Cerca de Documents Similars – Jocs de Prova

Per tots els casos de prova d'aquest document utilitzem un seguit de fitxers de novel·les obtinguts del Projecte Gutenberg, que es llisten a continuació a mode de referència. En tots els casos suposem que ja són carregats al sistema.

Title	Author	Creation Date	Modification Date	File Name
A Manual of Moral Philosophy	Andrew Preston Peabody	2022-12-22T21:02:28.976991	2022-12-22T21:02:29.396401	27531-0.txt
On the Origin of Species 1st Edition	Charles Darwin	2022-12-22T21:02:29.416471	2022-12-22T21:02:30.588579	DarwinOriginofSpecies.txt
A Christmas Carol	Charles Dickens	2022-12-22T21:02:30.613423	2022-12-22T21:02:30.867221	DickensAChristmasCarol.txt
Great Expectations	Charles Dickens	2022-12-22T21:02:30.897580	2022-12-22T21:02:32.138156	DickensGreatExpectations.txt
The Pickwick Papers	Charles Dickens	2022-12-22T21:02:32.172176	2022-12-22T21:02:34.361306	DickensThePickwickPapers.txt
The Jungle Book	Rudyard Kipling	2022-12-22T21:02:34.394443	2022-12-22T21:02:34.765512	KiplingJungleBook.txt
The Call of the Wild	Jack London	2022-12-22T21:02:34.802677	2022-12-22T21:02:35.044558	LondonCallofTheWild.txt
The Commission in Lunacy	Honore de Balzac	2022-12-22T21:02:35.495886	2022-12-22T21:02:35.712420	pg1410.txt
New York	James Fenimore Cooper	2022-12-22T21:02:38.663236	2022-12-22T21:02:38.767873	pg2482.txt
Moonbeams From the Larger Lunacy	Stephen Leacock	2022-12-22T21:02:44.734394	2022-12-22T21:02:45.102197	pg4064.txt
Giordano Bruno	Walter Horatio Pater	2022-12-22T21:02:45.132486	2022-12-22T21:02:45.213523	pg4228.txt
The Sea Fairies	L. Frank Baum	2022-12-22T21:02:46.726128	2022-12-22T21:02:47.020553	pg4358.txt
The Memoirs of General W. T. Sherm...	William T. Sherman	2022-12-22T21:02:47.049043	2022-12-22T21:02:49.568337	pg4361.txt
The Vampyre; A Tale	John William Polidori	2022-12-22T21:02:49.596291	2022-12-22T21:02:49.702780	pg6087.txt
Barford Abbey	Susannah Minific Gunning	2022-12-22T21:02:35.073007	2022-12-22T21:02:35.478418	pg13314.txt
A History of Trade Unionism in the U...	Selig Perlman	2022-12-22T21:02:35.747556	2022-12-22T21:02:36.392826	pg14458.txt
History Of Egypt, ChaldÃ¦a, Syria, Bab...	G. Maspero	2022-12-22T21:02:36.418815	2022-12-22T21:02:37.430122	pg17326.txt
Traditions of the North American Indi...	James Athearn Jones	2022-12-22T21:02:37.460644	2022-12-22T21:02:38.060971	pg20826.txt
The Song of the Exile--A Canadian Epic	Wilfred S. Skeats	2022-12-22T21:02:38.079250	2022-12-22T21:02:38.237201	pg20939.txt
The Emancipation Proclamation	Abraham Lincoln	2022-12-22T21:02:38.282029	2022-12-22T21:02:38.313296	pg22082.txt
Dorothy and the Wizard in Oz	L. Frank Baum	2022-12-22T21:02:38.348663	2022-12-22T21:02:38.643208	pg22566.txt
When Ghost Meets Ghost	William Frend De Morgan	2022-12-22T21:02:38.785765	2022-12-22T21:02:41.305709	pg30896.txt
The English Novel in the Time of Sha...	J. J. Jusserand	2022-12-22T21:02:41.323281	2022-12-22T21:02:42.074115	pg31151.txt
Logic, Inductive and Deductive	William Minto	2022-12-22T21:02:42.098500	2022-12-22T21:02:42.782963	pg31796.txt
Great Musical Composers: German, F...	George T. Ferris	2022-12-22T21:02:42.815283	2022-12-22T21:02:43.763038	pg34381.txt
Strange Stories from the Lodge of Lei...	Unknown	2022-12-22T21:02:43.789120	2022-12-22T21:02:44.018135	pg37766.txt
The Code of the Mountains	Charles Neville Buck	2022-12-22T21:02:44.055062	2022-12-22T21:02:44.706949	pg38498.txt
The Bible Of Bibles - Or Twenty-Seve...	Kersey Graves	2022-12-22T21:02:45.265025	2022-12-22T21:02:46.700855	pg43550.txt
The Works of Edgar Allan Poe - Volu...	Edgar Allan Poe	2022-12-22T21:02:49.729225	2022-12-22T21:02:50.371486	PoeWorksVol1.txt
The Works of Edgar Allan Poe - Volu...	Edgar Allan Poe	2022-12-22T21:02:50.400588	2022-12-22T21:02:51.068365	PoeWorksVol2.txt
A Princess of Mars	Edgar Rice Burroughs	2022-12-22T21:02:51.102132	2022-12-22T21:02:51.616355	RiceBurroughsAprincessofMars.txt
The Time Machine	H. G. Wells	2022-12-22T21:02:51.655745	2022-12-22T21:02:51.951742	WellsTimeMachine.txt
The War of the Worlds	H. G. Wells	2022-12-22T21:02:51.994789	2022-12-22T21:02:52.450337	WellsWarofTheWorlds.txt

Cas Estàndard

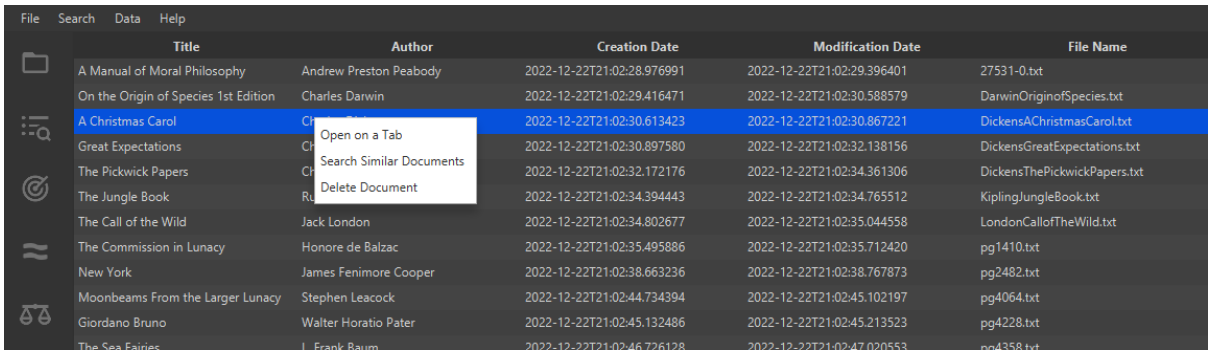
Objectius

Volem comprovar que els resultats obtinguts en cercar documents similars són raonables – és clar, es tracta d'una cerca aproximada, pel que ens conformem amb resultats que "tinguin sentit".

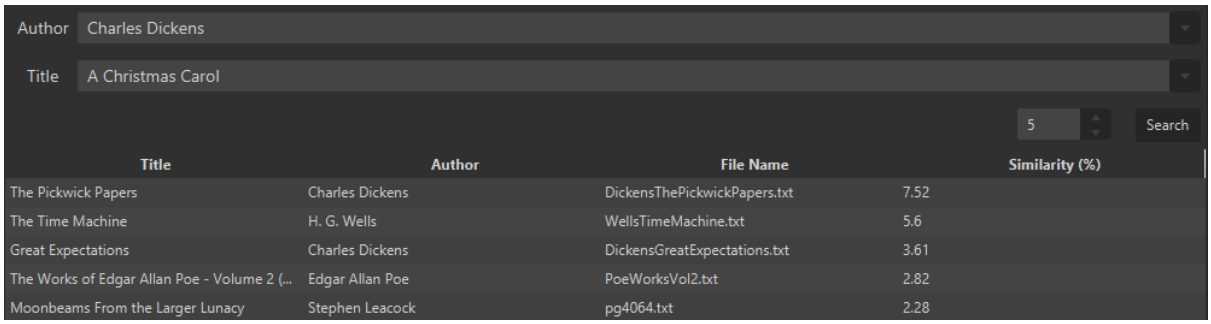
És esperable que novel·les escrites per un mateix autor s'assemblin més entre si que respecte a d'altres autors. Aprofitem que hi ha autors amb múltiples novel·les dins del corpus per a comprovar-ho.

Passos a seguir

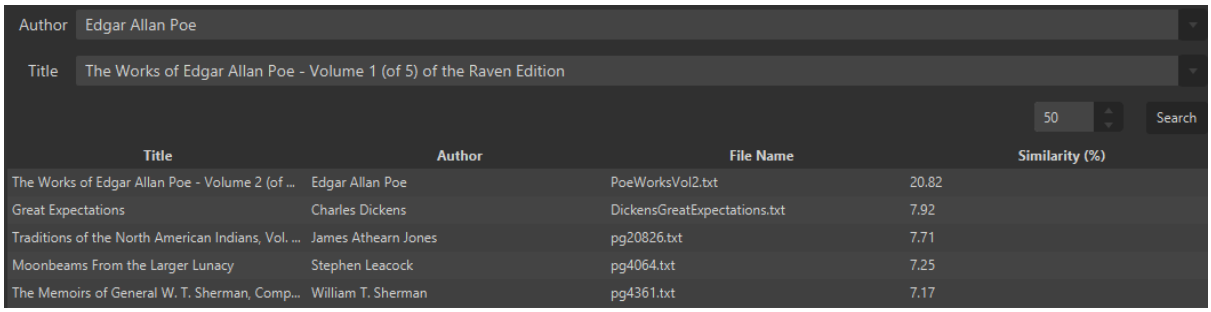
Seleccionem la novel·la “A Christmas Carol” de Charles Dickens. Click dret i premem “Search Similar Documents”.



Immediatament, s’obre la pestanya de cerques per semblança amb l’autor i títols fixats, i els resultats obtinguts: veiem que la novel·la més semblant és “The Pickwick Papers” (7.52% de semblança) i, en tercer lloc, la darrera novel·la d’aquest autor (3.61%).



Repetim el procés per “The Works of Edgar Allan Poe - Volume 1 (of 5) of the Raven Edition”, i obtenim el següent. Aquest cop el resultat és més il·lustratiu encara: l’altra document d’Edgar Allan Poe hi té un 20.82% de semblança – certament un valor molt elevat.



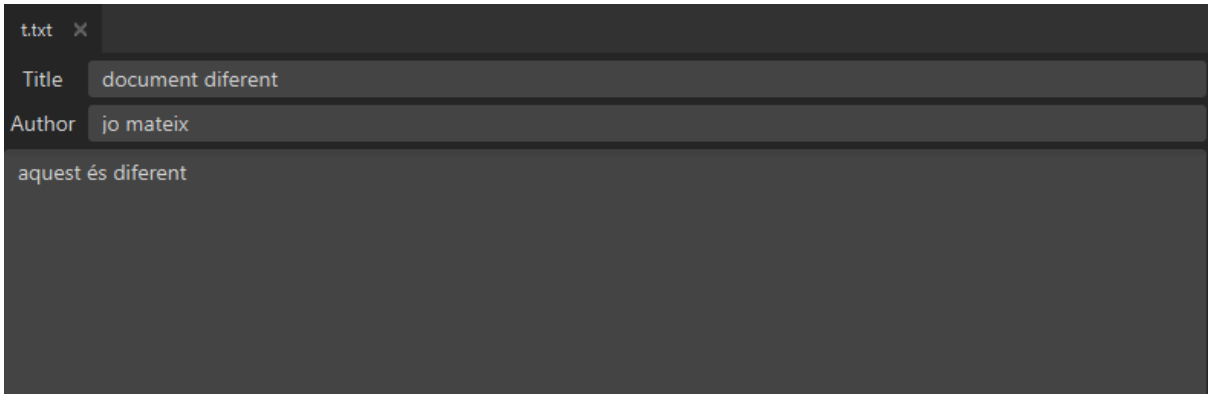
Cas Extrem – Document Completament Diferent

Objectius

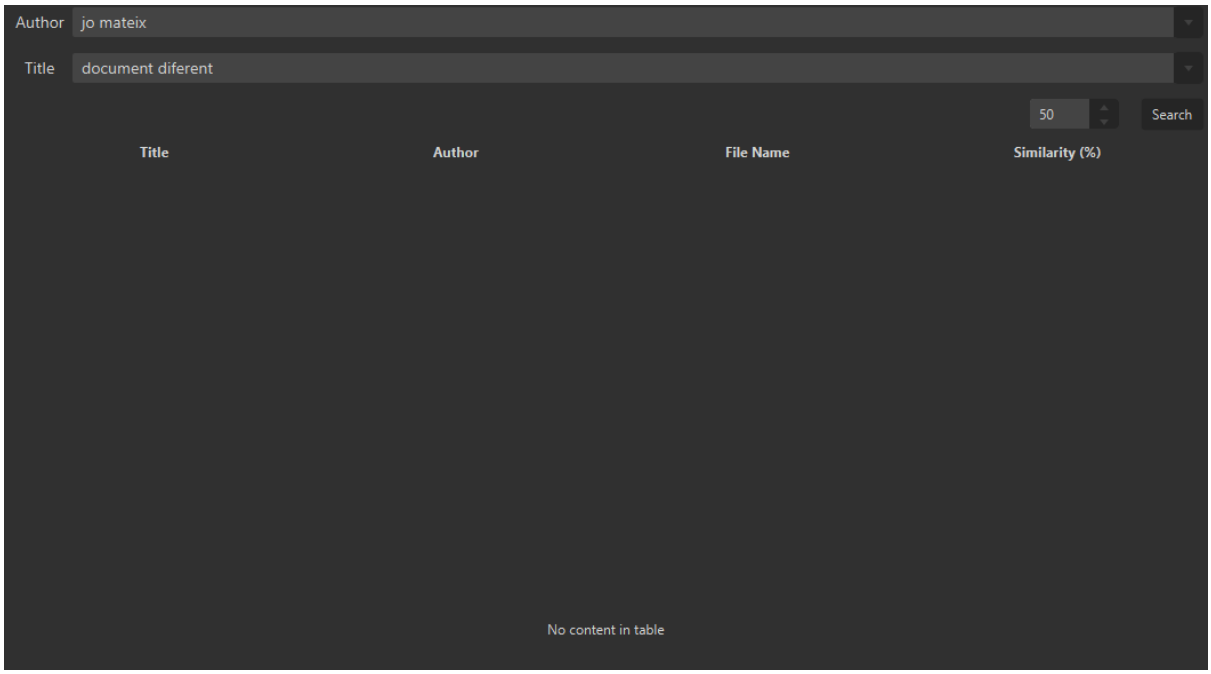
Comprovem que cercar els documents semblants a un que conté només paraules úniques (que només apareixen en aquest) no retorna cap resultat.

Passos a seguir

Creem el següent document, que estem segurs no contindrà paraules que ja existien doncs totes les novel·les són en anglès.



Executem el mateix procés per cercar-ne els similars: en efecte, no obtenim cap resultat.



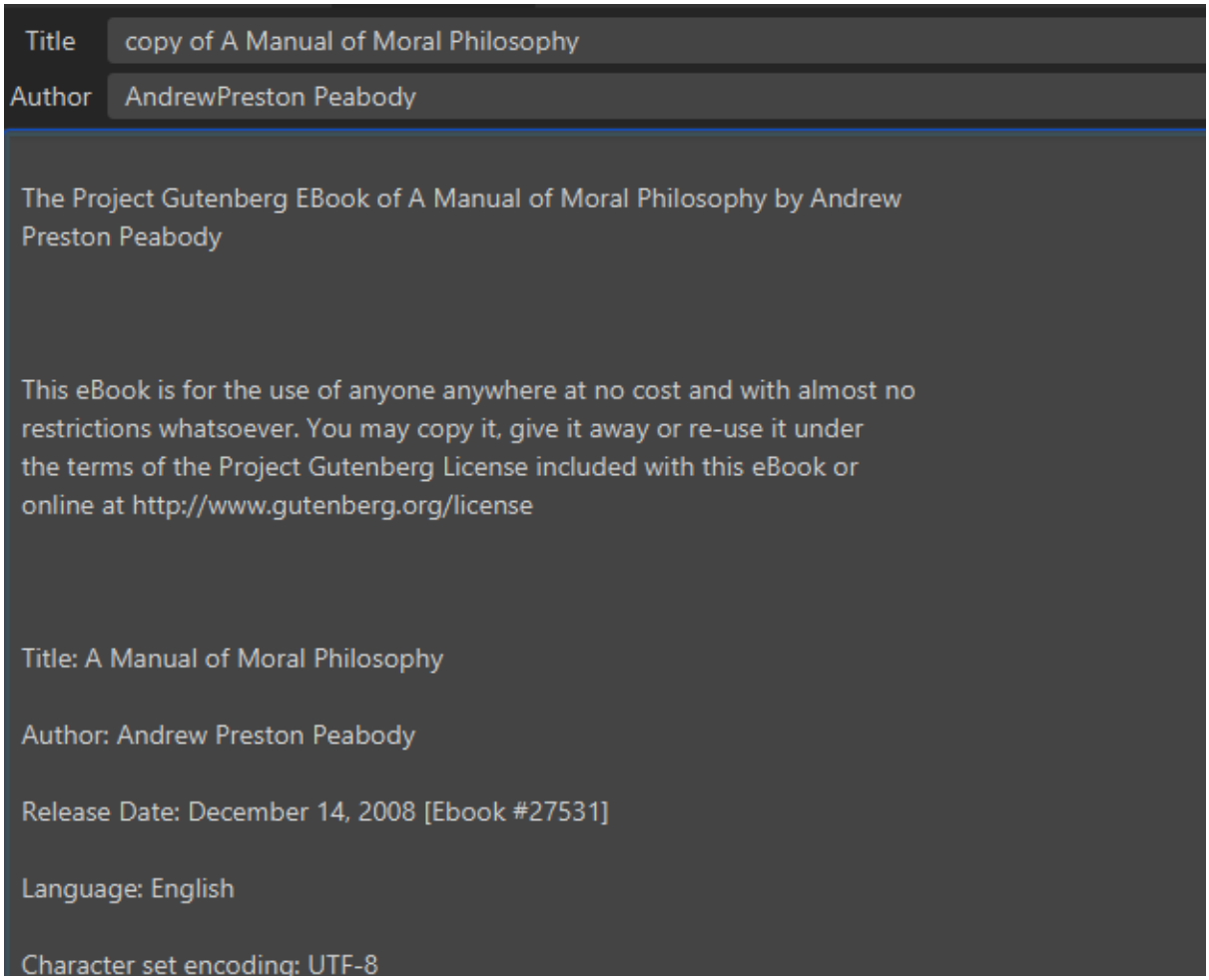
Documents Iguals

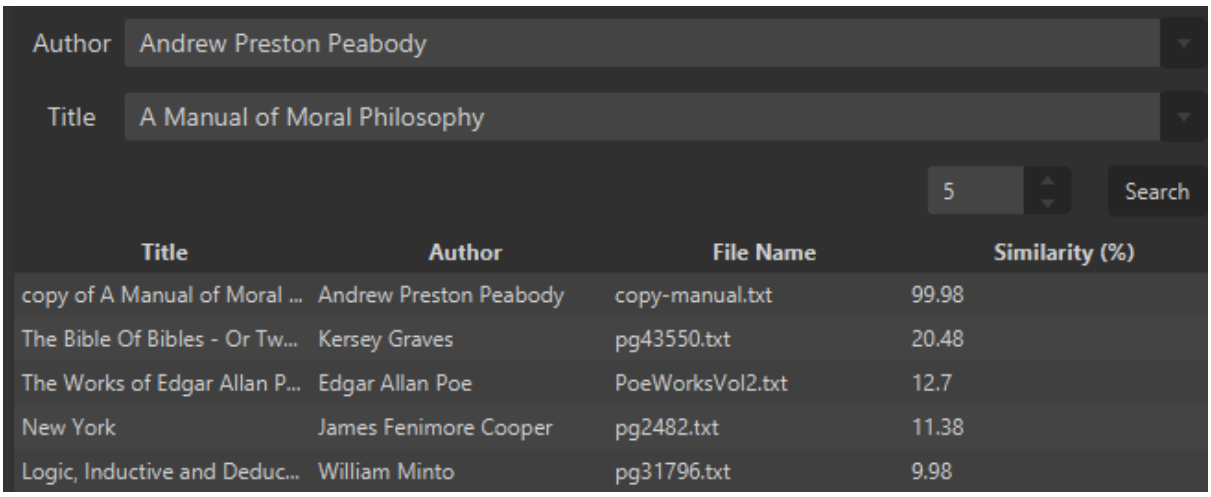
Objectius

Volem comprovar que, en efecte, documents iguals retornen una semblança del 100%.

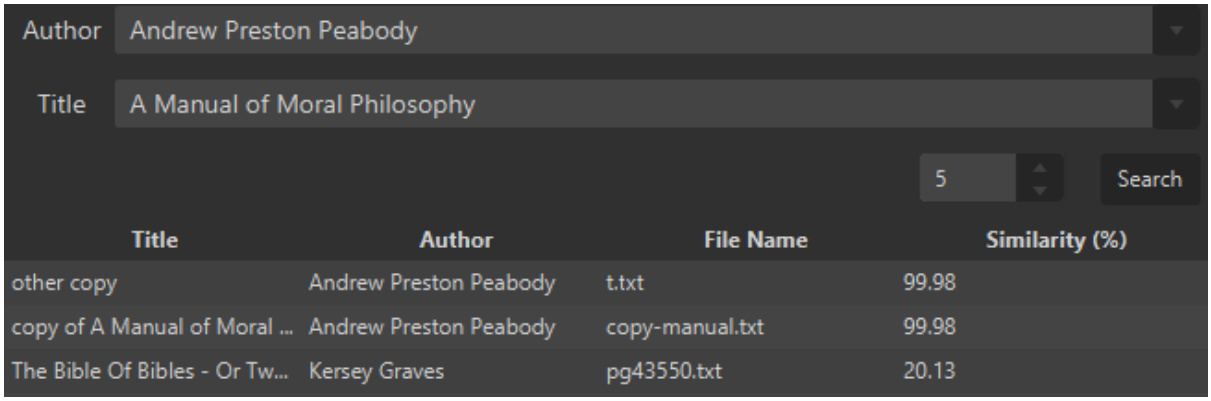
Passos a seguir

Copiem el contingut del fitxer “A Manual of Moral Philosophy” en un document nou, i el guardem en el sistema.





A continuació, cerquem els documents semblants: en efecte, la similaritat és del 99.98%, segurament degut a un error d'aproximació. En tot cas, per comprovar que és consistent – i no causa d'un error d'actualització – creem una altra còpia i repetim la cerca. Obtenim el resultat esperat.


Author
Andrew Preston Peabody
Title
A Manual of Moral Philosophy
5
Search
Title
Author
File Name
Similarity (%)
other copy
Andrew Preston Peabody
t.txt
99.98
copy of A Manual of Moral ...
Andrew Preston Peabody
copy-manual.txt
99.98
The Bible Of Bibles - Or Tw...
Kersey Graves
pg43550.txt
20.13