ALGORÍTMICA

Proyecto de prácticas: Búsqueda aproximada de cadenas (Informe)

Villanueva Tarazona, Jose
Fernández Sutil, Diego
Marcos Bravo, Pablo
Alexey Mengliev

Grupo 4CO21
Curso 2022/2023

0. Introducción

En el siguiente proyecto coordinado se ha implementado una ampliación respecto a lo trabajado en la asignatura de SAR, centrándonos en las distancia de edición de Levenshtein y su uso en la búsqueda aproximada de cadenas, que nos permita realizar una serie de sugerencias en base a las consultas aplicadas por el usuario.

Para ello, se ha dividido el trabajo de la siguiente manera:

- Reducción del coste espacial y umbralización del algoritmo Levenshtein. (Alexey y Jose)
- Recuperación del camino de Levenshtein. (Alexey)
- Implementación de la clase SpellSuggester. (Jose)
- Integración con el proyecto de SAR. (Alexey)
- Cálculo de la cota optimista. (Jose)
- Implementación de Damerau-Levenshtein restringido. (David y Pablo)
- Implementacion de Damerau-Levenshtein intermedio. (David y Pablo)
- Recuperación del camino de Damerau-Levenshtein restringido. (David y Pablo)

Se ha utilizado **git/github** como principal herramienta de control de versiones, y se puede hacer un seguimiento (no tomar como referencia para la carga de trabajo) en el siguiente repositorio:

https://github.com/losk4/ALT

1. Distancias

Como principal tarea, se ha implementado el algoritmo de distancias de *Levenshtein* en su versión iterativa con reducción de complejidad espacial y parada por umbral simple. Seguido, se ha ampliado este algoritmo a los distintos casos de transposición de carácteres, considerando las restricciones dadas, con la creación de las versiones para *Damerau-Levenshtein* restringida e intermedia.

Posteriormente se ha implementado una cota optimista basada en el modelo *bag of words* sugerida como ampliación. Dicha implementación se ha llevado a cabo para la distancia Levenshtein con umbral.

Por último, se han recuperado las operaciones de edición de cada algoritmo propuesto menos el de Damerau-Levenshtein intermedio.

Dichos métodos pueden encontrarse en **src/distancias.py** y han sido testeados tanto con **tests/test_distancias.py** como **tests/test_edición.py**, con los resultados volcados en sus respectivos ficheros de la carpeta **results**.

2. Sugerencias

A continuación, se requería completar **src/spellsuggest.py** (en concreto, el método *suggest(...))* para poder recuperar sugerencias en base a un término y un umbral de distancia de edición utilizando el diccionario generado por **corpora/miniquijote.txt**.

En nuestra versión se ha conseguido devolver las sugerencias correctas (ver **results/test_spellsuggester.txt**) y hemos evitado la ejecución de cualquier método de distancia con *threshold* si este era igual o menor a la diferencia absoluta entre la longitud de las dos cadenas comparadas (devolviendo por defecto, *threshold* + 1).

3. Adaptación al proyecto de SAR

A la hora de indexar, se ha habilitado una nueva funcionalidad que permitirá la construcción del vocabulario necesario para utilizar sugerencias. También existe esta opción a la hora de recuperar las consultas, con la que se le exigirá al buscador que recupere las noticias asociadas a las sugerencias de los términos que no contienen ningún resultado.

Este método solo devuelve el número de sugerencias, no siendo posible poder ver las sugerencias ofrecidas, en forma de caja negra.

Aprovechando el test proporcionado **test_spellsuggester.py**, podemos sacar la siguiente toma de tiempos:

levenshtein_m
--- 6.610319137573242 total seconds ---
levenshtein_r
--- 2.0275707244873047 total seconds ---
levenshtein
--- 1.0840988159179688 total seconds ---
levenshtein_o
--- 0.692176103591919 total seconds ---
damerau_rm
--- 6.96234655380249 total seconds ---
damerau_r
--- 1.2825655937194824 total seconds —
damerau_im
--- 7.842054128646851 total seconds ---
damerau_i
--- 1.587742567062378 total seconds ---

Considerando los resultados, hemos escogido como algoritmo principal de búsqueda de sugerencias el de Levesnhtein con reducción de coste espacial, umbralización y cota optimista.