INFORME DEL 50%

DESARROLLO DE UN ALGORITMO PARA DAR SOLUCIÓN A PQRS DE LA DIRECCIÓN DE INVESTIGACIONES Y TRANSFERENCIA (DIT).

SANTIAGO ANDRES ESPINAL MENDOZA

MATEMATICAS DISCRETAS II
ESTRUCTURAS DE DATOS
BASES DE DATOS I

JULIAN DARIO MIRANDA CALLE
ROSAURA GUTIERREZ ALMEYDA
JUAN SEBASTIAN GOMEZ ROSAS

UNIVERSIDAD PONTIFICIA BOLIVARIANA
ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2020

ADQUISICION DE DATOS:

Se inició con un barrido de información directamente de las oficinas del Departamento de Investigaciones y Transferencias (DIT), mediante preguntas directas con administrativos, asesores, docentes y estudiantes que hacen parte del departamento o que hayan tenido alguna experiencia directa con DIT. Por consiguiente, la información regida se administra para la adquisición de datos para tener respuestas más puntuales y naturales.

Otro proceso de adquisición de datos es con el método de web scraping, se puede definir como la técnica por la que un equipo de desarrolladores es capaz de rascar, escrapear o liberar datos de páginas web de gobiernos, instituciones públicas u organizaciones para acceder a datos privados o públicos que puedan ser publicados o distribuidos en formato abierto. El problema es que la mayoría de los datos de interés están en formatos no reutilizables y poco transparentes como un PDF si no en paginas con información encriptada y demás procesos de Proción de datos. [1]

Mediante fórmulas de ImportHTML: Dentro de las aplicaciones de Google, el gran buscador desarrolló su propio Excel llamado Google Spreadsheet (las hojas de cálculo de Google). Esta herramienta dispone de casi todas las características de Microsoft Excel, pero además dispone de algunas funcionalidades añadidas gracias al contenido indexado en internet por el buscador: lectura de feeds RSS, cambios en páginas web o extracción de datos. Todo esto es posible mediante el uso de fórmulas como ImportFeed, ImportHTML e ImportXML. Con la segunda de ellas, cualquier usuario puede extraer datos de tablas o listados de forma ordenada desde cualquier página web. Dependiendo de si es una tabla o una lista, el tipo de fórmula varia en uno de sus elementos. Dos ejemplos prácticos:

=IMPORTHTML("url página web", "table", 2)

=IMPORTXML(“url pagina web”, “consulta_xpath”)

A continuación, una demostración de la adquisición de datos por este método:

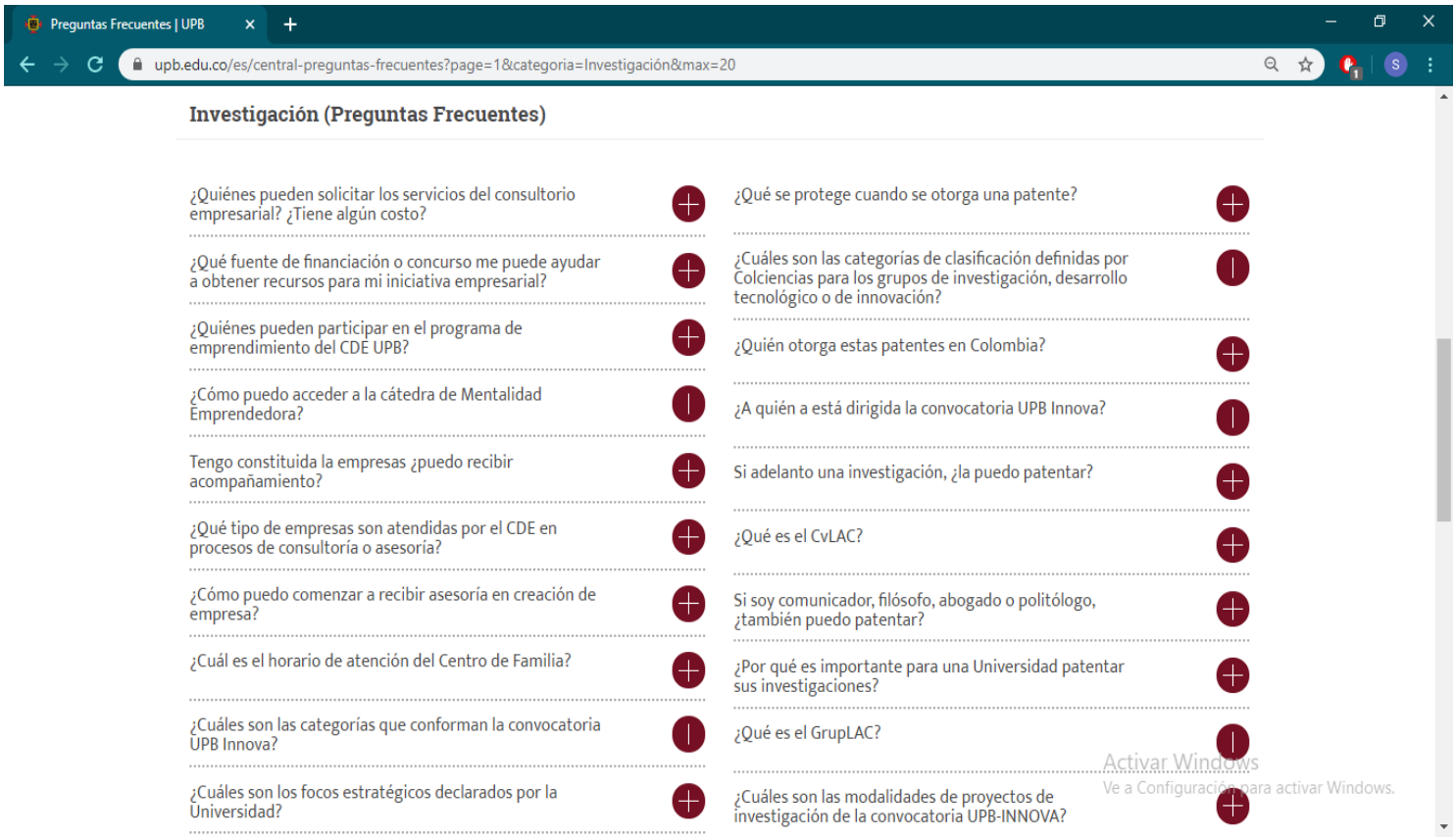


Imagen 1 tomada de: https://www.upb.edu.co/es/central-preguntas-frecuentes?page=1&categoria=Investigaci%c3%b3n&max=20

Como primer paso nos ubicamos en la página web de interés, en este caso es la pagina de la Universidad Pontificia Bolivariana en la cual se encuentra una sección de preguntas frecuentes incluyendo una búsqueda más directa en diferentes áreas de la universidad, que en nuestro caso nuestra área de interés es la de Investigación. Al seleccionar el área deseada, se despliega una serie de preguntas frecuentes con sus respuestas como se nota en la Imagen 1.

Como segundo paso seleccionamos una de las columnas para inspeccionar por HTML para copiar su Xpath como se ve en la siguiente imagen:

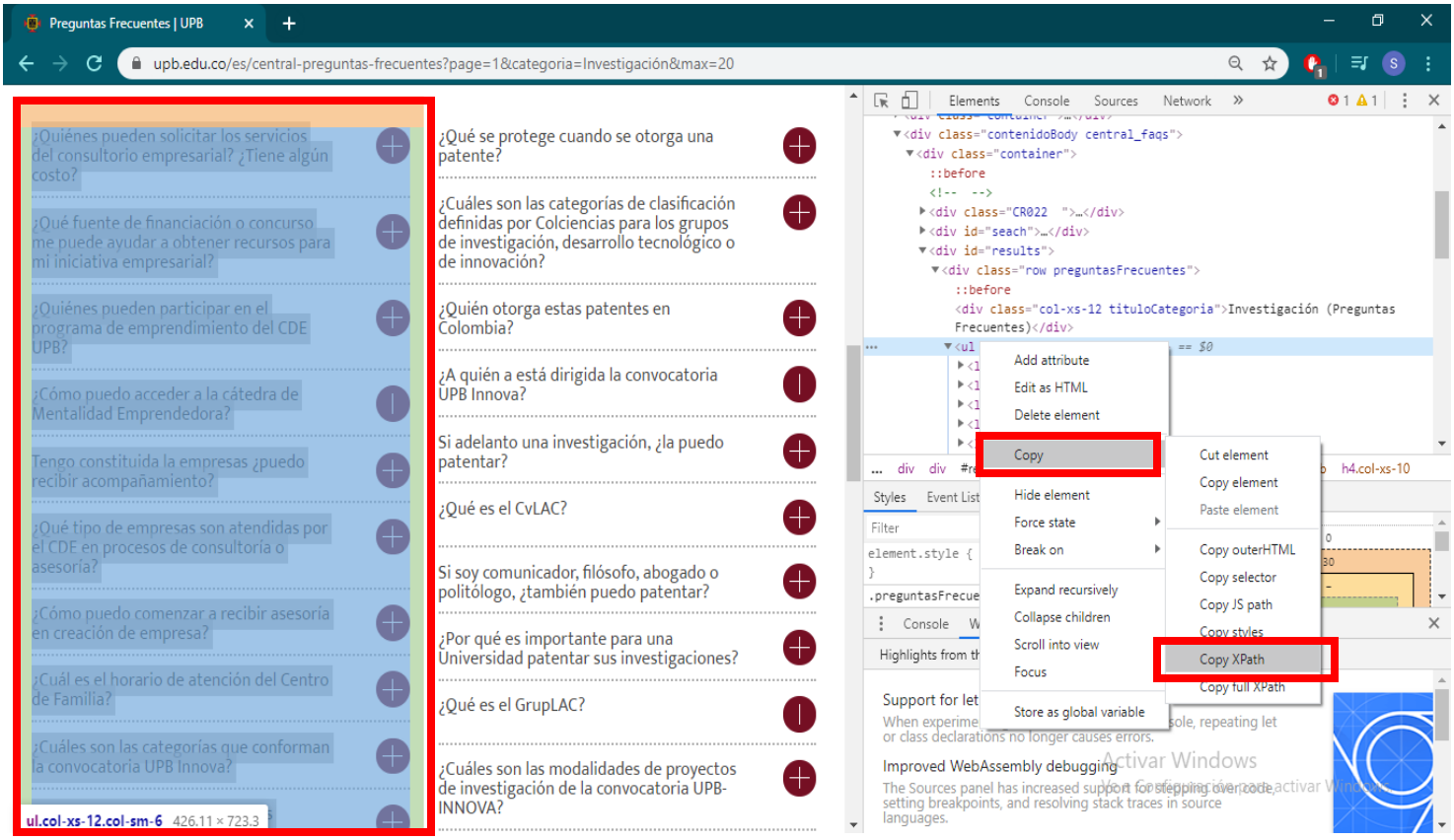


Imagen 2 tomada de: https://www.upb.edu.co/es/central-preguntas-frecuentes?page=1&categoria=Investigaci%c3%b3n&max=20

Como tercer paso, nos dirigimos a las herramientas de Google en su menú principal e ingresamos a la hoja de cálculo de Google:

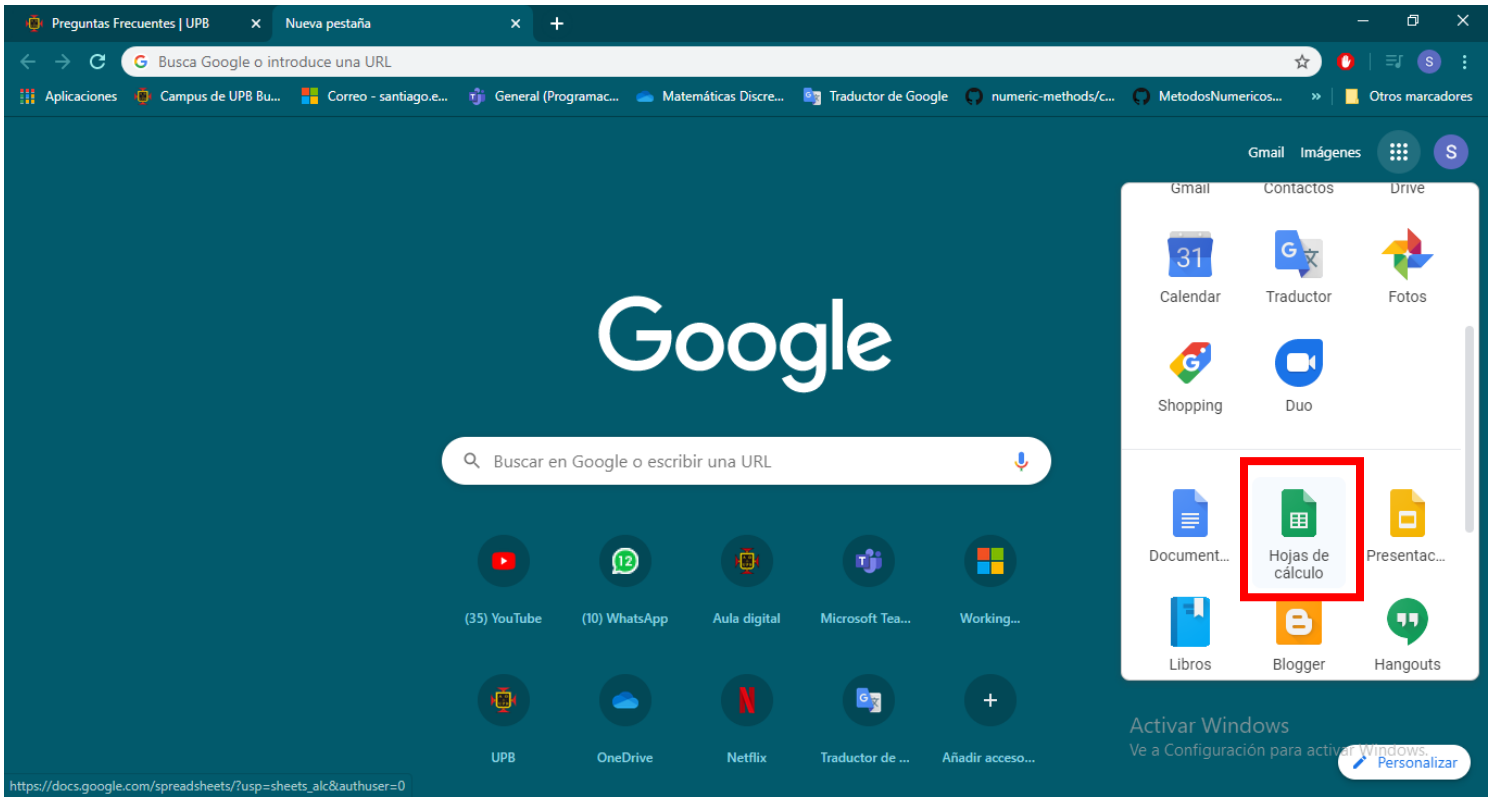


Imagen 3 tomada de: Nueva Pestaña de Google desplegando el menú principar de heramientas de Google

Luego abrimos una hoja nueva y le colocamos el comando = IMPORTXML () , dentro del cual le agregamos la url de la pagian donde consultmos y separado por punto y coma, colocamos la direccion Xpath.

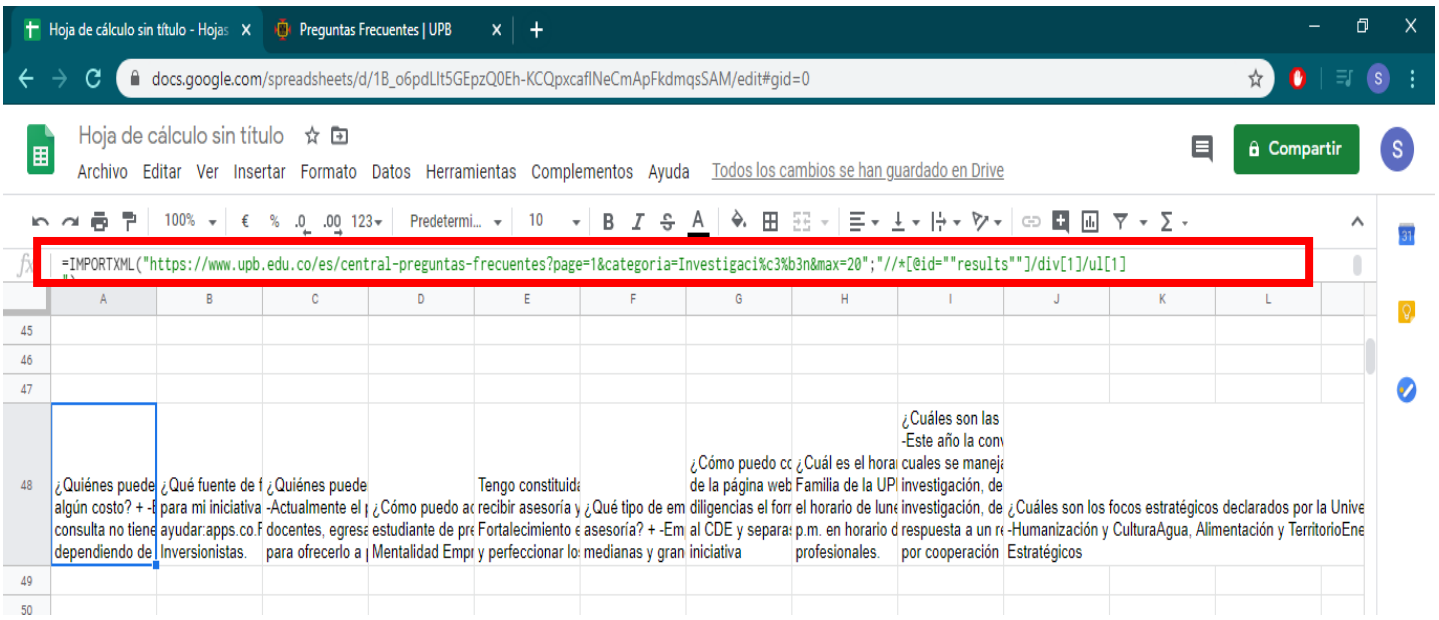


Imagen 4 tomada de: Hoja de calculo de Google

Para mayor comodidad al momento del manejo de datos, vamos a utilizar la funcion =TRANSPONER(), que al momento de utlizarla nos colocará de forma vertical los datos para que al momento de convertirlo en un archivo .csv sea una columnas con distintas filas.

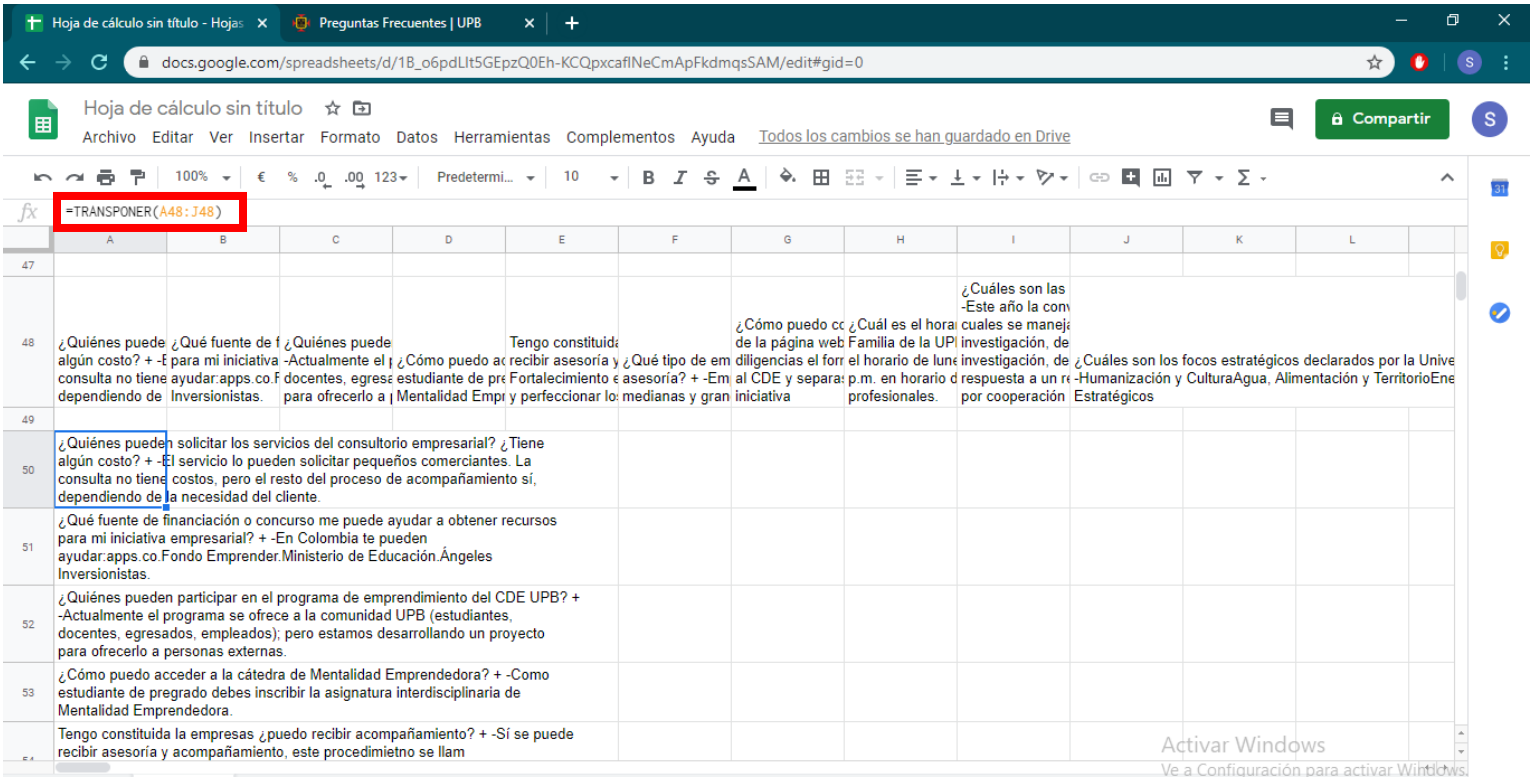
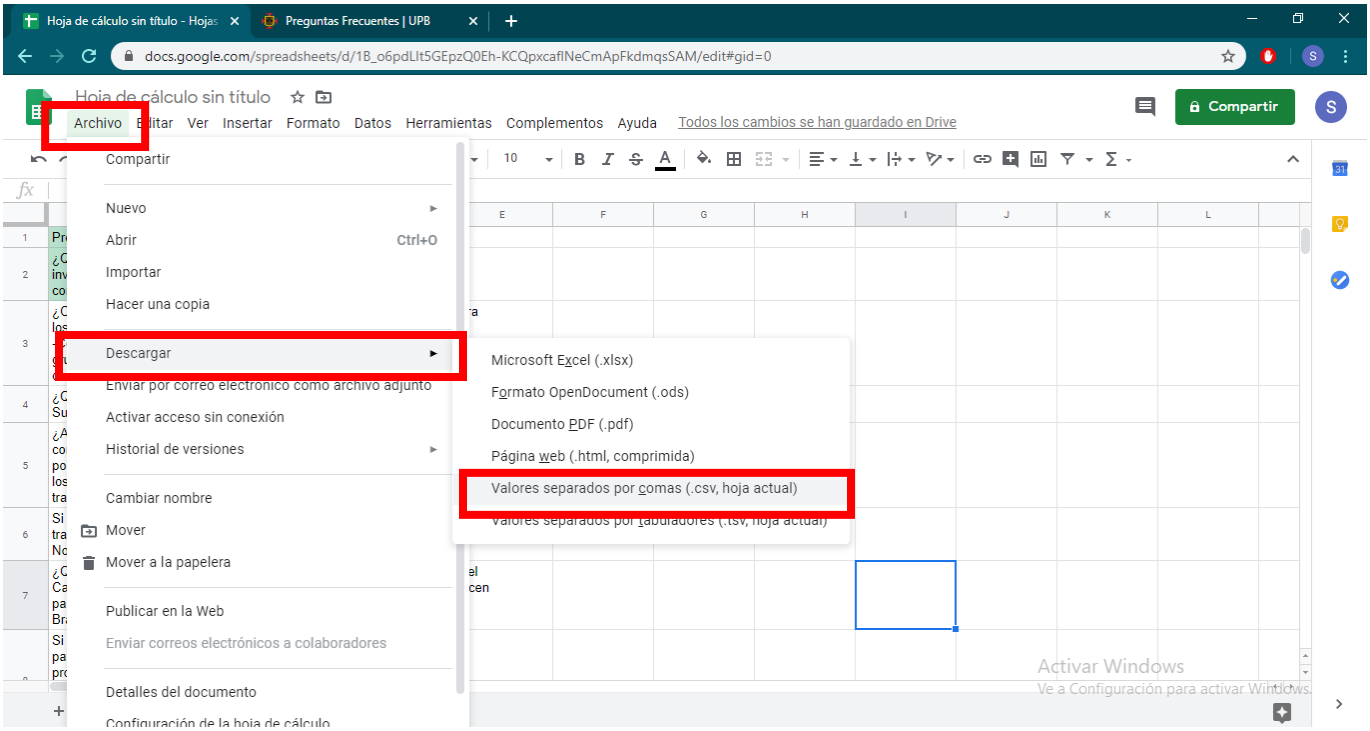


Imagen 5 tomada de: Hoja de calculo de Google

Despues de haber relizado este proseso, se repite con las columnas restantes en la pagina para posteriormente guardarlo como archivo .csv para poder consultarlo al moenteo de realizar la preparacion de los datos.



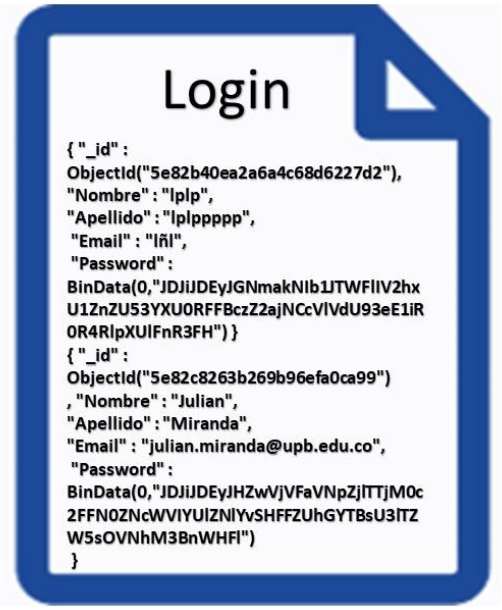
DIAGRAMAS:

En la siguiente parte del informe encontrará los modelos de la base y tambien el Diagrama de clases.

Diagrama de la base de datos:

La base de datos que implemetamos es MongoDB, una base de datos noSql, que es una conocida clase de sistemas de gestión de bases de datos que deiverge del modelo clásico de SGBDR (Sistema de Gestión de Bases de Datos Relacionales) en aspectos importantes, siendo el más destacado que no usan SQL como lenguaje principal de consultas. Los datos almacenados no requieren estructuras fijas como tablas, normalmente no soportan operaciones JOIN.[2]

En el algoritmo se utilizaran 2 documentos, en donde un docuemento almacenará los usuarios con la informacion solicitada en el algoritmo, como nombre, apellido, email y la contraseña para ingresar, dicha contraseña estará con cifrado para protección de los datos. En la siguiente imagen se muestra un ejemplo del documento con la información:



=

ID	Nombre	Apellido	Email	Password
ObjectId("5e82b40ea2a6a4c68d6227d2")	lplp	lplppppp	lñl	BinData(0,"JDJiJDEyJGNmakNlb1JTWFllV2hxU1ZnZU53YXU0RFFBczZ2ajNCcVlVdU93eE1iR0R4RlpXUlFnR3FH")
ObjectId("5e82c8263b269b96efa0ca99")	Julian	Miranda	julian.miranda@upb.edu.co	BinData(0,"JDJiJDEyJHZwVjVFaVNpZjlTTjM0c2FFN0ZNcWVlYUlZNlYvSHFFZUhGYTBsU3lTZW5sOVNhM3BnWHFl")

Vemos que se encuenta la formación de cada usuario la cual la representamos en una tabla, en donde se observa detalladamente la representacion de los datos en el domuento. En el ultimo docuento se encuentra los datos que fueron ya tratados en el proceso de preparación y preprocesamiento de los datos obtenidos de la adquisición de datos, en la siguiente imagen se muestra un ejemplo del documento con la informacion:

=

ID	Contenido
ObjectId("5e72c221600da6d43587a739	contenido
ObjectId("5e72c221600da6d43587a739	Contenido para respuestas
ObjectId("5e72c221600da6d43587a739	Contenido para manejoextra

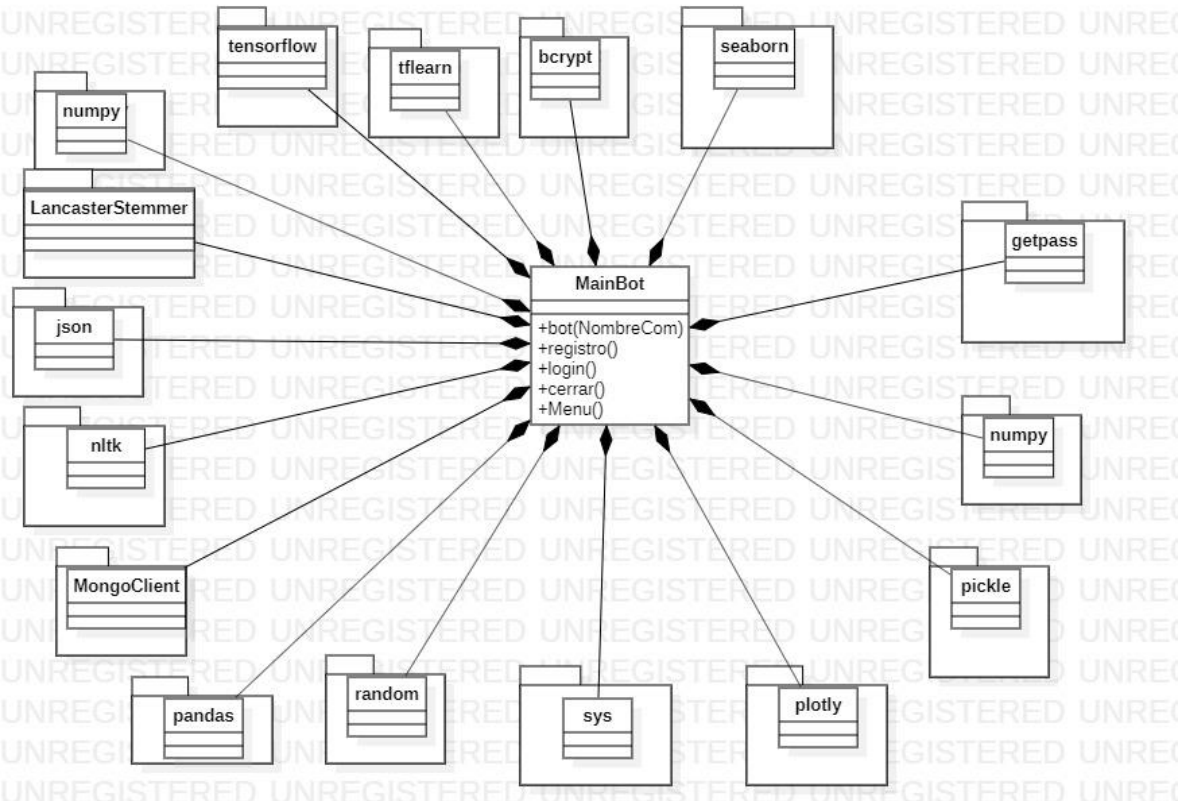
=

tag	Patrones	Respuestas
Saludo	Hola, Un saludo	Un gusto verte, Hola que tal!
Despedida	Adios, Nos vemos	Cuidate, Nos vemos pronto

La informacion que se encuentra es para la consultas para las respuestas para el algoritmo, como tambien tags y demas, la cual se encuentra tambien representada en las tablas.

Diagrama de clases:

Se identificaron los paquetes necesarios para el algoritmo, los cuales fueron estos:



Tenemos una clase principal que es MainBot en la cual se encuentra las funciones mas importantes que son dadas por cada uno de los paquetes.

BIBLIOGRAFIA:

[1] BBVAOpen4U. (2016). Herramientas de extracción de datos: para principiantes y profesionales. [online] Available at: https://bbvaopen4u.com/es/actualidad/herramientas-de-extraccion-de-datos-para-principiantes-y-profesionales [Accessed 5 Mar. 2020].

[2]"NoSQL", Es.wikipedia.org, 2019. [Online]. Available: https://es.wikipedia.org/wiki/NoSQL. [Accessed: 02- Apr- 2020].