

Esta práctica se inició en el marco de la asignatura **Tipología y ciclo de vida de los datos**, perteneciente al Máster en Ciencia de Datos de la Universitat Oberta de Catalunya (UOC).

Goal

Creation of a dataset from the data contained in a web. It means apply web scraping as data collection. Web scrapers are computer programs that extract information from web sites. The structure and content of a web page are encoded in Hypertext Markup Language (HTML), which you can see using your browser's 'view source' or 'inspect element' function. A scraper understands HTML, and is able to parse and extract information from it.

Creación de un dataset a partir de los datos contenidos en una web. Esto significa aplicar web scraping como recopilación de datos. Los web scrapers son programas informáticos que extraen información de sitios web. La estructura y el contenido de una página web están codificados en lenguaje de marcado de hipertexto (HTML), que se puede ver utilizando la función "ver fuente" o "inspeccionar elemento" de su navegador. Un raspador comprende HTML y es capaz de analizar y extraer información de él.

1. Contexto

La primera parte de cualquier proyecto de Inteligencia Artificial (Visión computacional y Procesamiento del Lenguaje Natural) es obtener una base de datos. De hecho, tener todo el conjunto de datos limpio y etiquetado solo se aplica en Kaggle, no en la vida real. Youtube tiene imágenes, videos, comentarios e información como likes, dislikes, número de visualizaciones y otras informaciones representan una gran fuente de datos infrautilizada.

Sabiendo todo esto, este proyecto tiene como objetivo comenzar a desarrollar una muestra de este conjunto de datos para ser explorado para diseñar aplicaciones de inteligencia artificial.

2- Título del dataset

SampleYoutubeTrendingPT

3 - Descripción

YouTube (el sitio web de fama mundial para compartir videos) mantiene una lista de los videos más populares de la plataforma. Según la [revista Variety] (<http://variety.com/2017/digital/news/youtube-2017-top-trending-videos-music-videos-1202631416/>), "Para determinar los videos de mayor tendencia del año, YouTube utiliza una combinación de factores que incluyen la medición de las interacciones de los usuarios (número de vistas, acciones, comentarios y me gusta) El dataset que resulta de nuestro web scraping se compone de dos secuencias de archivos extraídos diariamente llamados 'youtubetrending_tabular_YYYYMMDD.csv' y 'youtubetrending_YYYYMMDD.csv'. El primero contiene 18 columnas de características fundamentales de video; el segundo contiene 3 columnas sobre imágenes en miniatura de videos de YouTube.

Usamos scrapy porque es la biblioteca de código abierto más poderosa para recopilar datos web y Selenium. Después de todo, youtube tiene mucho javascript en su página, y esto hace con que Selenium sea interesante para recopilar información correctamente. El Selenium es considerablemente lento y será reemplazado en una versión futura para slash.

Nuestro algoritmo va a la página de tendencias en portugués de Portugal para recopilar la URL de los videos enumerados. Después de eso, va a cada página para recopilar la información relevante de cada video.

4 - Contenido

4A - Dataset tabular

Attribute Information:

- title = título del video de Youtube
- url = URL del video de Youtube
- views = vistas de videos de Youtube
- duration = duración del video de Youtube
- likes = "Me gusta" de videos de Youtube
- dislikes = "No me gusta" el video de Youtube
- channelName = nombre del canal
- subscribers = número de suscriptores
- description = descripción del video de Youtube
- keywords = palabras clave
- date_published = Fecha de publicación del video de Youtube
- date_scraped = fecha en que se recopiló la información
- tags = etiquetas
- comments = comentarios
- image_urls = URL de la imagen del video de Youtube
- path = ruta relativa a donde se almacenaron los datos jpg de la imagen
- checksum = un hash MD5 del contenido de la imagen
- status = la indicación del estado del archivo.

4B - Dataset de Imágenes

Es una colección de imágenes en miniatura de videos de YouTube. Cada imagen del conjunto de datos se serializó como una cadena Base64 y, por lo tanto, se representó en un archivo csv separado por tabulaciones, que consta de tres valores:

- El [Identificador único universal] (https://en.wikipedia.org/wiki/Universally_unique_identifier) (UUID) de la imagen.
- La ruta original a la imagen en el disco.
- La propia imagen serializada como una cadena Base64.

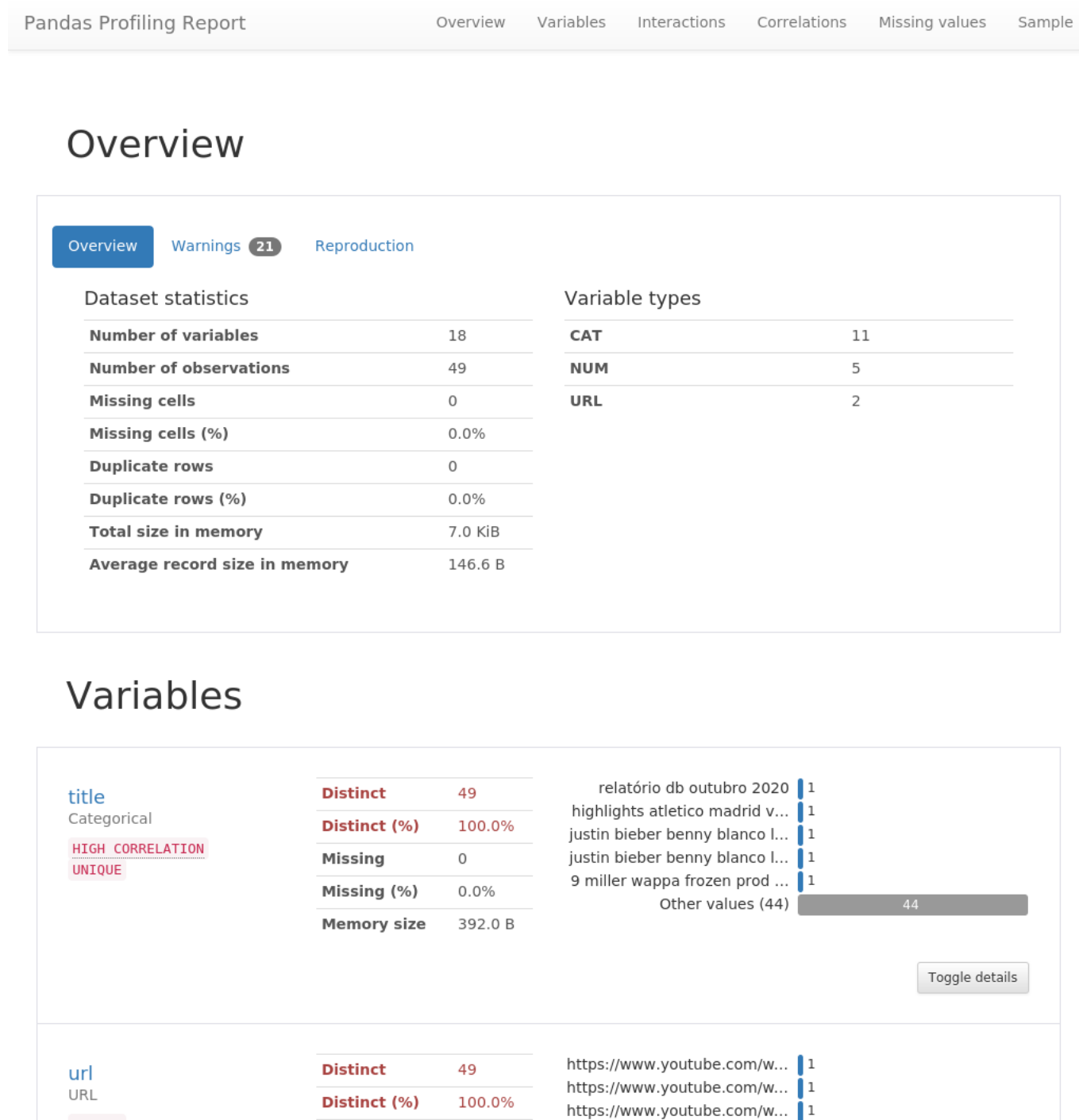
El archivo es más grande que las imagenes pero es útil para accesar a esta información de manera rápida y eficiente en nuestro Data Lake Hadoop. También mantenemos las imágenes en formato jpg.

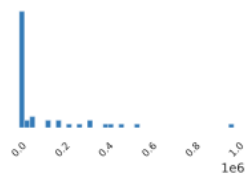
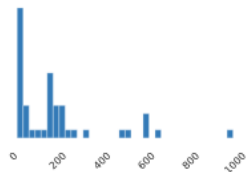
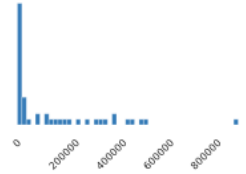
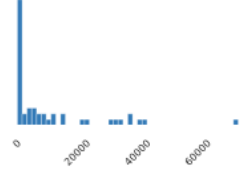
4C - Dataset de Texto

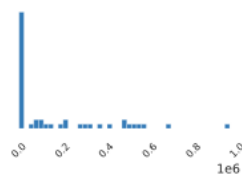
TODO

Esta parte se desarrollará pronto. Por ahora, eliminamos solo el primer comentario con fines de prueba.

5 - Representación gráfica



<div>UNIQUE</div>	<table><tr><td>Missing</td><td>0</td></tr><tr><td>Missing (%)</td><td>0.0%</td></tr><tr><td>Memory size</td><td>392.0 B</td></tr></table>	Missing	0	Missing (%)	0.0%	Memory size	392.0 B	<div>https://www.youtube.com/w...1</div> <div>https://www.youtube.com/w...1</div> <div>Other values (44)</div> <div>44</div> <div>Toggle details</div>																		
Missing	0																									
Missing (%)	0.0%																									
Memory size	392.0 B																									
<div>views</div> <div>Real number ($\mathbb{R}_{\geq 0}$)</div>	<table><tr><td>Distinct</td><td>41</td></tr><tr><td>Distinct (%)</td><td>83.7%</td></tr><tr><td>Missing</td><td>0</td></tr><tr><td>Missing (%)</td><td>0.0%</td></tr><tr><td>Infinite</td><td>0</td></tr><tr><td>Infinite (%)</td><td>0.0%</td></tr></table>	Distinct	41	Distinct (%)	83.7%	Missing	0	Missing (%)	0.0%	Infinite	0	Infinite (%)	0.0%	<table><tr><td>Mean</td><td>99512.16327</td></tr><tr><td>Minimum</td><td>1</td></tr><tr><td>Maximum</td><td>987243</td></tr><tr><td>Zeros</td><td>0</td></tr><tr><td>Zeros (%)</td><td>0.0%</td></tr><tr><td>Memory size</td><td>392.0 B</td></tr></table> <div></div> <div>Toggle details</div>	Mean	99512.16327	Minimum	1	Maximum	987243	Zeros	0	Zeros (%)	0.0%	Memory size	392.0 B
Distinct	41																									
Distinct (%)	83.7%																									
Missing	0																									
Missing (%)	0.0%																									
Infinite	0																									
Infinite (%)	0.0%																									
Mean	99512.16327																									
Minimum	1																									
Maximum	987243																									
Zeros	0																									
Zeros (%)	0.0%																									
Memory size	392.0 B																									
<div>duration</div> <div>Real number ($\mathbb{R}_{\geq 0}$)</div>	<table><tr><td>Distinct</td><td>36</td></tr><tr><td>Distinct (%)</td><td>73.5%</td></tr><tr><td>Missing</td><td>0</td></tr><tr><td>Missing (%)</td><td>0.0%</td></tr><tr><td>Infinite</td><td>0</td></tr><tr><td>Infinite (%)</td><td>0.0%</td></tr></table>	Distinct	36	Distinct (%)	73.5%	Missing	0	Missing (%)	0.0%	Infinite	0	Infinite (%)	0.0%	<table><tr><td>Mean</td><td>186.7346939</td></tr><tr><td>Minimum</td><td>15</td></tr><tr><td>Maximum</td><td>989</td></tr><tr><td>Zeros</td><td>0</td></tr><tr><td>Zeros (%)</td><td>0.0%</td></tr><tr><td>Memory size</td><td>392.0 B</td></tr></table> <div></div> <div>Toggle details</div>	Mean	186.7346939	Minimum	15	Maximum	989	Zeros	0	Zeros (%)	0.0%	Memory size	392.0 B
Distinct	36																									
Distinct (%)	73.5%																									
Missing	0																									
Missing (%)	0.0%																									
Infinite	0																									
Infinite (%)	0.0%																									
Mean	186.7346939																									
Minimum	15																									
Maximum	989																									
Zeros	0																									
Zeros (%)	0.0%																									
Memory size	392.0 B																									
<div>likes</div> <div>Real number ($\mathbb{R}_{\geq 0}$)</div> <div>UNIQUE</div>	<table><tr><td>Distinct</td><td>49</td></tr><tr><td>Distinct (%)</td><td>100.0%</td></tr><tr><td>Missing</td><td>0</td></tr><tr><td>Missing (%)</td><td>0.0%</td></tr><tr><td>Infinite</td><td>0</td></tr><tr><td>Infinite (%)</td><td>0.0%</td></tr></table>	Distinct	49	Distinct (%)	100.0%	Missing	0	Missing (%)	0.0%	Infinite	0	Infinite (%)	0.0%	<table><tr><td>Mean</td><td>142791.6327</td></tr><tr><td>Minimum</td><td>13</td></tr><tr><td>Maximum</td><td>935057</td></tr><tr><td>Zeros</td><td>0</td></tr><tr><td>Zeros (%)</td><td>0.0%</td></tr><tr><td>Memory size</td><td>392.0 B</td></tr></table> <div></div> <div>Toggle details</div>	Mean	142791.6327	Minimum	13	Maximum	935057	Zeros	0	Zeros (%)	0.0%	Memory size	392.0 B
Distinct	49																									
Distinct (%)	100.0%																									
Missing	0																									
Missing (%)	0.0%																									
Infinite	0																									
Infinite (%)	0.0%																									
Mean	142791.6327																									
Minimum	13																									
Maximum	935057																									
Zeros	0																									
Zeros (%)	0.0%																									
Memory size	392.0 B																									
<div>dislikes</div> <div>Real number ($\mathbb{R}_{\geq 0}$)</div>	<table><tr><td>Distinct</td><td>46</td></tr><tr><td>Distinct (%)</td><td>93.9%</td></tr><tr><td>Missing</td><td>0</td></tr><tr><td>Missing (%)</td><td>0.0%</td></tr><tr><td>Infinite</td><td>0</td></tr><tr><td>Infinite (%)</td><td>0.0%</td></tr></table>	Distinct	46	Distinct (%)	93.9%	Missing	0	Missing (%)	0.0%	Infinite	0	Infinite (%)	0.0%	<table><tr><td>Mean</td><td>10354.55102</td></tr><tr><td>Minimum</td><td>17</td></tr><tr><td>Maximum</td><td>73238</td></tr><tr><td>Zeros</td><td>0</td></tr><tr><td>Zeros (%)</td><td>0.0%</td></tr><tr><td>Memory size</td><td>392.0 B</td></tr></table> <div></div> <div>Toggle details</div>	Mean	10354.55102	Minimum	17	Maximum	73238	Zeros	0	Zeros (%)	0.0%	Memory size	392.0 B
Distinct	46																									
Distinct (%)	93.9%																									
Missing	0																									
Missing (%)	0.0%																									
Infinite	0																									
Infinite (%)	0.0%																									
Mean	10354.55102																									
Minimum	17																									
Maximum	73238																									
Zeros	0																									
Zeros (%)	0.0%																									
Memory size	392.0 B																									
<div>channelName</div> <div>Categorical</div> <div>HIGH CORRELATION</div> <div>UNIFORM</div>	<table><tr><td>Distinct</td><td>43</td></tr><tr><td>Distinct (%)</td><td>87.8%</td></tr><tr><td>Missing</td><td>0</td></tr></table>	Distinct	43	Distinct (%)	87.8%	Missing	0	<div>Gusttavo Lima Oficial2</div> <div>RTP2</div> <div>System Of A Down2</div> <div>Little Mix2</div> <div>NOW UNITED2</div>																		
Distinct	43																									
Distinct (%)	87.8%																									
Missing	0																									

	<table><tr><td>Missing (%)</td><td>0.0%</td></tr><tr><td>Memory size</td><td>392.0 B</td></tr></table> <div>Other values (38)<div>39</div></div> <div>Toggle details</div>	Missing (%)	0.0%	Memory size	392.0 B																					
Missing (%)	0.0%																									
Memory size	392.0 B																									
<div>subscribers</div> <div>Real number ($\mathbb{R}_{\geq 0}$)</div>	<table><tr><td>Distinct</td><td>43</td></tr><tr><td>Distinct (%)</td><td>87.8%</td></tr><tr><td>Missing</td><td>0</td></tr><tr><td>Missing (%)</td><td>0.0%</td></tr><tr><td>Infinite</td><td>0</td></tr><tr><td>Infinite (%)</td><td>0.0%</td></tr></table> <table><tr><td>Mean</td><td>157935.2857</td></tr><tr><td>Minimum</td><td>16</td></tr><tr><td>Maximum</td><td>965000</td></tr><tr><td>Zeros</td><td>0</td></tr><tr><td>Zeros (%)</td><td>0.0%</td></tr><tr><td>Memory size</td><td>392.0 B</td></tr></table> <div></div> <div>Toggle details</div>	Distinct	43	Distinct (%)	87.8%	Missing	0	Missing (%)	0.0%	Infinite	0	Infinite (%)	0.0%	Mean	157935.2857	Minimum	16	Maximum	965000	Zeros	0	Zeros (%)	0.0%	Memory size	392.0 B	
Distinct	43																									
Distinct (%)	87.8%																									
Missing	0																									
Missing (%)	0.0%																									
Infinite	0																									
Infinite (%)	0.0%																									
Mean	157935.2857																									
Minimum	16																									
Maximum	965000																									
Zeros	0																									
Zeros (%)	0.0%																									
Memory size	392.0 B																									
<div>description</div> <div>Categorical</div> <div>HIGH CORRELATION</div> <div>UNIFORM</div>	<table><tr><td>Distinct</td><td>48</td></tr><tr><td>Distinct (%)</td><td>98.0%</td></tr><tr><td>Missing</td><td>0</td></tr><tr><td>Missing (%)</td><td>0.0%</td></tr><tr><td>Memory size</td><td>392.0 B</td></tr></table> <div>siga nos http www facebook... 2 official video for pop smoke ... 1 now united presents one lov... 1 links no comentário afixado ... 1 bad bunny jhay cortez dákit... 1 Other values (43)<div>43</div></div> <div>Toggle details</div>	Distinct	48	Distinct (%)	98.0%	Missing	0	Missing (%)	0.0%	Memory size	392.0 B															
Distinct	48																									
Distinct (%)	98.0%																									
Missing	0																									
Missing (%)	0.0%																									
Memory size	392.0 B																									
<div>keywords</div> <div>Categorical</div> <div>HIGH CORRELATION</div>	<table><tr><td>Distinct</td><td>40</td></tr><tr><td>Distinct (%)</td><td>81.6%</td></tr><tr><td>Missing</td><td>0</td></tr><tr><td>Missing (%)</td><td>0.0%</td></tr><tr><td>Memory size</td><td>392.0 B</td></tr></table> <div>vídeo compartir teléfono co... 8 now united nowunited new ... 2 rtp rtp1 estação pública est... 2 you are so golden youre so ... 1 123go challenge portugues... 1 Other values (35)<div>35</div></div> <div>Toggle details</div>	Distinct	40	Distinct (%)	81.6%	Missing	0	Missing (%)	0.0%	Memory size	392.0 B															
Distinct	40																									
Distinct (%)	81.6%																									
Missing	0																									
Missing (%)	0.0%																									
Memory size	392.0 B																									
<div>date_published</div> <div>Categorical</div> <div>HIGH CORRELATION</div>	<table><tr><td>Distinct</td><td>18</td></tr><tr><td>Distinct (%)</td><td>36.7%</td></tr><tr><td>Missing</td><td>0</td></tr><tr><td>Missing (%)</td><td>0.0%</td></tr><tr><td>Memory size</td><td>392.0 B</td></tr></table> <div>2020-11-05 8 2020-11-06 8 2020-11-01 3 2020-11-07 3 2020-10-09 3 Other values (13)<div>24</div></div> <div>Toggle details</div>	Distinct	18	Distinct (%)	36.7%	Missing	0	Missing (%)	0.0%	Memory size	392.0 B															
Distinct	18																									
Distinct (%)	36.7%																									
Missing	0																									
Missing (%)	0.0%																									
Memory size	392.0 B																									
<div>date_scraped</div> <div>Categorical</div> <div>HIGH CORRELATION</div> <div>UNIQUE</div>	<table><tr><td>Distinct</td><td>49</td></tr><tr><td>Distinct (%)</td><td>100.0%</td></tr><tr><td>Missing</td><td>0</td></tr><tr><td>Missing (%)</td><td>0.0%</td></tr><tr><td>Memory size</td><td>392.0 B</td></tr></table> <div>2020-09-11 00:49:24 1 2020-09-11 00:34:59 1 2020-09-11 00:43:41 1 2020-09-11 00:20:21 1 2020-09-11 00:23:02 1 Other values (44)<div>44</div></div> <div>Toggle details</div>	Distinct	49	Distinct (%)	100.0%	Missing	0	Missing (%)	0.0%	Memory size	392.0 B															
Distinct	49																									
Distinct (%)	100.0%																									
Missing	0																									
Missing (%)	0.0%																									
Memory size	392.0 B																									
<div>tags</div>	<table><tr><td>Distinct</td><td>40</td></tr></table> <div>8 now united nowunited new ... 2</div> <div>Toggle details</div>	Distinct	40																							
Distinct	40																									

Categorical

HIGH CORRELATION

Distinct (%)	81.6%
Missing	0
Missing (%)	0.0%
Memory size	392.0 B

now united nowunited new ... 2

rtp rtp1 estação pública est... 2

jimmy jimmy kimmel jimmy... 1

justin bieber benny blanco l... 1

Other values (35) 35

Toggle details

comments

Categorical

HIGH CORRELATION

UNIQUE

Distinct	49
Distinct (%)	100.0%
Missing	0
Missing (%)	0.0%
Memory size	392.0 B

quien diria que the weeknd ... 1

vem curtir comigo o quarto ... 1

lolllllllll if even the drumme... 1

that correa assist wow 1

a desembrolhar o bollycao 1

Other values (44) 44

Toggle details

image_urls

URL

UNIQUE

Distinct	49
Distinct (%)	100.0%
Missing	0
Missing (%)	0.0%
Memory size	392.0 B

http://i4.ytimg.com/vi/Btk8... 1

http://i4.ytimg.com/vi/Xqmk... 1

http://i4.ytimg.com/vi/eh6x... 1

http://i4.ytimg.com/vi/ZHcZ... 1

http://i4.ytimg.com/vi/LOfM-... 1

Other values (44) 44

Toggle details

path

Categorical

HIGH CORRELATION

UNIQUE

Distinct	49
Distinct (%)	100.0%
Missing	0
Missing (%)	0.0%
Memory size	392.0 B

full/061dbe550c8d456fdd1d... 1

full/48db478b99a03be0eed... 1

full/a490a84dc82c9b8e1a4... 1

full/b486ef7347ed7169041... 1

full/2b3513f66d18b64da64... 1

Other values (44) 44

Toggle details

checksum

Categorical

HIGH CORRELATION

UNIQUE

Distinct	49
Distinct (%)	100.0%
Missing	0
Missing (%)	0.0%
Memory size	392.0 B

5a4991630e6548ff790dfb1... 1

1cedd350848bcfc5833a246... 1

4f5b46c311c23dc0fea67c8... 1

d23f1b5866c175323a094a0... 1

242f993afaebdd4acb0d329... 1

Other values (44) 44

Toggle details

status

Categorical

CONSTANT

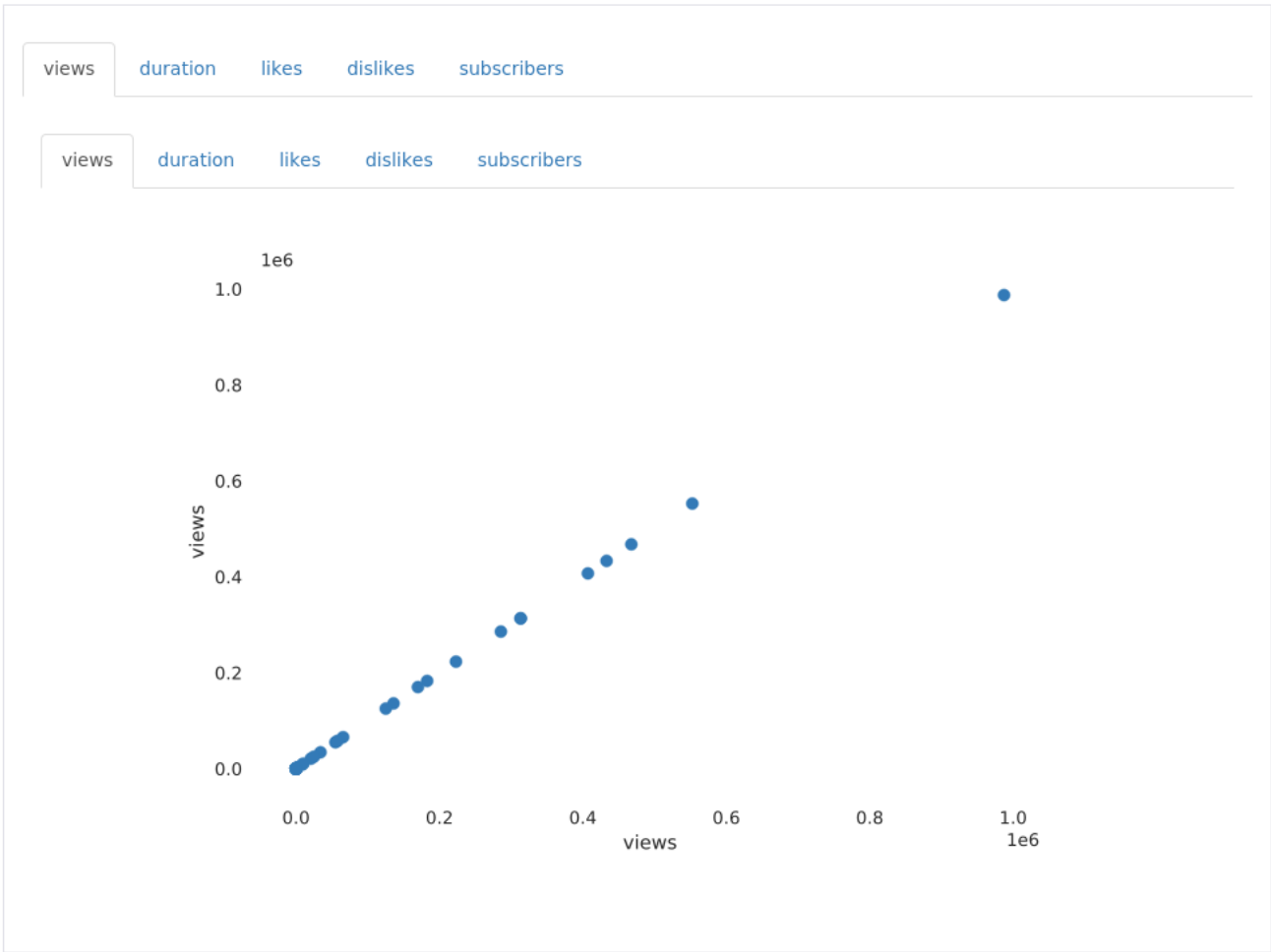
REJECTED

Distinct	1
Distinct (%)	2.0%
Missing	0
Missing (%)	0.0%
Memory size	392.0 B

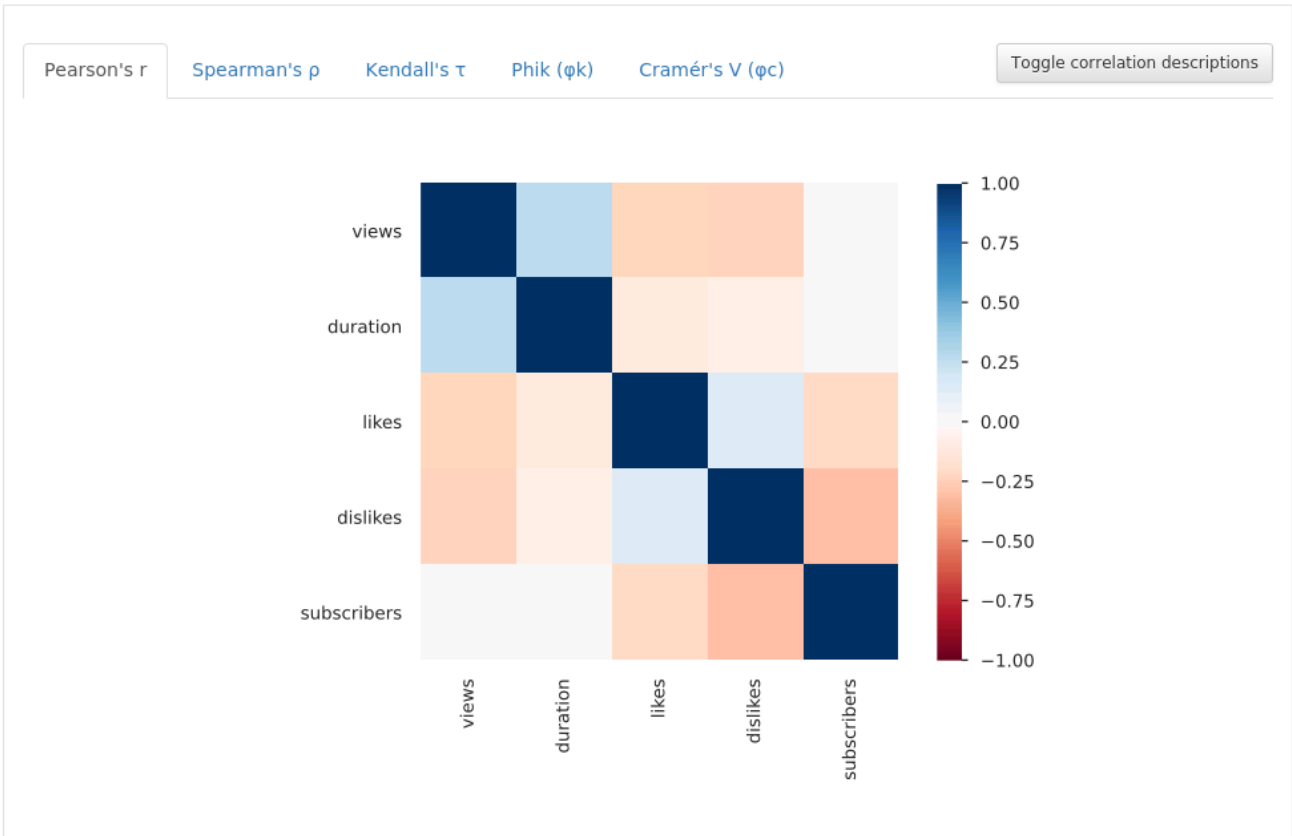
downloaded 49

Toggle details

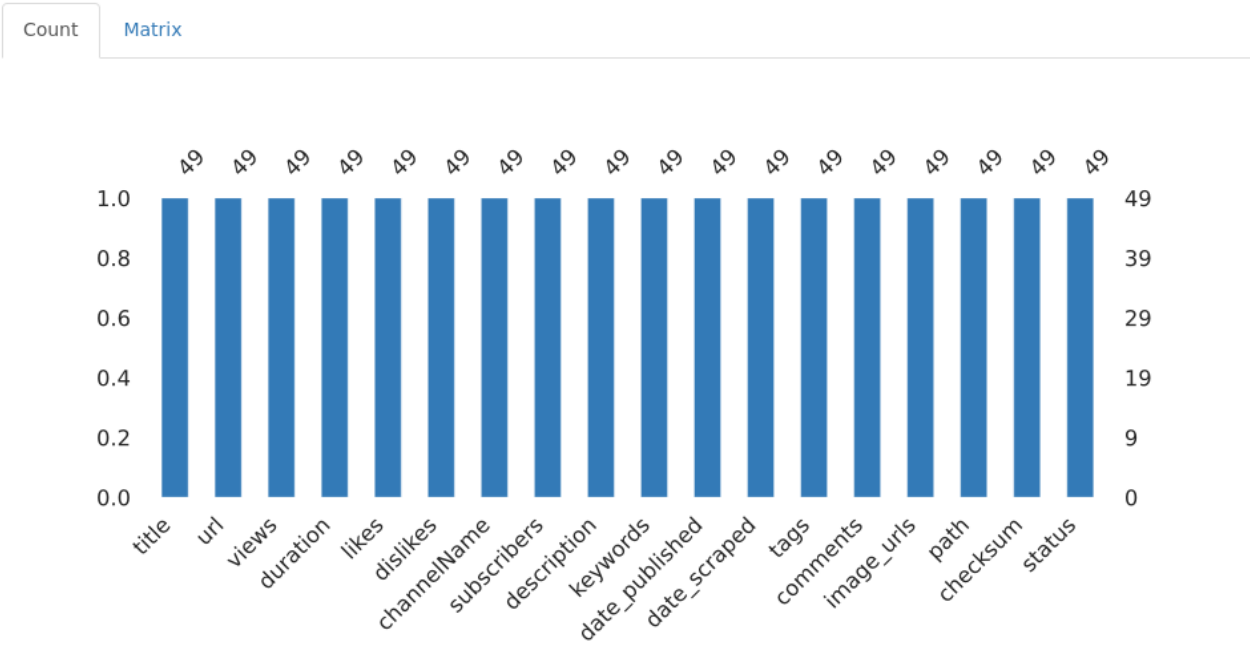
Interactions



Correlations



Missing values



Sample

First rows

	title	url
0	diogo piçarra vitor kley nada é para sempre	https://www.youtube.com/watch?v=8o41__F25
1	relatório db outubro 2020	https://www.youtube.com/watch?v=RV__bcAYc
2	maluma the weeknd hawái remix official video	https://www.youtube.com/watch?v=91vECNhv
3	now united r3hab one love official music video	https://www.youtube.com/watch?v=lv1VR5Jdg
4	mr carly ft irina barros se eu pudesse vídeo oficial	https://www.youtube.com/watch?v=1SLhz4vCl
5	nga só se vive uma vez por nós x dino d santiago pra nós	https://www.youtube.com/watch?v=m367oL35
6	highlights atletico madrid vs cádiz cf 4 0	https://www.youtube.com/watch?v=LOfM-T2jD
7	spezia 1 4 juventus ronaldo scores brace as juventus hit 4 serie a tim	https://www.youtube.com/watch?v=hOVDNSOI
8	22 22 official video gulab sidhu sidhu moose wala latest punjabi songs 2020	https://www.youtube.com/watch?v=_MqGhYiHl
9	minha nova fuc ing casaaa	https://www.youtube.com/watch?v=t82QuiDI-u

Last rows

	title	url
39	playstation 5 unboxing que maravilha	https://www.youtube
40	tipos de pessoas doce ou travessura	https://www.youtube

41	justin bieber benny blanco lonely official acoustic video	https://www.youtube
42	nelson freitas plena ft julinho ksd	https://www.youtube
43	giulia be e luan santana inesquecível video	https://www.youtube
44	justin bieber benny blanco lonely official music video	https://www.youtube
45	zé felipe só tem eu videoclipe oficial	https://www.youtube
46	comidas de uma só cor por 24 horas o desafio do alimento vermelho mukbang por 123 go challenge	https://www.youtube
47	9 miller wappa frozen prod mizzy miles	https://www.youtube
48	léo santana luísa sonza século 21	https://www.youtube

Report generated with [pandas-profiling](#).

6 - Reconocimiento

Me gustaría extender mi más sincero agradecimiento a:

[the Video Understanding group within Google Research](#).

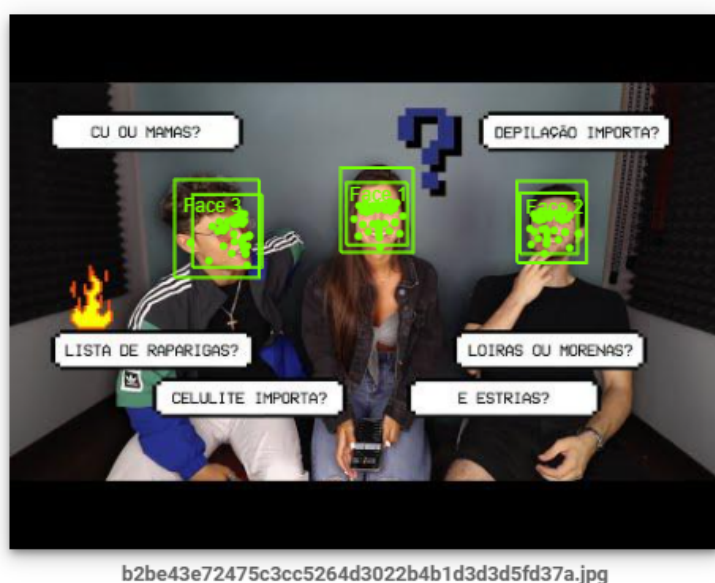
[Karan Murthy](#)

7 - Inspiration

Este conjunto de datos tiene como objetivo ser útil para análisis de Inteligencia Artificial como Predicciones de Series de Tiempo (número esperado de vistas en N días), Regresión (predecir la mejor duración) y tareas de Clasificación (automatizar etiquetas), ... Combinar imagen, texto y Los datos tabulares en el mismo modelo son una de las principales fuentes de inspiración para iniciar esta recopilación de datos.

Solo con las imágenes podemos:

- Faces detection



Face 1

Joy Very Likely

Sorrow Very Unlikely

Anger Very Unlikely

Surprise Very Unlikely

Exposed Very Unlikely

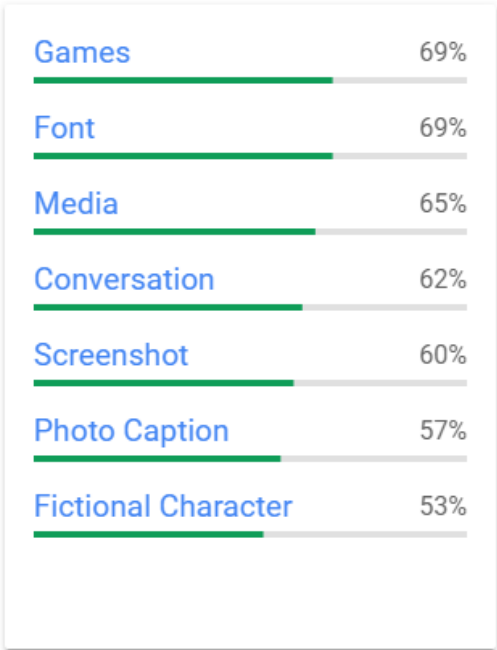
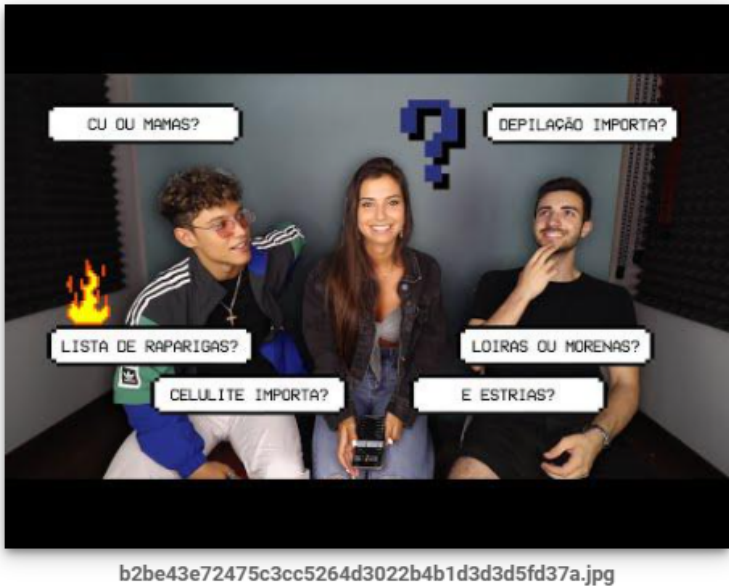
Blurred Very Unlikely

Headwear Very Unlikely

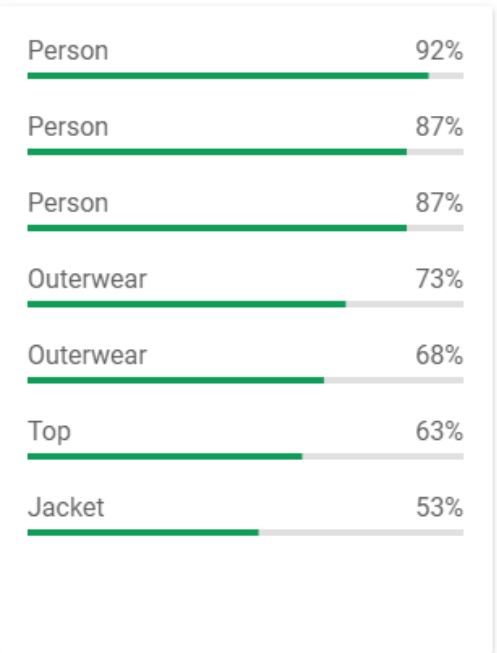
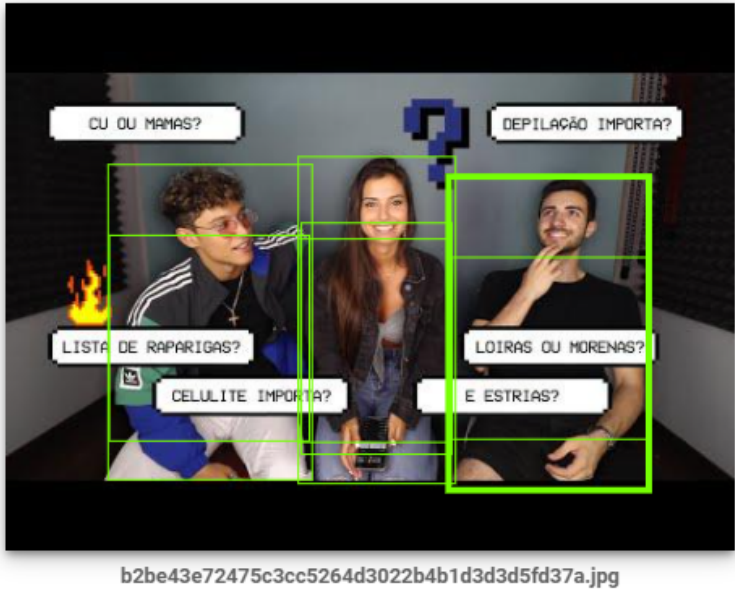
Roll: -3° Tilt: -3° Pan: -1°

Confidence 88%

- Label detection



- Object detection



- Subject detection

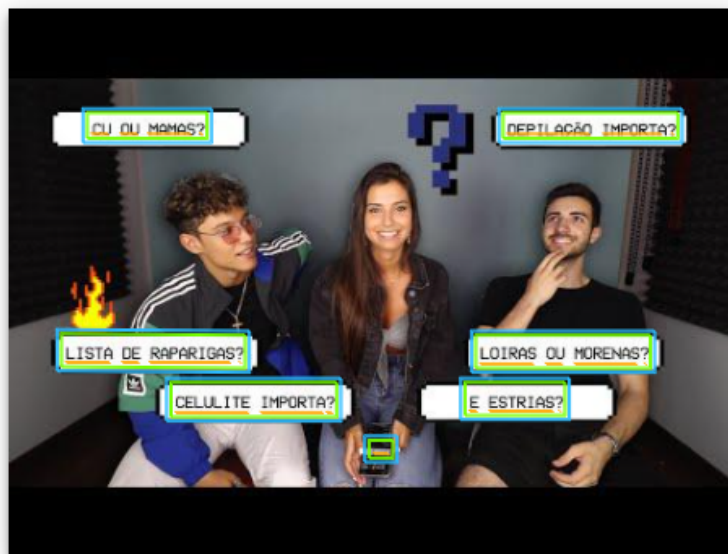


b2be43e72475c3cc5264d3022b4b1d3d3d5fd37a.jpg

Adult	<div><div></div><div></div><div></div><div></div><div></div></div>	Unlikely
Spoof	<div><div></div><div></div><div></div><div></div><div></div></div>	Likely
Medical	<div><div></div><div></div><div></div><div></div><div></div></div>	Very Unlikely
Violence	<div><div></div><div></div><div></div><div></div><div></div></div>	Unlikely
Racy	<div><div></div><div></div><div></div><div></div><div></div></div>	Unlikely

Likelihood values are Unknown, Very Unlikely, Unlikely, Possible, Likely, and Very Likely

- Text detection



b2be43e72475c3cc5264d3022b4b1d3d3d5fd37a.jpg

+Block 1

+Paragraph 1

CU OU MAMAS ?

+Block 2

+Paragraph 1

DEPILAÇÃO IMPORTA ?

+Block 3

+Paragraph 1

LISTA DE RAPARIGAS ?

8 - Licencia



Released Under CC0: Public Domain License

Elegimos esta licencia porque CC0 no impone ninguna obligación legal de proporcionar atribución, cortesía, buenas prácticas, normas y las expectativas de la comunidad a menudo significan que debe dar crédito de todos modos. Dar el crédito adecuado ayuda a otros a comprender el origen del texto para que puedan aprender más e identificar cualquier cambio que se haya realizado.

Steps

Estos pasos se realizaron manualmente con fines de prueba, pero se automatizarán mediante Apache Airflow /dags/youtube_dag.py

```
source /home/fernandovcb/virtualenvs/scrapy/bin/activate
```

```
(scrapy) fernandovcb@DESKTOP-5A84P9I:/mnt/d/BigDataLifeCycle/TrendingAnalytics$ scrapy list
YoutubeTrending
```

```
(scrapy) fernandovcb@DESKTOP-5A84P9I:/mnt/d/BigDataLifeCycle/TrendingAnalytics$ scrapy list
YoutubeTrending
```

```
(scrapy) fernandovcb@DESKTOP-5A84P9I:/mnt/d/BigDataLifeCycle/TrendingAnalytics$ scrapy crawl
YoutubeTrending -o
/mnt/d/BigDataLifeCycle/TrendingAnalytics/output/json/youtubetrending_$(date
+ "%Y%m%d").json
```

```
(scrapy) fernandovcb@DESKTOP-5A84P9I:/mnt/d/BigDataLifeCycle$ mkdir
/mnt/d/BigDataLifeCycle/TrendingAnalytics/output/images/$(date + "%Y%m%d")
```

```
(scrapy) fernandovcb@DESKTOP-5A84P9I:/mnt/d/BigDataLifeCycle$ mv -v
/mnt/d/BigDataLifeCycle/TrendingAnalytics/output/full/*
/mnt/d/BigDataLifeCycle/TrendingAnalytics/output/images/$(date + "%Y%m%d")
```

```
(scrapy) fernandovcb@DESKTOP-5A84P9I:/mnt/d/BigDataLifeCycle$
/home/fernandovcb/virtualenvs/scrapy/bin/python
/mnt/d/BigDataLifeCycle/dags/dataprocessing/prepare_image_dataset.py -d
/mnt/d/BigDataLifeCycle/TrendingAnalytics/output/images/$(date + "%Y%m%d") -
o
/mnt/d/BigDataLifeCycle/TrendingAnalytics/output/img_encoded/youtubetrendin
g_$(date + "%Y%m%d").csv
```

```
(scrapy) fernandovcb@DESKTOP-5A84P9I:/mnt/d/BigDataLifeCycle$
/home/fernandovcb/virtualenvs/scrapy/bin/python
/mnt/d/BigDataLifeCycle/dags/dataprocessing/prepare_tabular_dataset.py -d
/mnt/d/BigDataLifeCycle/TrendingAnalytics/output/json/youtubetrending_$(date
+ "%Y%m%d").json -o
```

```
/mnt/d/BigDataLifeCycle/TrendingAnalytics/output/csv/youtubetrending_tabular_$(date +"%Y%m%d").csv
```

```
(scrapy) fernandovcb@DESKTOP-5A84P9I:/mnt/d/BigDataLifeCycle$  
/home/fernandovcb/virtualenvs/scrapy/bin/python  
/mnt/d/BigDataLifeCycle/dags/dataprocessing/prepare_tabular_profiling.py -d  
/mnt/d/BigDataLifeCycle/TrendingAnalytics/output/csv/youtubetrending_tabular_$(date +"%Y%m%d").csv -o  
/mnt/d/BigDataLifeCycle/TrendingAnalytics/output/data_profiling/tabular_report_$(date +"%Y%m%d").html
```

Contribuciones	Firma
Investigación previa	Fernando Chafim
Redacción de las respuestas	Fernando Chafim
Desarrollo código	Fernando Chafim

DOI

DOI [10.5281/zenodo.4256746](https://doi.org/10.5281/zenodo.4256746)