

Hugging a Duck

Democratizing data access and
exploration with DuckDB and
Hugging Face Hub



About me

- Software Engineer at Hugging Face
- Data Scientist/ML Engineer in enterprise companies
- Degree in Computational Linguistics and Language Theory



About Hugging Face

- collaborative open-source and open-science ecosystem for ML

Our values:

- transparency
- openness
- collaboration

Our mission:

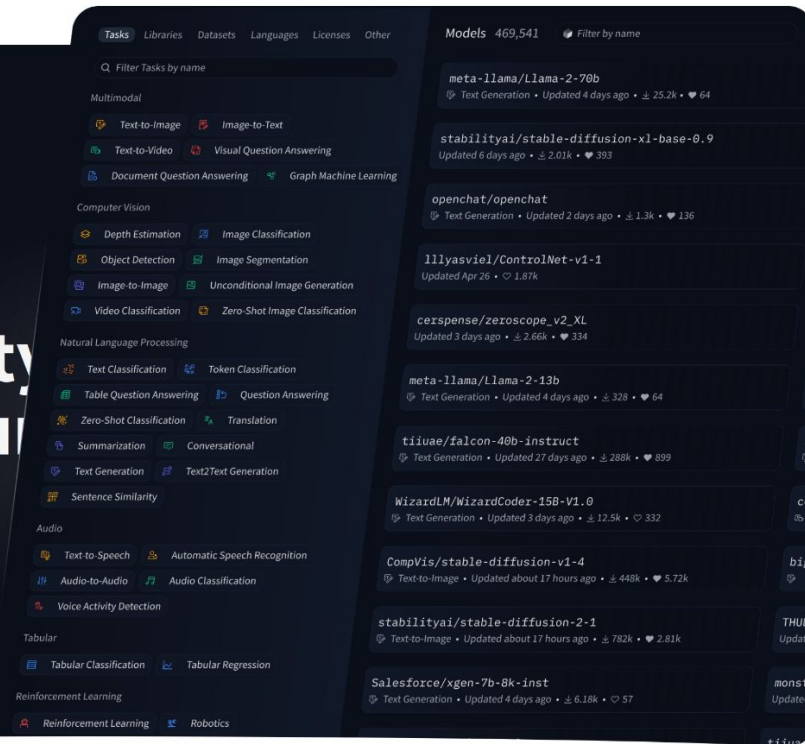
- to make good ML accessible





The AI community building the future

The platform where the machine learning community collaborates on models, datasets, and applications.





Tasks

Libraries

Datasets

Languages

Licenses

Other

Multimodal



Feature Extraction



Text-to-Image



Image-to-Text



Image-to-Video



Text-to-Video



Visual Question Answering



Document Question Answering



Graph Machine Learning



Text-to-3D



Image-to-3D

Computer Vision



Depth Estimation



Image Classification



Object Detection



Image Segmentation



Image-to-Image



Unconditional Image Generation



Video Classification



Zero-Shot Image Classification



Mask Generation



Zero-Shot Object Detection

Natural Language Processing



Text Classification



Token Classification



Table Question Answering



Question Answering



Zero-Shot Classification



Translation

Models 489,913

new Full-text search

Sort: Trending

mistralai/Mixtral-8x7B-Instruct-v0.1

Text Generation • Updated Dec 15, 2023 • ⬇ 1.21M • ❤ 2.53k

miqudev/miqu-1-70b

Text Generation • Updated 3 days ago • ❤ 234

vikhyatk/mondream1

Text Generation • Updated about 14 hours ago • ❤ 252

InstantX/InstantID

Text-to-Image • Updated 9 days ago • ⬇ 36.7k • ❤ 233

codellama/CodeLlama-70b-hf

Text Generation • Updated 2 days ago • ⬇ 550 • ❤ 147

codellama/CodeLlama-70b-Instruct-hf

Text Generation • Updated 1 day ago • ⬇ 719 • ❤ 112

MILVLG/imp-v1-3b

Text Generation • Updated about 7 hours ago • ⬇ 704 • ❤ 111

microsoft/phi-2

Text Generation • Updated 4 days ago • ⬇ 494k • ❤ 2.61k

stabilityai/stable-code-3b

Text Generation • Updated 2 days ago • ⬇ 7.46k • ❤ 446

RWKV/v5-Eagle-7B

Text Generation • Updated 2 days ago • ❤ 90

h94/IP-Adapter-FaceID

Text-to-Image • Updated 12 days ago • ⬇ 249k • ❤ 881

meta-llama/Llama-2-7b-chat-hf

Text Generation • Updated Nov 13, 2023 • ⬇ 984k • ❤ 2.58k

stabilityai/stable-diffusion-xl-base-1.0

Text-to-Image • Updated Oct 30, 2023 • ⬇ 2.98M • ❤ 4.27k

BAAI/bge-m3

Sentence Similarity • Updated about 15 hours ago • ⬇ 251 • ❤ 73

mistralai/Mixtral-8x7B-v0.1

Text Generation • Updated 10 days ago • ⬇ 146k • ❤ 1.18k

defog/sqlcoder-70b-alpha

Text Generation • Updated about 12 hours ago • ❤ 63

openai/whisper-large-v3

Automatic Speech Recognition • Updated about 12 hours ago • ⬇ 1.4M • ❤ 1.54k

mistralai/Mistral-7B-v0.1

Text Generation • Updated Dec 11, 2023 • ⬇ 753k • ❤ 2.7k

facebook/w2v-bert-2.0

Feature Extraction • Updated 6 days ago • ⬇ 2.97k • ❤ 82

codellama/CodeLlama-70b-Python-hf

Text Generation • Updated 1 day ago • ⬇ 353 • ❤ 55



Tasks

Sizes

Sub-tasks

Languages

Licenses

Other

Multimodal



Feature Extraction



Text-to-Image



Image-to-Text



Image-to-Video



Text-to-Video



Visual Question Answering



Graph Machine Learning



Text-to-3D



Image-to-3D

Computer Vision



Depth Estimation



Image Classification



Object Detection



Image Segmentation



Image-to-Image



Unconditional Image Generation



Video Classification



Zero-Shot Image Classification



Mask Generation



Zero-Shot Object Detection

Natural Language Processing



Text Classification



Token Classification



Table Question Answering



Question Answering



Zero-Shot Classification



Translation



Summarization



Conversational

Datasets 102,638

new Full-text search

Sort: Trending

litagin/moe-speech

Updated about 15 hours ago • 69 • 162

PleIAs/French-PD-Newspapers

Viewer • Updated 5 days ago • 23 • 43

math-ai/AutoMathText

Viewer • Updated about 3 hours ago • 13 • 31

glaiveai/glaive-function-calling-v2

Viewer • Updated Sep 27, 2023 • 1.23k • 137

maywell/korean_textbooks

Viewer • Updated 21 days ago • 3.24k • 61

Intel/orca_dpo_pairs

Viewer • Updated Nov 29, 2023 • 12.8k • 159

math-ai/StackMathQA

Viewer • Updated 18 days ago • 96 • 33

hieunguyenminh/roleplay

Viewer • Updated 12 days ago • 1.85k • 14

wikimedia/wikipedia

Viewer • Updated 22 days ago • 145k • 305

ai4bharat/indic-instruct-data-v0.1

Viewer • Updated 2 days ago • 105 • 11

fka/awesome-chatgpt-prompts

Viewer • Updated Mar 7, 2023 • 5.17k • 4.54k

nampdn-ai/tiny-strange-textbooks

Viewer • Updated 4 days ago • 473 • 73

PleIAs/French-PD-Books

Viewer • Updated 5 days ago • 35 • 22

HuggingFaceM4/WebSight

Viewer • Updated 15 days ago • 7.22k • 184

Locutusque/hercules-v1.0

Viewer • Updated 2 days ago • 41 • 16

argilla/distilabel-intel-orca-dpo-pairs

Viewer • Updated 9 days ago • 14k • 96

NumbersStation/NSText2SQL

Viewer • Updated 6 days ago • 238 • 48

uonlp/CulturaX

Viewer • Updated 6 days ago • 3.39k • 306

joujiboi/japanese-anime-speech

Viewer • Updated Dec 19, 2023 • 27 • 21

KBlueLeaf/danbooru2023-webp-2Mpixel

Viewer • Updated about 23 hours ago • 10

Datasets on the Hugging Face Hub



```
from datasets import load_dataset
```

```
# load any dataset from the HF Hub
```

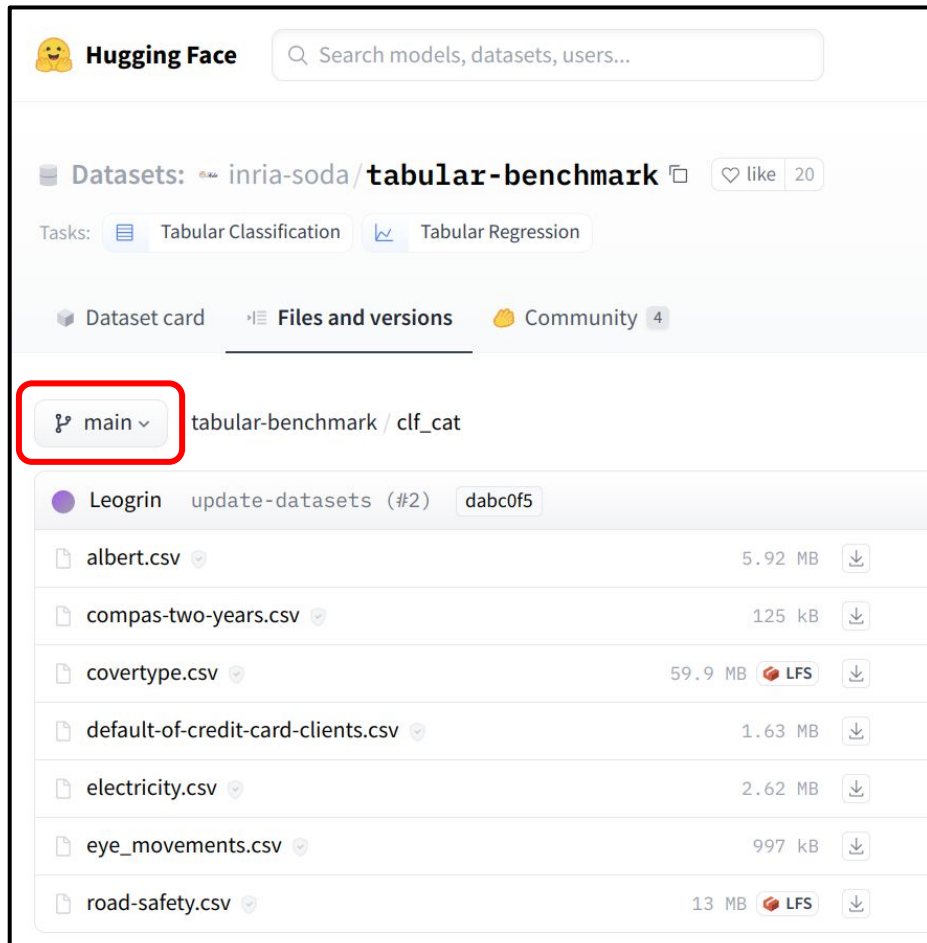
```
dataset = load_dataset("google/boolq")
```

```
# then do whatever you need:
```

```
# preprocess data and train/evaluate a model
```


Parquet export

- each dataset on the Hub is essentially a git repository containing data files
- these files can be of different formats (csv, jsonl, mp3, jpg, ...)

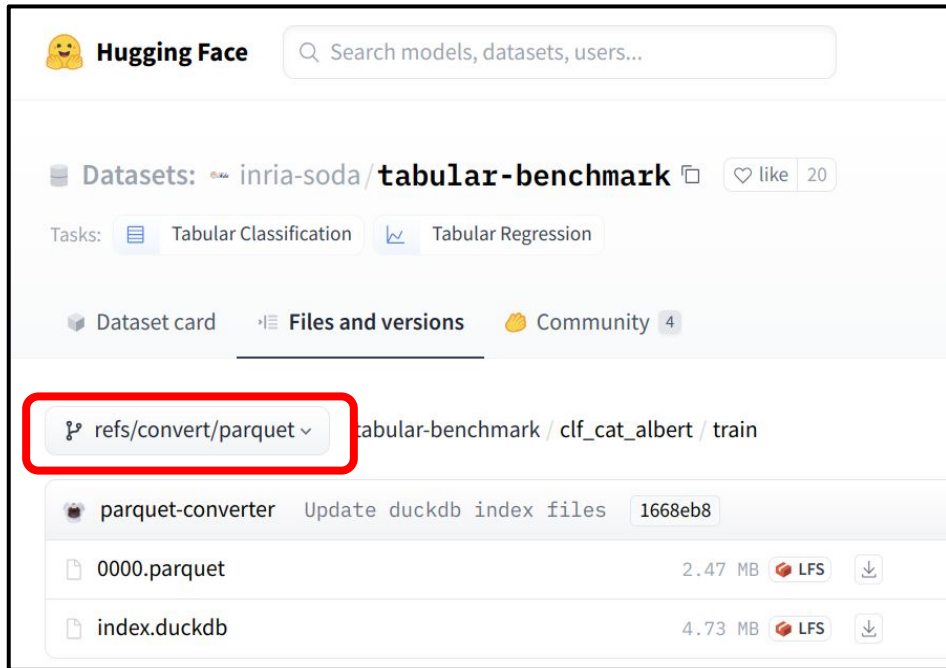


The screenshot shows the Hugging Face interface for the dataset 'tabular-benchmark' by 'inria-soda'. The page includes a search bar, dataset title, task tags (Tabular Classification, Tabular Regression), and tabs for 'Dataset card', 'Files and versions', and 'Community'. A red box highlights the 'main' branch selector. Below, a table lists files from the 'update-datasets (#2)' commit, including 'albert.csv', 'compas-two-years.csv', 'covertypes.csv', 'default-of-credit-card-clients.csv', 'electricity.csv', 'eye_movements.csv', and 'road-safety.csv'.

| File | Size | Format | Download |
|------------------------------------|---------|--------|----------|
| albert.csv | 5.92 MB | | |
| compas-two-years.csv | 125 kB | | |
| covertypes.csv | 59.9 MB | LFS | |
| default-of-credit-card-clients.csv | 1.63 MB | | |
| electricity.csv | 2.62 MB | | |
| eye_movements.csv | 997 kB | | |
| road-safety.csv | 13 MB | LFS | |

Parquet export

- each public dataset is converted to parquet format (up to first 5Gb)
- parquet files are hosted in the same dataset repository under a different branch
- parquet files are publicly accessible



```
In [1]: import duckdb

In [2]: from huggingface_hub import HfFileSystem

In [3]: duckdb.register_filesystem(HfFileSystem())

In [4]: dataset_name = "alfredodeza/wine-ratings"

In [5]: url = f"hf://datasets/{dataset_name}@~parquet/**/*.*parquet"

In [6]: # then do whatever you want with duckdb as usual

In [7]: duckdb.sql("SET enable_progress_bar = false;")

In [8]: duckdb.sql(f"FROM '{url}' LIMIT 3;")
Out[8]:
```

| name varchar | region varchar | variety varchar | rating float | notes varchar |
|-----------------------|-----------------------|--------------------|-----------------|--|
| Jim Barry Lodge Hi... | Clare Valley, Sout... | Red Wine | 90.0 | This wine has deep, dark red/black hues, lifted aromas of mulberry, raspberry... |
| Argyle Reserve Pin... | Willamette Valley,... | Red Wine | 91.0 | This Pinot Noir has a seductively rich, super dark ruby-violet color. Aromas ... |
| Cambria Katherine'... | Central Coast, Cal... | White Wine | 87.0 | The 1997 Katherines Vineyard Chardonnay is a rich, mouthfilling wine, with lu... |

Garbage in?

- thousands of datasets (and the number is growing)
- but are they garbage or not?



- goal: to help people **make informed decisions**



How to hug a duck?

- we use duckdb to power no-code data exploration on the Hub with:
 1. full-text search
 2. filtering
 3. simple statistics



1. Full text search



Tasks Sizes Sub-tasks Languages Licenses Other

Filter Tasks by name

Multimodal

Feature Extraction Text-to-Image
Image-to-Text Image-to-Video
Text-to-Video Visual Question Answering
Graph Machine Learning Text-to-3D
Image-to-3D

Computer Vision

Depth Estimation Image Classification
Object Detection Image Segmentation
Image-to-Image
Unconditional Image Generation
Video Classification
Zero-Shot Image Classification
Mask Generation
Zero-Shot Object Detection

Natural Language Processing

Text Classification Token Classification
Table Question Answering
Question Answering
Zero-Shot Classification Translation
Summarization Conversational
Text Generation Text2Text Generation
Fill-Mask Sentence Similarity

Datasets 280,341 Filter by name

new Full-text search

Sort: Trending

litagin/moe-speech

Updated about 17 hours ago • 69 • 162

PleIAs/French-PD-Newspapers

Viewer • Updated 5 days ago • 23 • 43

nampdn-ai/tiny-strange-textbooks

Viewer • Updated 4 days ago • 473 • 73

glaiveai/glaive-function-calling-v2

Viewer • Updated Sep 27, 2023 • 1.23k • 137

maywell/korean_textbooks

Viewer • Updated 21 days ago • 3.24k • 61

math-ai/StackMathQA

Viewer • Updated 18 days ago • 96 • 34

NumbersStation/NSText2SQL

Viewer • Updated 6 days ago • 238 • 48

hieunguyenminh/roleplay

Viewer • Updated 12 days ago • 1.85k • 14

wikimedia/wikipedia

Viewer • Updated 22 days ago • 145k • 305

KBlueLeaf/danbooru2023-webp-2Mpixel

Viewer • Updated 1 day ago • 10

biglam/europeana_newspapers

Viewer • Updated about 9 hours ago • 51 • 11

fka/awesome-chatgpt-prompts

Viewer • Updated Mar 7, 2023 • 5.17k • 4.54k

math-ai/AutoMathText

Viewer • Updated about 6 hours ago • 13 • 32

PleIAs/French-PD-Books

Viewer • Updated 5 days ago • 35 • 22

HuggingFaceM4/WebSight

Viewer • Updated 15 days ago • 7.22k • 184

Locutusque/hercules-v1.0

Viewer • Updated 2 days ago • 41 • 16

argilla/distilabel-intel-orca-dpo-pairs

Viewer • Updated 9 days ago • 14k • 96

Intel/orca_dpo_pairs

Viewer • Updated Nov 29, 2023 • 12.8k • 159

uonlp/CulturaX

Viewer • Updated 6 days ago • 3.39k • 306

ai4bharat/indic-instruct-data-v0.1

Viewer • Updated 2 days ago • 105 • 11

argilla/distilabel-capybara-dpo-7k-binar...

Viewer • Updated about 1 hour ago • 24 • 10

joujiboi/japanese-anime-speech

Viewer • Updated Dec 19, 2023 • 27 • 21

1. Full text search

- for each dataset create FTS index with duckdb 'PRAGMA create_fts_index' command
- 'index.duckdb' file containing FTS index is hosted on the Hub in the same utility branch with parquet files

The screenshot shows the Hugging Face interface for the dataset 'inria-soda/tabular-benchmark'. The 'Files and versions' tab is selected, displaying a list of files. The file 'index.duckdb' is highlighted with a red box. The file size is 4.73 MB and it is stored on LFS. The path shown is 'refs/convert/parquet' for 'tabular-benchmark / clf_cat_albert / train'.

Hugging Face

Search models, datasets, users...

Datasets: inria-soda / **tabular-benchmark** like 20

Tasks: Tabular Classification Tabular Regression

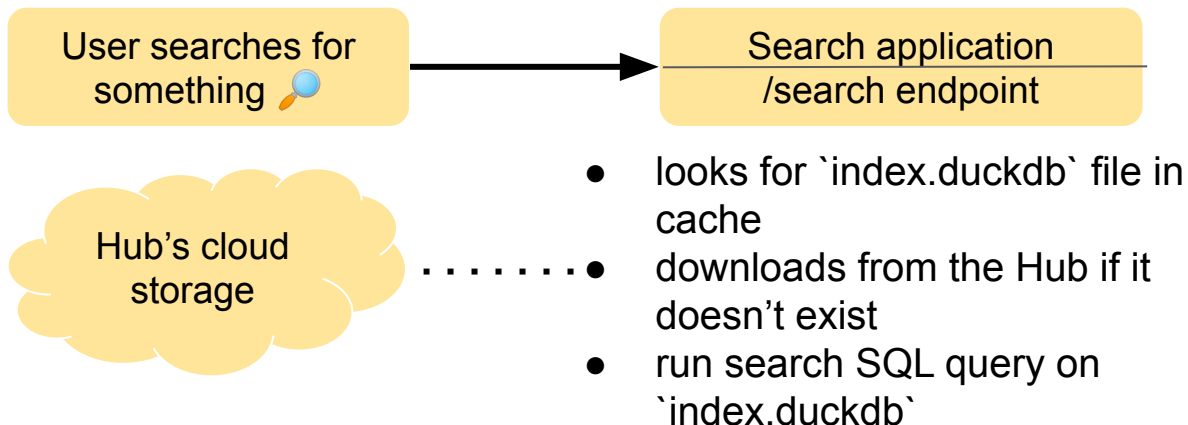
Dataset card Files and versions Community 4

refs/convert/parquet tabular-benchmark / clf_cat_albert / train

parquet-converter Update duckdb index files 1668eb8

| | | | |
|---------------------|---------|-----|----------|
| 0000.parquet | 2.47 MB | LFS | Download |
| index.duckdb | 4.73 MB | LFS | Download |

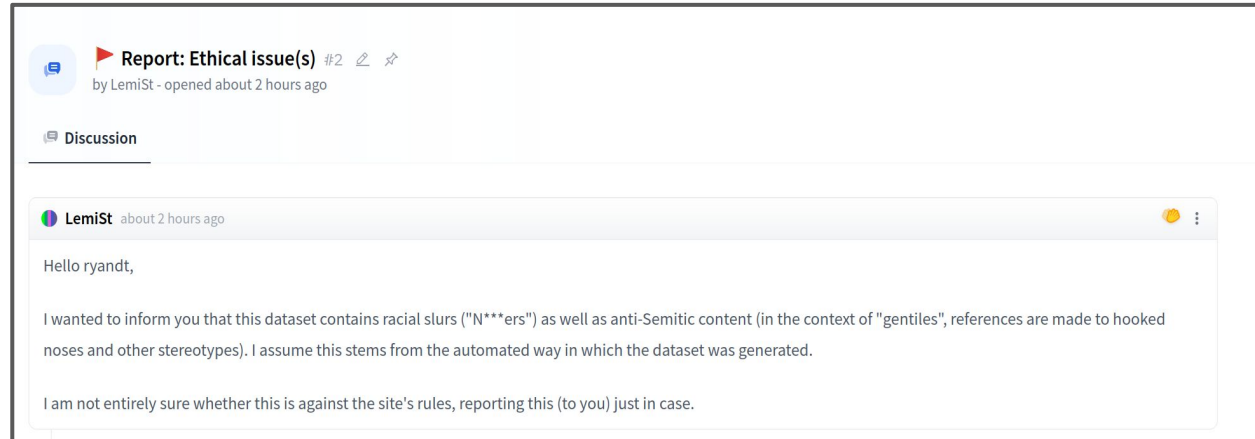
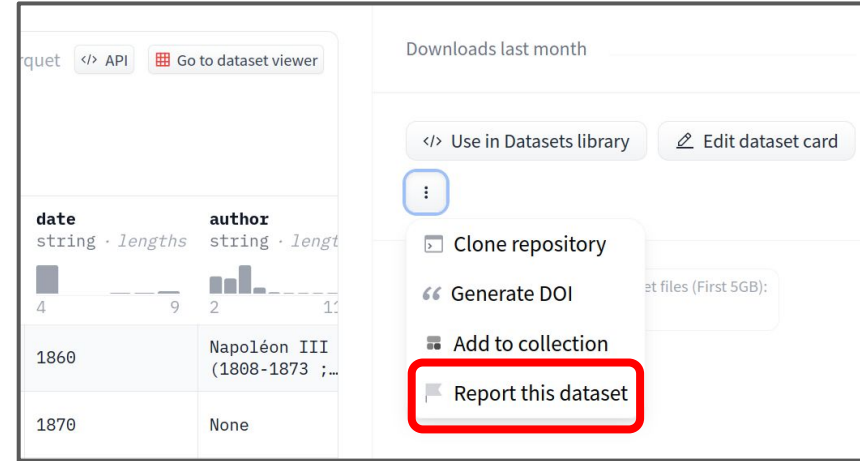
1. Full text search



```
SELECT * EXCLUDE ( __hf_fts_score )
FROM
    (SELECT *, fts_main_data.match_bm25 ( __hf_index_id, ?) AS __hf_fts_score
    FROM data) AS A
WHERE A.__hf_fts_score IS NOT NULL
ORDER BY A.__hf_fts_score DESC;
```

1. FTS for content moderation

- Duckdb-powered search makes it easier to prove claims about inappropriate content **inside** datasets




2. Filtering





Datasets: **imdb**   like 157

Tasks:  Text Classification Sub-tasks: sentiment-classification Languages:  English Multilinguality: monolingual Size Categories: 10K<n<100K


Language Creators: expert-generated Annotations Creators: expert-generated Source Datasets: original License:  other

 Dataset card  Files and versions  Community 6  Settings

Dataset Viewer

 Auto-converted to Parquet  API

Split

train (25k rows) 

 Search this dataset

text

string · lengths



label

class: label

2 classes

I rented I AM CURIOUS-YELLOW from my video store because of all the controversy that surrounded it when it was first released in 1967. I also heard that at first it was seized by U.S. customs if it ever tried to enter this country, therefor...

0 neg.

"I Am Curious: Yellow" is a risible and pretentious steaming pile. It doesn't matter what one's political views are because this film can hardly be taken seriously on any level. As for the claim that frontal male nudity is an automatic NC-17, that...

0 neg.

If only to avoid making this type of film in the future. This film is interesting as an experiment but tells no cogent story.

One might feel virtuous for sitting thru it because it touches on so many IMPORTANT issues but it does so...

0 neg.

This film was probably inspired by Godard's Masculin, féminin and I urge you to see that film instead.

The film has two strong elements and those are, (1) the realistic acting (2) the impressive, undeservedly good, photo. Apart from...

0 neg.

Oh, brother...after hearing about this ridiculous film for umpteen years all I can think of is that old Peggy Lee song..

"Is that all there is?" ...I was just an early teen when this smoked fish hit the U.S. I was too young to get in...

0 neg.

I would put this at the top of my list of films in the category of unwatchable trash! There are films that are bad, but the worst kind are the ones that are unwatchable but you are suppose to like them because they are supposed to be good for you!...

0 neg.

Whoever wrote the screenplay for this movie obviously never consulted any books about Lucille Ball, especially her autobiography. I've never seen so many mistakes in a biopic, ranging from her early years in Celoron and Jamestown to her...

0 neg.

When I first saw a glimpse of this movie, I quickly noticed the actress who was playing the role of Lucille Ball. Rachel York's portrayal of Lucy is absolutely awful. Lucille Ball was an astounding comedian with incredible talent. To think about...

0 neg.

Who are these "They"- the actors? the filmmakers? Certainly couldn't be the audience- this is among the most air-puffed

Datasets: inria-soda / tabular-benchmark

like 20

Tasks: Tabular Classification Tabular Regression

Dataset card Files and versions Community 4 Settings

Dataset Viewer

Auto-converted to Parquet API

Subset

clf_cat_covertypes (424k rows)

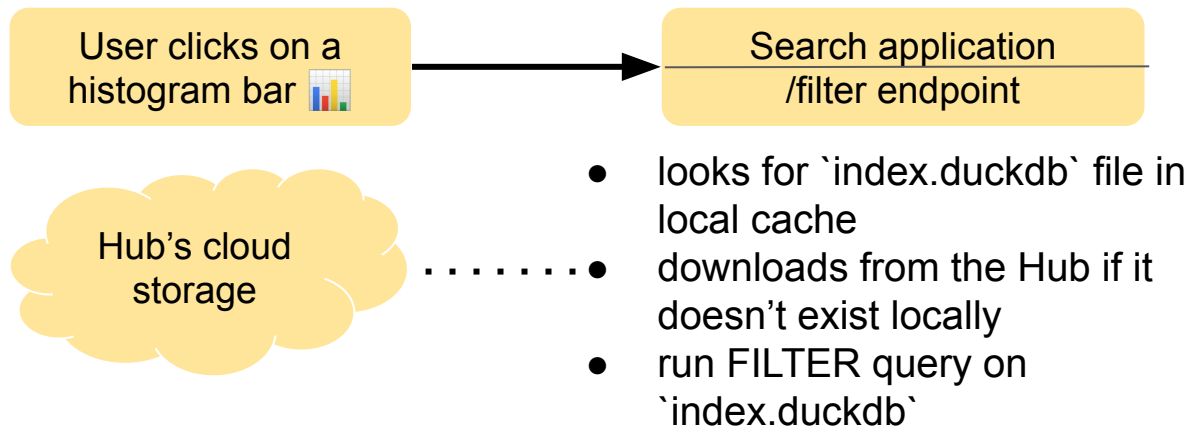
Split

train (424k rows)

Search this dataset

| Elevation float64 | Aspect float64 | Slope int64 | Horizontal_Distance_To_Hydrology float64 | Vertical_Distance_To_Hydrology float64 | Horizontal_Distance_To_Roadways float64 | Hillshade_9am int64 |
|----------------------|-------------------|----------------|---|---|--|------------------------|
| 2,14k | 3,69k | 0 | 1,39k | -166 | 7,12k | 254 |
| 3,156 | 45 | 15 | 212 | 39 | 5,208 | 223 |
| 3,164 | 346 | 2 | 295 | 33 | 3,114 | 215 |
| 2,839 | 136 | 13 | 190 | 28 | 3,000 | 240 |
| 2,924 | 324 | 14 | 60 | 11 | 4,699 | 183 |
| 3,090 | 45 | 22 | 430 | 20 | 4,108 | 220 |
| 2,912 | 151 | 2 | 60 | -2 | 4,050 | 222 |
| 3,215 | 11 | 17 | 268 | 38 | 1,719 | 201 |
| 3,117 | 27 | 7 | 351 | 8 | 4,168 | 217 |
| 2,527 | 45 | 5 | 474 | 14 | 983 | 222 |
| 2,919 | 23 | 12 | 60 | -2 | 4,470 | 213 |
| 2,991 | 66 | 8 | 228 | -18 | 4,026 | 229 |
| 2,533 | 129 | 16 | 95 | 15 | 272 | 245 |
| 2,931 | 256 | 27 | 277 | 118 | 1,566 | 146 |
| 2,632 | 59 | 17 | 30 | 5 | 1,613 | 230 |
| 3,079 | 252 | 2 | 319 | 40 | 3,267 | 214 |

2. Filtering



```
SELECT {columns} FROM data WHERE {where};
```

2. Filtering: texts lengths



Datasets: Open-Orca Open0rca like 1.04k

Tasks: Conversational Text Classification Token Classification + 7 Languages: English Size Categories: 10M<n<100M

ArXiv: arxiv:2306.02707 arxiv:2301.13688 License: mit

Dataset card Files and versions Settings

Dataset Viewer (First 5GB)

Auto-converted to Parquet API

Split

train (2.91M rows)

Search this dataset

| id | system_prompt | question | response |
|------------------|---|---|--|
| string - lengths | string - classes | string - lengths | string - lengths |
| | | | |
| 4 | 17 values | 12 40.6k | 0 15k |
| niv.242684 | | You will be given a definition of a task first, then some input of... | [["AFC Ajax (amateurs)", "has ground", "Sportpark De Toekomst"], ["Ajax Youth Academy", "plays at", "Sportpark De Toekomst"]] |
| flan.564327 | You are an AI assistant. You will be given a task. You must... | Generate an approximately fifteen-word sentence that... | Midsummer House is a moderately priced Chinese restaurant with a 3/5 customer rating, located near All Bar One. |
| flan.1875913 | You are a helpful assistant, who always provide explanation... | What happens next in this paragraph? She then rubs a needl... | C. She then dips the needle in ink and using the pencil to draw a design on her leg, rubbing it off with a rag in the end. In this option, she is... |
| t0.408370 | You are an AI assistant. You will be given a task. You must... | Please answer the following question: I want to test the... | Based on the passage, discuss the primary motivations and outcomes of the 1901 Federation of Australia, including the roles and responsibilities of... |
| cot.86217 | You are an AI assistant that helps people find information. | James runs a TV show and there are 5 main characters and 4 mino... | James pays the minor characters \$15,000 each episode. Since there are 4 minor characters, he pays them a total of $4 * \$15,000 = \$60,000$ per... |
| cot.18180 | You are an AI assistant that helps people find information. | Given the stream of consciousness rationale, provide a reasonable... | Question: What is the proper technique for a female beach volleyball player to serve the ball effectively in a game? Answer: To serve the ball... |
| flan.2136716 | You are an AI assistant. User will you give you a task. Your... | Multi-choice question: What is the sentiment of the following... | To determine the sentiment of the tweet, we need to analyze it thoroughly. Tweet: @nikkigreen I told you Step 1: Identify the words or phrases that... |
| cot.84626 | You are an AI assistant that helps people find information... | John was a terrible writer. To practice, his teacher suggest... | Step 1: Analyze the situation - John is a terrible writer and needs practice to improve his skills. His teacher gives him advice on how to... |

3. Statistics



Dataset Viewer

 Auto-converted to Parquet  API

Subset




clf_cat_covertypes (424k rows)

Split

train (424k rows)

 Search this dataset

| Elevation float64 | Aspect float64 | Slope int64 | Horizontal_Distance_To_Hydrology float64 | Vertical_Distance_To_Hydrology float64 | Horizontal_Distance_To_Roadways float64 | Hillshade_9am int64 |
|---|---|---|---|---|--|---|
|  |  |  |  |  |  |  |
| 2.14k 3.69k | 0 360 | 0 66 | 0 1.39k | -166 598 | 0 7.12k | 0 254 |
| 3,156 | 45 | 15 | 212 | 39 | 5,208 | 223 |
| 3,164 | 346 | 2 | 295 | 33 | 3,114 | 215 |
| 2,839 | 136 | 13 | 190 | 28 | 3,000 | 240 |
| 2,924 | 324 | 14 | 60 | 11 | 4,699 | 183 |
| 3,090 | 45 | 22 | 430 | 20 | 4,108 | 220 |
| 2,912 | 151 | 2 | 60 | -2 | 4,050 | 222 |
| 3,215 | 11 | 17 | 268 | 38 | 1,719 | 201 |
| 3,117 | 27 | 7 | 351 | 8 | 4,168 | 217 |
| 2,527 | 45 | 5 | 474 | 14 | 983 | 222 |
| 2,919 | 23 | 12 | 60 | -2 | 4,470 | 213 |
| 2,991 | 66 | 8 | 228 | -18 | 4,026 | 229 |
| 2,533 | 129 | 16 | 95 | 15 | 272 | 245 |

 Datasets: **glue**   like 275

Tasks:  Text Classification Sub-tasks: **acceptability-classification** **natural-language-inference** **semantic-similarity-scoring** + 2 Languages:  English

Multilinguality: **monolingual** Size Categories: **10K<n<100K** Language Creators: **other** Annotations Creators: **other** Source Datasets: **original** ArXiv:  arxiv:1804.07461

Tags: **qa-nli** **coreference-nli** **paraphrase-identification** License:  other

 **Dataset card**  Files and versions  Community **19**  Settings

Dataset Viewer

 Auto-converted to Parquet  API

Subset

cola (10.7k rows)

Split

train (8.55k rows)

 Search this dataset

sentence

string · lengths



label

class label



idx

int32



Our friends won't buy this analysis, let alone the next one we propose.

1 acceptable

0

One more pseudo generalization and I'm giving up.

1 acceptable

1

One more pseudo generalization or I'm giving up.

1 acceptable

2

The more we study verbs, the crazier they get.

1 acceptable

3

Day by day the facts are getting murkier.

1 acceptable

4

I'll fix you a drink.

1 acceptable

5

Fred watered the plants flat.

1 acceptable

6

Bill coughed his way out of the restaurant.

1 acceptable

7

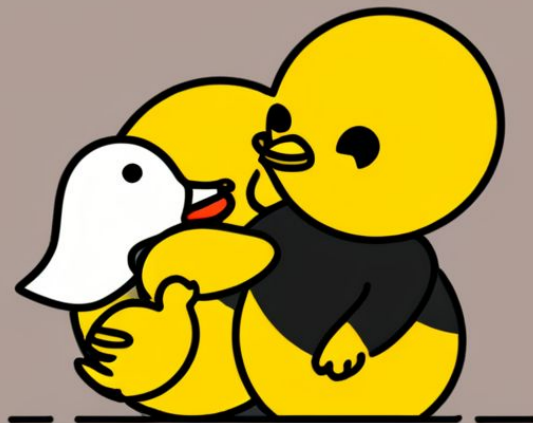
We're dancing the night away.

1 acceptable

8

3. Statistics

- download dataset's parquet files from the Hub
- load them with `duckdb`
- run SQL queries to compute different measurements
- store results in a database of precomputed responses (MongoDB)



Our experience

- super easy and lightweight (no extra complicated setup for search lol)
- reliable
- **versatile**
- DuckDB team support 💜



Check it out!

- play with the [dataset viewer](#)
- open [feature requests and bug reports](#)
- check out the [API docs](#)

contact me:

- polina@huggingface.co
- [LinkedIn](#)



Congratulations!

You are now
hugging 🤗
a duck 🦆



