**Proyecto unidad 2:**

**Elasticsearch**

|  |  |
| --- | --- |
| Asignatura: | Big Data |
| Integrantes: | Ignacio Arancibia  Eduardo Silva |
| Profesor: | Edio Mardones |
|  |  |

**Índice General**

[1. Introducción 2](#_Toc181149353)

[2. Desarrollo 3](#_Toc181149354)

[3. Conclusión 11](#_Toc181149355)

[4. Bibliografía 12](#_Toc181149356)

# Introducción

Wikipedia es la plataforma de colaboración abierta más activa de internet, este sitio se caracteriza por su gran cantidad de artículos en constante actualización. En pocos segundos miles de usuarios pueden editar un contenido de esta vasta enciclopedia, así generando un historial gigante de modificaciones en sus artículos.

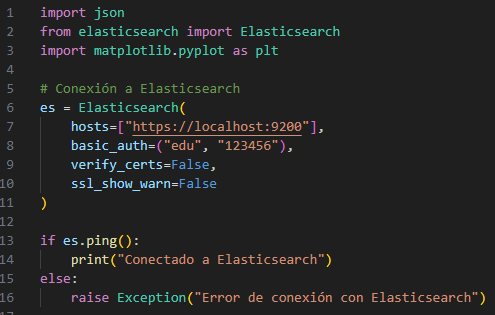
Cada dato que se genera luego de una edición origina una gran cantidad de datos no estructurados que pueden ser usados para llevar a cabo un análisis sobre las conductas de los usuarios influyentes en este sitio web. El problema que se produce al trabajar con esta gran cantidad de datos resulta en el uso de complejos motores de búsqueda, como en este caso, Elasticsearch.

Elasticsearch, como ya fue nombrado anteriormente, es un motor de búsqueda avanzado, destinado a indexar y analizar datos. Para este proyecto el uso de este buscador fue esencial para el análisis del historial de ediciones para la página de Wikipedia. A través de este sistema de búsqueda se implementaron el uso de consultas avanzadas para permitir extraer información que podrá ser visualizada mediante gráficos que facilitan la interpretación de los datos, todo esto usando una biblioteca como matplotlib y utilizando el lenguaje python.

# Desarrollo

Código utilizado para desarrollar la actividad:

* Conexión a Elasticsearch (más importaciones)



Primero tenemos las importaciones necesarias para el futuro funcionamiento del código. Lo sigue la conexión a Elasticsearch donde se utiliza el host general para la conexión al cluster, además se añade una función para verificar el estado de la conexión hacia Elasticsearch.

* Creación de un índice "wikipedia\_edits" y definición de mapeo:



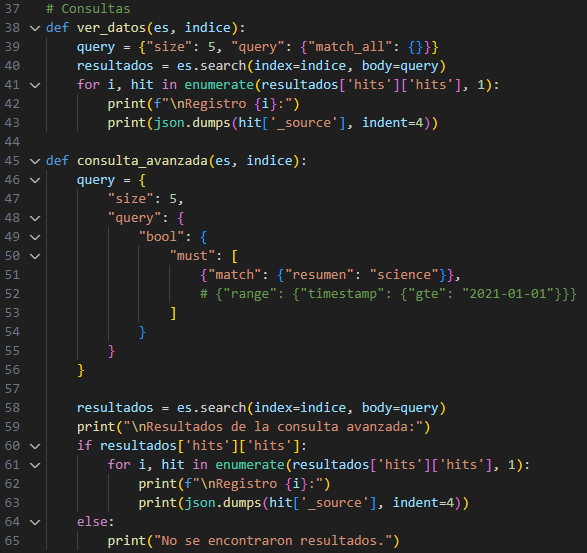
Al inicio se crea el índice junto a su nombre “Wikipedia\_edits”, luego se define el mapeo para el índice, esto significa la creación de la estructura que especifica los tipos de datos para cada campo, en este caso se requieren los títulos, resumen, nombres de los editores y la fecha de edición. Finalmente está la creación del índice donde se utiliza una función para crear el índice si este no existe y dar aviso cuando este índice ya exista, además se especifica “cuerpo” utilizado por el índice.

* Ingesta de datos:



La primera función que se nos muestra en el código nos permite consultar con la API de Wikipedia para obtener el conjunto de ediciones recientes de una página, luego almacena los datos obtenidos en el índice. Luego está la función que permite convertir la respuesta a un formato JSON. Ya casi terminando tenemos la iteración para las revisiones donde se imprime el documento con enumeración. Finalmente tenemos la última función que permite automatizar la obtención de las ediciones.

* Consultas avanzadas:



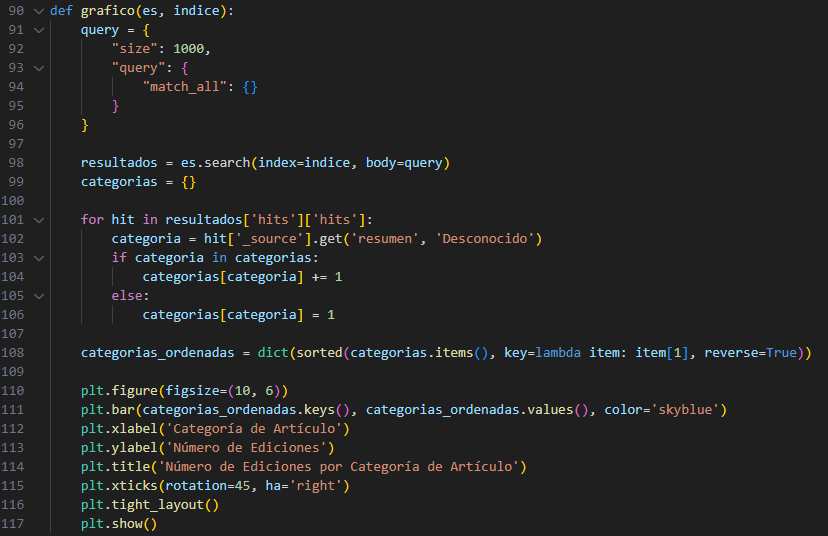
Empezamos con función para visualizar los 5 primeros datos del índice, luego tenemos la función que permite realizar una consulta especifica buscando en cada documento que coincida con el requisito, en este caso que “resumen” coincida con “science”

* Consulta ponderada:

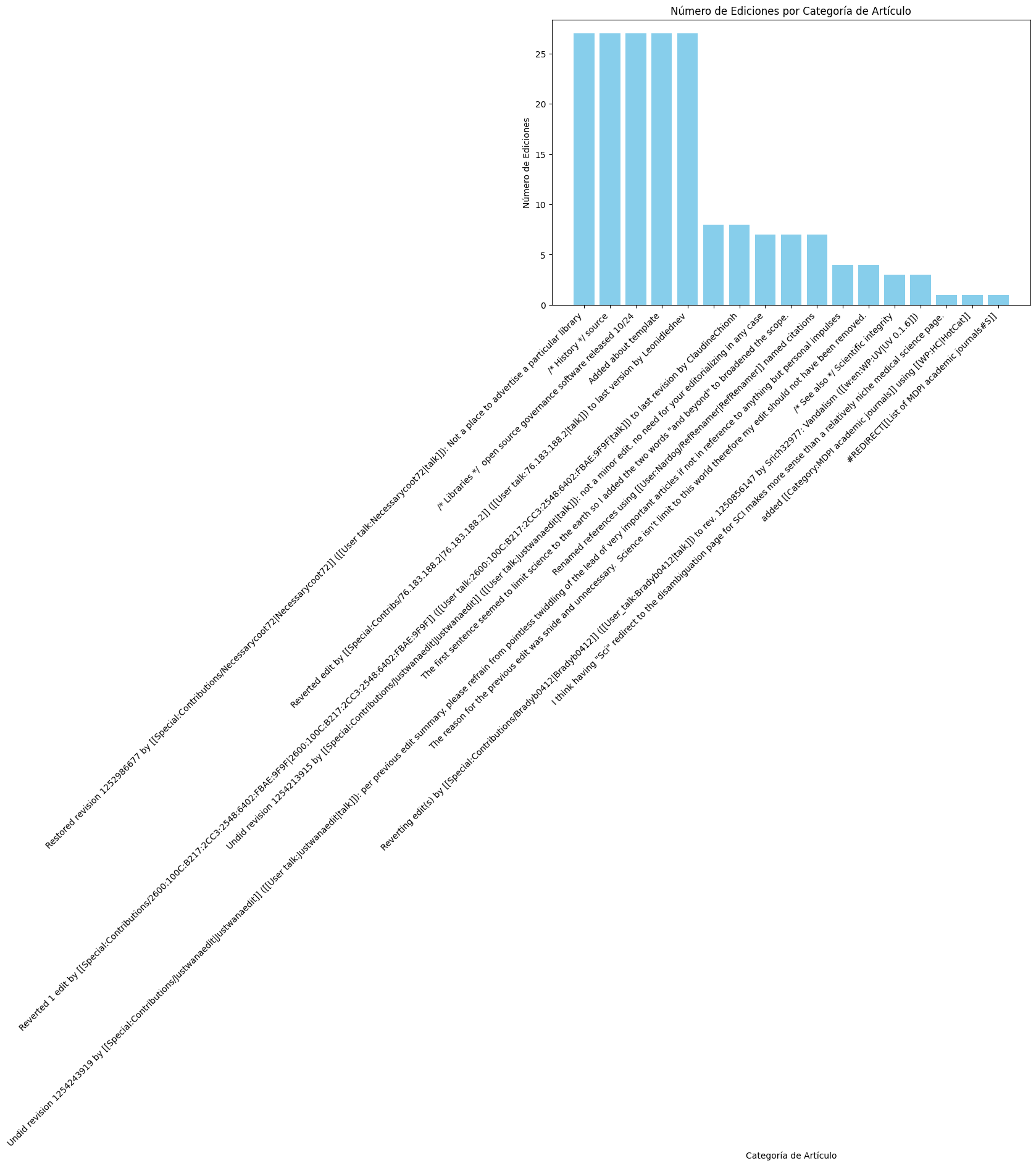


La función para la consulta ponderada nos permite visualizar datos para términos que están con un “boost” o con una importancia mayor que los demás términos dentro de la misma búsqueda. Finalmente tenemos la función que nos permita mostrar los resultados, donde, si se encuentran datos los muestra y de lo contrario indica que no se han encontrado resultados.

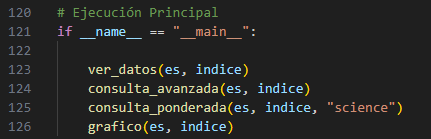
* Visualización de datos:



Finalmente nos encontramos con la visualización de los datos donde se crea el gráfico que recopila y genera un gráfico de barras que muestra el número de ediciones agrupadas por categoría, dentro de esta función se iteran los resultados obtenidos y se categorizan para finalmente visualizarse en un gráfico mostrado a continuación:



* Ejecución principal:



Como último tenemos la función que nos permite ejecutar y llevar a cabo todo este código inicializando una a una.

# Conclusión

Pese a las dificultades técnicas, para este proyecto se lograron completar todos los requisitos necesarios para abarcar el manejo de grandes volúmenes de datos dentro de la gran enciclopedia como es Wikipedia. A través de variadas consultas se extrajeron valiosos patrones de datos de colaboración y se comprendió de mejor manera el ajuste de los parámetros enfocados en la relevancia de las búsquedas, lo cual forma una parte esencial en el uso y manejo de resultados con un enfoque en la precisión de estas herramientas.

La visualización de los datos mediante las gráficas enseñadas al final del documento permitió una observación clara de la distribución de las ediciones por artículo, las cuales estuvieron centradas en la reversión de las mismas y en correcciones que fueron transformadas en poco tiempo.

# Bibliografía

Index modules | Elasticsearch Guide [8.15] | Elastic. (s. f.). Elastic. https://www.elastic.co/guide/en/elasticsearch/reference/current/index-modules.html