ComparisonTokenisation

August 25, 2020

[1]:
```
!pip uninstall texthero --yes
!pip install git+https://github.com/SummerOfCode-NoHate/
 ↪texthero@decorator_for_parallelization
```

```
Found existing installation: texthero 1.0.9
Uninstalling texthero-1.0.9:
  Successfully uninstalled texthero-1.0.9
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-r0tizlnr
  Running command git clone -q https://github.com/SummerOfCode-NoHate/texthero
/private/var/folders/ff/v8q71qfn4hbdkzbpmf28ymsr0000gn/T/pip-req-build-r0tizlnr
  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: 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) (1.1.1)
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
```

```
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=315f8b739e51a1241334a87367e4ec823fb3ee44e32d915d717059f9c394a851
  Stored in directory:
/private/var/folders/ff/v8q71qfn4hbdkzbpmf28ymsr0000gn/T/pip-ephem-wheel-cache-8
9ll7kr4/wheels/e7/d1/60/88628de1662df5ddf78097e355a7bea59be0a1e213f5f636e2
Successfully built texthero
Installing collected packages: texthero
Successfully installed texthero-1.0.9
```

[2]:

```python
import texthero as hero
import pandas as pd
```

0.1 Set up functions

[6]:

```python
import spacy

nlp = spacy.load("en_core_web_sm", disable=["ner", "tagger", "parser"])


def tokenize_with_spacy(s: pd.Series) -> pd.Series:


    tokenized = []
    for doc in nlp.pipe(s, n_process=8):
        tokenized.append(list(map(str, doc)))

    return pd.Series(tokenized, index=s.index)
```

[13]:

```python
def _tokenize_with_spacy_own_parallelization(s: pd.Series) -> pd.Series:

    tokenized = []
    for doc in nlp.pipe(s):
        tokenized.append(list(map(str, doc)))

    return pd.Series(tokenized, index=s.index)

def tokenize_with_spacy_own_parallelization(s: pd.Series) -> pd.Series:
    return hero.parallel(s, _tokenize_with_spacy_own_parallelization)
```

0.2 Load Data

```
[7]: data_small = pd.read_csv("https://raw.githubusercontent.com/jbesomi/texthero/
      ↪master/dataset/bbcsport.csv")
     data_big = pd.DataFrame([text for _ in range(200) for text in␣
      ↪data_small["text"].values], columns=["text"])
     print("Big dataset has {} texts".format(len(data_big)))
```

```
Big dataset has 147400 texts
```

1 Speed Comparison

We now compare: 1. current implementation without parallelization 2. current implementation with parallelization (see #162) 3. tokenize_with_spacy with spacy built-in parallelization through n_process 4. tokenize_with_spacy with our custom parallelization

Results below.

We can see that

- our current implementation is much faster than spaCy (22 vs 51 seconds with both parallelized)
- as shown in #162, our parallelization works better than spaCy's.

Thus, our options:

1. keep everything as proposed in #162 (-> multiprocessing applied to current solution)
2. option 1, but we give users a parameter `use_spacy` that works like our `tokenize_with_spacy_own_parallelization` above, and explain to them that this might give them better results but takes about 3x as long.

We don't really have a preference.

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

     %timeit -r 1 -n 1 x = hero.tokenize(data_big["text"])
```

```
34.1 s ± 0 ns per loop (mean ± std. dev. of 1 run, 1 loop each)
```

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

      %timeit -r 1 -n 1 x = hero.tokenize(data_big["text"])
```

```
21.4 s ± 0 ns per loop (mean ± std. dev. of 1 run, 1 loop each)
```

```
[12]: %timeit -r 1 -n 1 x = tokenize_with_spacy(data_big["text"])
```

```
7min 5s ± 0 ns per loop (mean ± std. dev. of 1 run, 1 loop each)
```

```
[18]: %timeit -r 1 -n 1 x = tokenize_with_spacy_own_parallelization(data_big["text"])
```

```
55.6 s ± 0 ns per loop (mean ± std. dev. of 1 run, 1 loop each)
```

```
[ ]:
```