MultiProcessing Implementation in Texthero-Compare With Mondin

September 1, 2020

We now compare our paralisation with the approche 4 from PR #162 (where we convert the input to modin, do calculation using the modin paralisation and convert the output back to pandas).

0.1 Result:

We can now clearly see below, that our implementation is twice as fast for big datasets. The issue is the overhead of converting the DF to and from modin.DF

[1]:
```
!pip uninstall textero --yes
!pip install git+https://github.com/SummerOfCode-NoHate/
 ↪texthero@decorator_for_parallelization
!pip install modin
import modin.pandas as mpd
import pandas as pd
import numpy as np
import texthero as hero
```

```
WARNING: Skipping textero as it is not installed.
Collecting git+https://github.com/SummerOfCode-
NoHate/texthero@decorator_for_parallelization
  Cloning https://github.com/SummerOfCode-NoHate/texthero (to revision
decorator_for_parallelization) to
/private/var/folders/ff/v8q71qfn4hbdkzbpmf28ymsr0000gn/T/pip-req-build-a06dt0wq
  Running command git clone -q https://github.com/SummerOfCode-NoHate/texthero
/private/var/folders/ff/v8q71qfn4hbdkzbpmf28ymsr0000gn/T/pip-req-build-a06dt0wq
  Running command git checkout -b decorator_for_parallelization --track
origin/decorator_for_parallelization
  Switched to a new branch 'decorator_for_parallelization'
  Branch 'decorator_for_parallelization' set up to track remote branch
'decorator_for_parallelization' from 'origin'.
Requirement already satisfied (use --upgrade to upgrade): texthero==1.0.9 from
git+https://github.com/SummerOfCode-
NoHate/texthero@decorator_for_parallelization in
/opt/anaconda3/lib/python3.7/site-packages
Requirement already satisfied: numpy>=1.17 in /opt/anaconda3/lib/python3.7/site-
packages (from texthero==1.0.9) (1.18.1)
Requirement already satisfied: scikit-learn>=0.22 in
```

```
/opt/anaconda3/lib/python3.7/site-packages (from texthero==1.0.9) (0.22.1)
Requirement already satisfied: spacy>=2.2.2 in
/opt/anaconda3/lib/python3.7/site-packages (from texthero==1.0.9) (2.3.2)
Requirement already satisfied: tqdm>=4.3 in /opt/anaconda3/lib/python3.7/site-
packages (from texthero==1.0.9) (4.42.1)
Requirement already satisfied: nltk>=3.3 in /opt/anaconda3/lib/python3.7/site-
packages (from texthero==1.0.9) (3.4.5)
Requirement already satisfied: plotly>=4.2.0 in
/opt/anaconda3/lib/python3.7/site-packages (from texthero==1.0.9) (4.9.0)
Requirement already satisfied: pandas>=1.0.2 in
/opt/anaconda3/lib/python3.7/site-packages (from texthero==1.0.9) (1.0.5)
Requirement already satisfied: wordcloud>=1.5.0 in
/opt/anaconda3/lib/python3.7/site-packages (from texthero==1.0.9) (1.7.0)
Requirement already satisfied: unidecode>=1.1.1 in
/opt/anaconda3/lib/python3.7/site-packages (from texthero==1.0.9) (1.1.1)
Requirement already satisfied: gensim>=3.6.0 in
/opt/anaconda3/lib/python3.7/site-packages (from texthero==1.0.9) (3.8.3)
Requirement already satisfied: matplotlib>=3.1.0 in
/opt/anaconda3/lib/python3.7/site-packages (from texthero==1.0.9) (3.1.3)
Requirement already satisfied: joblib>=0.11 in
/opt/anaconda3/lib/python3.7/site-packages (from scikit-
learn>=0.22->texthero==1.0.9) (0.14.1)
Requirement already satisfied: scipy>=0.17.0 in
/opt/anaconda3/lib/python3.7/site-packages (from scikit-
learn>=0.22->texthero==1.0.9) (1.4.1)
Requirement already satisfied: setuptools in /opt/anaconda3/lib/python3.7/site-
packages (from spacy>=2.2.2->texthero==1.0.9) (46.0.0.post20200309)
Requirement already satisfied: blis<0.5.0,>=0.4.0 in
/opt/anaconda3/lib/python3.7/site-packages (from spacy>=2.2.2->texthero==1.0.9)
(0.4.1)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in
/opt/anaconda3/lib/python3.7/site-packages (from spacy>=2.2.2->texthero==1.0.9)
(2.22.0)
Requirement already satisfied: srsly<1.1.0,>=1.0.2 in
/opt/anaconda3/lib/python3.7/site-packages (from spacy>=2.2.2->texthero==1.0.9)
(1.0.2)
Requirement already satisfied: catalogue<1.1.0,>=0.0.7 in
/opt/anaconda3/lib/python3.7/site-packages (from spacy>=2.2.2->texthero==1.0.9)
(1.0.0)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in
/opt/anaconda3/lib/python3.7/site-packages (from spacy>=2.2.2->texthero==1.0.9)
(3.0.2)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in
/opt/anaconda3/lib/python3.7/site-packages (from spacy>=2.2.2->texthero==1.0.9)
(2.0.3)
Requirement already satisfied: thinc==7.4.1 in
/opt/anaconda3/lib/python3.7/site-packages (from spacy>=2.2.2->texthero==1.0.9)
(7.4.1)
```

```
Requirement already satisfied: plac<1.2.0,>=0.9.6 in
/opt/anaconda3/lib/python3.7/site-packages (from spacy>=2.2.2->texthero==1.0.9)
(1.1.3)
Requirement already satisfied: wasabi<1.1.0,>=0.4.0 in
/opt/anaconda3/lib/python3.7/site-packages (from spacy>=2.2.2->texthero==1.0.9)
(0.7.1)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in
/opt/anaconda3/lib/python3.7/site-packages (from spacy>=2.2.2->texthero==1.0.9)
(1.0.2)
Requirement already satisfied: six in /opt/anaconda3/lib/python3.7/site-packages
(from nltk>=3.3->texthero==1.0.9) (1.14.0)
Requirement already satisfied: retrying>=1.3.3 in
/opt/anaconda3/lib/python3.7/site-packages (from plotly>=4.2.0->texthero==1.0.9)
(1.3.3)
Requirement already satisfied: python-dateutil>=2.6.1 in
/opt/anaconda3/lib/python3.7/site-packages (from pandas>=1.0.2->texthero==1.0.9)
(2.8.1)
Requirement already satisfied: pytz>=2017.2 in
/opt/anaconda3/lib/python3.7/site-packages (from pandas>=1.0.2->texthero==1.0.9)
(2019.3)
Requirement already satisfied: pillow in /opt/anaconda3/lib/python3.7/site-
packages (from wordcloud>=1.5.0->texthero==1.0.9) (7.0.0)
Requirement already satisfied: smart-open>=1.8.1 in
/opt/anaconda3/lib/python3.7/site-packages (from gensim>=3.6.0->texthero==1.0.9)
(2.1.0)
Requirement already satisfied: kiwisolver>=1.0.1 in
/opt/anaconda3/lib/python3.7/site-packages (from
matplotlib>=3.1.0->texthero==1.0.9) (1.1.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in
/opt/anaconda3/lib/python3.7/site-packages (from
matplotlib>=3.1.0->texthero==1.0.9) (2.4.6)
Requirement already satisfied: cycler>=0.10 in
/opt/anaconda3/lib/python3.7/site-packages (from
matplotlib>=3.1.0->texthero==1.0.9) (0.10.0)
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in
/opt/anaconda3/lib/python3.7/site-packages (from
requests<3.0.0,>=2.13.0->spacy>=2.2.2->texthero==1.0.9) (3.0.4)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in
/opt/anaconda3/lib/python3.7/site-packages (from
requests<3.0.0,>=2.13.0->spacy>=2.2.2->texthero==1.0.9) (1.25.8)
Requirement already satisfied: idna<2.9,>=2.5 in
/opt/anaconda3/lib/python3.7/site-packages (from
requests<3.0.0,>=2.13.0->spacy>=2.2.2->texthero==1.0.9) (2.8)
Requirement already satisfied: certifi>=2017.4.17 in
/opt/anaconda3/lib/python3.7/site-packages (from
requests<3.0.0,>=2.13.0->spacy>=2.2.2->texthero==1.0.9) (2019.11.28)
Requirement already satisfied: importlib-metadata>=0.20; python_version < "3.8"
in /opt/anaconda3/lib/python3.7/site-packages (from
catalogue<1.1.0,>=0.0.7->spacy>=2.2.2->texthero==1.0.9) (1.5.0)
Requirement already satisfied: boto in /opt/anaconda3/lib/python3.7/site-
packages (from smart-open>=1.8.1->gensim>=3.6.0->texthero==1.0.9) (2.49.0)
Requirement already satisfied: boto3 in /opt/anaconda3/lib/python3.7/site-
packages (from smart-open>=1.8.1->gensim>=3.6.0->texthero==1.0.9) (1.14.23)
Requirement already satisfied: zipp>=0.5 in /opt/anaconda3/lib/python3.7/site-
packages (from importlib-metadata>=0.20; python_version <
"3.8"->catalogue<1.1.0,>=0.0.7->spacy>=2.2.2->texthero==1.0.9) (2.2.0)
Requirement already satisfied: s3transfer<0.4.0,>=0.3.0 in
/opt/anaconda3/lib/python3.7/site-packages (from boto3->smart-
open>=1.8.1->gensim>=3.6.0->texthero==1.0.9) (0.3.3)
Requirement already satisfied: botocore<1.18.0,>=1.17.23 in
/opt/anaconda3/lib/python3.7/site-packages (from boto3->smart-
open>=1.8.1->gensim>=3.6.0->texthero==1.0.9) (1.17.23)
Requirement already satisfied: jmespath<1.0.0,>=0.7.1 in
/opt/anaconda3/lib/python3.7/site-packages (from boto3->smart-
open>=1.8.1->gensim>=3.6.0->texthero==1.0.9) (0.10.0)
Requirement already satisfied: docutils<0.16,>=0.10 in
/opt/anaconda3/lib/python3.7/site-packages (from
botocore<1.18.0,>=1.17.23->boto3->smart-
open>=1.8.1->gensim>=3.6.0->texthero==1.0.9) (0.15.2)
Building wheels for collected packages: texthero
  Building wheel for texthero (setup.py) … done
  Created wheel for texthero: filename=texthero-1.0.9-py3-none-any.whl
size=43898
sha256=405353ec98c0a791403a36711bbb5ddc4d87e869ca1758e3f7634af632b3b7bc
  Stored in directory:
/private/var/folders/ff/v8q71qfn4hbdkzbpmf28ymsr0000gn/T/pip-ephem-wheel-cache-8
bn_s4k8/wheels/e7/d1/60/88628de1662df5ddf78097e355a7bea59be0a1e213f5f636e2
Successfully built texthero
Requirement already satisfied: modin in /opt/anaconda3/lib/python3.7/site-
packages (0.8.0)
Requirement already satisfied: pandas==1.0.5 in
/opt/anaconda3/lib/python3.7/site-packages (from modin) (1.0.5)
Requirement already satisfied: packaging in /opt/anaconda3/lib/python3.7/site-
packages (from modin) (20.1)
Requirement already satisfied: numpy>=1.13.3 in
/opt/anaconda3/lib/python3.7/site-packages (from pandas==1.0.5->modin) (1.18.1)
Requirement already satisfied: pytz>=2017.2 in
/opt/anaconda3/lib/python3.7/site-packages (from pandas==1.0.5->modin) (2019.3)
Requirement already satisfied: python-dateutil>=2.6.1 in
/opt/anaconda3/lib/python3.7/site-packages (from pandas==1.0.5->modin) (2.8.1)
Requirement already satisfied: six in /opt/anaconda3/lib/python3.7/site-packages
(from packaging->modin) (1.14.0)
Requirement already satisfied: pyparsing>=2.0.2 in
/opt/anaconda3/lib/python3.7/site-packages (from packaging->modin) (2.4.6)
```

```
UserWarning: The Dask Engine for Modin is experimental.
```

[10]:

```
# this function prepares as modin series from a given pandas series;
# performs a texthero remove diacritics and converts the result back

def modin_remove_diacritics(s: pd.Series) -> pd.Series:
    hero.config.PARALLELIZE = False
    s_modin = mpd.Series(s)

    s_result_modin = hero.remove_diacritics(s_modin)

    s_result = s_result_modin._to_pandas()

    return s_result
```

[3]:

```
# init dataset for speed comparison

data_small = pd.read_csv("https://raw.githubusercontent.com/jbesomi/texthero/
 ↪master/dataset/bbcsport.csv")
data_big = pd.DataFrame([text for _ in range(100) for text in␣
 ↪data_small["text"].values], columns=["text"])
print("Big dataset has {} texts".format(len(data_big)))
```

```
Big dataset has 73700 texts
```

[4]:

```
data_huge = pd.DataFrame([text for _ in range(1000) for text in␣
 ↪data_small["text"].values], columns=["text"])
print("Huge dataset has {} texts".format(len(data_huge)))
```

```
Huge dataset has 737000 texts
```

[5]:

```
# turn on multithreading in texthero and take time of multi threads
hero.config.PARALLELIZE = True

%time hero.remove_diacritics(data_big["text"])
```

```
CPU times: user 224 ms, sys: 329 ms, total: 553 ms
Wall time: 7.19 s
```

[5]:

```
0        Claxton hunting first major medal\n\nBritish h…
1        O'Sullivan could run in Worlds\n\nSonia O'Sull…
2        Greene sets sights on world title\n\nMaurice G…
3        IAAF launches fight against drugs\n\nThe IAAF …
4        Dibaba breaks 5,000m world record\n\nEthiopia'…
                                  …
73695    Agassi into second round in Dubai\n\nFourth se…
73696    Mauresmo fights back to win title\n\nWorld num…
73697    Federer wins title in Rotterdam\n\nWorld numbe…
73698    GB players warned over security\n\nBritain's D…
```

```
73699    Sharapova overcomes tough Molik\n\nWimbledon c…
Name: text, Length: 73700, dtype: object
```

[11]:
```
%time modin_remove_diacritics(data_big["text"])
```

```
Finished fst conversion!
Finished calculation
Finished converting back to pandas
CPU times: user 1.54 s, sys: 375 ms, total: 1.92 s
Wall time: 20.9 s
```

[11]:
```
0        Claxton hunting first major medal\n\nBritish h…
1        O'Sullivan could run in Worlds\n\nSonia O'Sull…
2        Greene sets sights on world title\n\nMaurice G…
3        IAAF launches fight against drugs\n\nThe IAAF …
4        Dibaba breaks 5,000m world record\n\nEthiopia'…
                               …
73695    Agassi into second round in Dubai\n\nFourth se…
73696    Mauresmo fights back to win title\n\nWorld num…
73697    Federer wins title in Rotterdam\n\nWorld numbe…
73698    GB players warned over security\n\nBritain's D…
73699    Sharapova overcomes tough Molik\n\nWimbledon c…
Name: text, Length: 73700, dtype: object
```

[14]:
```
hero.config.PARALLELIZE = True

%time hero.remove_diacritics(data_huge["text"])
```

```
CPU times: user 5.71 s, sys: 3.59 s, total: 9.3 s
Wall time: 1min 14s
```

[14]:
```
0         Claxton hunting first major medal\n\nBritish h…
1         O'Sullivan could run in Worlds\n\nSonia O'Sull…
2         Greene sets sights on world title\n\nMaurice G…
3         IAAF launches fight against drugs\n\nThe IAAF …
4         Dibaba breaks 5,000m world record\n\nEthiopia'…
                               …
736995    Agassi into second round in Dubai\n\nFourth se…
736996    Mauresmo fights back to win title\n\nWorld num…
736997    Federer wins title in Rotterdam\n\nWorld numbe…
736998    GB players warned over security\n\nBritain's D…
736999    Sharapova overcomes tough Molik\n\nWimbledon c…
Name: text, Length: 737000, dtype: object
```

[15]:
```
hero.config.PARALLELIZE = False

%time modin_remove_diacritics(data_huge["text"])
```

```
CPU times: user 11 s, sys: 3.97 s, total: 15 s
Wall time: 2min 22s
```

```
[15]: 0         Claxton hunting first major medal\n\nBritish h…
      1         O'Sullivan could run in Worlds\n\nSonia O'Sull…
      2         Greene sets sights on world title\n\nMaurice G…
      3         IAAF launches fight against drugs\n\nThe IAAF …
      4         Dibaba breaks 5,000m world record\n\nEthiopia'…
                                           …
      736995    Agassi into second round in Dubai\n\nFourth se…
      736996    Mauresmo fights back to win title\n\nWorld num…
      736997    Federer wins title in Rotterdam\n\nWorld numbe…
      736998    GB players warned over security\n\nBritain's D…
      736999    Sharapova overcomes tough Molik\n\nWimbledon c…
      Name: text, Length: 737000, dtype: object
```

```
[15]: %timeit x = mpd.Series(data_huge["text"])
      %timeit x._to_pandas()
```

```
129 ms ± 2.26 ms per loop (mean ± std. dev. of 7 runs, 10 loops each)
126 ms ± 2.33 ms per loop (mean ± std. dev. of 7 runs, 10 loops each)
```