demo_usage_sherlock

May 11, 2021

[1]:
```
import sys
sys.path.append("..")
```

[2]:
```
import pandas as pd
from sherlock.features.preprocessing import convert_string_lists_to_lists,␣
↪extract_features
```

[3]:
```
test_samples = pd.read_parquet('../data/data/raw/test_values.parquet').head(5)
test_labels = pd.read_parquet('../data/data/raw/test_labels.parquet').head(5)
y_test_processed = pd.read_parquet('../data/data/processed/y_test.parquet').
↪head(5)
X_test_processed = pd.read_parquet('../data/data/processed/X_test.parquet').
↪head(5)

test_samples_converted, y_test = convert_string_lists_to_lists(test_samples,␣
↪test_labels, "values", "type")
X_test_extracted = extract_features(test_samples_converted)
```

```
100%|      | 5/5 [00:00<00:00, 6521.00it/s]

Preparing feature extraction by downloading 2 files:

 ../sherlock/features/glove.6B.50d.txt and
 ../sherlock/features/par_vec_trained_400.pkl.docvecs.vectors_docs.npy.

All files for extracting word and paragraph embeddings are present.

/home/ycrouin/Desktop/lab/schema-matching/tete/sherlock-
project/venv/lib/python3.8/site-packages/pandas/core/strings/object_array.py:90:
FutureWarning: Possible nested set at position 1
  regex = re.compile(pat, flags=flags)
```

[4]:
```
X_test_extracted[["n_values", "length-agg-mean", "frac_numcells",␣
↪"length-agg-kurtosis"]]
```

[4]:

	n_values	length-agg-mean	frac_numcells	length-agg-kurtosis
0	7	11.142857	0.000000	2.166667
1	19	2.000000	0.368421	-3.000000

```
2          9        13.714286       0.111111           -1.323362
3          5         8.400000       0.000000           -0.667223
4        249        30.000000       0.000000            0.244243
```

[5]:
```
X_test_processed[["n_values", "length-agg-mean", "frac_numcells",␣
↪"length-agg-kurtosis"]]
```

[5]:
```
   n_values  length-agg-mean  frac_numcells  length-agg-kurtosis
0         7           13.008          0.000            -0.909080
1      1000            2.058          1.000            12.302950
2         9           13.007          0.120            -0.488313
3      1000            9.245          0.000            -0.728836
4      1000           24.078          0.018             6.642598
```

[6]:
```
y_test_processed
```

[6]:
```
              label
index
511600  affiliation
146358       weight
665579       jockey
148486     religion
3546        company
```

[7]:
```
test_labels
```

[7]:
```
               type
20368   affiliation
664102       weight
366813       jockey
530567     religion
176253      company
```