Universitat Oberta de Catalunya

uoc.edu

Tipología y ciclo de vida de los datos

Práctica 1

Alumnos:

Alexis Germán Arroyo Peña

Gabriel Pulido de Torres

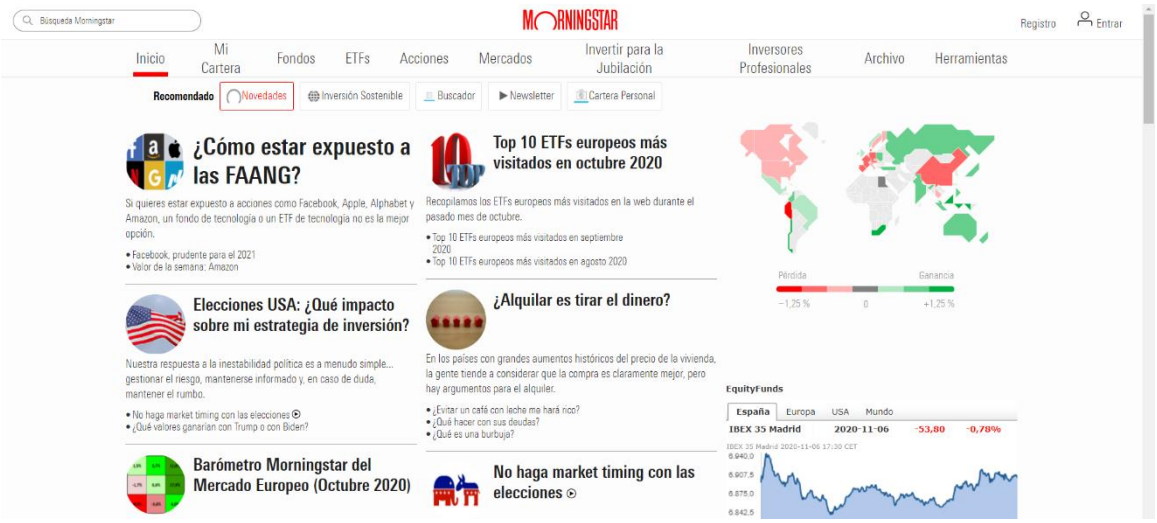
Contenido

1.- Contexto

Para cumplimentar los objetivos de esta práctica se decidió realizar un web scraping de la página web de MorningStar (https://www.morningstar.es/es/), una referencia para el análisis financiero.

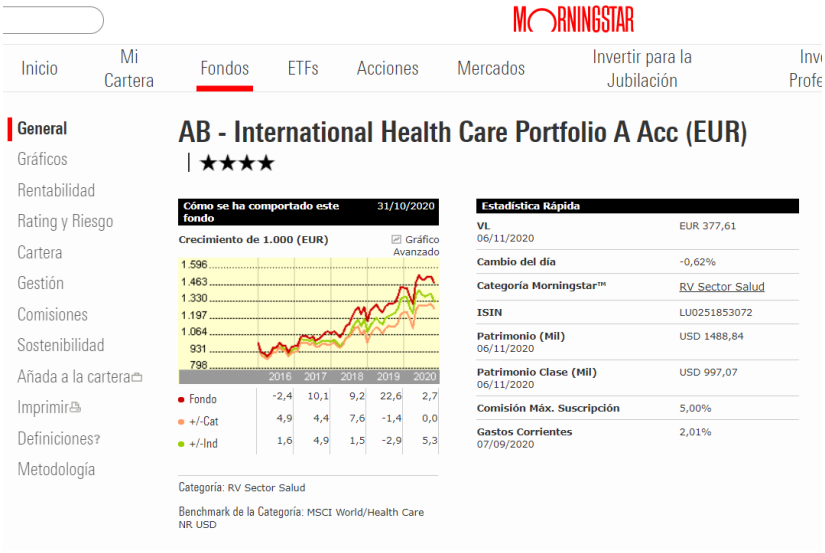
Morningstar es un proveedor de análisis y evaluaciones de productos financieros, que inició su actividad en 1984 con la idea de generar herramientas y productos que pudieran ayudar al inversor final. Por lo que, aglutina, clasifica y organiza en una misma plataforma información –accesible y comprensible– sobre todos los productos de inversión, de manera que cualquier usuario pueda gestionar mejor sus inversiones y alcanzar sus objetivos financieros. En la actualidad, posee una base de datos con más de 500.000 activos, a los que se puede acceder desde su página web.

En esta, podemos encontrar diversa información económica que tiene como objetivo potenciar el éxito de los inversores y ayudarles a tomar mejores decisiones de inversión. Para ello, incluyen ratings cuantitativos, ratings de sostenibilidad, informes de fondos y acciones, y noticias económicas que pueden influir en los diversos productos.



Morningstar ha desarrollado un sistema de clasificación que evalúa la calidad de los fondos de inversión, agrupándolos por categorías, en función de su rentabilidad (en tres, cinco y diez últimos años) y teniendo en cuenta las comisiones de cada producto (no solo las de gestión y de depósito, sino también las de suscripción que influyen en la rentabilidad final que obtendrás), así como el riesgo del mismo (a misma rentabilidad es preferible un fondo que presente menos riesgos, es decir, que registre menos fluctuaciones en sus rentabilidades). De acuerdo con todo esto, elabora su particular rating, por el que otorga una calificación de entre una y cinco estrellas a cada producto.

En la siguiente imagen podemos observar la información que nos muestra este sitio web para uno de los fondos buscados:



El proyecto se encarga de escanear la estructura de los fondos presentes en el sitio web. Esto lo realizamos utilizando tres técnicas:

1- Mediante una llamada a su API para conseguir la lista de identificadores.
2- Mediante el uso de peticiones para parsear los datos estáticos.
3- Mediante selenium, para parsear dinámicamente la página con los valores de sostenibilidad.

En los siguientes apartados iremos ampliando la información sobre el proceso de extracción realizado, así como un análisis del conjunto de datos extraído.

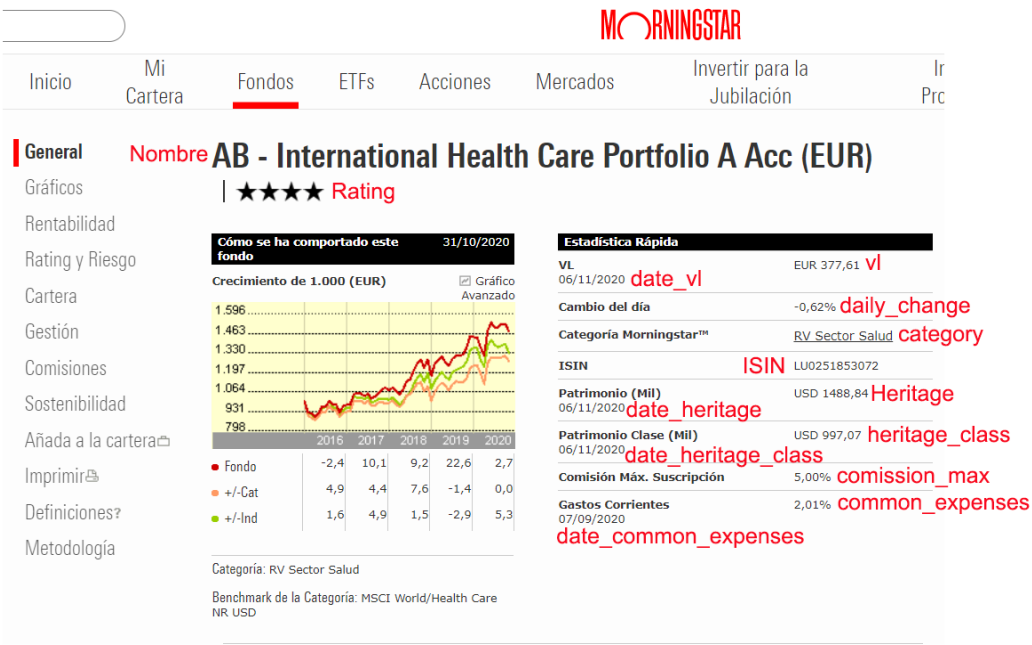
2.- Título y descripción del conjunto de datos

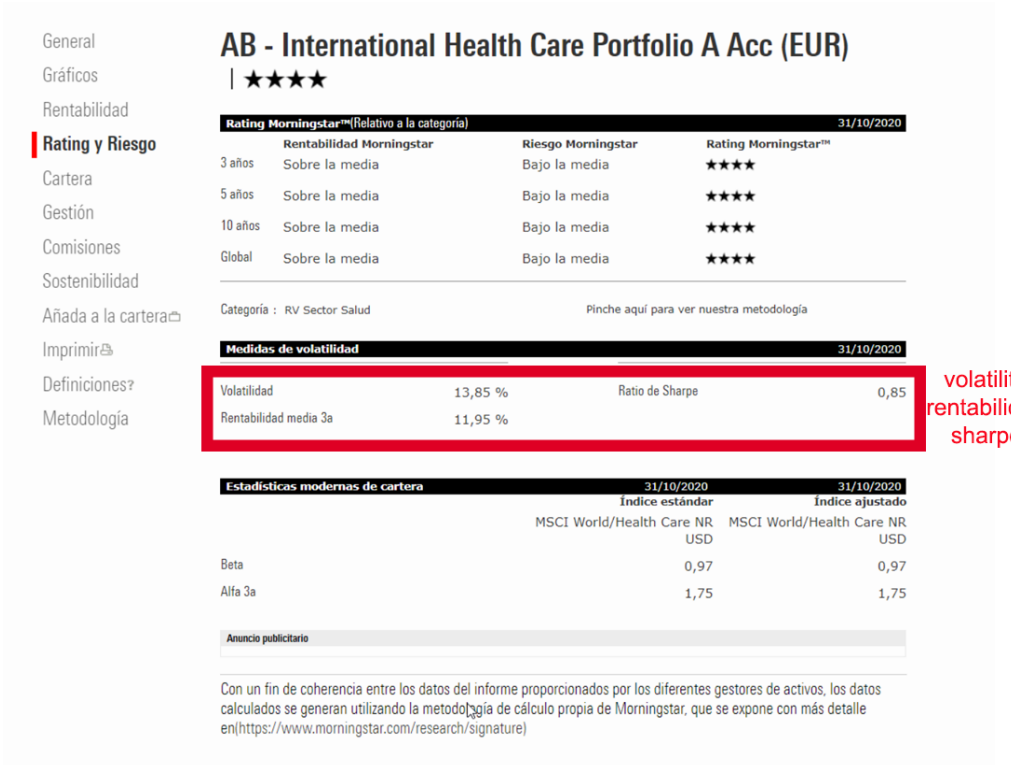
Título: Fondos de Morning Star con altos valores de rating.

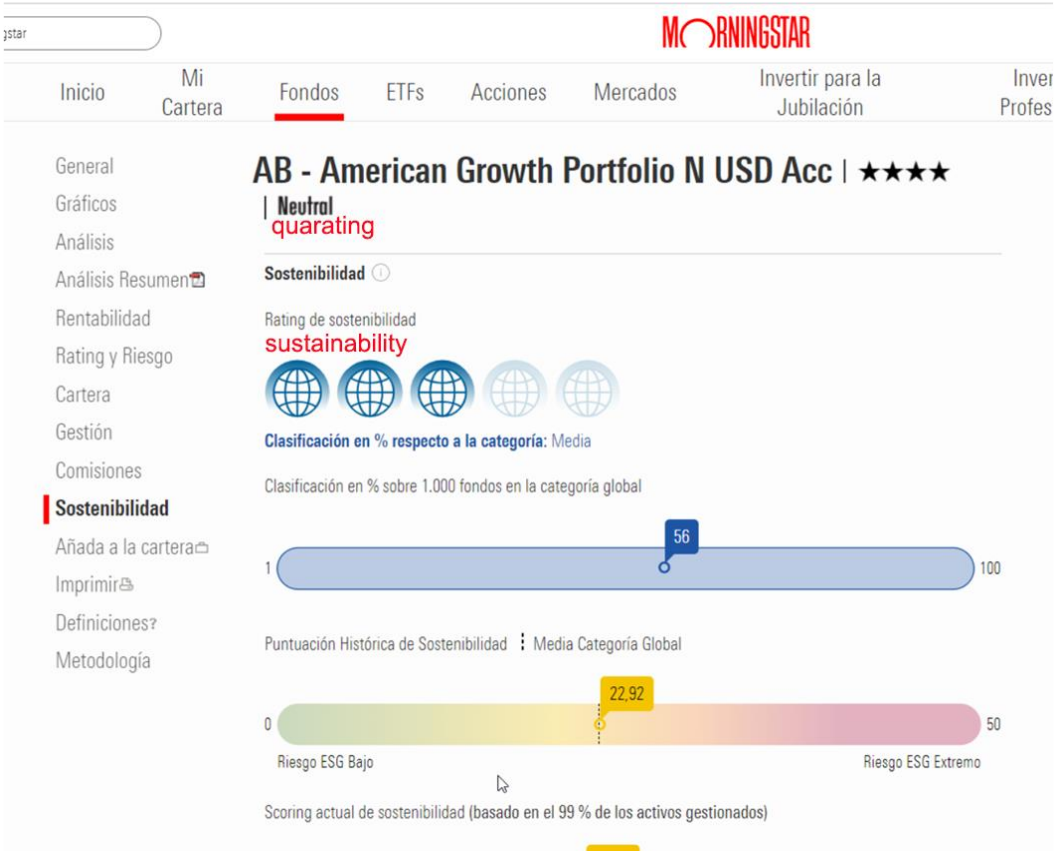
Descripción: El presente conjunto de datos contienen 10307 fondos cuyos valores de rentabilidad, según el ranking de Morning Star, son altos y, por lo tanto, sobre los cuales se puede plantear una inversión. Se han seleccionados aquellos cuyos valores están entre 3 y 5 estrellas de puntuación.

3.- Representación gráfica del conjunto de datos

A continuación, mostraremos gráficamente los campos obtenidos para nuestro conjunto de datos. Todos estos campos tienen sentido y significado para un usuario inversor, y son de fácil observación con una interfaz clara y sencilla.




General
Gráficos
Rentabilidad
Rating y Riesgo
Cartera
Gestión
Comisiones
Sostenibilidad
Añada a la cartera
Imprimir
Definiciones?
Metodología
AB - International Health Care Portfolio A Acc (EUR)
| ★★★★
Rating Morningstar™(Relativo a la categoría) 31/10/2020
Rentabilidad Morningstar
Riesgo Morningstar
Rating Morningstar™
3 años Sobre la media Bajo la media ★★★★
5 años Sobre la media Bajo la media ★★★★
10 años Sobre la media Bajo la media ★★★★
Global Sobre la media Bajo la media ★★★★
Categoría : RV Sector Salud
Pinche aquí para ver nuestra metodología
Medidas de volatilidad 31/10/2020
Volatilidad 13,85 %
Rentabilidad media 3a 11,95 %
Ratio de Sharpe 0,85
volatility
rentabilidad
sharpe
Estadísticas modernas de cartera 31/10/2020 31/10/2020
Índice estándar
Índice ajustado
MSCI World/Health Care NR USD
MSCI World/Health Care NR USD
Beta 0,97 0,97
Alfa 3a 1,75 1,75
Anuncio publicitario
Con un fin de coherencia entre los datos del informe proporcionados por los diferentes gestores de activos, los datos calculados se generan utilizando la metodología de cálculo propia de Morningstar, que se expone con más detalle en(https://www.morningstar.com/research/signature)


MORNINGSTAR
Inicio
Mi Cartera
Fondos
ETFs
Acciones
Mercados
Invertir para la Jubilación
General
Gráficos
Análisis
Análisis Resumen
Rentabilidad
Rating y Riesgo
Cartera
Gestión
Comisiones
Sostenibilidad
Añada a la cartera
Imprimir
Definiciones?
Metodología
AB - American Growth Portfolio N USD Acc | ★★★★
| Neutral
quarating
Sostenibilidad
Rating de sostenibilidad
sustainability
Clasificación en % respecto a la categoría: Media
Clasificación en % sobre 1.000 fondos en la categoría global
1
56
100
Puntuación Histórica de Sostenibilidad
Media Categoría Global
0
22,92
50
Riesgo ESG Bajo
Riesgo ESG Extremo
Scoring actual de sostenibilidad (basado en el 99 % de los activos gestionados)

4.- Contenido del conjunto de datos

El conjunto de datos contiene 10307 fondos. Los campos que hemos incluido son los enumerados a continuación:

- **MSID:** Identificador del fondo en la base de datos el cual podemos utilizar para obtener sus datos mediante búsqueda directa. Se componen de un código de 10 **dígitos alfanumérico**.

- **ISIN:** El estándar ISIN se usa en todo el mundo para identificar títulos específicos como valores, acciones (comunes y preferentes), futuros, órdenes judiciales, pagarés y opciones. Se componen de un código de 12 **dígitos alfanuméricos** y actúan como unificadores de distintos símbolos de mercado para el mismo título.

- **Name:** Nombre del fondo, compuesto por una **cadena de texto**.

- **Rating:** El Rating Morningstar mide la rentabilidad ajustada por el riesgo de un determinado fondo respecto a su categoría Morningstar utilizando datos de rentabilidad de los últimos 36 meses o 3 años. En función de estos criterios clasificamos los fondos de la siguiente manera. El 10% de los mejores fondos reciben 5 estrellas, el 22,5% siguiente 4 estrellas, el 35% siguiente 3 estrellas, el 22,5% siguiente 2 estrellas y el último 10% 1 estrella. Este campo se compone de un **campo entero** de un dígito.

- **Quarating:** Es una valoración que hace Morningstar sobre los gestores del fondo. Traducimos sus valores como "negative", "neutral", "bronze","silver" y "gold". Este campo es una **cadena de texto**.

- **Sustainability:** Rating de sostenibilidad. Mide la manera en la que las compañías que el fondo tiene en cartera gestionan sus riesgos ESG respecto a las carteras de la categoría correspondiente. Este campo se compone de un **campo entero** de un dígito.

- **vl:** El valor liquidativo de un fondo de inversión corresponde al precio de cada participación. Se calcula dividiendo el patrimonio total del fondo por el número de participaciones. En este campo introducimos el valor en euros. Este campo es una **cadena de texto** que incluye el valor y la divisa.

- **date_vl:** Fecha en que el campo VL tiene ese valor. Se corresponde a un campo **fecha en formato dd/mm/yyyy**.

- **daily_change:** Variación diaria del precio, en porcentaje, pudiendo este ser positivo (crecimiento) o negativo (decrecimiento). Es un **campo porcentual**.

- **Category:** Los fondos se clasifican en función de su estilo de gestión actual y no simplemente en función de lo que diga el reglamento de gestión, ni tampoco en función del hecho de haber obtenido una determinada rentabilidad. Para que las categorías sean homogéneas, Morningstar clasifica los fondos en función de los activos que tienen en cartera. Se analizan varias carteras para asegurar que se toma en cuenta la política de inversión real del fondo. Este campo es una **cadena de texto**.

- **Heritage**: Es el patrimonio del fondo medido en millones de la divisa en la que está denominado el fondo (millones de euros para un fondo denominado en euros). Este campo es una **cadena de texto** que incluye el valor y la divisa.

- **date_heritage:** Fecha en que el patrimonio (heritage) tiene ese valor. Se corresponde a un campo **fecha en formato dd/mm/yyyy**.

- **heritage_class:** Un fondo puede tener varias clases. Cada clase tiene su propio patrimonio y la suma de los patrimonios de todas las clases es el patrimonio del fondo. Este campo es una **cadena de texto** que incluye el valor y la divisa.

- **date_heritage_class:** Fecha en que el patrimonio de clase (heritage_class) tiene ese valor. Se corresponde a un campo **fecha en formato dd/mm/yyyy**.

- **comission_max:** La comisión máxima de suscripción sólo se cobra al inversor cuando éste compra o vuelve a comprar participaciones del fondo. Esta comisión generalmente se cobra como un porcentaje sobre la cantidad invertida. Este campo corresponde a un porcentaje. Es un **campo porcentual**.

- **common_expenses:** Gastos totales producidos en la gestión del fondo, indicado como un porcentaje. Es un **campo porcentual**.

- **date_common_expenses:** Fecha a partir de la cual los gastos corrientes (common_expenses) tienen ese valor. Se corresponde a un campo **fecha en formato dd/mm/yyyy**.

- **Volatility:** La volatilidad de un fondo es un dato interesante en dos sentidos. En primer lugar, porque cuanto mayores fluctuaciones presenta la rentabilidad de un fondo, más arriesgado será el fondo en cuestión. Además, permite comparar todos los fondos entre ellos, desde los fondos monetarios hasta los fondos de

acciones emergentes. En segundo lugar, porque los fondos que han sido más volátiles en el pasado también tienen tendencia a ser los más volátiles en el futuro. Se expresa como un porcentaje y se calcula basándose en las rentabilidades de los últimos 36 meses. Es un **campo porcentual**.

- **Rentabilidad:** Esta mide cómo una inversión se ha revalorizado (o caído) en un determinado periodo de tiempo. Los inversores pueden comparar la rentabilidad obtenida por su fondo con otros fondos de similares características para ver si éste lo ha hecho bien o mal. Este campo se expresa como un porcentaje. Es un **campo porcentual**.

- **Sharpe:** Uno está dispuesto a invertir en un fondo o en un activo con riesgo si la rentabilidad esperada es mayor que la del activo sin riesgo. Sharpe toma como riesgo la volatilidad de las rentabilidades pasadas. Representa al exceso de rentabilidad respecto al activo sin riesgo, en relación con el riesgo asumido. Este campo se representa mediante **números decimales**.

Los datos obtenidos se guardan en un fichero con formato CSV en cuyo nombre irá indicado la fecha de extracción. De esta forma se puede seguir un histórico de los valores de estos campos a lo largo del tiempo.

Para extraer estos datos hemos utilizado:

1- Llamada a la API de la página para conseguir la lista de identificadores de los fondos, siendo esta de forma filtrada por determinados campos. Al no ser pública ni encontrarse documentada se decidió utilizar otras técnicas de extracción de información, siendo además estos objetivos claros de la práctica.
2- Para cada fondo de los extraídos, se han realizado peticiones requests para parsear los datos estáticos correspondientes a la página impresión.
3- Para cada fondo de los extraídos, se ha utilizado de selenium para parsear de forma dinámica la página de sostenibilidad de los fondos.

Además, se ha creado una línea de comandos, user-friendly, que facilite el uso de la aplicación para generar más conjuntos de datos. Se pueden utilizar parámetros como:

- **"--rating":** Acepta valores como "negative", "neutral", "bronze","silver" y "gold".
- **"--star":** Acepta valores del 1 al 5. Siendo el valor introducido como un valor mínimo. Por ejemplo, si el valor es igual a 3 se recuperarán aquellos fondos con valor 3, 4 y 5.

- **"--qualitative":** Acepta valores del 1 al 5. Siendo el valor introducido como un valor mínimo. Por ejemplo, si el valor es igual a 3 se recuperarán aquellos fondos con valor 3, 4 y 5.
- **"--max":** Limita el número de fondos totales que queremos obtener.

5.- Agradecimientos

Los datos se han obtenido de la web de MorningStar. Este es un proveedor de análisis y evaluaciones de productos financieros, que inició su actividad en 1984 con la idea de generar herramientas y productos que pudieran ayudar al inversor final. Por lo tanto, es una página conocida por muchos inversores a nivel mundial.

Para observar las condiciones legales que incluyera la página web, observamos el archivo robots.txt y el apartado de condiciones legales.

El archivo robots.txt, indica las restricciones a tener en cuenta cuando se pretende rastrear la página. Aunque estas restricciones son solo una sugerencia y nunca una obligación, es recomendable tenerlas en cuenta, principalmente con el objetivo de reducir las posibilidades de ser bloqueados. Pudimos observar que las restricciones incluían una serie específica de directorios:

```
User-agent: *
Disallow: /admin/
Disallow: /virtual/
Disallow: /*.axd
Disallow: /WebResource.axd
Disallow: /ScriptResource.axd
Disallow: /App_Data/
Disallow: /dynimg/
Disallow: /img/
Disallow: /common/
Disallow: /*/util/
Disallow: /*/admin/
Disallow: /*/virtual/
Disallow: /*/*.axd
Disallow: /*/WebResource.axd
Disallow: /*/ScriptResource.axd
Disallow: /*/App_Data/
Disallow: /*/dynimg/
Disallow: /*/img/
Disallow: /*/common/
Disallow: /*p_snapshot.aspx
Disallow: /*p_article.aspx
Disallow: /*PrintArticle.aspx
Disallow: /*/fundscreener/results.aspx
Disallow: /*/snapshot/snapshot.aspx?id=FHUSA04EZM
Disallow: /*/snapshot/snapshot.aspx?id=F00000QHZB
Disallow: /*/snapshot/snapshot.aspx?id=VAUSA06622
Disallow: /errors/*
Disallow: /static/UploadManager/Image
```

Disallow: /static/UploadManager/Other

Respecto a las condiciones legales y de uso de la página web, presentes en https://www.morningstar.es/es/Disclaimer/Disclaimer.aspx?id=TermsofUse, no se ha incumplido ninguna de ellas.

Siendo las restricciones exclusivamente para:

- Copiar de modo sistemático (bien imprimiendo en papel, almacenando en disco o de cualquier otra manera) partes sustanciales de la página.
- Cambiar, alterar o esconder de cualquier manera los contenidos de la página o utilizar cualquier material de la página excepto en los términos explicitados en las Condiciones de Uso.
- Incluir o crear hipervínculos a esta página o a cualquier material contenido en ella.
- Utilizar la página y cualquier elemento disponible en ella para hacer una publicación o cualquier otro servicio que compita después con ella (bien on-line bien a través de cualquier otro medio de comunicación).
- Vulneración de la ley presente en la legislación vigente.

6.- Inspiración

La elección de este conjunto de datos viene inspirada por el interés, que los autores tienen, por el mundo de la inversión. Tener un conjunto de datos que responda de forma rápida sobre la rentabilidad de un fondo determinado, así como poder guardar un histórico sobre los mismos, permite al inversor tener un cuadro que le inspire confianza a la hora de invertir.

Gracias a estos datos, y al historial que se puede crear de los mismos, se pueden desarrollar modelos predictivos sobre los valores de estos fondos. Estos modelos permitirían deducir respuestas a preguntas como: ¿qué rentabilidad futura puede tener este fondo? ¿es ahora el momento de invertir? ¿qué correlaciones se pueden encontrar entre las variables de los fondos dentro de una misma categoría?

7.- Licencia

Se ha optado por escoger la licencia Released Under CC BY-NC-SA 4.0 License (https://creativecommons.org/licenses/by-nc-sa/4.0/).

Consideramos que los datos no deben deban usarse comercialmente sin permiso previo de MorningStar, sería incumplir las condiciones de uso que vienen explícitas en su página web. Cualquier persona o entidad que desee usarlos comercialmente puede

pedir permiso a MorningStar y generar un conjunto de datos personal utilizando el código disponible.

Además, proporcionamos atribución para poder limitar el uso comercial. Permitimos a su vez que la gente utilice este conjunto de datos y lo adapte a necesidades propias. Respecto al sharing, lo permitimos, pero en las mismas condiciones, evitando así una readaptación para uso comercial.

8.- Código

El código puede encontrarse en el siguiente enlace a GitHub:

https://github.com/gpulido/msfundscrap

Es preciso para la correcta ejecución de este proyecto, tener la herramienta Chromedriver. Además, se debe verificar que se tienen instaladas las librerías: requests, bs4, selenium, webdriver_manager, marsmallow, marsmallow-dataclass.

Se puede ejecutar, utilizando los filtros necesarios, mediante ejecutar la siguiente instrucción:

python ms_scrapper.py filter --rating silver --star 4 --qualitative 3 (valores de ejemplo).

El código, presente en GitHub, escrito en Python, está comentado para su correcto seguimiento.

9.- Dataset

Se ha publicado el dataset en formato CSV en Zenodo.

DOI 10.5281/zenodo.4263256

URL https://doi.org/10.5281/zenodo.4263256

10. Tabla contribuciones

Contribuciones	Firma
Investigación previa	GPT; AGAP
Redacción de las respuestas	GPT; AGAP
Desarrollo del código	GPT; AGAP