era: 824
23) marius: 785
24) le: 768
25) dijo: 689
26) como: 688
27) mas: 664
28) para: 664
29) pero: 647
30) senor: 630
31) jean: 597
32) valjean: 573
33) si: 547
34) cosette: 545
35) me: 531
36) esta: 523
37) sin: 507
38) hombre: 499
39) sus: 461
40) estaba: 456
41) todo: 398
42) mi: 395
43) ya: 363
44) cuando: 361
45) dos: 357
46) thenardier: 355
47) os: 347
48) tenia: 343
49) hacia: 316
50) yo: 289
51) despues: 285
52) este: 276
53) fue: 264
54) javert: 252
55) casa: 246
56) muy: 244
57) padre: 239
58) aquel: 237
59) ella: 237
60) sobre: 234
61) ese: 230
62) donde: 228
63) nada: 224
64) vez: 220
65) ni: 220
66) ha: 217
67) quien: 216
68) voz: 215
69) dia: 215
70) calle: 214
71) momento: 213
72) mismo: 210
73) bien: 210
74) ser: 203
75) solo: 202
76) noche: 201
77) mano: 197
78) tiempo: 197
79) puerta: 193
80) hay: 193
81) volvio: 182

82) esto: 180
83) ojos: 179
84) esa: 177
85) tan: 177
86) todos: 176
87) hasta: 176
88) poco: 174
89) ahora: 171
90) aqui: 170
91) magdalena: 170
92) 0: 169
93) alli: 166
94) anos: 165
95) uno: 164
96) aquella: 161
97) cabeza: 159
98) francos: 158
99) porque: 157
100) mujer: 155

IV. 100 palabras que menos se repiten

1) pusieras: 1
2) endeble: 1
3) enojado: 1
4) transcribimos: 1
5) adulto: 1
6) resumen: 1
7) traza: 1
8) renuncia: 1
9) manifesto: 1
10) causados: 1
11) ahorro: 1
12) muchachito: 1
13) soltura: 1
14) envueltos: 1
15) piensas: 1
16) manada: 1
17) figuro: 1
18) ailly: 1
19) oprime: 1
20) morland: 1
21) contesta: 1
22) tremendamente: 1
23) asaltantes: 1
24) espadin: 1
25) aturde: 1
26) silenciosa: 1
27) cambiaremos: 1
28) grunon: 1
29) dciles: 1
30) divisa: 1
31) chaume: 1
32) infames: 1
33) escapatorias: 1
34) coronada: 1
35) resoplar: 1
36) nacia: 1
37) destinadas: 1
38) aix: 1
39) rechazaban: 1
40) pillo: 1
41) sobresaltada: 1
42) ganais: 1
43) rayaba: 1
44) imprevision: 1
45) estupefactos: 1
46) pirueta: 1
47) docenas: 1
48) vena: 1
49) armarlo: 1
50) deberian: 1
51) diversa: 1
52) apretadas: 1
53) pagados: 1
54) prevista: 1
55) rompa: 1
56) prolongar: 1
57) templaba: 1
58) absorbido: 1
59) desconfiaban: 1
60) bergante: 1
61) explicaros: 1
62) zapatilla: 1
63) penara: 1
64) creaciones: 1
65) moribundos: 1
66) recojan: 1
67) guarida: 1
68) deshojado: 1
69) atormenta: 1
70) disculpara: 1
71) sexta: 1
72) furtiva: 1
73) rina: 1
74) ocultais: 1
75) alojamiento: 1
76) movido: 1
77) obedecemos: 1
78) disculpa: 1
79) pisandome: 1
80) repartidas: 1
81) conservar: 1
82) ahogara: 1
83) desparramo: 1
84) presentad: 1
85) paradero: 1
86) sorprendia: 1
87) denotaba: 1
88) chasco: 1
89) espesor: 1

90) precios: 1
91) dislocado: 1
92) amenazaba: 1
93) guantes: 1
94) escuchame: 1
95) innegables: 1
96) maravilla: 1
97) relatara: 1
98) cheminvert: 1
99) burguesa: 1
100) cortaplumas: 1

Experiencia

Me fue divertido debido a que me recordó mucho al ejercicio de movielens el cual me gusto mucho y me sirio para seguir practicando Go, ya que lo habíamos abandonado un poco después de empezar en esta parte de Python.