CARRERA: TECNICATURA SUPERIOR DE CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL

ESPACIO CURRICULAR: PROYECTO INTEGRADOR I

PROFESORA: SILVIA PEROTTI

PROFESOR: HÉCTOR PRADO

COHORTE: 2022

PROYECTO TECNOLÓGICO INTEGRADOR

2022

PROYECTO ALFA

OBJETIVO

El objetivo principal del Proyecto Tecnológico Integrador es que puedan vivenciar una experiencia real de trabajo colaborativo en Ciencia de Datos, en un entorno Ágil.

Por eso además del trabajo en equipo y lo aprendido durante la cursada del presente año en los diferentes espacios curriculares, es fundamental la investigación, la creatividad y el aprovechamiento de los diferentes perfiles que hay en cada grupo para alcanzar el éxito de esta experiencia.

ANTES DE COMENZAR

Les sugerimos que lean muy bien las consignas, investiguen sobre el material adicional y sobre todo, busquen información que les sea de utilidad para llevarlo adelante, de igual manera aprovechen de compartir, no solo las dudas, si no los avances dentro y fuera de los grupos que están realizando el Proyecto Integrador.

Como podrán ver en la estructura del proyecto, hay una serie de documentos/hitos a cumplimentar/cumplir, en un plazo acotado, les sugerimos que dividan las actividades y no permitan que el atraso en alguna consigna, frene las presentaciones de las restantes.

FUNDAMENTACIÓN

Una tarea habitual en Ciencia de Datos es obtener datos de múltiples fuentes.

Una de estas técnicas suele ser el “Scraping”, el cual permite obtener datos de páginas web para manipularlas desde la aplicación desarrollada.

El web scraping es una técnica que permite extraer datos e información de una web. Recomendamos que realicen el web scraping con Python, utilizando para ello la librería Beautiful Soup.

PROYECTO

La consigna consiste en desarrollar una aplicación en Python, que tome los datos de alguna página web, haciendo Web Scraping, guardando los datos recolectados en un archivo csv, para luego ser mostrados en una tabla por consola.

Por último, con dicha información, realizar un informe sobre el procesamiento de la misma. Este informe puede ser simple a través de la comparación entre elementos recolectados, estadísticos sobre las series de los datos o cualquier tipo de análisis del tipo Big Data o Machine Learning, como regresiones, proyecciones, etc.

EJEMPLO

Solo a los fines de proponer un ejemplo que amplíe la consigna, imaginemos que tomamos páginas como https://pypi.org/project/investpy/ ó https://es.finance.yahoo.com/quote/TEF?p=TEF&.tsrc=fin-srch ó https://www.bolsamadrid.es/esp/aspx/Mercados/Precios.aspx?indice=ESI100000000 ó de Gobiernos Abiertos.

Como bien sabemos las páginas web son documentos estructurados formados por una jerarquía de elementos. Así que luego de elegir la página, el siguiente paso consiste en identificar correctamente el elemento o elementos que contienen la información deseada.

Luego hay que descargar el contenido de la página utilizando la librería requests.

El contenido de la página obtenido en el paso anterior será el que utilicemos para crear la «sopa», esto es, el árbol de objetos Python que representan al documento HTML.

Ahora estamos en condiciones de buscar en el árbol y obtener los objetos que contienen la información y datos que necesitamos.

Una vez obtenidos los datos los guardarán en un archivo CSV.

Por último, mostrar los datos en una tabla en consola y realizar algún informe.

GESTIÓN DEL PROYECTO

La gestión del proyecto se llevará adelante, con las herramientas utilizadas, por lo tanto, tendrán que crear un repositorio público en GitHub con el nombre del grupo, dentro del https://github.com/ispc-programador2022/. En el mismo subirán los documentos de trabajo, enlaces con documentación utilizada para encaminar el proyecto, más los códigos y las bases utilizadas.

De igual manera crearán un tablero en Trello, con acceso para ispc.programador2022@gmail.com en donde compartiran como van trabajando e interactuando entre Uds.

ENTREGABLES Y PLAZOS

Los entregables y plazos del proyecto son los siguientes

i) ***Documento de KickOff del proyecto***. El mismo se ira completando sobre el siguiente documento colaborativo https://docs.google.com/spreadsheets/d/13yc5EIMPd_KinnN1uaM0laSg48Qzvu3a1bEMR1rchpM/edit?usp=sharing y las tres columnas principales (Nombre del Grupo, Tema a desarrollar y Pagina Web Origen de Datos) tienen que estar **completas por el equipo antes del 26/10/22**.

ii) ***Ejecución del Proyecto***. Esto hace referencia a la codificación del WebScrapping, las bases de datos, los informes y el documento que contenga el resumen del proyecto. Y tiene que estar **finalizado antes del 16/11/22**.

iii) ***Presentación del proyecto***. Consiste en la realización de un video o grabación de un meet interno del equipo de proyecto, con una duración no mayor a 15 minutos, en donde el equipo presente los resultados obtenidos y comenten sobre el proceso de ejecución del proyecto. Esto tiene que ser **presentado antes del 23/11/22.**

iv) Cabe aclarar que cada una de estas etapas y generación de estos documentos, tiene que ser cargado en el documento de KickOff. El que

también es un desafío, puesto al ser colaborativo sirve no solo para que todos los equipos estén al corriente del avance de los otros equipos, si no que quien lo edita para agregar información debe ser muy cuidadoso y respetuoso para que no se vaya a perder información de ningún otro equipo compañero.

RUBRICA

A continuación, les compartimos un cuadro, que será tenido en cuenta para la valoración final de cada uno de los proyectos. Como podrán ver en el mismo, hay competencias que se complementan y otras que no se acumulan, por tal motivo es muy importante que, si alguna actividad no la llegan a cumplimentar en tiempo y/o forma, no dejen de presentar el resto.

Evaluación del Proyecto Integrador		
Gestión del Proyecto	**Desde Puntos**	**Hasta Puntos**
Idea/Creatividad/innovación	0	10
GitHub	0	15
Trello	0	15
Gestion del Tiempo	0	10
Ejecución del Proyecto	**Desde Puntos**	**Hasta Puntos**
Phyton	0	30
Bases de Datos	0	20
Otras Tecnologías	0	20
Presentación del Proyecto	**Desde Puntos**	**Hasta Puntos**
Presentación Sincrónico	0	15
Presentación Virtual	0	15

MATERIAL DE SOPORTE

Como el proceso es autogestionado y el equipo docente no estará disponible para guiarlos en temas puntuales, es fundamental que vean todo el material de referencia que les dejamos a continuación, lo mismo que los invitamos a que busquen Uds. todo lo que necesiten en bibliotecas educativas y la web en su conjunto, para llevar el proyecto adelante.

Github

https://youtu.be/eQMcIGVc8N0

Web Scrapping

https://j2logo.com/python/web-scraping-con-python-guia-inicio-beautifulsoup/

Web Scraping con Python, utilizando para ello la librería Beautiful Soup

https://youtu.be/kPNHKrOqedI

https://youtu.be/SuwCyiCLe8Y (ejercicio scraping, comienza minuto 35 - muy importante minuto 55 a 1:05)

Html - etiquetas básicas

https://youtu.be/RafelMz450g

Kagle

https://youtu.be/NhHTWGIglRI

Conexión a BBDD desde Python - MySql y MongoDB

https://github.com/narcisoperez/21SemanaProgramacion_210521

https://github.com/narcisoperez/7SemanaBasedeDatos_300621

https://github.com/narcisoperez/8vaSemanaBase_de_Datos_070721

BD MySQL Phyton

https://youtu.be/Ch7xsRmaJCs

BIBLIOTECA DEL PROYECTO

Como todo proyecto colaborativo, buscamos fortalecernos e incrementar nuestras competencias entre todos y por ese motivo, los invito a que en conjunto vayamos creando la biblioteca del proyecto, es decir cada que un equipo encuentra un documento, código o video que consideren les ha resultado de interés para llevar adelante el proyecto, lo compartiremos en el siguiente documento colaborativo

https://docs.google.com/spreadsheets/d/1ofo1QjRsTy0m038SJyxAdGpgdsNBCDDNBkpg5slrEug/edit?usp=sharing

MANOS A LA OBRA

Estamos en el tramo final, así que manos a la obra y no demorar para alcanzar el objetivo en los plazos definidos.

Éxitos !!!