

Práctica 2

Vamos a desarrollar una implementación de la estructura de datos árbol de sufijos vista en clase junto con los algoritmos de string matching y “problema del substring”.

Se trata de una parte del contenido del tema 4 visto en clase.

1.1. El problema del substring

El problema del substring consiste en dados un patrón p y N textos t_1, \dots, t_N , encontrar todos los textos en los que aparece el patrón, es decir, devolver un conjunto $I \subseteq \{1, \dots, N\}$ tal que $i \in I$ sii p aparece en el texto t_i .

Para resolver el problema del substring usando árboles de sufijos, se construye el árbol compacto correspondiente al string $t_1 t_2 \dots t_N$ (es decir, los N textos concatenados con $\$$ entre cada dos). Puede ser conveniente modificar las hojas del árbol y añadir información adicional a los nodos.

1.2. ¿Qué hay que hacer?

Hacer un programa que construya un árbol de sufijos compacto para los siguientes casos:

- un único string t ,
- una colección de strings t_1, \dots, t_N (es decir, un string $t_1 t_2 \dots t_N$ como en la sección anterior).

Se puede utilizar el método visto en clase o uno más eficiente (idealmente con coste $O(n \log n)$). El alfabeto de entrada puede asumirse un conjunto de caracteres ASCII o bien darse como parte de la entrada.

Además deben implementarse algoritmos que resuelvan los siguientes problemas:

- String matching,
- El problema del substring.

El programa se llamará **SuffixTree** y tendrá al menos las siguientes opciones:

1. Leer dos strings, patrón y texto, y encontrar todas las ocurrencias del patrón en el texto.
2. Leer un string patrón p , un número N y N strings textos t_1, \dots, t_N y encontrar todos los textos que contienen el patrón.

Deben realizarse un número razonable de pruebas con suficientes strings de suficiente longitud, en concreto correspondientes a textos suficientemente largos en español. Para ello es recomendable buscar y utilizar un corpus en español, por ejemplo el de la RAE.

1.3. Entrega

Deberá entregarse hasta el 26 de enero de 2018.