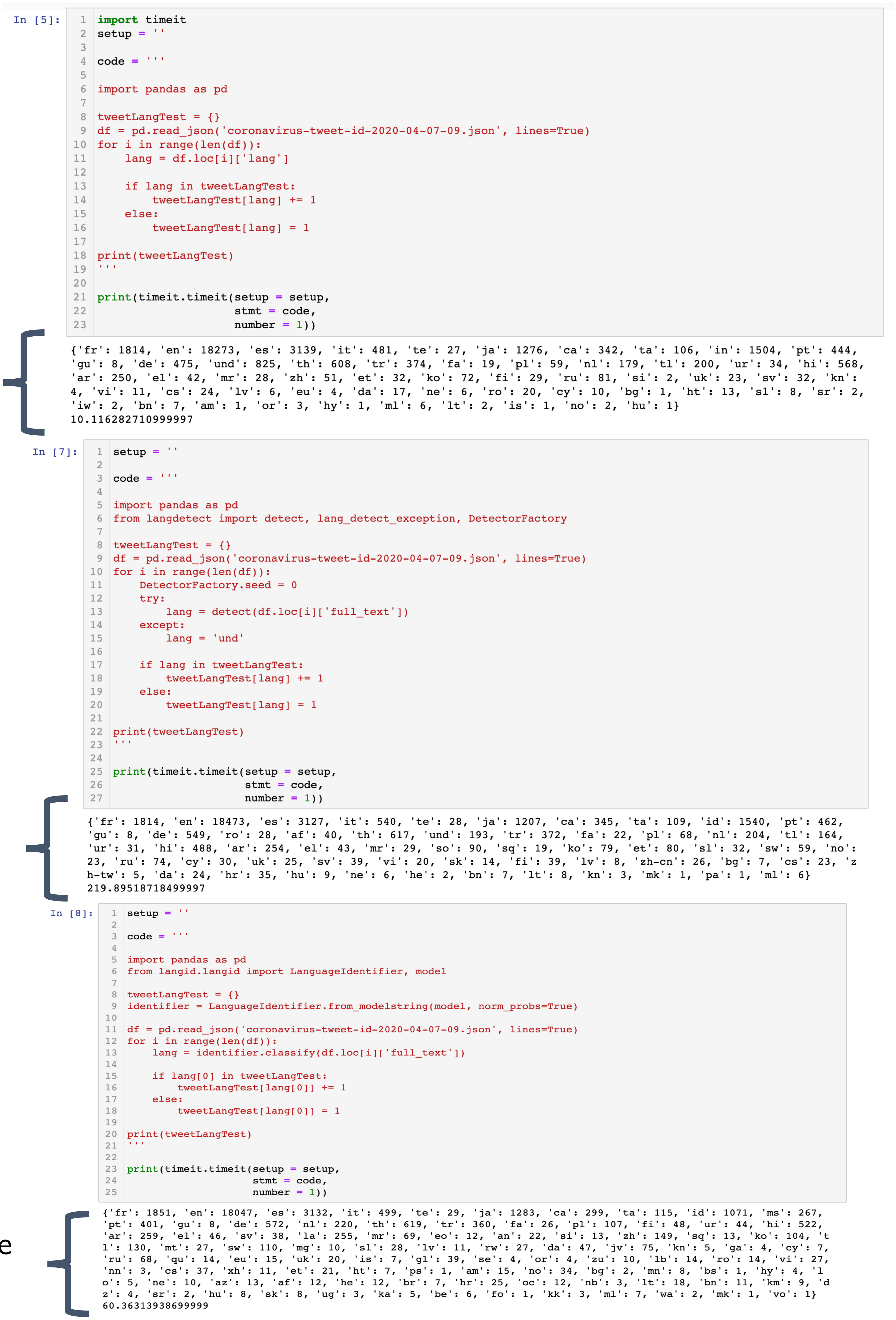

In [5]:

```
import timeit
setup = ''

code = '''

import pandas as pd

tweetLangTest = {}
df = pd.read_json('coronavirus-tweet-id-2020-04-07-09.json', lines=True)
for i in range(len(df)):
    lang = df.loc[i]['lang']

    if lang in tweetLangTest:
        tweetLangTest[lang] += 1
    else:
        tweetLangTest[lang] = 1

print(tweetLangTest)
'''

print(timeit.timeit(setup = setup,
                    stmt = code,
                    number = 1))
```

Twitter
Language
10.12 seconds

```
{'fr': 1814, 'en': 18273, 'es': 3139, 'it': 481, 'te': 27, 'ja': 1276, 'ca': 342, 'ta': 106, 'in': 1504, 'pt': 444,
'gu': 8, 'de': 475, 'und': 825, 'th': 608, 'tr': 374, 'fa': 19, 'pl': 59, 'nl': 179, 'tl': 200, 'ur': 34, 'hi': 568,
'ar': 250, 'el': 42, 'mr': 28, 'zh': 51, 'et': 32, 'ko': 72, 'fi': 29, 'ru': 81, 'si': 2, 'uk': 23, 'sv': 32, 'kn':
4, 'vi': 11, 'cs': 24, 'lv': 6, 'eu': 4, 'da': 17, 'ne': 6, 'ro': 20, 'cy': 10, 'bg': 1, 'ht': 13, 'sl': 8, 'sr': 2,
'iw': 2, 'bn': 7, 'am': 1, 'or': 3, 'hy': 1, 'ml': 6, 'lt': 2, 'is': 1, 'no': 2, 'hu': 1}
10.116282710999997
```

In [7]:

```
setup = ''

code = '''

import pandas as pd
from langdetect import detect, lang_detect_exception, DetectorFactory

tweetLangTest = {}
df = pd.read_json('coronavirus-tweet-id-2020-04-07-09.json', lines=True)
for i in range(len(df)):
    DetectorFactory.seed = 0
    try:
        lang = detect(df.loc[i]['full_text'])
    except:
        lang = 'und'

    if lang in tweetLangTest:
        tweetLangTest[lang] += 1
    else:
        tweetLangTest[lang] = 1

print(tweetLangTest)
'''

print(timeit.timeit(setup = setup,
                    stmt = code,
                    number = 1))
```

langdetect
Language
219.90 seconds

```
{'fr': 1814, 'en': 18473, 'es': 3127, 'it': 540, 'te': 28, 'ja': 1207, 'ca': 345, 'ta': 109, 'id': 1540, 'pt': 462,
'gu': 8, 'de': 549, 'ro': 28, 'af': 40, 'th': 617, 'und': 193, 'tr': 372, 'fa': 22, 'pl': 68, 'nl': 204, 'tl': 164,
'ur': 31, 'hi': 488, 'ar': 254, 'el': 43, 'mr': 29, 'so': 90, 'sq': 19, 'ko': 79, 'et': 80, 'sl': 32, 'sw': 59, 'no':
23, 'ru': 74, 'cy': 30, 'uk': 25, 'sv': 39, 'vi': 20, 'sk': 14, 'fi': 39, 'lv': 8, 'zh-cn': 26, 'bg': 7, 'cs': 23, 'z
h-tw': 5, 'da': 24, 'hr': 35, 'hu': 9, 'ne': 6, 'he': 2, 'bn': 7, 'lt': 8, 'kn': 3, 'mk': 1, 'pa': 1, 'ml': 6}
219.89518718499997
```

In [8]:

```
setup = ''

code = '''

import pandas as pd
from langid.langid import LanguageIdentifier, model

tweetLangTest = {}
identifier = LanguageIdentifier.from_modelstring(model, norm_probs=True)

df = pd.read_json('coronavirus-tweet-id-2020-04-07-09.json', lines=True)
for i in range(len(df)):
    lang = identifier.classify(df.loc[i]['full_text'])

    if lang[0] in tweetLangTest:
        tweetLangTest[lang[0]] += 1
    else:
        tweetLangTest[lang[0]] = 1

print(tweetLangTest)
'''

print(timeit.timeit(setup = setup,
                    stmt = code,
                    number = 1))
```

langid Language
60.36 seconds

```
{'fr': 1851, 'en': 18047, 'es': 3132, 'it': 499, 'te': 29, 'ja': 1283, 'ca': 299, 'ta': 115, 'id': 1071, 'ms': 267,
'pt': 401, 'gu': 8, 'de': 572, 'nl': 220, 'th': 619, 'tr': 360, 'fa': 26, 'pl': 107, 'fi': 48, 'ur': 44, 'hi': 522,
'ar': 259, 'el': 46, 'sv': 38, 'la': 255, 'mr': 69, 'eo': 12, 'an': 22, 'si': 13, 'zh': 149, 'sq': 13, 'ko': 104, 't
l': 130, 'mt': 27, 'sw': 110, 'mg': 10, 'sl': 28, 'lv': 11, 'rw': 27, 'da': 47, 'jv': 75, 'kn': 5, 'ga': 4, 'cy': 7,
'ru': 68, 'qu': 14, 'eu': 15, 'uk': 20, 'is': 7, 'gl': 39, 'se': 4, 'or': 4, 'zu': 10, 'lb': 14, 'ro': 14, 'vi': 27,
'nn': 3, 'cs': 37, 'xh': 11, 'et': 21, 'ht': 7, 'ps': 1, 'am': 15, 'no': 34, 'bg': 2, 'mn': 8, 'bs': 1, 'hy': 4, 'l
o': 5, 'ne': 10, 'az': 13, 'af': 12, 'he': 12, 'br': 7, 'hr': 25, 'oc': 12, 'nb': 3, 'lt': 18, 'bn': 11, 'km': 9, 'd
z': 4, 'sr': 2, 'hu': 8, 'sk': 8, 'ug': 3, 'ka': 5, 'be': 6, 'fo': 1, 'kk': 3, 'ml': 7, 'wa': 2, 'mk': 1, 'vo': 1}
60.36313938699999
```