read_predictions

September 23, 2025

[1]:
```
import pandas as pd
import json
```

[2]:
```
# load JSONL
records = []
with open("predictions.tsv") as f:
    for line in f:
        records.append(json.loads(line))
```

[3]:
```
df = pd.DataFrame(records)
df.head()
```

[3]:
```
                                                left  \
0  COL cictt_make VAL BOMBARDIER COL cictt_model …
1  COL cictt_make VAL BEECH COL cictt_model VAL 3…
2  COL cictt_make VAL BEECH COL cictt_model VAL 9…
3  COL cictt_make VAL BEECH COL cictt_model VAL 2…
4  COL cictt_make VAL SCHWEIZER COL cictt_model V…

                                               right  match  match_confidence
0  COL make VAL TIMM COL model VAL COLEGT COL ser…      0          0.999727
1  COL make VAL NORWST COL model VAL 35 COL serie…      0          0.999738
2  COL make VAL BE COL model VAL 400 COL series V…      0          0.999860
3  COL make VAL HS COL model VAL 125 COL series V…      0          0.999929
4  COL make VAL WACO COL model VAL U COL series V…      0          0.999934
```

[4]:
```
gold = pd.read_csv("data/er_magellan/Structured/ditto_aircraft/all_pairs.txt",␣
 ↪sep="\t", header=None, names=["left", "right", "gold"])
print(gold.head())
```

```
                                                left  \
0  COL cictt_make VAL BOMBARDIER COL cictt_model …
1  COL cictt_make VAL BEECH COL cictt_model VAL 3…
2  COL cictt_make VAL BEECH COL cictt_model VAL 9…
3  COL cictt_make VAL BEECH COL cictt_model VAL 2…
4  COL cictt_make VAL SCHWEIZER COL cictt_model V…

                                               right  gold
```

```
0  COL make VAL TIMM COL model VAL COLEGT COL ser…     0
1  COL make VAL NORWST COL model VAL 35 COL serie…     0
2  COL make VAL BE COL model VAL 400 COL series V…     0
3  COL make VAL HS COL model VAL 125 COL series V…     0
4  COL make VAL WACO COL model VAL U COL series V…     0
```

[5]:

```
df["gold"] = gold["gold"]
```

[6]:

```
df
```

[6]:

```
                                                   left  \
0     COL cictt_make VAL BOMBARDIER COL cictt_model …
1     COL cictt_make VAL BEECH COL cictt_model VAL 3…
2     COL cictt_make VAL BEECH COL cictt_model VAL 9…
3     COL cictt_make VAL BEECH COL cictt_model VAL 2…
4     COL cictt_make VAL SCHWEIZER COL cictt_model V…
…                                                 …
7349  COL cictt_make VAL CALLAIR COL cictt_model VAL…
7350  COL cictt_make VAL AEROSPATIALE COL cictt_mode…
7351  COL cictt_make VAL COLUMBIA COL cictt_model VA…
7352  COL cictt_make VAL PIPER COL cictt_model VAL P…
7353  COL cictt_make VAL ROCKWELL COL cictt_model VA…

                                                  right  match  \
0     COL make VAL TIMM COL model VAL COLEGT COL ser…      0
1     COL make VAL NORWST COL model VAL 35 COL serie…      0
2     COL make VAL BE COL model VAL 400 COL series V…      0
3     COL make VAL HS COL model VAL 125 COL series V…      0
4     COL make VAL WACO COL model VAL U COL series V…      0
…                                                 …    …
7349  COL make VAL B COL model VAL 75 COL series VAL…      0
7350  COL make VAL AS COL model VAL 350C COL series …      1
7351  COL make VAL XJL COL model VAL 1 COL series VAL 1      1
7352  COL make VAL PA COL model VAL 25 COL series VA…      1
7353  COL make VAL GA COL model VAL 112 COL series V…      1

      match_confidence  gold
0             0.999727     0
1             0.999738     0
2             0.999860     0
3             0.999929     0
4             0.999934     0
…                  …   …
7349          0.999865     0
7350          0.999933     1
7351          0.999952     1
7352          0.999929     1
```

```
7353          0.999950     1

[7354 rows x 5 columns]
```

[7]:

```
from sklearn.metrics import classification_report, confusion_matrix,␣
 ↪accuracy_score

y_true = df["gold"]
y_pred = df["match"]

print("Accuracy:", accuracy_score(y_true, y_pred))
print("\nClassification report:\n", classification_report(y_true, y_pred))
print("\nConfusion matrix:\n", confusion_matrix(y_true, y_pred))
```

```
Accuracy: 0.9961925482730487

Classification report:
               precision    recall  f1-score   support

           0       1.00      0.99      1.00      3677
           1       0.99      1.00      1.00      3677

    accuracy                           1.00      7354
   macro avg       1.00      1.00      1.00      7354
weighted avg       1.00      1.00      1.00      7354


Confusion matrix:
 [[3655   22]
 [   6 3671]]
```

[8]:

```
errors = df[df["gold"] != df["match"]]
print(errors[["left","right","gold","match","match_confidence"]].head(20))
```

```
                                                   left  \
410   COL cictt_make VAL NAVION COL cictt_model VAL ...
497   COL cictt_make VAL BEECH COL cictt_model VAL 1...
644   COL cictt_make VAL AERONCA COL cictt_model VAL...
1554  COL cictt_make VAL AMERICAN BLIMP COL cictt_mo...
1697  COL cictt_make VAL BOEING COL cictt_model VAL ...
2002  COL cictt_make VAL BEECH COL cictt_model VAL 2...
2108  COL cictt_make VAL STEARMAN COL cictt_model VA...
3192  COL cictt_make VAL BOEING COL cictt_model VAL ...
3819  COL cictt_make VAL CONVAIR COL cictt_model VAL...
4080  COL cictt_make VAL CURTISS WRIGHT COL cictt_mo...
4336  COL cictt_make VAL ERCOUPE COL cictt_model VAL...
4391  COL cictt_make VAL AEROSPATIALE COL cictt_mode...
4539  COL cictt_make VAL EMBRAER COL cictt_model VAL...
```

```
4834  COL cictt_make VAL AERONCA COL cictt_model VAL…
4933  COL cictt_make VAL CURTISS WRIGHT COL cictt_mo…
5031  COL cictt_make VAL NAVION COL cictt_model VAL …
5460  COL cictt_make VAL BELL COL cictt_model VAL 47…
5488  COL cictt_make VAL BELLANCA COL cictt_model VA…
5496  COL cictt_make VAL BEECH COL cictt_model VAL 1…
5660  COL cictt_make VAL LOCKHEED COL cictt_model VA…

                                                  right  gold  match  \
410   COL make VAL BL COL model VAL 1413 COL series …     0      1
497   COL make VAL BE COL model VAL 17 COL series VAL R     1      0
644   COL make VAL BHT COL model VAL 206 COL series …     0      1
1554  COL make VAL ABC COL model VAL A COL series VA…     1      0
1697  COL make VAL B COL model VAL 737 COL series VA…     0      1
2002  COL make VAL MOONEY COL model VAL 22 COL serie…     0      1
2108  COL make VAL HAWKER COL model VAL 750 COL seri…     0      1
3192  COL make VAL B COL model VAL 727 COL series VA…     0      1
3819  COL make VAL CV COL model VAL 640 COL series V…     1      0
4080  COL make VAL GLASFL COL model VAL KESTRL COL s…     0      1
4336  COL make VAL VIZOLA COL model VAL A21 COL seri…     0      1
4391  COL make VAL AS COL model VAL 350D COL series …     0      1
4539  COL make VAL AERORS COL model VAL J2 COL serie…     0      1
4834  COL make VAL SCHLER COL model VAL ASW12 COL se…     0      1
4933  COL make VAL GA COL model VAL 690 COL series V…     0      1
5031  COL make VAL AETNA COL model VAL 2SA COL serie…     0      1
5460  COL make VAL BHT COL model VAL 47 COL series V…     1      0
5488  COL make VAL SWALOW COL model VAL SWALOW COL s…     0      1
5496  COL make VAL BE COL model VAL 17 COL series VAL L     1      0
5660  COL make VAL DH COL model VAL 114 COL series V…     0      1

      match_confidence
410           0.986928
497           0.936396
644           0.677968
1554          0.891472
1697          0.978735
2002          0.918447
2108          0.999843
3192          0.773282
3819          0.999891
4080          0.981872
4336          0.688484
4391          0.999923
4539          0.753194
4834          0.846940
4933          0.728909
5031          0.990620
5460          0.449317
```

```
5488          0.999676
5496          0.566662
5660          0.986106
```

[10]:
```
errors
```

[10]:
```
                                                  left  \
410   COL cictt_make VAL NAVION COL cictt_model VAL ...
497   COL cictt_make VAL BEECH COL cictt_model VAL 1...
644   COL cictt_make VAL AERONCA COL cictt_model VAL...
1554  COL cictt_make VAL AMERICAN BLIMP COL cictt_mo...
1697  COL cictt_make VAL BOEING COL cictt_model VAL ...
2002  COL cictt_make VAL BEECH COL cictt_model VAL 2...
2108  COL cictt_make VAL STEARMAN COL cictt_model VA...
3192  COL cictt_make VAL BOEING COL cictt_model VAL ...
3819  COL cictt_make VAL CONVAIR COL cictt_model VAL...
4080  COL cictt_make VAL CURTISS WRIGHT COL cictt_mo...
4336  COL cictt_make VAL ERCOUPE COL cictt_model VAL...
4391  COL cictt_make VAL AEROSPATIALE COL cictt_mode...
4539  COL cictt_make VAL EMBRAER COL cictt_model VAL...
4834  COL cictt_make VAL AERONCA COL cictt_model VAL...
4933  COL cictt_make VAL CURTISS WRIGHT COL cictt_mo...
5031  COL cictt_make VAL NAVION COL cictt_model VAL ...
5460  COL cictt_make VAL BELL COL cictt_model VAL 47...
5488  COL cictt_make VAL BELLANCA COL cictt_model VA...
5496  COL cictt_make VAL BEECH COL cictt_model VAL 1...
5660  COL cictt_make VAL LOCKHEED COL cictt_model VA...
5743  COL cictt_make VAL DOUGLAS COL cictt_model VAL...
5796  COL cictt_make VAL CESSNA COL cictt_model VAL ...
5837  COL cictt_make VAL SIKORSKY COL cictt_model VA...
5875  COL cictt_make VAL BELLANCA COL cictt_model VA...
6047  COL cictt_make VAL BOEING COL cictt_model VAL ...
6275  COL cictt_make VAL BELLANCA COL cictt_model VA...
6368  COL cictt_make VAL TAYLORCRAFT COL cictt_model...
7037  COL cictt_make VAL EMBRAER COL cictt_model VAL...

                                                 right  match  \
410   COL make VAL BL COL model VAL 1413 COL series ...      1
497     COL make VAL BE COL model VAL 17 COL series VAL R      0
644   COL make VAL BHT COL model VAL 206 COL series ...      1
1554  COL make VAL ABC COL model VAL A COL series VA...      0
1697  COL make VAL B COL model VAL 737 COL series VA...      1
2002  COL make VAL MOONEY COL model VAL 22 COL serie...      1
2108  COL make VAL HAWKER COL model VAL 750 COL seri...      1
3192  COL make VAL B COL model VAL 727 COL series VA...      1
3819  COL make VAL CV COL model VAL 640 COL series V...      0
4080  COL make VAL GLASFL COL model VAL KESTRL COL s...      1
```

```
4336  COL make VAL VIZOLA COL model VAL A21 COL seri…      1
4391  COL make VAL AS COL model VAL 350D COL series …      1
4539  COL make VAL AERORS COL model VAL J2 COL serie…      1
4834  COL make VAL SCHLER COL model VAL ASW12 COL se…      1
4933  COL make VAL GA COL model VAL 690 COL series V…      1
5031  COL make VAL AETNA COL model VAL 2SA COL serie…      1
5460  COL make VAL BHT COL model VAL 47 COL series V…      0
5488  COL make VAL SWALOW COL model VAL SWALOW COL s…      1
5496  COL make VAL BE COL model VAL 17 COL series VAL L      0
5660  COL make VAL DH COL model VAL 114 COL series V…      1
5743   COL make VAL DC COL model VAL 2 COL series VAL 2      1
5796  COL make VAL NAVION COL model VAL NAVION COL s…      1
5837  COL make VAL SK COL model VAL 58 COL series VA…      1
5875  COL make VAL HELIO COL model VAL 500 COL serie…      1
6047  COL make VAL B COL model VAL 737 COL series VA…      1
6275  COL make VAL BL COL model VAL 149 COL series V…      0
6368  COL make VAL CURTIS COL model VAL TRVAIR COL s…      1
7037  COL make VAL EMB COL model VAL 145 COL series …      1
```

```
      match_confidence  gold
410           0.986928     0
497           0.936396     1
644           0.677968     0
1554          0.891472     1
1697          0.978735     0
2002          0.918447     0
2108          0.999843     0
3192          0.773282     0
3819          0.999891     1
4080          0.981872     0
4336          0.688484     0
4391          0.999923     0
4539          0.753194     0
4834          0.846940     0
4933          0.728909     0
5031          0.990620     0
5460          0.449317     1
5488          0.999676     0
5496          0.566662     1
5660          0.986106     0
5743          0.999898     0
5796          0.710229     0
5837          0.998830     0
5875          0.771497     0
6047          0.828460     0
6275          0.573050     1
6368          0.663645     0
```

```
7037          0.868521     0
```

[9]:

```
errors.to_csv("errors_review.csv", index=False)
```

[10]:

```
import re

def parse_record(record: str):
    """Parse Ditto serialized record into a dict of {field: value}."""
    parts = re.split(r"COL |VAL ", record.strip())
    parts = [p for p in parts if p]  # drop empties
    return {parts[i].strip(): parts[i+1].strip() for i in range(0, len(parts),␣
 ↪2)}
```

[11]:

```
parsed = []

for _, row in errors.iterrows():
    left = parse_record(row["left"])
    right = parse_record(row["right"])
    parsed.append({
        "cictt_make": left.get("cictt_make"),
        "make": right.get("make"),
        "cictt_model": left.get("cictt_model"),
        "model": right.get("model"),
        "cictt_series": left.get("cictt_series"),
        "series": right.get("series"),
        "gold": row["gold"],
        "predicted": row["match"],
        "confidence": row["match_confidence"]
    })

aligned = pd.DataFrame(parsed)
```

[12]:

```
aligned.to_csv("aligned_errors_review.csv", index=False)
```