Licenciatura en Ciencia de Datos
Datos Masivos I

Tarea 1

Fecha de entrega**: 1 de Marzo a las 09:00 hrs**

Datos sobre el reporte:
- Escribir el código en el reporte (copy-paste).
- Adjuntar los archivos .py dentro de la carpeta enviada al correo (zip).
- La discusión y los resultados deben estar presentes en el informe.
- Enviar reporte en archivo PDF.

1. Principio Bonferroni (20%)

- Supongamos que creemos que ciertos grupos de terroristas se reúnen ocasionalmente en hoteles para tramar hacer el mal.
- Queremos encontrar personas (no relacionadas) que al menos dos veces se hayan hospedado en el mismo hotel el mismo día.

Hay **10 millones** de personas
Cada persona va a un hotel cada **50** días
Un hotel hospeda **40** personas y hay **10,000** hoteles
Para determinar un terrorista buscamos **tres** personas que en **dos días** distintos en una ventana de **300** días fueron al mismo hotel.

a) Aplicar el principio de Bonferroni.

b) Discutir el resultado obtenido.

Nota. Este ejercicio puede realizarse a mano y agregarlo como foto al reporte.

2. Map – Reduce (30%)

Dados los siguientes párrafos

- Problem solving skills are not something that can be distilled down into a single step-by-step process
- Testing involves finding, designing, and developing test cases: actual instances of the problem that can be used to test your solution.
- Computer hardware usually refers to the physical components in a computing system which includes input devices such as a mouse/touchpad, keyboard, or touchscreen, output devices such as monitors, storage devices such as hard disks and solid-state drives, as well as the electronic components such as graphics cards, main memory, motherboards, and chips that make up the Central Processing Unit (CPU)
- A program may contain many functions and pieces of code, but this special function is defined as the one that gets invoked when a program starts.
- Good code is not just functional, it is also beautiful, good code is organized, easy to read, and well documented, organization can be achieved by separating code into useful functions and collecting functions into modules or libraries, good organization means that at any one time, we only need to focus on a small part of a program
- It would be difficult to read an essay that contained random line breaks, paragraphs were not indented, it contained different spacing or different fonts, etcetera, likewise, code should be legible, well written code is consistent and makes good use of whitespace and indentation, code within the same code block should be indented at the same level, nested blocks should be further indented just like the outline of an essay or table of contents.

A) Aplicar Map – Reduce. Realizar el conteo de palabras y graficar en un histograma las 5 palabras con mayor frecuencia en los párrafos.
B) Discutir los resultados. Por ejemplo, ¿con las 5 palabras con mayor frecuencia es posible saber de qué se está hablando en los párrafos? ¿Con 10 palabras? ¿Con 20 palabras? ¿Cuántas palabras consideras necesarias para conocer de qué se está hablando en los párrafos?

3. NLTK (50%)

Utilizando el archivo “amazon_alexa-1.csv”, extrae las primeras 1,000 opiniones.
Realiza el preprocesamiento (eliminar puntuación, palabras vacías, etc.)
Utilizar SentimentIntensityAnalyzer

a) Para cada una de las 1,000 opiniones: determinar si son positivas, negativas o neutras.
b) Graficar en un histograma el número (conteo) de opiniones positivas, negativas y neutras.
c) Discutir de acuerdo con los resultados, ¿los clientes están satisfechos con los productos de Amazon?

Para todos los Ejercicios:
Nota General.
En los problemas de programación, aplicar buenas prácticas de programación.
Nombrado de variables. Ejemplo: variable_uno, first_variable
Nombrado de funciones. Ejemplo: removerPuntuación(), remove_punctuation()
Nombrado para clases. ClaseExtraDos(), SecondExtraClass().
Agregar explicación de lo que hace cada función o variable de importancia para la resolución de problemas. Ejemplo: #Calcula las potencias del parámetro que recibe. # Ejecuta el análisis de sentimientos.

Opcional.
Siéntete libre de utilizar otras librerías (además de las solicitadas) para realizar los ejercicios, lo cual puede darte puntos extras sobre la calificación de la tarea.