## TRABAJO FINAL

**Curso:** Lenguaje de Programación II.

**Semestre:** 2021-2

**Tema:** webscrapeo a UniProt (https://www.uniprot.org/).

**INDICACIONES:**

* El trabajo es grupal (integrantes de grupos ya indicados)
* Utilizar *Jupyter notebook*.
* Trabajar en equipo vía *GitHub*. El coordinador creará un repositorio del proyecto e invitará los demás integrantes a colaborar en el proyecto o los participantes solicitarán contribuir (pull request)

**PROCEDIMIENTO REFERENCIAL:**

1. Descargar y explorar el archivo ***02 Ch Genes ordenados con vecinos.xlsx*** compuesta por varias hojas las cuales tienen una columna con etiqueta: GenAbrev.
2. Para cada nombre del gen en abreviado (GenAbrev), abrir la página <https://www.uniprot.org> y en el buscador solicitar la búsqueda de cada una de la forma siguiente: colocar el nombre del gen y la palabra *Capra Hircus.*
3. Se mostrará una lista de genes con el nombre buscado. Extraer información de ***entry*** y ***entry name***.
4. Seleccionar la primera opción del gen de la primera columna “entry” (clickear), la cual llevará a una pantalla de la cual extraer la información de **Protein**, **Gene**, **Organism** y **Status**.
5. Extarer la infomación que corresponde a: **GO - Biological process**
6. La información debe estructurarse en un archivo **.csv** o **.xlsx** que muestre el código abreviado del gen (GenAbrev) y la información organizada de keywords (diferenciando cada tipo de función).

Sugerencia de estructura de reporte: **Reporte.xlsx**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **GenAbrev** | **Entry** | **Entry name** | **Protein** | **Gene** | **Organism** | **Status** | **GO - Biological process** |
| **HSPB7** | [A0A452G662](https://www.uniprot.org/uniprot/A0A452G662) | A0A452G662\_CAPHI | Submitted name:  **Heat shock protein family B (small) member 7** | **HSPB7** | Capra hircus (Goat) | Unreviewed | [heart development](https://www.ebi.ac.uk/QuickGO/term/GO:0007507) |