El proyecto está divido en 3 partes, una por cada notebook mostrado en el articulo original <https://databricks.com/blog/2020/07/10/a-data-driven-approach-to-environmental-social-and-governance.html>

La primera parte corresponde al notebook

<https://databricks.com/notebooks/esg_notebooks/01_esg_report.html>

Donde se analizan pdf’s de reportes de sustentabilidad, para comparar empresas entre si.

La segunda parte corresponde al notebook

<https://databricks.com/notebooks/esg_notebooks/02_esg_scoring.html>

Donde se desarrolla una herramienta para generar un puntaje ESG complejo, que toma en cuenta las relaciones entre empresas

La tercera y ultima parte corresponde al notebook

<https://databricks.com/notebooks/esg_notebooks/03_esg_market.html>

Donde se utiliza el puntaje ESG desarrollado anteriormente para correlacionarlo con el desempeño de mercado de las empresas.

**Sobre la primera parte:**

La base de datos GDELT es el ‘core’ de este proyecto. La manera en que se extraen datos de esa base es mediante un archivo llamado masterfilelist.txt, que es un archivo de texto cuyas líneas corresponden a enlaces de descarga de archivos csv, correspondientes a “pedazos” de la base de datos. Lo que se hace es descargar y juntar todos esos pedazos.

El archivo mastefilelist se encuentra en <https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/> tanto la versión que contiene todos los enlaces de la base de datos, como también la versión que se actualiza en tiempo real, para contener el csv generado en los últimos 15 minutos (útil para mejorar el proyecto para actuar en tiempo real).

#### indicar preparación de datos

1. Filter\_masterfilelist.ipynb

Lo primero a realizar antes de la segunda y tercera sección del proyecto, es filtar y limpiar el archivo masterfilelist.txt para poder descargar fácilmente los csv de interés. Esto pues el archivo masterfilelist incluye un listado recurrente de 3 tipos de archivos distintos, y solo nos interesa uno, el csv de datos gkg.

Para ese fin se ejecuta este notebook.

1. Entities.ipynb

El siguiente paso es obtener un listado de todas las entidades reconocidas por gdelt, con el fin de explorar e identificar de que manera están mencionadas las compañías de interés.

Para ello se descargan todos los csv anteriormente filtrados y uno a uno se leen para rescatar las entidades presentes en cada archivo.

Para obtener el archivo .txt correspondiente al listado de entidades basta con ejecutar el notebook en orden, solo se necesita cambiar la segunda celda de código, en caso de cambiar el archivo "english.gkg.oct2020-feb2021.txt" por otro listado de links (de csv gkg).

**Sobre la segunda y tercera parte**

1. Continuación

Teniendo los pasos anteriores el notebook correspondiente a la segunda y tercera parte del proyecto se encuentran en databricks, con el fin de aprovechar el poder de la computación distribuida para el procesamiento de la cantidad masiva de datos que tenemos.