Proyecto I: Búsqueda en texto

Descripción

El problema básico de búsqueda en texto consiste en encontrar todas las ocurrencias de un patrón *P* de largo *m* en un texto *T* de largo *n*. Existen diferentes soluciones a este problema que tratan de mejorar el algoritmo de fuerza bruta (complejidad *O(nm)*) asintóticamente o en la práctica. En este proyecto se implementará como *baseline* el algoritmo de fuerza bruta. Además, se implementarán algunos de los algoritmos más conocidos para resolver el problema (Boyer-Moore, Knuth-Morris-Prat, Rabin-Karp, Horspool, etc.).

Objetivos específicos

La entrega de la tarea (un informe y código fuente) debe satisfacer los siguientes objetivos específicos:

- Describir las características principales de los algoritmos implementados, incluyendo pseudocódigo.
- Describir las decisiones de implementación más importantes.
- Evaluar experimentalmente los algoritmos seleccionados. Dado que los algoritmos son relativamente sencillos y conocidos, la carga principal del proyecto reside en esta evaluación experimental. Recomendaciones:
  - Emplear diferentes *datasets*, tanto sintéticos como reales (en http://pizzachili.dcc.uchile.cl/texts.html se pueden encontrar datasets reales con diferentes características, así como un generador de texto aleatorio).
  - Estudiar la influencia del tamaño del texto y del patrón.
  - Diseñar experimentos que fuercen peores/mejores casos, etc.
  - Incluir en el informe gráficos de tiempo vs el parámetro en estudio (largo de patrón, largo de texto, etc.).
  - Se deben realizar al menos 5 experimentos en los que se varíe alguna configuración (tipo de texto, características de los patrones buscados, largos del texto/patrón, etc.). Por cada experimento se debe documentar el propósito, los resultados obtenidos y una discusión acerca de los mismos.

Condiciones

- El proyecto se realizará en grupos de dos o tres estudiantes. Excepcionalmente se puede realizar de forma individual.
- Se deben implementar y evaluar tantos algoritmos como miembros tenga el grupo, además del *baseline*. Es decir, un grupo de X estudiantes, debe implementar X+1 algoritmos (incluyendo entre ellos el algoritmo de fuerza bruta).
- Todas las implementaciones del grupo deben ser en el mismo lenguaje de programación y deben emplear el mismo tipo de optimizaciones.
- Si se toma como base un código descargado de Internet, debe indicarse explícitamente la fuente

y el código debe estar documentado con suficiente detalle para mostrar que se entendió. En caso de que se utilice un código disponible y no se indique esto, se considerará copia acarreando NCR en la asignatura.

- Los lenguajes de programación aceptados son C, C++, Java y Python.
- El *baseline* debe ser implementado entre todos los miembros del grupo. Lo mismo sucede para la evaluación experimental y para la redacción del informe.