Mario Bru Barrero
Marcos Gómez Vázquez

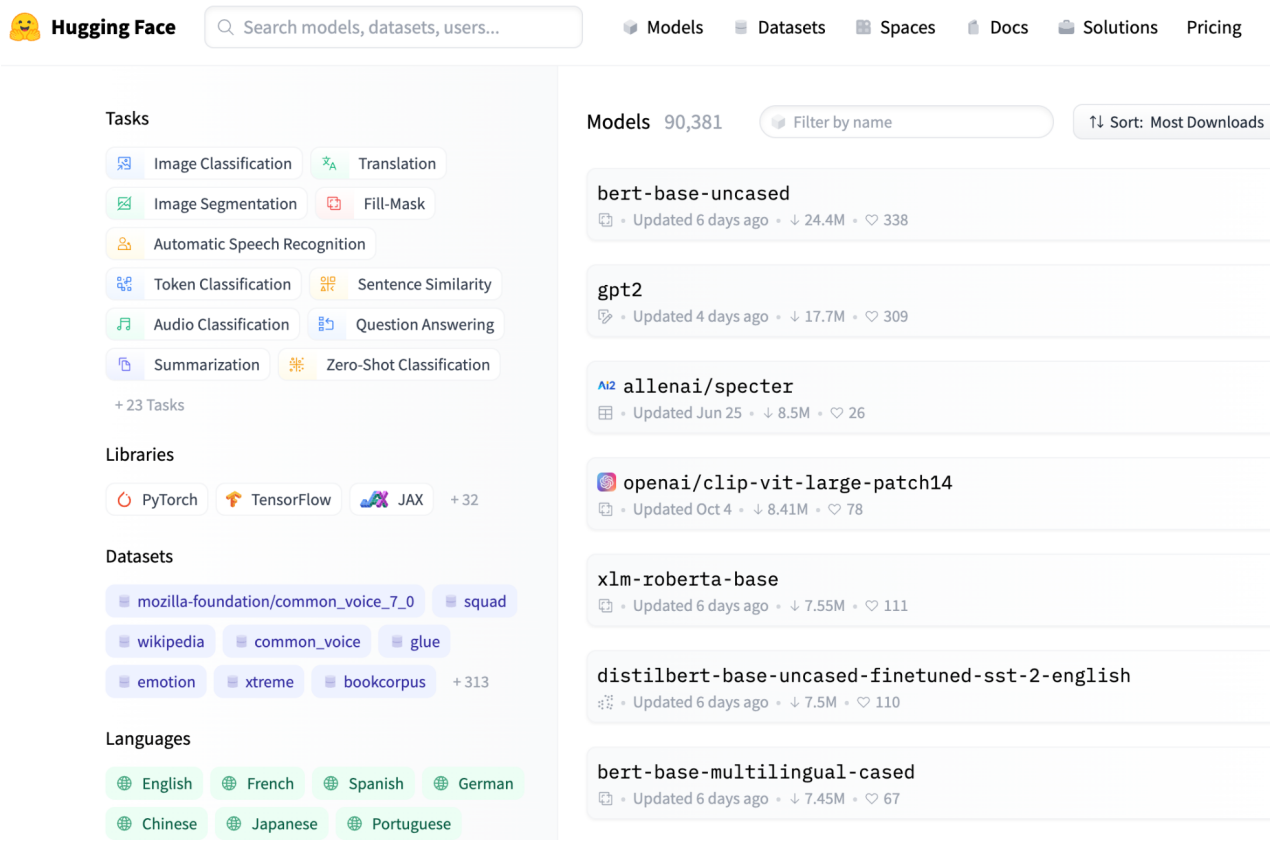
Tipología y ciclo de vida de los datos

Práctica 1

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información. Indicar la dirección del sitio web.

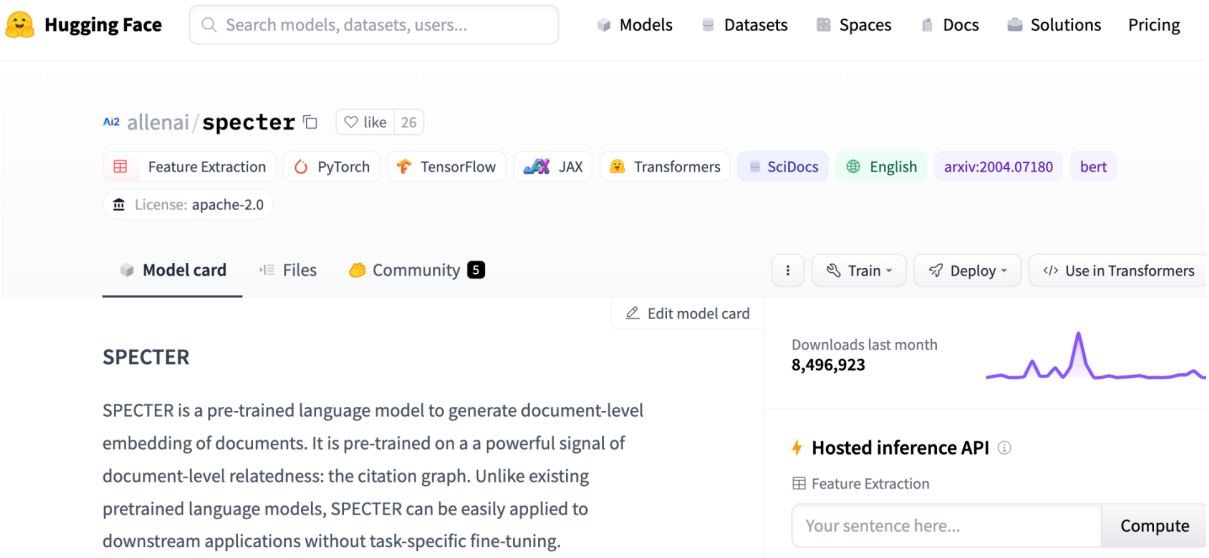
https://huggingface.co/models

El proyecto consiste en la creación de un dataset sobre los modelos de Machine Learning (ML) publicados en **Hugging Face**. Hugging Face es una plataforma open source, creada en 2016, que permite publicar modelos de ML a cualquier persona u organización. El propósito de esta plataforma es democratizar el ML. Inicialmente se publicaban solamente modelos de Natural Language Processing (language models), pero a día de hoy se puede encontrar otro tipo de modelos, como por ejemplo de Computer Vision.



Además, gracias a la librería Transformers de Hugging Face, se pueden desplegar e incluso entrenar o refinar (con nuevos datos) multitud de modelos con diferentes tipos de arquitecturas de forma muy sencilla, lo cual permite la rápida creación de modelos por parte de la comunidad que son capaces de desempeñar un amplio abanico de tareas tanto de dominio abierto como de dominios muy específicos, como por ejemplo, traducción de texto de un lenguaje A a un lenguaje B, clasificación de texto por categorías, resumir texto, etc.

Hugging Face también cuenta con un apartado 'Datasets' donde se publican datasets de forma abierta, y 'Spaces', donde la comunidad despliega aplicaciones (Spaces) que ejecutan modelos listos para ser utilizados. Pero en este proyecto el objetivo es centrarnos solamente en la categoría 'Models' y crear un dataset con meta-datos sobre la gran cantidad de modelos que hay en la plataforma. Este dataset permitirá a otros desarrolladores crear muchos tipos de análisis sobre el estado de la tecnología actual.



2. Título. Definir un título que sea descriptivo para el dataset.

huggingface_models_dataset.csv

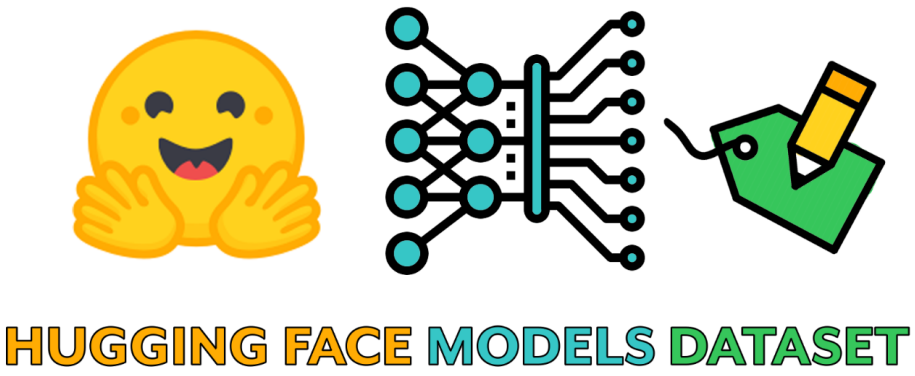
3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido.

El dataset contiene información relevante sobre los modelos de Machine Learning publicados en Hugging Face. Cada modelo es representado como una fila en el dataset, y cada columna es un atributo del modelo.

A la fecha de *scrapping* de los datos, se han registrado **89919 modelos** y un total de **15 atributos**. Aunque algunos modelos tienen vacíos algunos atributos, ya que no están presentes en la página web, y por lo tanto, su valor es nulo.

El tamaño del archivo huggingface_models_dataset.csv es de **14.4 MB.**

4. Representación gráfica. Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.



5. Contenido. Explicar los campos que incluye el dataset y el periodo de tiempo de los datos.

Campo	Descripción	Tipo	Ejemplo
id	identificador del modelo (único), compuesto por el nombre del autor seguido del nombre del modelo	string	facebook/bart-base
author	autor del modelo, puede ser nulo (i.e. publicado por Hugging Face)	string	facebook
downloads_last_month	número de descargas del modelo registradas en el últimos mes	integer	1124008
lastModified	fecha de última modificación del modelo	date	2022-11-16T23:23:10.000Z
likes	número de likes por parte de los usuarios que tiene el modelo	integer	33
cardExists	indica si el modelo tiene o no model card (un documento descriptivo del modelo)	boolean	True
pipeline_tag	indica el tipo de tarea (o tareas) a la que está destinado el modelo	list[string]	['text-classification']

subType	indica el sub-tipo de la tarea a la que está destinado el modelo (depende de pipeline_tag)	string	nlp
library	indica en las que está disponible el modelo (a través de las cuales se puede desplegar)	list[string]	['pytorch', 'tf', 'jax', 'transformers']
dataset	indica los datasets que han sido utilizados para crear (entrenar) el modelo	list[string]	['dataset:cnn_dailymail', 'dataset:xsum']
language	indica los lenguajes para los que el modelo está entrenado	list[string]	['en', 'fr', 'ro', 'de']
arxiv	papers citados en el modelo	list[string]	['arxiv:2010.11430', 'arxiv:2006.11477']
other	otras etiquetas del modelo	list[string]	['gptj', 'causal-lm', 'has_space']
license	licencia o licencias del modelo	list[string]	['license:mit']
doi	Digital Object Identifier (DOI) del modelo	list[string]	['doi:10.57967/hf/0100']

Dado que en Hugging Face se publican modelos a diario y la comunidad es muy activa, este dataset se desactualiza muy rápido. Por ejemplo, lastModified (fecha de última actualización) es un atributo que probablemente cambie a diario, al igual que downloads_last_month o likes.

6. Propietario. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

Los datos provienen de la plataforma Hugging Face, y están publicados de forma abierta y pública en Hugging Face Hub. Se pueden obtener mediante la API de Hugging Face Hub https://huggingface.co/docs/hub/api, aunque también se pueden obtener mediante web scraping.

Los modelos de los cuales se obtienen los datos para el dataset, en su mayoría, son creados por la comunidad, pero a su vez son publicados de forma abierta en la plataforma. Por lo tanto, al realizar web scraping sobre esta página web, no se vulnera la privacidad ni la propiedad de ningún autor. De hecho, como ya se ha mencionado, estos datos se pueden obtener también de forma “oficial” mediante la propia API de Hugging Face.

Para no correr el riesgo de realizar web scraping sobre datos propietarios, el proceso se ha realizado sin ningún tipo de login en la plataforma. De esta manera, nos aseguramos de que todo lo que se obtiene es 100% público y accesible para todo el mundo.

Un trabajo similar al realizado en este proyecto tuvo lugar en junio de 2021. El dataset se llamó huggingface-modelhub y se encuentra en Kaggle y en Hugging Face (https://huggingface.co/datasets/dk-crazydiv/huggingface-modelhub). Hugging Face publicó un artículo promocionando el dataset y utilizándolo para realizar una serie de análisis sobre el estado de la plataforma y el desarrollo de la comunidad: https://observablehq.com/@huggingface/dataset-huggingface-modelhub.

7. Inspiración. Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

En el artículo mencionado en el apartado anterior se visualizaron las tareas (pipeline tag) más populares, las librerías más utilizadas, las fechas de publicación de los modelos, las etiquetas (tags) más utilizadas, la longitud de las model cards, los autores más activos, y las descargas de los modelos.

Este artículo es un ejemplo de los análisis que se pueden realizar con este tipo de datos. Pero lo cierto es que son muy genéricos, y se pueden llegar a crear análisis y visualizaciones más profundas sobre problemas más concretos. Además, el número de atributos del dataset citado es menor que el del dataset de este proyecto, lo cual da lugar a otro tipo de análisis. Algunos ejemplos que se podrían realizar mediante este proyecto son:

- ¿Qué tareas son las más populares para un idioma concreto?
- ¿Qué librerías son las más utilizadas para cada tarea de NLP?
- ¿Cuál es el nivel de madurez de los modelos en catalán en comparación con los modelos en inglés?
- Relación likes-descargas: ¿Tienen valor los likes? ¿Afectan a la popularidad de un modelo?
- ¿Los modelos respaldados por papers científicos son los más efectivos?

El problema con este tipo de datasets es que, como ya se ha mencionado, se desactualizan muy rápido. El dataset citado registró 10.354 modelos. 17 meses después, contamos con 89919 modelos y un estado del arte mucho más avanzado. Dado que el crecimiento de esta tecnología es exponencial (al menos por ahora, que es relativamente nueva), es necesario disponer de este tipo de datos de forma frecuente y actualizada para poder realizar análisis lo más fieles posibles al estado actual de la tecnología.

8. Licencia. Seleccionar una licencia adecuada para el dataset resultante y justificar el motivo de su elección. Ejemplos de licencias que pueden considerarse:

- Released Under CC0: Public Domain License.
- Released Under CC BY-NC-SA 4.0 License.
- Released Under CC BY-SA 4.0 License.
- Database released under Open Database License, individual contents under Database Contents License.
- Otra (especificar cuál).

El dataset resultante tendrá una **licencia CC0 de dominio público universal**, mediante la cual renunciamos a cualquier tipo de propiedad intelectual sobre el mismo y lo hacemos accesible a todo el mundo, pudiendo ser de utilidad a estudiantes, investigadores y científicos de datos (entre otros) tanto a nivel profesional como personal.

Entendemos que al tratarse de un ejercicio práctico dentro del contexto de la asignatura, y de datos públicos accesibles a través de la web de Hugging Face, esta es la mejor elección. Por tanto, se podrá copiar, modificar, distribuir el dataset y hacer comunicación pública del mismo, incluso para fines comerciales, sin pedir permiso.

9. Código. Código con el que se ha obtenido el dataset, preferiblemente en Python o, alternativamente, en R.

El código consta de 2 archivos:

- `source/script.py:` ejecutable, que cuenta con un método main
- `source/scraper.py:` funciones necesarias para realizar el web scraping

Para ejecutar el código debemos introducir el siguiente comando en nuestra terminal (desde dentro de la carpeta source):

```
python3 script.py
```

Opcionalmente, podemos especificar 2 argumentos al ejecutar el script:

- `p`: El número de páginas que queremos scrapear (por defecto se scrapean todas)
- `s`: El tiempo de espera entre requests (que por defecto es 0), es decir, el tiempo de espera entre cada lectura de la página web de un modelo.

Este segundo argumento puede llegar a ser necesario si se satura el servidor debido al alto número de requests que el script realiza. También puede ser importante especificar el número de páginas ya que a fecha de creación del dataset, Hugging Face contiene 2998 páginas y el proceso entero duró más de 15 horas. Un ejemplo de comando de ejecución con los 2 argumentos:

```
python3 script.py -p 10 -s 1
```

El método main de `script.py`, una vez parseados los argumentos, preparado el logger y definido los headers, ejecuta el método **get_model_urls**, que obtiene todas las URLs de modelos que se encuentran en las páginas de huggingface.co/models. Para realizarlo, es necesario navegar por la web ejecutando un request para cada página, ya que en cada una se encuentran 30 modelos.

Una vez obtenida esta lista de URLs, para cada una de ellas se ejecuta el método **get_model_attributes**, que realiza un request a la URL del modelo, obteniendo el html fuente donde se encuentra toda su información (pública). No todos los atributos que queremos extraer se encuentran en el mismo elemento html, por lo que es necesario extraerlos de diferentes lugares y unificarlos en un diccionario.

Finalmente, este diccionario se almacena en un DataFrame como una nueva fila. Este DataFrame es convertido a formato CSV y exportado al directorio dataset.

10. Dataset. Publicar el dataset obtenido en formato CSV en Zenodo, incluyendo una breve descripción. Obtener y adjuntar el enlace del DOI del dataset (https://doi.org/…). El dataset también deberá incluirse en la carpeta /dataset del repositorio.

https://zenodo.org/record/7347175#.Y30UzOzMIeb

`10.5281/zenodo.7347175`

11. Vídeo. Realizar un breve vídeo explicativo de la práctica (máximo 10 minutos), que deberá contar con la participación de los dos integrantes del grupo. En el vídeo se deberá realizar una presentación del proyecto, destacando los puntos más relevantes, tanto de las respuestas a los apartados como del código utilizado para extraer los datos. Indicar el enlace del vídeo (https://drive.google.com/…), que deberá ubicarse en el Google Drive de la UOC

https://drive.google.com/file/d/1xBEok0xfo7J4A6gGQa1PvtmOhPw5pT_B/view?usp=share_link

CONTRIBUCIONES

Contribuciones	Firma
Investigación previa	MGV, MBB
Redacción de las respuestas	MGV, MBB
Desarrollo del código	MGV, MBB
Participación en el vídeo	MGV, MBB

BIBLIOGRAFÍA

-Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.