Approximate Strig Matching

David Barbas, Victoria Beltrán, Xiangyang Ge y Alberto Romero
Universitat Politècnica de València

24th December 2019

Contents

1 Introducción

El objetivo de esta segunda parte del proyecto conjunto SAR-ALT es añadir al motor de recuperación de información que se desarrolló en la primera parte el proyecto (prácticas de SAR) la capacidad de hacer búsquedas aproximadas en nuestro índice. Para lograr este objetivo antes deberemos elegir un algoritmo eficiente para hacer la búsquedas aproximada de una cadena respecto de todas las cadenas del diccionario de términos.

2 Tiempos

Los tiempos a continuación están mostrados en segundos.

Tabla 1: Umbral = 1

	de	lluis	capital	antiheroes	iluminacion
Levenshtein	5.03128910	5.35392833	5.36292601	4.90795302	5.00696898
Damerau	4.31206417	4.16383457	4.06347084	4.15784192	4.18845391
Levenshtein vs Trie	2.98550224	2.95879936	3.05793285	3.07298493	3.24075222
Damerau vs Trie	4.01207042	3.81526494	3.79269266	3.77744532	3.70296049
Leve. vs Trie Ramif.	0.00601864	0.00705242	0.00502920	0.00598741	0.00498319
Dame. vs Trie Ramif.	0.00897408	0.01096749	0.01105928	0.01063442	0.00897527

Tabla 2: Umbral = 2

	de	lluis	capital	antiheroe	iluminacion
Levenshtein	8.04722452	7.67783976	7.00475097	7.12829351	7.39022589
Damerau	7.00074434	6.71333075	6.39508319	6.66620183	6.61127305
Levenshtein vs Trie	3.03989577	3.24728751	3.23534560	3.26725912	3.26925111
Damerau vs Trie	3.57942605	3.69910264	3.40887856	3.57846999	3.57838511
Leve. vs Trie Ramif.	0.08277988	0.06980944	0.04587555	0.04591560	0.04487967
Dame. vs Trie Ramif.	0.09275222	0.06482601	0.06482649	0.06482768	0.06483412

Tabla 3: Umbral = 3

	de	lluis	capital	antiheroe	iluminacion
Levenshtein	9.28715301	9.76675820	10.27002764	9.63193750	9.32997274
Damerau	8.57201171	8.32871294	8.28161287	8.21538162	8.74404025
Levenshtein vs Trie	3.27728176	3.24509192	2.85989761	2.87847567	2.86529946
Damerau vs Trie	3.33403277	3.29447365	3.28810000	3.32370090	3.18736148
Leve. vs Trie Ramif.	0.36581922	0.39285994	0.39687109	0.41774774	0.35357618
Dame. vs Trie Ramif.	0.38653493	0.41804981	0.41996884	0.44471216	0.42907381

Tabla 4: Umbral = 4

	de	lluis	capital	antiheroe	iluminacion
Levenshtein	10.67627335	10.59732533	10.50507832	11.63274789	10.48344803
Damerau	9.82672524	10.02748084	9.25132751	10.05506134	9.56260061
Levenshtein vs Trie	3.15892959	3.32909369	3.25429344	2.87335563	2.71872258
Damerau vs Trie	3.05482626	3.06280255	3.40067077	3.11229753	3.89898896
Leve. vs Trie Ramif.	3.11221290	2.77910852	2.88697791	3.19445300	3.11071181
Dame. vs Trie Ramif.	3.42883062	3.50462461	3.48966408	2.97010922	3.08887887

Tabla 5: Umbral = 5

	de	lluis	capital	antiheroe	iluminacion
Levenshtein	11.89118648	12.37489629	12.69403934	11.85332346	12.12190580
Damerau	11.18398070	11.79294491	11.32167721	11.36089039	10.77300572
Levenshtein vs Trie	2.81646633	2.79850054	2.85936141	3.11067843	2.76261020
Damerau vs Trie	3.14658141	3.21639061	3.23235250	3.26227689	3.34105730
Leve. vs Trie Ramif.	10.82204556	11.68170834	10.74622560	10.94368196	10.95166373
Dame. vs Trie Ramif.	11.64480710	11.07266140	12.20837164	14.12351036	14.66676688

Tras medir todo los tiempos hemos llegado a la conclusión que para umbrales menores a 6, el mejor método es el de ramificación y poda pero pasado ese umbral, sin duda, es mejor el método de comparación contra el TRIE completo. Además, para palabras cortas la variante de Damerau-Levenshtein resulta más rápida pero con longitudes superiores a aproximadamente 7 letras, Levenshtein simple es más efectivo. Creemos que esto es dado por el hecho que Damerau-Levenshtein ha de realizar más comparaciones y por tanto es más costoso.