|  |
| --- |
| [Nombre de la compañía] |
| TALLER 2 |
| Minería de Datos & Big Data: “Web Scraping” |

|  |
| --- |
| Integrantes: Cristian Álvarez – Jorge González  Profesor Mauricio Sepúlveda |

****

# 

[**Introducción**](#_wn6n9bdxzsp2) **3**

[**Desarrollo**](#_e8jfkxst3e26) **3**

[Parte 1. Implementación de Código](#_8fq6l9fwsskz) 3

[Parte 2. Análisis LDA](#_x6xjl2d439jk) 4

[Parte 3. Modificación Web Scraping](#_isasafxakm2o) 5

[Parte 4. Serie de Tiempo y red LSTM](#_5t6kfa152tc6) 5

[Parte 5. Clasificaciones y Agrupaciones.](#_w2scbllcafum) 5

[**Conclusión**](#_qj5uvxgjxdzu) **5**

[**Anexos**](#_9usu94t8za7g) **5**

[**Referencias**](#_jj0qta3sx0wd) **5**

# 

# 

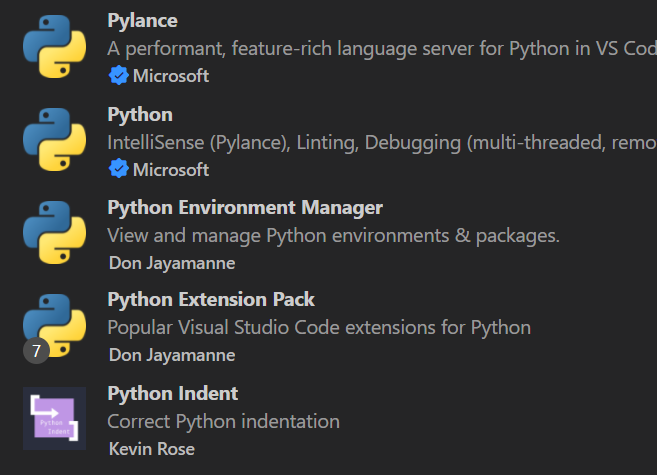
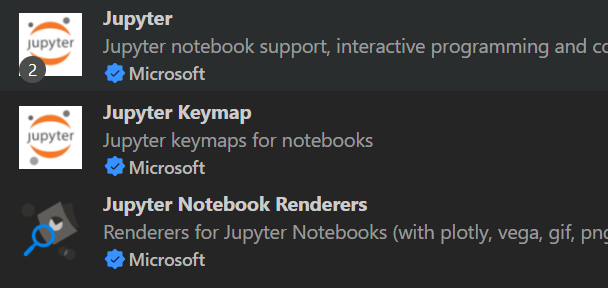
Como bien sabemos dentro de toda la red de internet existe una cantidad inimaginable de información y datos de todo tipo, donde estos datos nos sirven tanto para realizar múltiples funcionalidades que a su vez intervienen muchos factores, pero en este caso nos interesa el tema de extracción y análisis de estos datos. Para poder realizar y facilitar la extracción de grandes cantidades de datos que nos resultan importantes para nuestros trabajos o análisis posteriores es que existe el Web Scraping, una metodología el cual consiste en extraer datos de manera automática y programada.

El Web Scraping, como se comentaba, es una técnica utilizada para extraer grandes cantidades de información desde sitios web, como datos de contacto, correos electrónicos o como en nuestro caso términos de búsqueda. Para luego poder guardar nuestros datos en bases de datos locales con el fin de procesarlos.

En este informe utilizaremos esta técnica y recopilaremos información desde sitios de publicaciones científicas relacionados a Security TI, tema que decidimos enfocarnos a investigar.

# **Desarrollo**

Para empezar con el desarrollo del “Taller 2”, indicaremos las herramientas/software que nosotros utilizamos, siendo uno de ellos es Visual Studio Code(VSC) para trabajar a través de la herramienta que ofrece Anaconda Navigator, ya que es donde mas cómodos nos sentíamos para trabajar. Además, mencionar que adaptamos el VSC para utilizar Jupyter, como también algunas extensiones para trabajar con Python. Le adjunto imagen de las extensiones que utilizamos para hacer posible el uso de Jupyter y Python.

Otro punto a considerar es que en todos los archivos copiamos las mismas importaciones, pero no todas fueron utilizadas, las colocamos en caso de que fuera necesario, por ello algunas no fueron requeridas, pero están en los documentos igualmente. Agregar que en las imágenes que colocaremos se puede apreciar un color mas oscuro en las importaciones, el cual indica que no están siendo utilizadas en ese documento, por lo que no hace falta llamarlas. Ejemplo:



## **Parte 1. Implementación de Código**

Para esta primera etapa se nos pide implementar un código para extraer la información desde un sitio de publicaciones científicas, en específico “Web Of Science” y buscar artículos sobre un tema, que en nuestro caso quisimos seleccionar “Security TI” entre los años 2010 y 2021.

Por ello a través de web Scraping debemos lograr que se pueda llegar a la pagina destinatario con los filtros especificados para posteriormente extraer la información.

En cuanto a la recopilación de datos estos se almacenaron en un archivo Excel con las siguientes columnas: “ID”, “TÍTULO”, “MES”, “AÑO”, “AUTORES”, “REVISTA”, “CITACIONES” y “REFERENCIAS”. Para el tópico de “Security TI”, el resultado de la extracción de datos fue de un total de 750 registros.

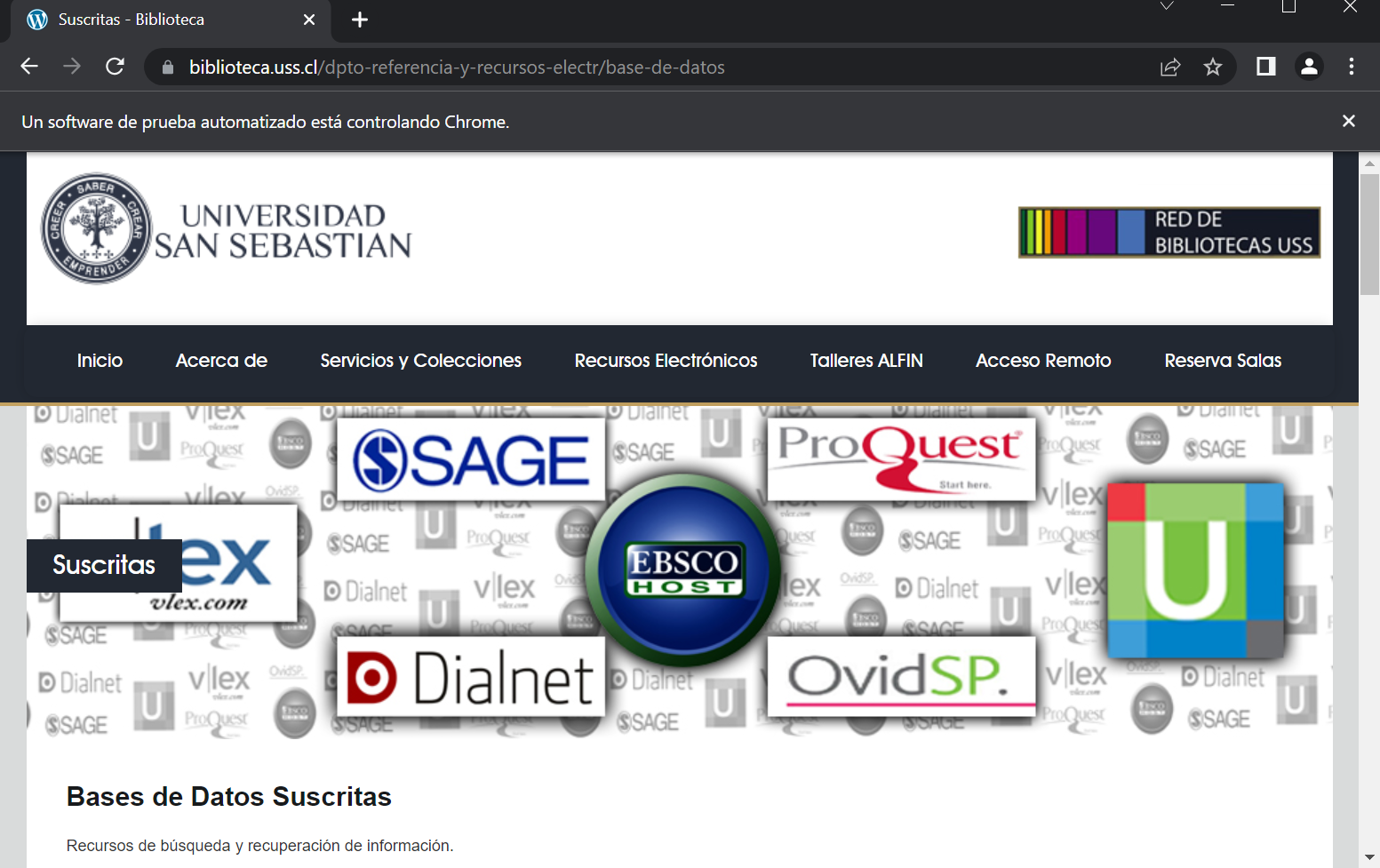
### Código

A continuación, le dejamos el link del GitHub para que puedan ver todo el código que se realizo para cumplir con lo requerido

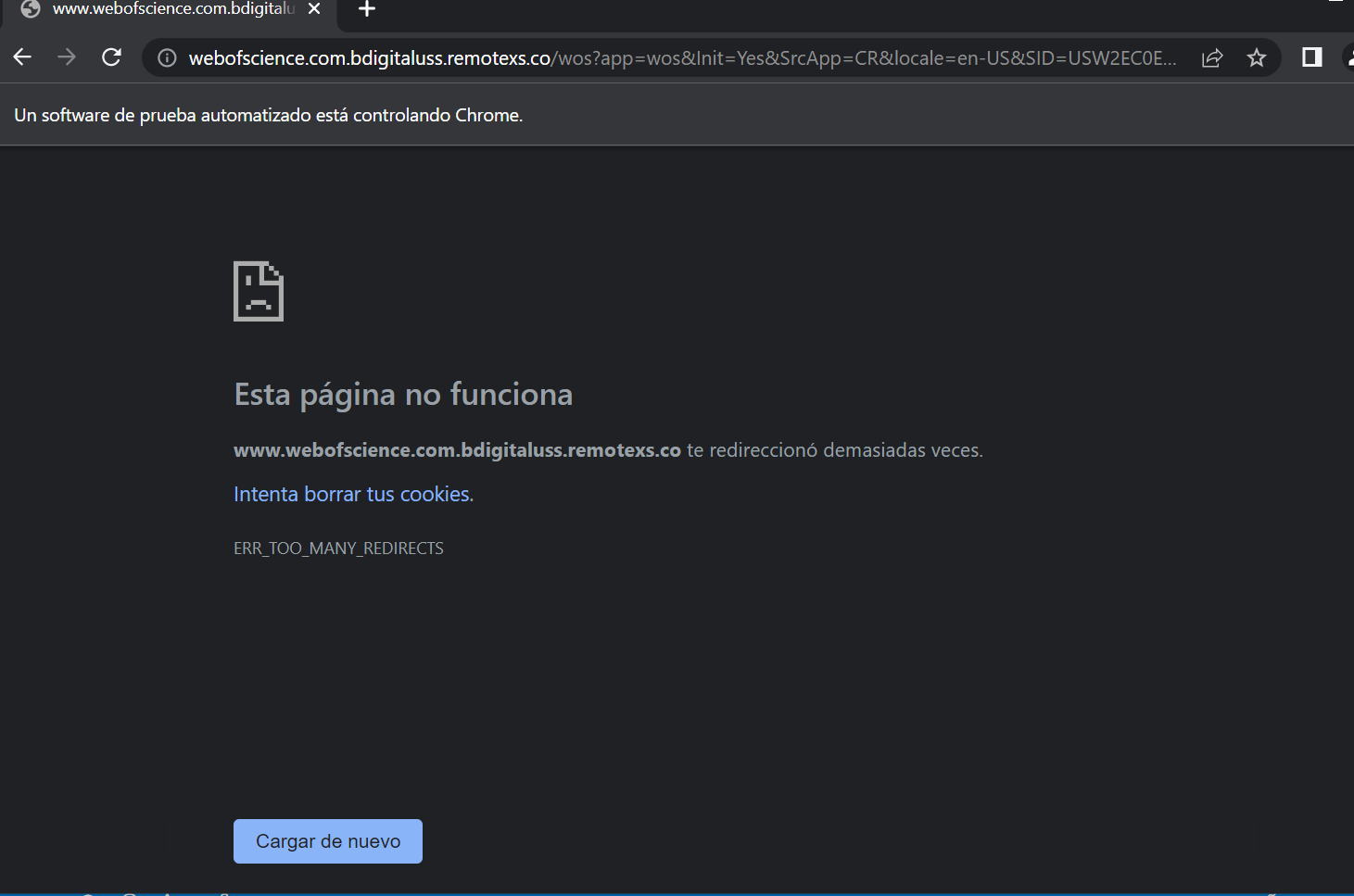
<https://github.com/jorgeigna/USS-DMDB-Taller2/blob/main/Parte1.ipynb>

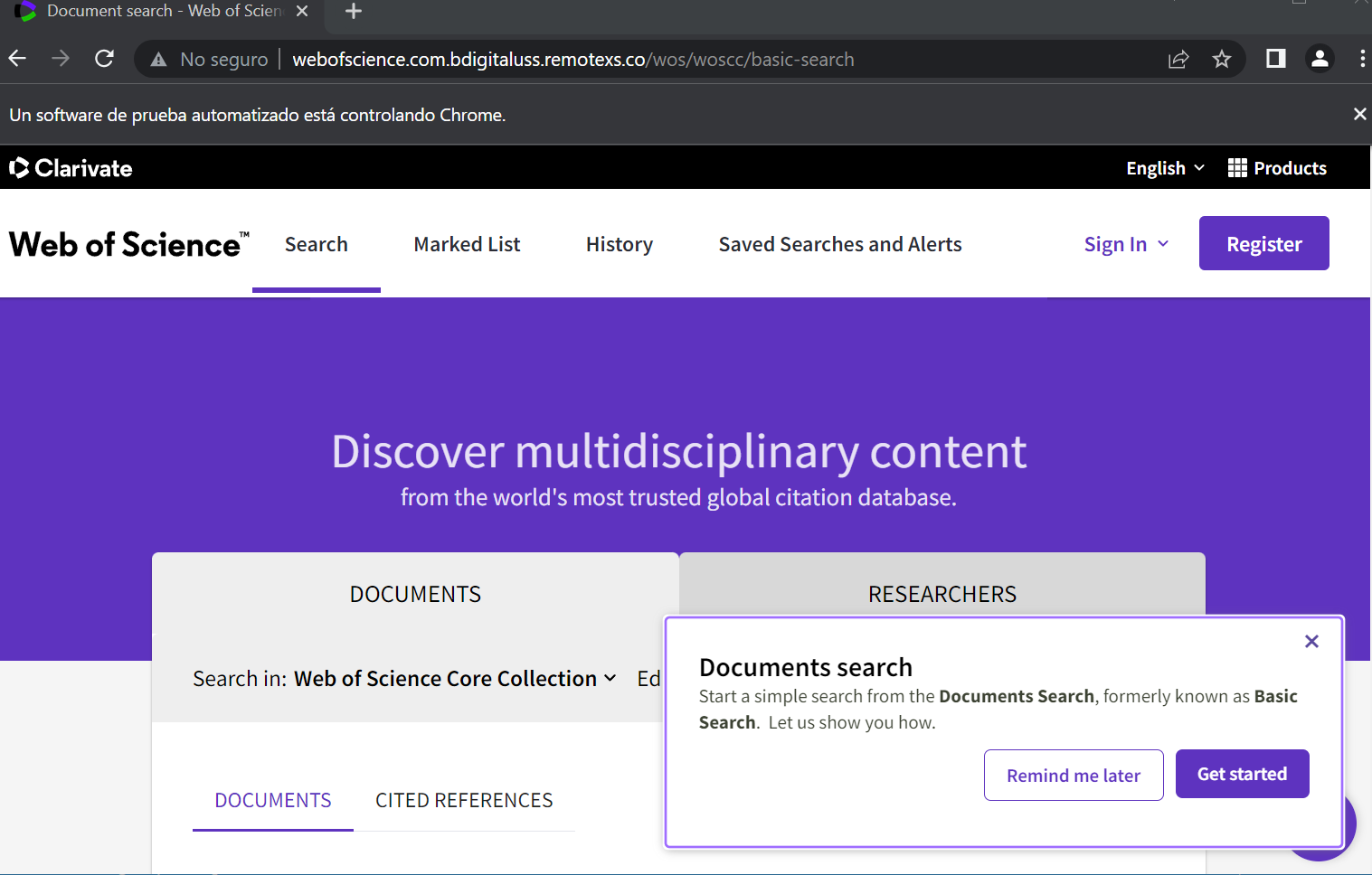
### Resultados

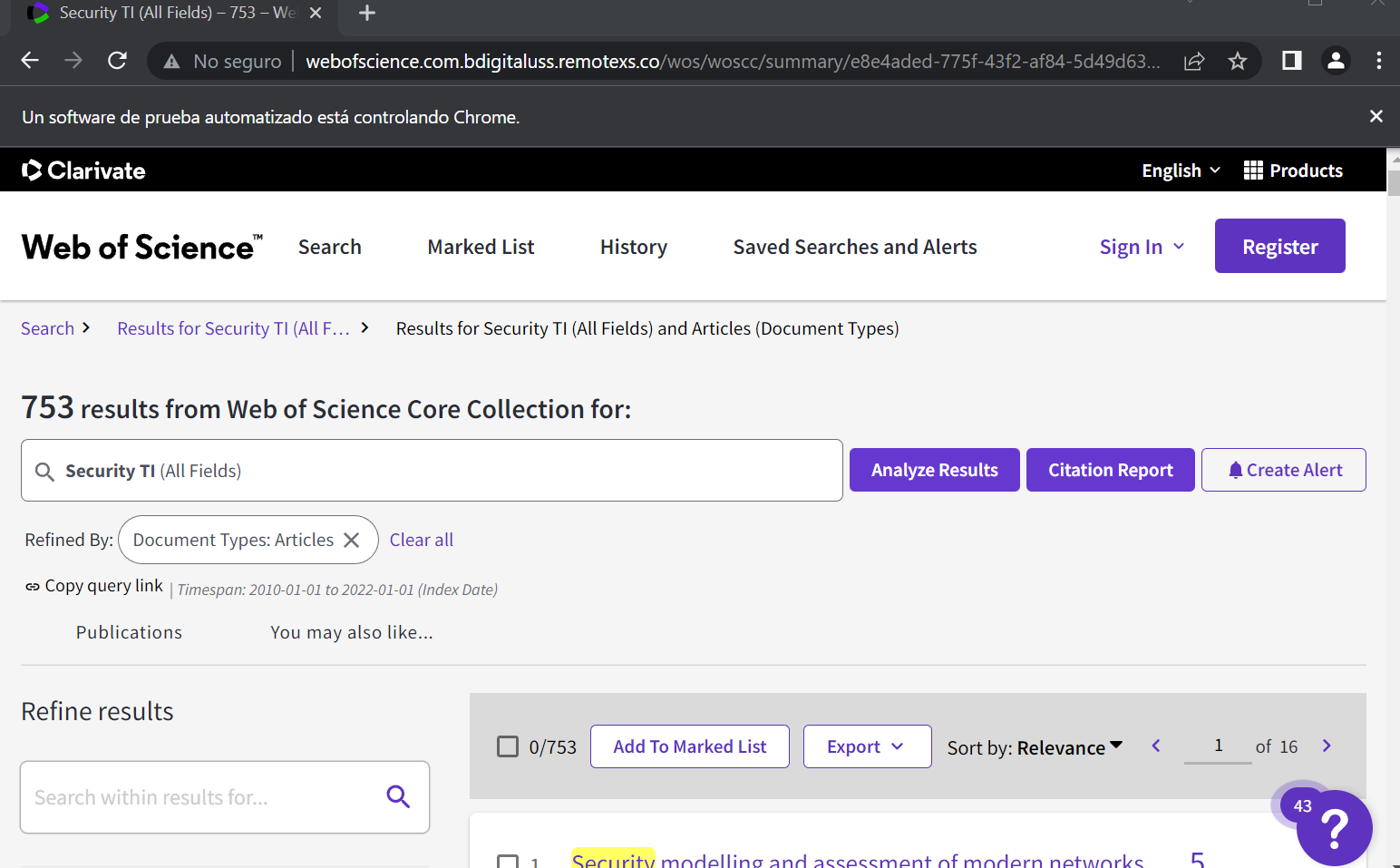
Iniciamos en el link entregado el cual se selecciona el apartado “Web Of Science”, luego se inicia sesión en la universidad ingresando usuario y contraseña

 (*Imagen Pagina Inicial*)

Debido a un mal rendimiento general de la pagina esta no carga, por lo que volvemos a cargar la página y así poder continuar con el ingreso de la búsqueda “Security TI”, abrimos la opción de filtrar por fecha “2010-01-01” hasta “2022-01-01”, aceptamos los parámetros ingresados y en la siguiente pagina hacemos filtro para ver solo los artículos. Así acaba la primera parte de este punto.

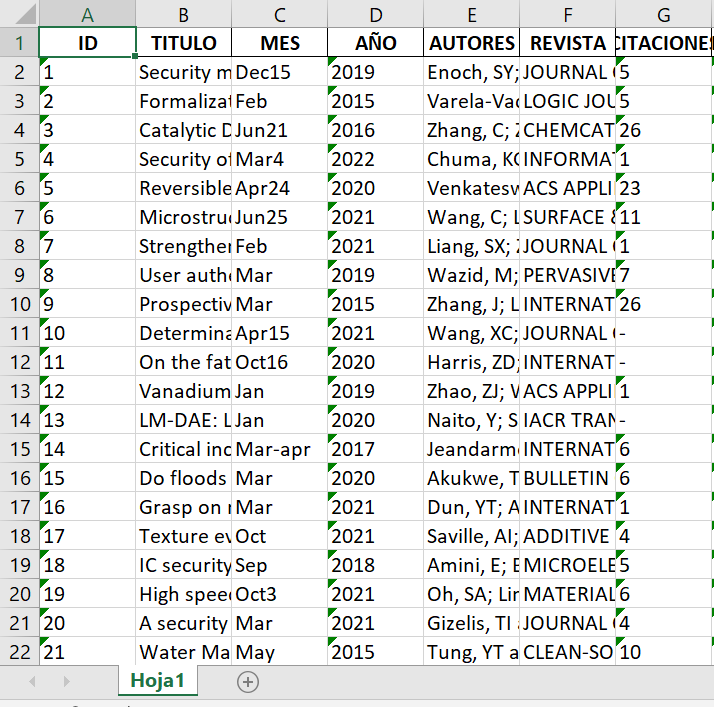






Ahora es cuando comienza la extracción de datos el cual extraemos “ID”, “TÍTULO”, “MES”, “AÑO”, “AUTORES”, “REVISTA”, “CITACIONES” y “REFERENCIAS” de cada uno de los artículos y recorriendo uno a uno de ellos, como también todas las páginas de la búsqueda que estamos haciendo.

Esta información la almacenamos un archivo Excel llamado “DatosInvestigacion.xlsx”. donde podrán encontrar no solo ese archivo sino la copia llamada “**DatosInvestigacion1-CargaCompleta.xlsx**” que contiene todos los datos almacenados para ver el resultado.



## **Parte 2. Análisis LDA**

Para esta etapa, se nos pide implementar un análisis LDA, el cual nos permite clasificar un texto y así generar Topics para cada revista y tener un análisis de ello.

En este caso utilizamos el archivo Excel donde se obtuvieron los datos previos, eliminamos las columnas, quedándonos solo con las de “TÍTULO” y “REVISTA”, lo realizamos mediante un procedimiento llamado Wordcloud. Luego hacemos esta llamada nube de palabras para determinar la relevancia de estas.

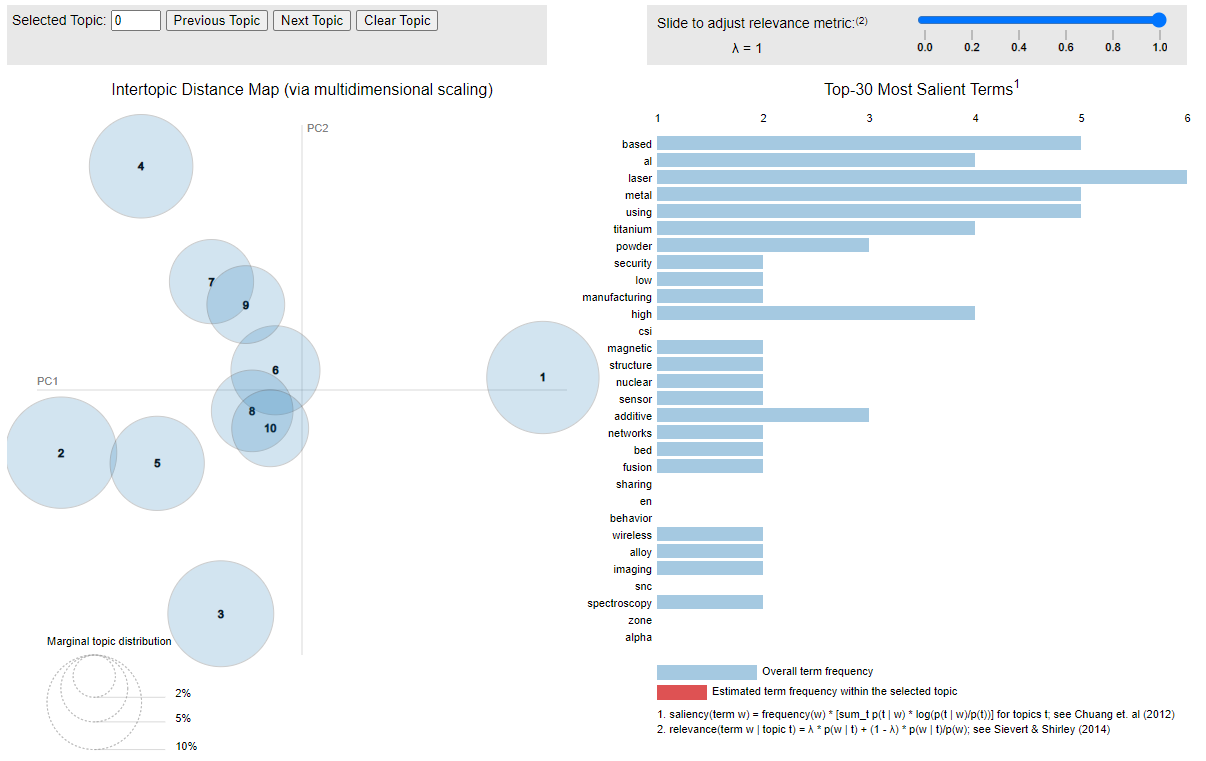
Para finalizar con la ayuda de la librería “pyLDAvis”, visualizamos de una manera interactiva los clústeres que se crearon con el análisis LDA.

### Código

El código lo puedes encontrar en el siguiente link

<https://github.com/jorgeigna/USS-DMDB-Taller2/blob/main/Parte2.ipynb>

Resultados

Como resultados podemos ver el siguiente panel de análisis que nos proporciona la librería de “pyLDAvis” donde podemos seleccionar los Topic definidos en el código, pero debemos tener en cuenta que la librería no puede mostrar más de 30 Topic. También este panel nos proporciona la frecuencia de cada palabra obteniendo así el top global de frecuencia, pero si seleccionamos uno de los topic podemos ver la frecuencia especifica de las palabras del topic.

## **Parte 3. Modificación Web Scraping**

Como su título lo dice, en esta etapa realizamos una modificación del Web Scraping, para obtener información adicional de “Author Keywords”, “Keyword Plus”, “Categorie” y “Research Areas”, estas columnas las añadimos a nuestro archivo previo de Excel. Y para ver los resultados dejamos un archivo Excel con la carga completa.

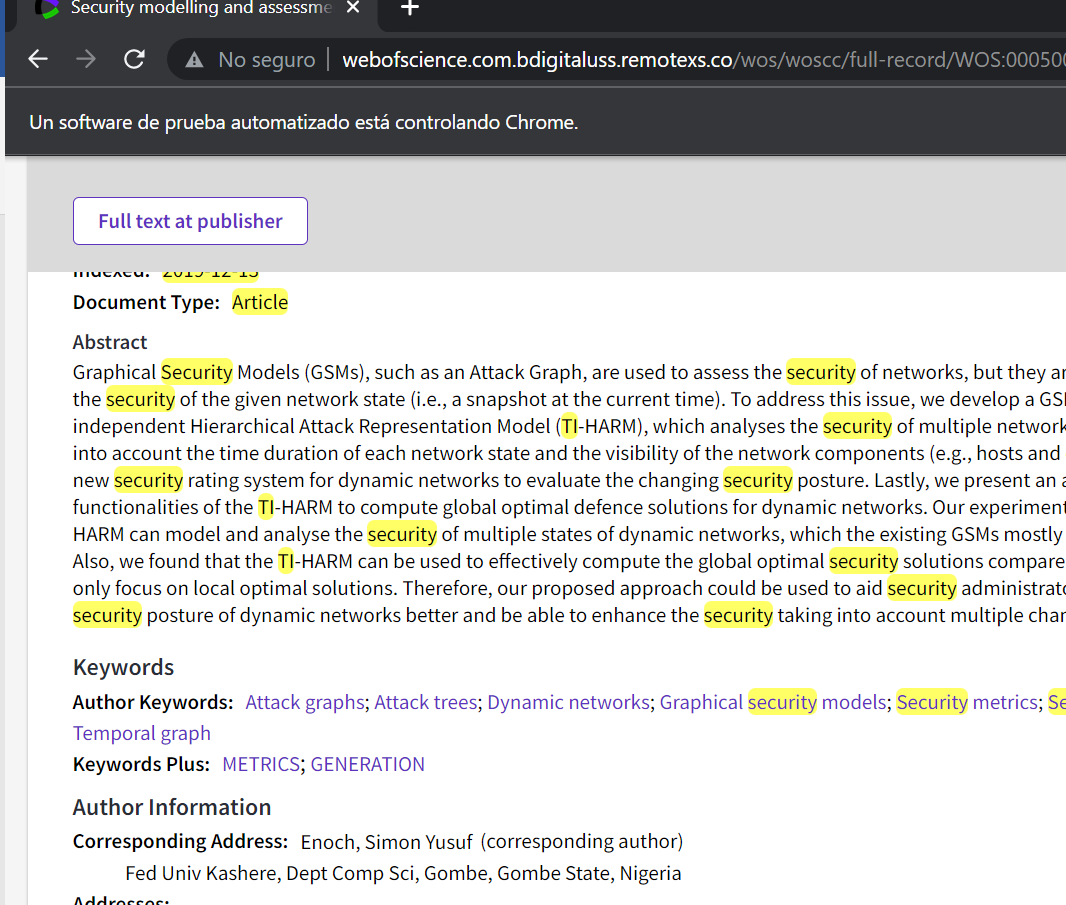
### Código

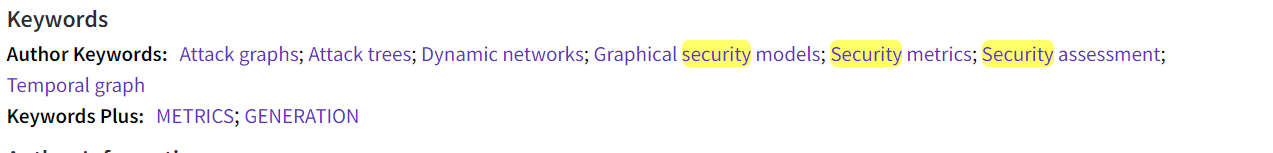
El código lo puedes encontrar en el siguiente link

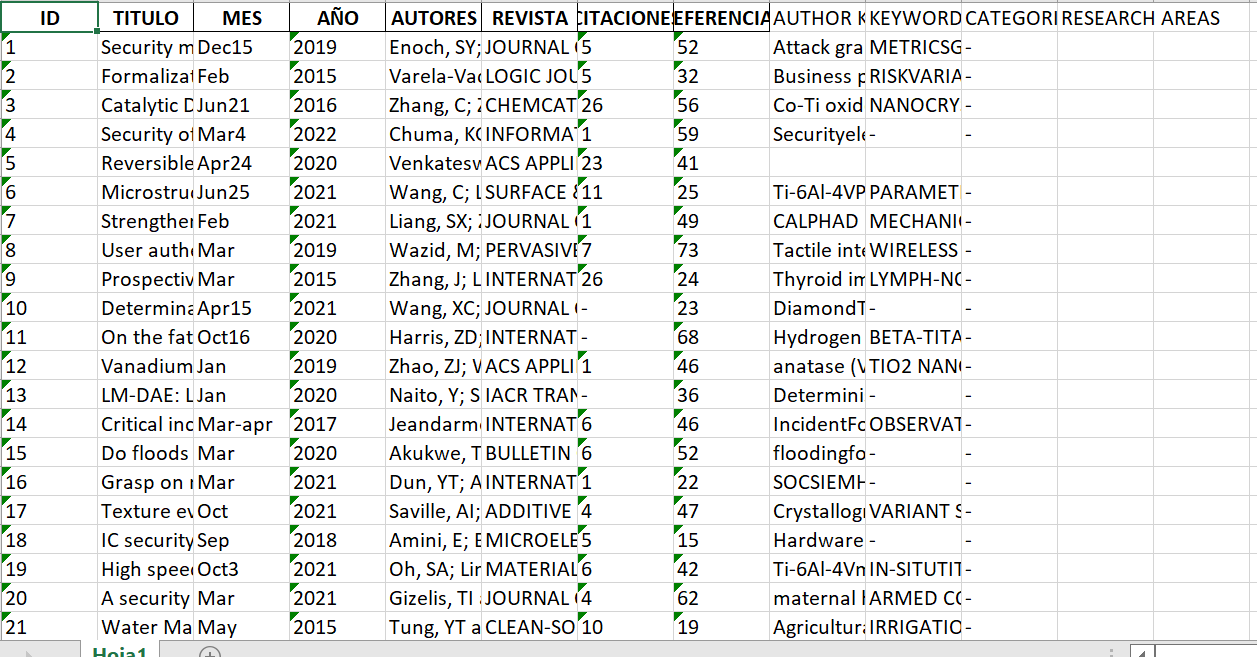
<https://github.com/jorgeigna/USS-DMDB-Taller2/blob/main/Parte3.ipynb>

### Resultados

Como resultado obtenemos la información de la pagina especifica de cada artículo por lo que debemos ir de articulo a articulo capturando los datos y almacenándolos en el archivo mencionado





## **Parte 4. Serie de Tiempo y red LSTM**

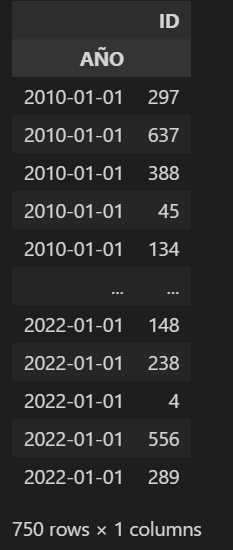
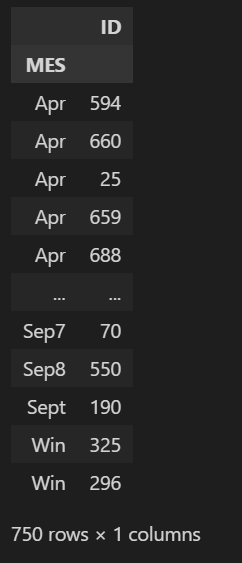
En este caso, se nos pide entrenar una red LSMT, generando series de tiempo para así predecir o pronosticar el número de publicaciones que se van a registrar en el año seleccionado. De esta forma tener un análisis un pronóstico tanto de periodos de año como de mes.

### Código

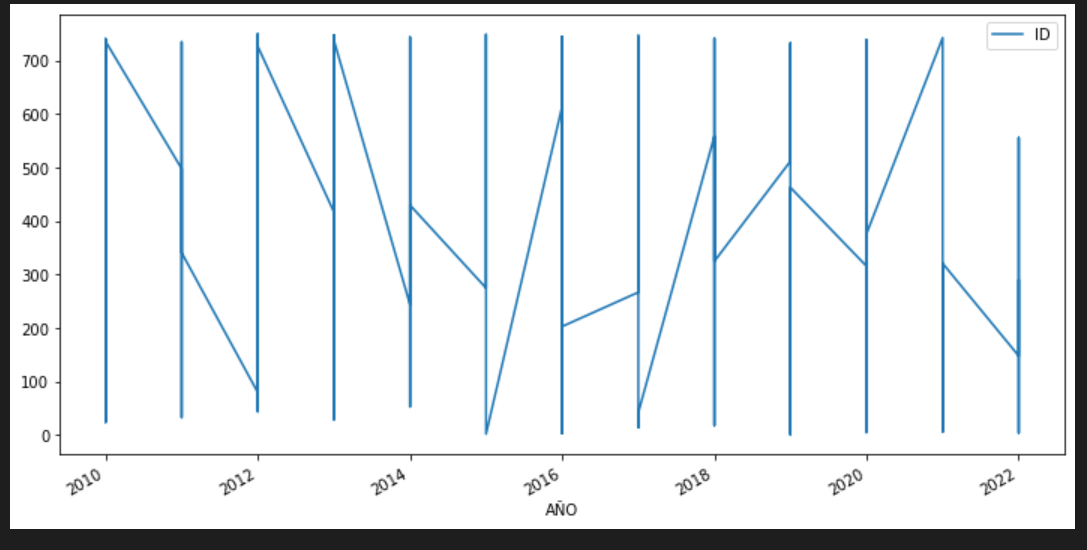
<https://github.com/jorgeigna/USS-DMDB-Taller2/blob/main/Parte%204.ipynb>

### Resultados

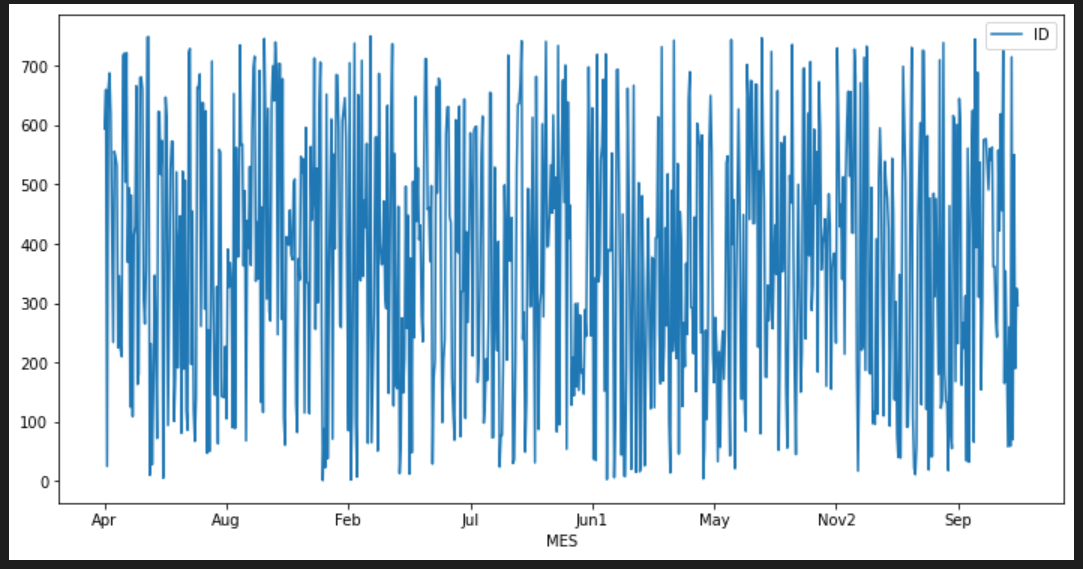
Como resultados obtenidos es la clasificación de los artículos ordenados tanto por año como por mes, además de ello ver como se manifiestan a lo largo del tiempo en relación con la cantidad de artículos existentes por fecha. Lo que quiere decir que se mostrara cada año y la cantidad de artículos realizados en ese periodo, por otro lado, en el caso de los meses podemos ver la cantidad de artículos realizados en cada uno de los meses.

Años vs Artículos



Meses vs Artículos



Debido a un problema con la librería Tensorflow, no se nos fue posible testear con la LSMT para ver el pronóstico esperado, pero en teoría si ejecuta el código teniendo la libraría podría ver resultados junto al pronóstico.

Así poder ver y apreciar que en los próximos meses se obtendrá cantidades aproximadas a la media de los datos que ya se tienen. Siendo una media de 50 artículos en el próximo año y una media de 7 artículos por mes. Mencionar que para esta búsqueda de “Security TI” obtuvimos un total de 750 artículos en 10 años, por lo que podemos considerar lógico que sean un número reducido los artículos nuevos que se pronostiquen a próximos años o meses.

## **Parte 5. Clasificaciones y Agrupaciones.**

En esta última etapa nos piden realizar clasificaciones y agrupaciones con nuevos datos extraídos, por ellos realizamos los mismos procedimientos que en la parte1 con la diferencia que filtramos solo para un año en específico.

### Código

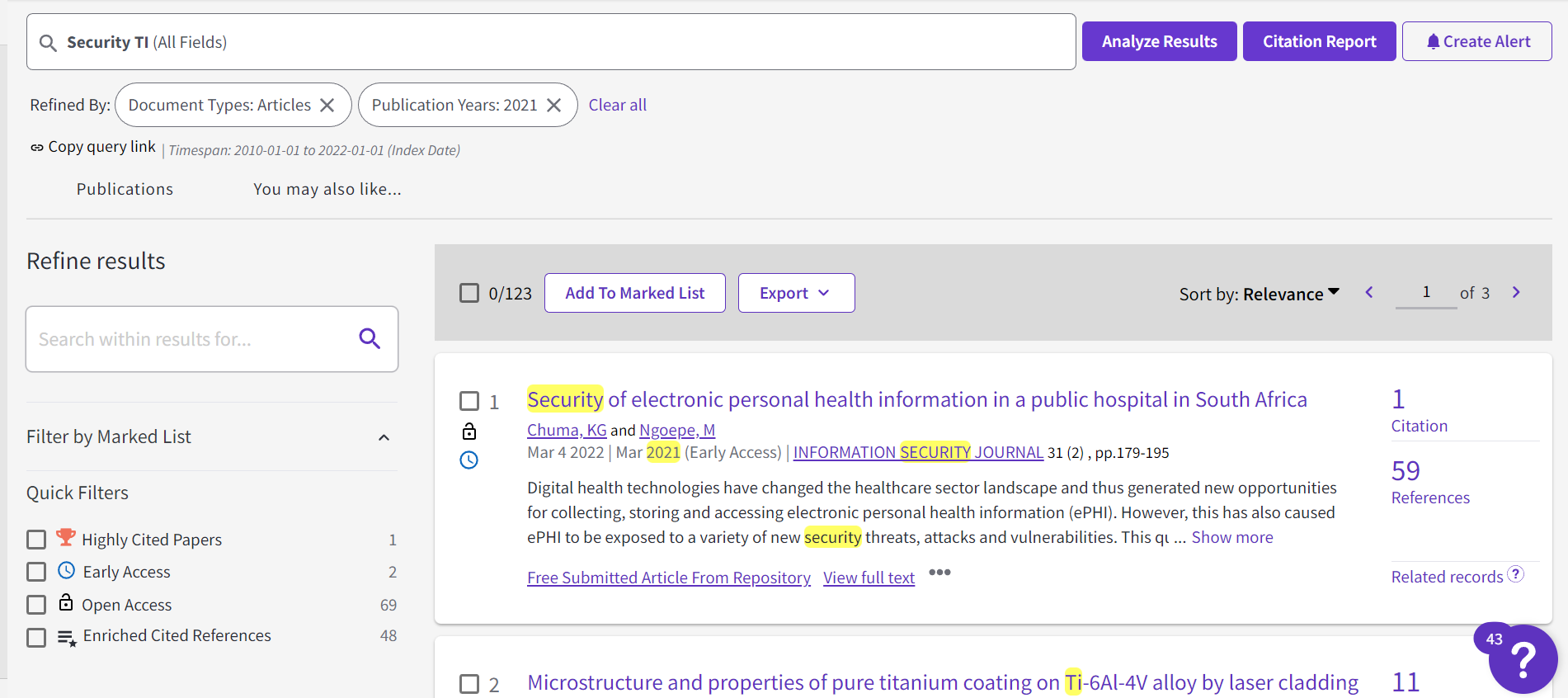
<https://github.com/jorgeigna/USS-DMDB-Taller2/blob/main/Parte5.ipynb>

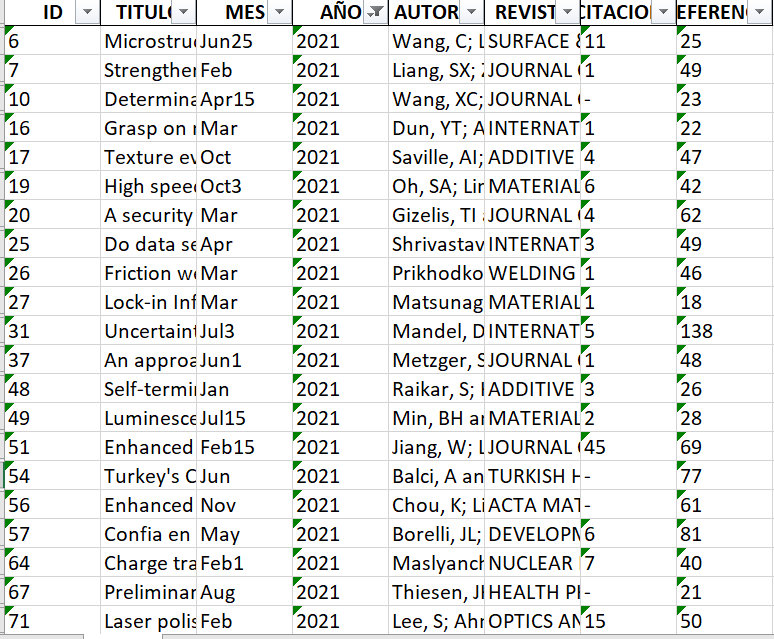
### Resultados

Una vez realizado los mismos procedimientos que en la parte 1, con la diferencia que consideramos realizar un filtro de un año en particular, continuamos con la extracción misma y obtenemos resultados.

Como resultado obtuvimos un nuevo archivo con los datos solo de un año anteriormente filtrado, además de ello recopilamos las palabras mas utilizadas en cada articulo

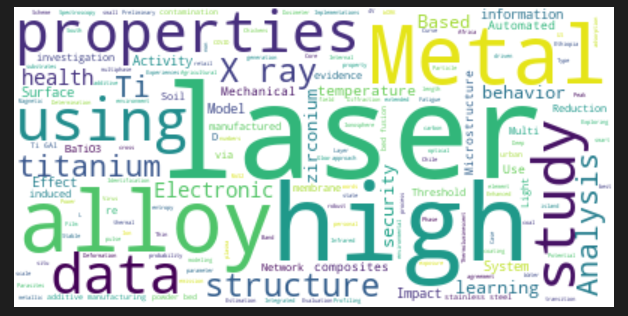
Extracción de datos:





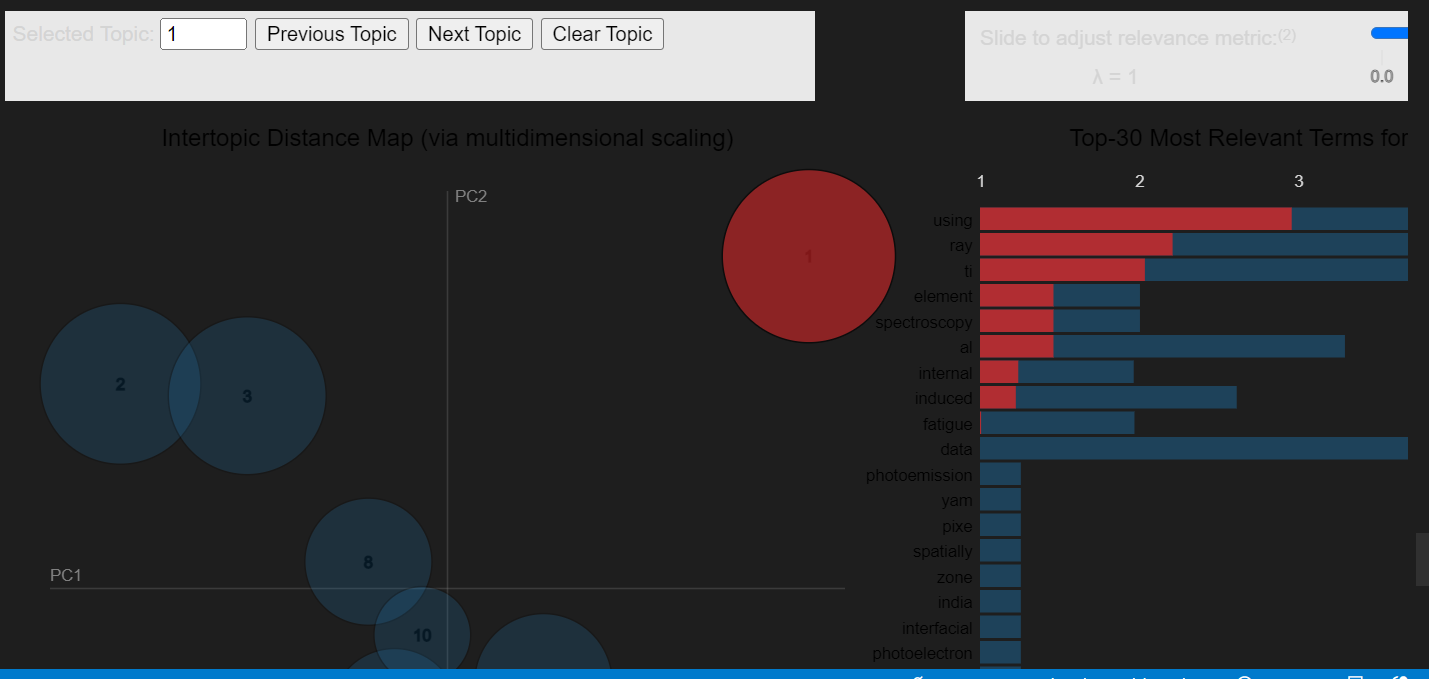
Realizamos el análisis y procedimiento muy similar al punto 2 para ver las palabras que contienen cada articulo

Palabras más utilizadas ejemplo:



Obtenemos las palabras más utilizadas con la biblioteca “pyLDAvis”

Resultados de top de palabras



En los siguientes pasos debiésemos almacenar las 15 palabras que muestra el panel y escanear los títulos de los artículos para ver si contienen estas palabras, además en el mismo proceso eliminamos quien no las tenga, de esta forma hacemos un tipo de filtro quedándonos con los artículos que se nos solicitan. Dado tales resultados podemos ver que el número de artículos es reducido en comparación a la cantidad inicial y que no todos tienen los mismos temas tratados a pesar de que hicimos una búsqueda detallada y especifica en un comienzo. Además, con este procedimiento podemos encontrar un grupo de artículos que coinciden y podemos tener una mejor referencia de que tratan estos artículos.

# **Conclusión**

Como se pudo apreciar en el trabajo que hicimos y en este informe, el Web Scraping nos permite obtener una gran cantidad de información para poder analizarla de una forma más eficaz considerando la gran cantidad de información que se maneja y se requiere buscar.

Este método de extracción de datos, puede llegar a ser muy complejo si es que no cuentas con las herramientas necesarias o información clara de lo que necesitas o tienes que hacer, es por ello que hay que el proceso de filtrado o agrupación es indispensable para poder encontrar información útil o que mínimamente nos sirva de algo.

El Web Scraping, se puede utilizar para examinar los distintos productos o alguna comparación de precios, y así analizar los datos dentro de una empresa que está relacionada con estos productos, por ejemplo, procesar estos datos.

# 

# **Anexos**

# **Referencias**

*How to Get Month Name from Month Number in Python - Studytonight*. (s. f.). Study Tonight. Recuperado 27 de junio de 2022, de https://www.studytonight.com/python-howtos/how-to-get-month-name-from-month-number-in-python

*Biswal, A. (2022, 22 junio). Recurrent Neural Network (RNN) Tutorial: Types, Examples, LSTM and More. Simplilearn.Com. Recuperado 27 de junio de 2022, de https://www.simplilearn.com/tutorials/deep-learning-tutorial/rnn*

# 

## Bibliografía / Links

<https://elmundodelosdatos.com/topic-modeling-gensim-asignacion-topicos/>

<https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0>

<https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0>

<https://datascienceplus.com/topic-modeling-and-latent-dirichlet-allocation-lda/>

Para el punto 3

<https://guru99.es/rnn-tutorial/#5>

[How to Get Month Name from Month Number in Python - Studytonight](https://www.studytonight.com/python-howtos/how-to-get-month-name-from-month-number-in-python)

Códigos Pt4

<https://drive.google.com/drive/folders/16BstwgmzTAGLXpFG7_nWNtOb7-InaFBd?usp=sharing>