Conceptos fundamentales de Python

Objetivos

El objetivo de esta actividad es afianzar los conocimientos esenciales de Python que permitirán implementar con éxito las herramientas bioinformáticas propuestas en los sucesivos temas.

Pautas de elaboración

Para esta actividad necesitarás un ordenador con Python y Conda instalados, además de un archivo .txt y dos archivos .csv, que serán proporcionados en el momento de empezar esta actividad.

Accede al siguiente enlace para instalar Python:

<https://www.python.org/downloads/>

Accede al siguiente enlace para instalar Conda:

<https://docs.conda.io/projects/conda/en/latest/user-guide/install/windows.html>

**Actividades**

1. **Preparación del entorno de trabajo (2 puntos)**

En esta primera actividad será necesaria la creación de un ambiente en Conda y la descarga de paquetes adecuados. Indica el comando o comandos necesarios para:

* + A. Crear un ambiente de Conda que tenga el nombre «actividad1» e instalar en él las librerías de: Pandas, Matplotlib y Seaborn, además del IDE Spyder. (0,8)
  + B. Instala la versión 1.25.0 de NumPy. (0,6)
  + C. Abre una consola. ¿Cómo comprobarías los ambientes creados en Conda? ¿Y las librerías instaladas en el ambiente que acabas de crear? (0,6)

1. **Manipulación de archivos de texto plano (3,5 puntos)**

La manipulación de archivos de texto plano es una tarea recurrente en bioinformática. Por ello es importante saber utilizar el administrar contextos *(with)*. El formato .pdb es un ejemplo de formato de texto, plano el cual contiene información estructural de una proteína (en este caso, la P53) 1TUP por su identificador en el Protein Data Bank. Completa las siguientes tareas:

* + A. Examina manualmente el archivo con un editor de texto de tu elección (bloc de notas, Notepad, etc.). ¿Cuál es el título descriptivo de esta estructura (TITLE)? ¿Quiénes son los autores de este archivo (AUTHOR)? (0,25)
  + B. Mediante el administrador de contextos *with*, lee el documento .pdb y extrae solo la secuencia de aminoácidos (SEQRES) de la P53 y guárdalos en una lista. Ten cuidado de no incluir nada no sea un aminoácido en dicha lista. (1,5)
  + C. Toma la lista que has creado en el apartado B y crea un programa para contar cuántos aminoácidos hay de cada tipo. Guarda este conteo en un diccionario donde la clave sea el aminoácido y el valor el número de veces que se encuentra repetido. No definas las claves del diccionario manualmente. (0,5)
  + D. Crea un histograma con Seaborn con la frecuencia de cada aminoácido. No olvides incluir el título del gráfico, así como el título de los ejes. Asigna a cada uno un color distinto. (1,5)

1. **Manipulación de conjuntos de datos (4,5 puntos)**

Pandas es una de las librerías más populares de Python debido a las herramientas que ofrece a la hora de manipular conjuntos de datos. Toma ambos .csv y lleva a cabo las siguientes tareas:

* + A. Lee el archivo «nombre\_archivo» y cambia el nombre de las columnas por los siguientes: id, dieta, pulsaciones, tiempo y actividad. (0,25)
  + B. Determina la presencia de celdas vacías con uno de los métodos de la clase DataFrame y elimina dichas filas. (0,5)
  + C. ¿Cuántos niveles hay en la columna «dieta» y cuál es su frecuencia? Lleva a cabo esta tarea con un solo método de la clase DataFrame. (0,25)
  + D. Ayúdate del método *groupby* para agrupar los datos por el nivel de actividad. Genera una lista a partir del *groupby* de forma que puedas observar cómo se ha llevado a cabo la agrupación. Esta lista debería contener tres elementos, ¿por qué tres? ¿Qué contienen cada uno de estos elementos de la lista? (1)
  + E. Toma el objeto *groupby* (no la lista que hayas generado) y ayúdate del método agg para calcular la frecuencia cardíaca media y la desviación estándar para cada nivel de actividad. (0,75)
  + F. Los pacientes de este estudio, cuyo nombre ha sido sustituido por un ID, provenían de distintas ciudades. Utiliza el método *merge* para añadir esta información y completar el conjunto de datos. (1)
  + G. Utiliza Matplotlib/Seaborn para hacer un gráfico en el que se pueda ver la relación entre las variables pulsaciones y tiempo según el tipo de actividad y la dieta. No olvides dar un título apropiado al gráfico así como reflejar en cada eje el nombre de la variable correspondiente. Recibirá más puntuación si se consigue hacerlo en una única figura multi-facetada (una faceta por estrato). (0,75)

Extensión y formato

* Los resultados deberán ser entregados como un *script* de Python (formato .py).
* Incluye los comentarios pertinentes para explicar el código.
* No hay una extensión máxima del script.

Rúbrica

|  |  |  |  |
| --- | --- | --- | --- |
| Conceptos fundamentales de Python | Descripción | Puntuación máxima  (puntos) | Peso  % |
| Preparación del entorno de trabajo | Preparación del entorno de trabajo. | 2 | 20 % |
| Manipulación de archivos de texto plano | Manipulación de archivos de texto plano | 3,5 | 35 % |
| Manipulación de conjuntos de datos | Manipulación de conjuntos de datos. | 4,5 | 45 % |
|  |  | **10** | **100 %** |