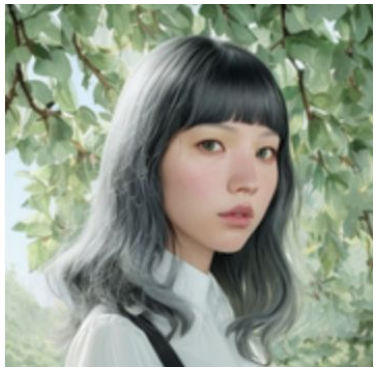
情報ゼロから Polars を触ってみた

Polars Data Crunch #1

自己紹介



- 翡翠（ひすい）X アカウント @kiro_anna
- 某教育系企業に所属
- 元 SE から AI エンジニア
- 現在、AI・機械学習エンジニア（自称）、データサイエンスをやる人
- Kaggle 挑戦中

初 LT 

よろしくお願いします。

最近よく目にする Polars って何？

Pandas よりも高速に処理できるライブラリ？

使う場面がぱっとしない

最近よく目にする Polars って何？

Pandas よりも高速に処理できるライブラリ？

使う場面がぱっとしない

というわけで、使ってみた！！

Polars vs Pandas

▼ 検証内容

- Pandas でよく使う処理を用いてPolars とのちがいを比較
- メモリの…という話ではなく、今回はぱっと見てわかりやすい処理時間で比較
- データ量の多いファイルを扱ってみる（約3GB）

データの読み込み

```
start_time = time.time()
df_pandas = pd.read_csv(file_path)
pandas_read_time = time.time() - start_time

print(f"Pandas read time: {pandas_read_time:.2f} seconds")
```

✓ 49.8s

```
Pandas read time: 49.69 seconds
```

```
start_time = time.time()
df_polars = pl.read_csv(file_path)
polars_read_time = time.time() - start_time

print(f"Polars read time: {polars_read_time:.2f} seconds")
```

✓ 7.7s

```
Polars read time: 7.76 seconds
```

データのフィルタリング

```
start_time = time.time()
pandas_filtered = df_pandas[df_pandas["passenger_count"] > 100]
pandas_filter_time = time.time() - start_time

print(f"Pandas filter time: {pandas_filter_time:.2f} seconds")
```

```
Pandas filter time: 0.20 seconds
```

```
start_time = time.time()
polars_filtered = df_polars.filter(pl.col("passenger_count") > 100)
polars_filter_time = time.time() - start_time

print(f"Polars filter time: {polars_filter_time:.2f} seconds")
```

```
Polars filter time: 0.42 seconds
```

データのグループ化と集計

```python
start_time = time.time()
pandas_grouped = pandas_filtered.groupby("VendorID")["total_amount"].sum().reset_index()
pandas_groupby_time = time.time() - start_time

print(f"Pandas groupby time: {pandas_groupby_time:.2f} seconds")
```

```
Pandas groupby time: 0.02 seconds
```

```python
start_time = time.time()
polars_grouped = polars_filtered.group_by("VendorID").agg(pl.sum("total_amount"))
polars_groupby_time = time.time() - start_time

print(f"Polars groupby time: {polars_groupby_time:.2f} seconds")
```

```
Polars groupby time: 0.01 seconds
```

データの並び替え

```
start_time = time.time()
pandas_sorted = df_pandas.sort_values("trip_distance", ascending=False)
pandas_sort_time = time.time() - start_time

print(f"Pandas sort time: {pandas_sort_time:.2f} seconds")
```

```
Pandas sort time: 33.08 seconds
```

```
start_time = time.time()
polars_sorted = df_polars.sort("trip_distance")
polars_sort_time = time.time() - start_time

print(f"Polars sort time: {polars_sort_time:.2f} seconds")
```

```
Polars sort time: 12.50 seconds
```

ユニークな値の取得

```
start_time = time.time()
pandas_unique = df_pandas["DOLocationID"].unique()
pandas_unique_time = time.time() - start_time

print(f"Pandas unique time: {pandas_unique_time:.2f} seconds")
```

```
Pandas unique time: 0.18 seconds
```

```
start_time = time.time()
polars_unique = df_polars["DOLocationID"].unique()
polars_unique_time = time.time() - start_time

print(f"Polars unique time: {polars_unique_time:.2f} seconds")
```

```
Polars unique time: 1.59 seconds
```

データのサンプリング

```
start_time = time.time()
pandas_sampled = df_pandas.sample(frac=0.1)
pandas_sample_time = time.time() - start_time

print(f"Pandas sample time: {pandas_sample_time:.2f} seconds")
```

```
Pandas sample time: 21.85 seconds
```

```
start_time = time.time()
polars_sampled = df_polars.sample(n=int(0.1 * len(df_polars)), seed=random.randint(0,
10000))
polars_sample_time = time.time() - start_time

print(f"Polars sample time: {polars_sample_time:.2f} seconds")
```

```
Polars sample time: 9.36 seconds
```

欠損値の処理

```
start_time = time.time()
pandas_filled = df_pandas["tpep_dropoff_datetime"].fillna(0, inplace=False)
pandas_fill_time = time.time() - start_time

print(f"Pandas fill time: {pandas_fill_time:.2f} seconds")
```

```
Pandas fill time: 2.07 seconds
```

```
start_time = time.time()
polars_filled = df_polars.with_columns([pl.col("tpep_dropoff_datetime").fill_null(0).alias
("tip_amount")])    "tpep": Unknown word.
polars_fill_time = time.time() - start_time

print(f"Polars fill time: {polars_fill_time:.2f} seconds")
```

Python

```
Polars fill time: 0.03 seconds
```

使ってみた感想

▼ Kaggle

- 今まで Pandas を使っていたが、Polars の方が良いかもしれない

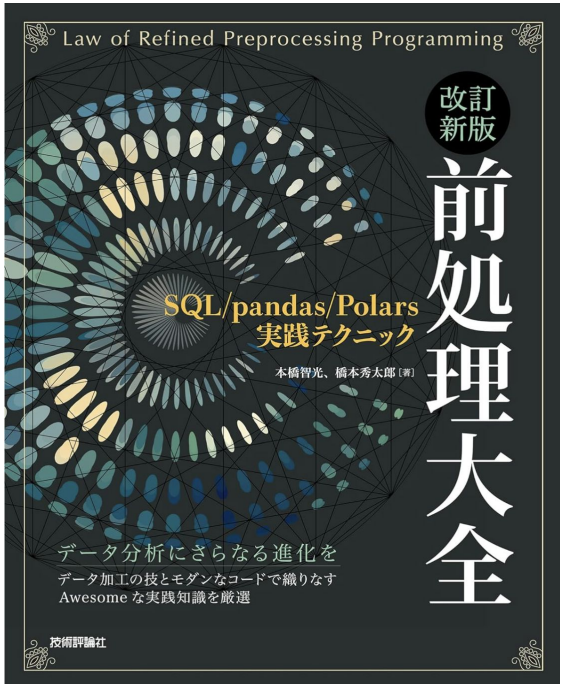
▼ PyTorchLightning との相性は？（要調査）

- import pytorch_lightning as pl vs import polars as pl

▼ 新しい技術を取り入れるタイミングの難しさ

- 業務で扱う実装で Polars を取り入れるタイミング、難しい

その他情報について



▼ 改訂新版 前処理大全～SQL/pandas/Polars実践テクニック

amzn.to/3USSV5P @amazon

読んでみたい本リストに追加！

改定版
2024/5/22 発売日

タイムリー！！