INFD0003 - Técnicas Avanzadas de Programación
Tarea 1: Uso de pySpark y librerias cómputo paralelo distribuido

Profesor: Jorge R. Vergara
Primer Semestre 2025

NOTA: Todos los análisis deben ser realizados con pyspark. Utilice todas las librerías que considere necesarias. En el caso que utilice realice pruebas en el cluster, escriba sus rutinas en un script y envíe a procesar mediante el sistema de colas. Evite realizar experimentos que requieran mucho cómputo en su sesión del cluster.

Problema 1

En el área de la bioinformática, el alineamiento de secuencias es una forma de organizar las secuencias de ADN, ARN o proteínas para identificar regiones de similitud que pueden ser consecuencia de relaciones funcionales, estructurales o evolutivas entre las secuencias. Una de las principales dificultades de este proceso es el costo computacional en identificar los calces o matching, ya que las secuencias son de gran tamaño y que en ciertos casos el calce ocurre solo con algunas partes entre las secuencias, escalando el problema en forma exponencial. El objetivo de esta actividad es que usted pueda realizar una identificación de secuencias con los datos entregados en esta tarea. En este caso debe identificar las secuencias (o parte de estas) en el genoma de referencia. Para esto utilice pyspark para optimizar y distribuir la búsqueda. A continuación se detalla una lista de actividades que debe realizar con los datos entregados :

1. Utilice expresiones regulares para realizar la búsqueda. Identifique la cantidad de calces encontrados y la posición donde ocurrió el calce.
2. Varíe el largo y forma de la secuencia a buscar estableciendo una métrica de comparación para analizar los tiempos y numero de paralelizaciones utilizando pyspark.
3. Realice un gráfico de autosimilaridad a través de un gráfico de recurrencia considerando distintas partes de secuencias. Repita el experimento para ver la similaridad entre secuencias y secuencias-referencia.
4. Considerando que el diccionario de elementos se compone solo de las letras A,T,G y C (4 posibilidades) codifique los datos como enteros de 8 bits para generar un diccionario de 256 elementos. Con esta referencia realice un análisis a través de convoluciones en 1D y genere series de tiempo con los resultados de esta.
5. Realice una búsqueda jerárquica de convoluciones en 1D como lo hace una red neuronal. Utilice alguna estrategia de red neuronal convolucional para realizar esto. Una vez entrenada la red realice una visualización de los diferentes filtros encontrados y sus resultados obtenidos.

Problema 2

Para las secuencias de la carpeta «data», diseñe una estrategia de codificación y genere un autoencoder para mapear los datos en un espacio 2D.

Resultados e Informe

Se debe entregar un informe con los puntos solicitados anteriormente y el respectivo análisis de los resultados obtenidos. El análisis debe incluir comparaciones entre lo obtenido en la práctica y la teoría. Respalde su análisis con tablas y gráficos, procurando que sus figuras sean legibles, auto explicativas y pertinentes. Un exceso de gráficos sin un análisis adecuado será calificado negativamente.

- La tarea es individual.
- La entrega del informe (en formato PDF) y códigos es hasta el Miercoles 11 de Junio de 2025 hasta las 23:59:59 hrs. a través de Canvas.
- Los atrasos se penalizarán con 0.5 puntos menos por día de atraso.
- Se recomienda que el informe no exceda las 10 páginas. En el caso que posea más graficas incluyalas como anexos.
- La estructura del informe debe incluir:
    1. Introducción.
    2. Descripción concisa de los métodos y algoritmos usados.
    3. Resultados.
    4. Análisis y discusión de resultados.
    5. Conclusiones.
    6. Referencias.