Proyecto primer corte

Johan Villalba
Escuela de Ciencias Exactas e Ingeniería
Universidad Sergio Arboleda-Bogotá, Colombia
johan.villalba01@correo.usa.edu.co

Sebastián Merchán
Escuela de Ciencias Exactas e Ingeniería
Universidad Sergio Arboleda-Bogotá, Colombia
sebastian.merchan01@correo.usa.edu.co

Miguel Angel Rippe
Escuela de Ciencias Exactas e Ingeniería
Universidad Sergio Arboleda-Bogotá, Colombia
miguel.rippe01@correo.usa.edu.co

Resumen

Para este informe se muestra la solución a una aplicación la cual muestra en tiempo real la cantidad de contagios de Covid-19, la cantidad de recuperados y los respectivos decesos en el país. Dicha aplicación se realizo en el lenguaje python.

1. Marco teórico

1.1. Python

Python es un lenguaje de programación de propósito general, este lenguaje se define comúnmente como un lenguaje de secuencias de comandos las cuales están orientado a objetos, pero no solo se utiliza para este paradigma de programación, ya que es un lenguaje multiparadigmas, python trabaja con los paradigmas procedimentales, funcionales y oriendo a objetos, por eso es un lenguaje muy conocido y ulitizado. [1]

1.2. Covid-19

Los corona virus son una familia de virus que pueden generar enfermedades tanto en humanos, como en animales. Esta enfermedad se manifiesta en los humanos en su mayoria como infecciones, respiratorias las cuales pueden ir desde resfriados leves hasta enfermedades graves. El covid-19 es el tipo de corona virus más reciente que se a encontrado en la ciudad de Wuhan en China y actualmente se a convertido en una pandemia que afecta a muchos países en el mundo. [2]

1.3. Web scraping

El web scraping es el termino comúnmente usado para la extracción de datos con base a una pagina web. Existen diferentes métodos de extracción de datos con base al software que sera usado, estos datos pueden ser usados en varios softwares, tales como: lenguajes de programación, bases de datos, archivos de tipo excel, entre otros.

2. Recursos y software

Para realizar este proyecto y llegar a lo requerido por el docente se utilizaron diferentes recursos los cuales ayudaron con el buen funcionamiento del proyecto como lo son:

- **Pycharm:** Se utilizo este interprete de python para la construcción del código sobre el cual se desarrollara el presente proyecto.
- **BeautifulSoup:** Es una librería propia del lenguaje de programación python, la cual permite analizar documentos de tipo html, es decir paginas web que hayan diseñadas en html, mediante esta librería se es posible buscar información y exportar datos sobre una pagina web dada.
- **requests:** Es una librería propia de python la cual permite generar conexiones, avisos y permisos sobre la url de una pagina web, esto para permitir el acceso sobre esta.

- **matplotlib:** Es una biblioteca para generar gráficos a partir de datos conglomerados en arrays en lenguaje de programación Python y con una extensión matemática NumPy.

- **Mysql:** Es un sistema de gestión de bases de datos, esta se basa en el funcionamiento de los datos de manera relacional

3. Metodologia

3.1. Extracción de datos

Para la extracción de los datos hicimos uso de la pagina web Wikipedia, esto debido a que esta documentando la situación con base al día a día, y además de esto, la extracción de datos sobre esta resulto ser muy eficiente, permite extraer todo el contenido e información de la pagina, y no esta protegida contra el web scraping. Esta fuente es usada por Google Noticias para dar las actualizaciones acerca del comportamiento del Covid-19 en Colombia. Nos puede proporcionar datos muy específicos, como la población de hombres, mujeres, menores de edad y mayores de edad que han contraído la enfermedad,también divide los casos por municipios, y por ciudades, nos brinda información de los mapas de calor del covid, así podrán ser usados en posteriores entregas y a lo largo del desarrollo de este proyecto.

		Grupo Etario (Años)[4]										Total [4][8][9]	
		0 a 9	10 a 19	20 a 29	30 a 39	40 a 49	50 a 59	60 a 69	70 a 79	80 a 89	≥90	Hoy	Acumulado
Estado[4]	Recuperado[1]	26747	48217	164434	177229	121713	95743	51773	24088	10581	2011	*11064*	**722536**
	En Casa	1825	3713	12204	12367	8609	6859	3814	1824	819	162	*-6004*	**52196**
	En Hospital	702	351	1110	1572	1922	2646	2694	1950	1070	188	*-25*	**14205**
	En UCI	41	15	62	154	247	431	495	322	129	11	*-63*	**1907**
	Fallecido[1]	40	33	289	745	1660	3514	5944	6588	5386	1442	*153*	**25641**
Total [4][8][9]	Hoy	*182*	*369*	*1099*	*1165*	*861*	*643*	*456*	*223*	*124*	*25*	***5147***	
	Acumulado	**29392**	**52357**	**178171**	**192127**	**134270**	**109387**	**65086**	**35170**	**18325**	**3918**		**818203**

Figura 1: Tabla de datos de los contagios proporcionada por Wikipedia.

Con base a la tabla de datos, al posicionarse sobre algún dato en concreto e inspeccionar la fuente teníamos acceso al código en html sobre la estructura del articulo, y en si, del valor en concreto que se selecciono previamente.

Figura 2: Código fuente con base a un dato

Una vez ubicados los datos a ser usados, mediante el lenguaje de programación python y a un interpretador de nombre ”pycharm”se iba a realizar la extracción de datos. Para la extracción de datos era necesario hacer uso de varias librerías provistas por python, las cuales serian: BeautifulSoup y requests.

Una vez importadas las librerías se hará uso de un request que permita dar ingreso al usuario sobre la pagina web, hecho esto se indicara una variable bajo los parámetros de la librería de BeautifulSoup la cual se encargara de determinar que se trabajara sobre un archivo de tipo de html.

```
fecha = datetime.datetime.now()
req = requests.get('https://es.wikipedia.org/wiki/Pandemia_de_enfermedad_por_coronavirus_de_2020_en_Colombia')
soup = BeautifulSoup(req.content,'html.parser')
```

Figura 3: Declaracion de variables

Ahora se buscaran los datos provistos en la tabla, para esto con base a la etiqueta en la cual se están almacenando los datos th (el cual indica una tabla en html), de esta forma, al indicar la etiqueta se extraerán todos los datos que se encuentran ubicados en la tabla(figura 1). Finalmente se creara una lista, al cual se encargara se almacenar todos los datos de la tabla, esto mediante un ciclo for.

```
datos=soup.find_all('th')
total=list()

for i in datos:
   total.append(i.text)
```

Figura 4: Búsqueda de datos en la tabla de la pagina

Una vez se han ingresado los datos en el arreglo se hará extracción de los datos necesarios, dado que hay datos que no útiles y hay otros de los cuales no se hará uso, debido a esto se indicaran los datos que van a ser usados, esto se hará sobre la lista, con base a la posición de esta se buscaran los datos que serán usados. Estos datos serán convertidos a un formato de tipo entero, debido a que para la creación de las gráficas es necesario que estos se definan como enteros.

```
#HOY
Recuperados=int(total[49])
Casa=int(total[52])
Hospital=int(total[55])
UCI=int(total[58])
Fallecidos=int(total[61])
Contagios=int(total[75])

#TOTAL
Recuperados1=int(total[50])
Casa1=int(total[53])
Hospital1=int(total[56])
UCI1=int(total[59])
Fallecidos1=int(total[62])
Contagios1=int(total[89])
```

Figura 5: Datos extraídos que son almacenados en variables

3.2. Creación de la base de datos

La creación de la base de datos se realizo con el programa MySql Workbench, el nombre que se le dio fue covid, y se usara después para que sea conectada con Python, la base de datos consta de una tabla llamada casos, la tabla tendrá atributos que serán la mayoría de tipo entero, y solo el idcasos sera de tipo varchar, el id sera la llave primaria, y en nuestro caso se utiliza la fecha para que no se repitan los datos en un mismo día. Las sentencias utilizadas fueron las que se muestran a continuación:

```
drop database if exists covid;
create database covid;
use covid;

create table casos(
id_casos varchar(15) primary key,
numCasos int,
numCasosHoy int,
numHospi int,
numHospiHoy int,
numFalle int,
numFalleHoy int,
numCasa int,
```

Figura 6: Sentencias utilizadas para la creación de la base de datos

3.3. Conexión a la base de datos

Para la conexión sobre la base de datos se hizo uso de una librería de nombre: pymysql, la cual permite generar una conexión entre un programa de python y una base de datos de tipo mysql.

Con base a la base de datos creada anteriormente se generara la conexión entre el script de python y esta, para ello es necesario declarar una variable con los atributos de la librería previamente mencionada y indicar los datos sobre la base de datos, tales como: el nombre del host, el usuario, el nombre asignado a la base de datos, la contraseña de la misma, el formato de los datos que se exportaran a la base de datos y el cursor el cual nos permitirá ejercer operaciones sobre la base de datos desde el script de python.

```
connection = pymysql.connect(host='localhost',
                             user='root',
                             password='1234',
                             db='covid',
                             charset='utf8mb4',
                             cursorclass=pymysql.cursors.DictCursor)
```

Figura 7: Conexión y parámetros usados para la conexión

Finalmente podemos hacer uso de la base de datos, le enviaremos la información de los datos que fueron extraídos por python, esto se hace a través de sentencias y querys que son ejecutadas y enviadas a la base de datos, para mantener constante comunicación con la misma. Un ejemplo de una sentencia para saber un dato en especifico se hace de la siguiente forma:

```
with connection.cursor() as cursor:
    # Read a single record
    sql = "select id_casos from casos where id_casos="+fecha+";"
    cursor.execute(sql)
    result = cursor.fetchone()
```

Figura 8: Ejemplo de Query ejecutado por el programa a la base de datos

4. Resultados

En la parte de resultados tendremos dividido en dos partes significativas, primero los datos que fueron obtenidos con el procedimiento del código, van a ser almacenados en una base de dato. En este caso usamos MySQL para que todos los

datos de contagios, recuperados y personas fallecidas queden guardados en la base de datos.
Los datos van a ser almacenados cada día que se haga la ejecución y compilación del programa, es decir si se compila dos veces el programa en el día solo va almacenar en la base de datos la información correspondiente a ese mismo día. Asi mismo los datos son almacenados en una tabla que se llama casos, y que contiene los siguientes datos

- Numero de contagiados totales en Colombia
- Numero de nuevos contagios en el día de hoy
- Numero de recuperados
- Numero de personas que se están recuperando en casa tanto en el día de hoy como el acumulado
- Numero de personas en el hospital tanto en el día de hoy como el acumulado
- Numero de personas que están en la UCI tanto en el día de hoy como el acumulado

La tabla queda representada en la base de datos de la siguiente forma:

id_casos	numCasos	numCasosHoy	numHospi	numHospiHoy	numFalle	numFalleHoy	numCasa	numCasaHoy	numRecupe	numRecupeHoy	numUci	numUciHoy
1982	818203	5147	14205	-25	25641	153	52196	-6004	722536	11064	1907	-63
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

Figura 9: Tabla de la base de datos

A medida que el programa sea ejecutado y se requieran los datos de un día en especifico se irán guardando en esta tabla, y en la base de datos. Estos datos pueden ser utilizados después para hacer un estudio de los casos a lo largo del tiempo.

Luego de especificar los datos, se quiere mostrar un cambio en el comportamiento que han tenido desde el primer día de contagio mostrando un acumulado y por otra parte un cambio más reciente como lo es mostrar una cantidad del día. Es importante destacar que en las gráficas se muestran valores negativos, los cuales muestran un comportamiento inverso al de contagio, en otras palabras las personas que se han recuperado de Covid-19, estos valores negativos se comparan con un conjunto de personas las que poseen la enfermedad para lograr ver la velocidad a la que se desocupan las UCI y bajan los contagios.

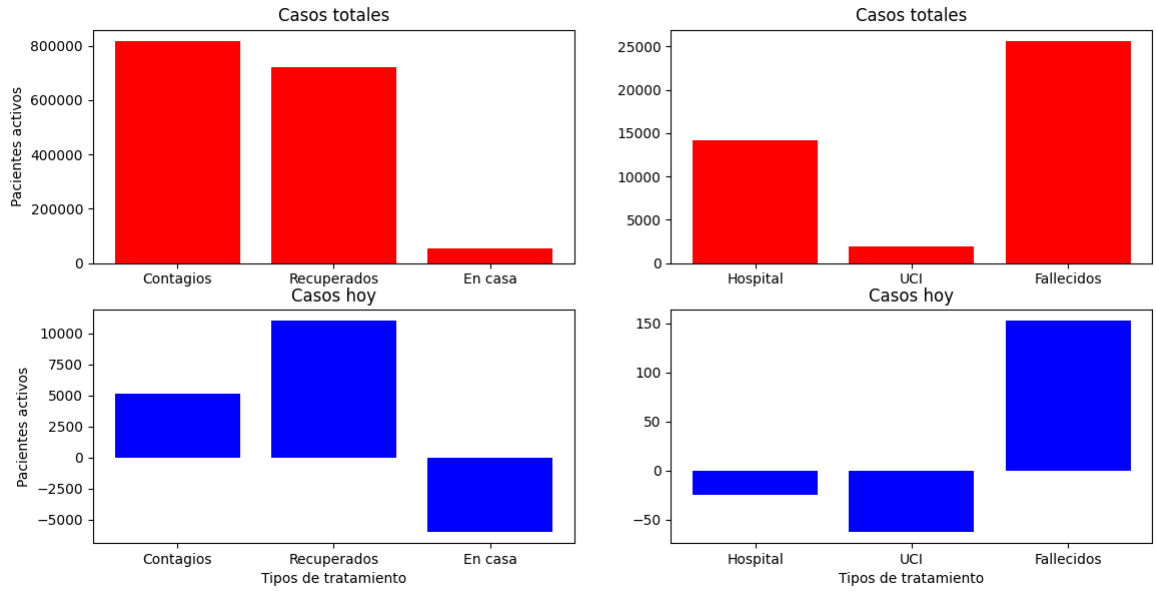

Figura 10: Gráficas de barra

Para tener una mejor visualización del cambio de estos datos se presenta una gráfica de torta en la que se muestran los valores de porcentaje en cada una de estas, para esto se hacen 2 gráficas de torta, la primera evidencia el cambio de pacientes activos de Covid-19 y el segundo gráfico, muestra la cantidad de personas que se encuentran recuperando en hospitales, como se muestra en las figuras 11 y 12 respectivamente.

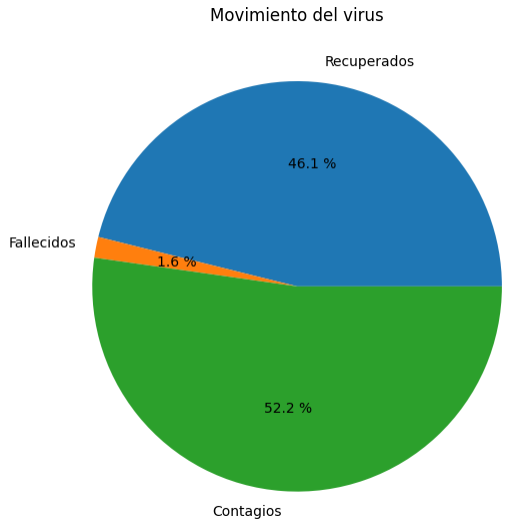

Figura 11: Porcentaje contagio

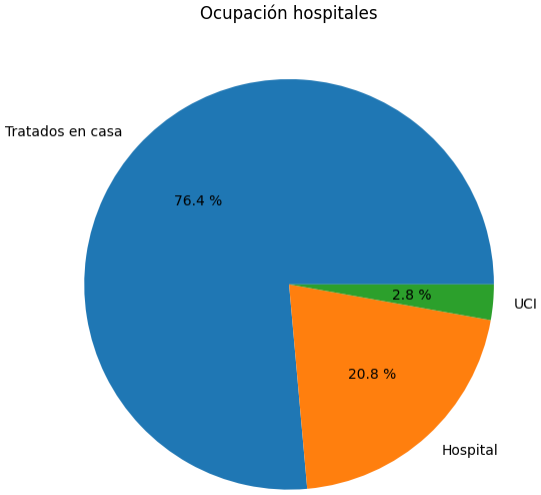

Figura 12: Ocupación hospitales

5. Conclusiones

- Existen distintos tipos de web scraping, además de distintas herramientas provistas para poder realizarlo, esto puede variar con base al software que se usara para la extracción de datos

- El uso adecuado de las librerías que ofrece python facilita la realización del programa y su eficiencia.
- Python es un lenguaje sumamente flexible y dinámico, provee una amplia cantidad de herramientas que buscan ofrecer efectividad sobre los procesos, y debido a esto es un lenguaje sumamente fácil de usar y su interacción con distintos tipos de datos suele ser muy flexible y simple.

- Almacenar los datos en una Base de Datos permite que sean mas accesibles y fáciles de adquirir posteriormente. MySql permite crear estas bases de datos, y facilita la conectividad entre los programas.

- Este proyecto nos dará una visión mas limpia y concreta acerca de los casos de Covid-19 en Colombia, no solo porque trataremos la información sensible, si no que permitirá que otras personas puedan informarse con nuestro programa.

- Antes de realizar una gráfica es importante ver que valores se están comparando ya que el cambio de estos datos se podrá ver mejor si se selecciona en pequeños grupos que están directamete relacionados entre sí.
- La tarea principal de una gráfica en mostar al lector un valor o porcentaje de interés para el.

Referencias

[1] M. Lutz, *Learning Python*, 2013. [Online]. Available: https://cfm.ehu.es/ricardo/docs/python/Learning_Python.pdf

[2] OMS, *Preguntas y respuestas sobre la enfermedad por coronavirus (COVID-19)*, 2020. [Online]. Available: https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019/advice-for-public/q-a-coronaviruses?gclid=CjwKCAjw8MD7BRArEiwAGZsrBSQWuE0uPnTjyhm1ghBMWQywIOcoj5NtEs1lrriQ-ZPT_TbVxjWfDxoCh-sQAvD_BwE

[3] Varios, *Relacion señal-ruido*. [Online]. Available: https://la.mathworks.com/help/signal/ref/snr.html

[4] G. Signal, *Extracting Data from HTML with BeautifulSoup*, 2019. [Online]. Available: https://www.pluralsight.com/guides/extracting-data-html-beautifulsoup

[5] I. Naokil, *PyMySQL*. [Online]. Available: https://github.com/PyMySQL/PyMySQL

[6] Varios, *Pandemia de enfermedad por coronavirus de 2020 en ColombiaL*. [Online]. Available: https://es.wikipedia.org/wiki/Pandemia_de_enfermedad_por_coronavirus_de_2020_en_Colombia