Narzędzie do rozpoznawania języka oparte na bibliotece fastText

Autorzy:

Tomasz Macałka

Michał Pieczara

FastText

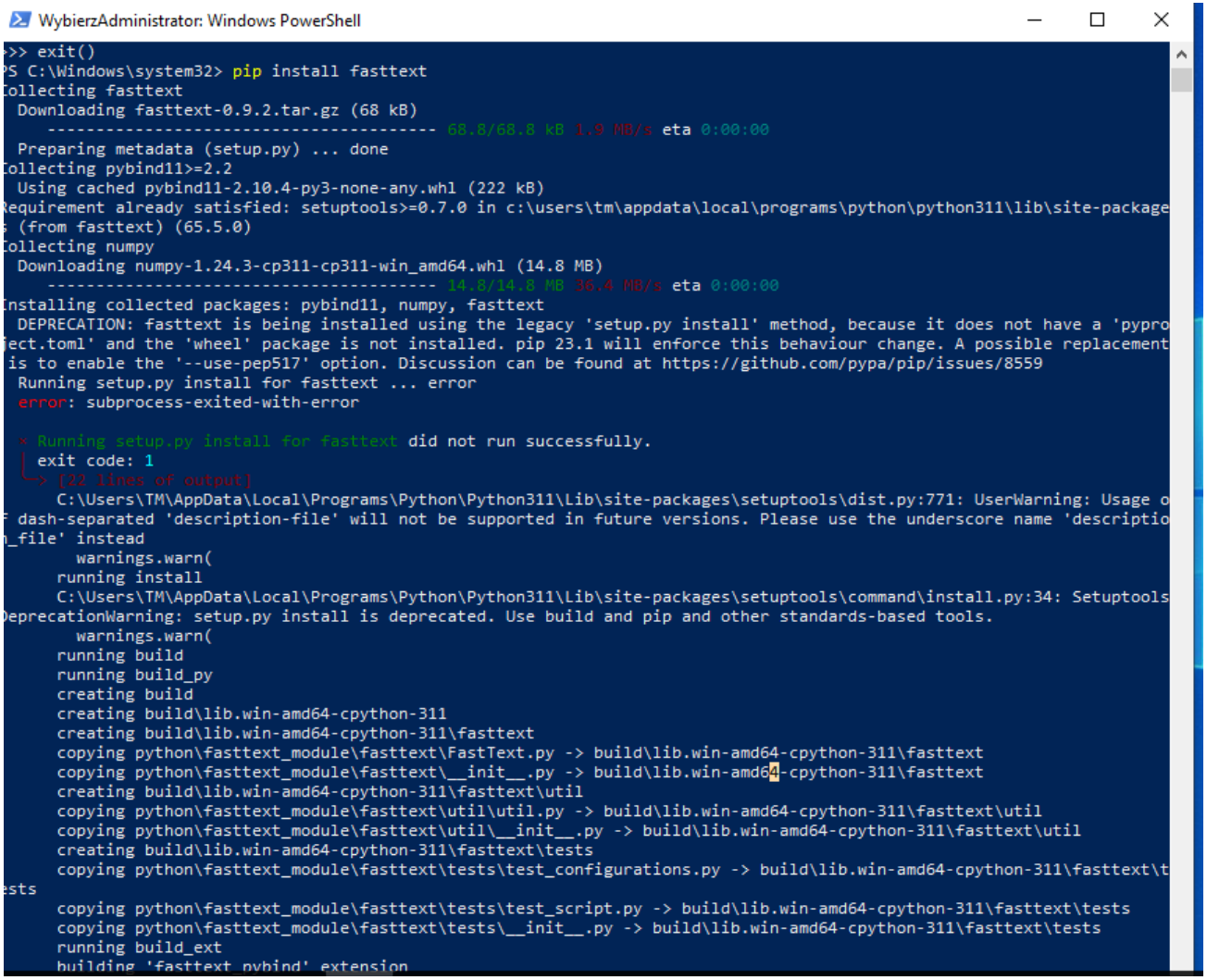
- https://fasttext.cc/
- biblioteka Facebook AI Research
- umożliwia naukę osadzania słów i klasyfikację tekstu
- może być używana do generowania wektorowych reprezentacji słów
- przewidywanie etykiet na podstawie wcześniej nauczonych wzorców
- wydajna i skalowalna
- znajduje zastosowanie w wielu dziedzinach przetwarzania języka naturalnego.

Rozpoznawanie języka

ma wiele praktycznych zastosowań w różnych dziedzinach, takich jak:

- tłumaczenie
- analiza sentymentu
- personalizacja treści
- pozwala lepiej zrozumieć i przetwarzać różnorodne treści dostępne w różnych językach

Początki


WybierzAdministrator: Windows PowerShell
>>> exit()
PS C:\Windows\system32> pip install fasttext
Collecting fasttext
Downloading fasttext-0.9.2.tar.gz (68 kB)
68.8/68.8 kB 1.9 MB/s eta 0:00:00
Preparing metadata (setup.py) ... done
Collecting pybind11>=2.2
Using cached pybind11-2.10.4-py3-none-any.whl (222 kB)
Requirement already satisfied: setuptools>=0.7.0 in c:\users\tm\appdata\local\programs\python\python311\lib\site-package
(from fasttext) (65.5.0)
Collecting numpy
Downloading numpy-1.24.3-cp311-cp311-win_amd64.whl (14.8 MB)
14.8/14.8 MB 36.4 MB/s eta 0:00:00
Installing collected packages: pybind11, numpy, fasttext
DEPRECATION: fasttext is being installed using the legacy 'setup.py install' method, because it does not have a 'pypro
ject.toml' and the 'wheel' package is not installed. pip 23.1 will enforce this behaviour change. A possible replacement
is to enable the '--use-pep517' option. Discussion can be found at https://github.com/pypa/pip/issues/8559
Running setup.py install for fasttext ... error
error: subprocess-exited-with-error
× Running setup.py install for fasttext did not run successfully.
exit code: 1
[22 lines of output]
C:\Users\TM\AppData\Local\Programs\Python\Python311\Lib\site-packages\setuptools\dist.py:771: UserWarning: Usage o
f dash-separated 'description-file' will not be supported in future versions. Please use the underscore name 'descriptio
n_file' instead
warnings.warn(
running install
C:\Users\TM\AppData\Local\Programs\Python\Python311\Lib\site-packages\setuptools\command\install.py:34: Setuptools
DeprecationWarning: setup.py install is deprecated. Use build and pip and other standards-based tools.
warnings.warn(
running build
running build_py
creating build
creating build\lib.win-amd64-cpython-311
creating build\lib.win-amd64-cpython-311\fasttext
copying python\fasttext_module\fasttext\FastText.py -> build\lib.win-amd64-cpython-311\fasttext
copying python\fasttext_module\fasttext__init__.py -> build\lib.win-amd64-cpython-311\fasttext
creating build\lib.win-amd64-cpython-311\fasttext\util
copying python\fasttext_module\fasttext\util\util.py -> build\lib.win-amd64-cpython-311\fasttext\util
copying python\fasttext_module\fasttext\util__init__.py -> build\lib.win-amd64-cpython-311\fasttext\util
creating build\lib.win-amd64-cpython-311\fasttext\tests
copying python\fasttext_module\fasttext\tests\test_configurations.py -> build\lib.win-amd64-cpython-311\fasttext\t
ests
copying python\fasttext_module\fasttext\tests\test_script.py -> build\lib.win-amd64-cpython-311\fasttext\tests
copying python\fasttext_module\fasttext\tests__init__.py -> build\lib.win-amd64-cpython-311\fasttext\tests
running build_ext
building 'fasttext_pybind' extension

Aplikacja

- dostępna pod **http://macalka.duckdns.org:5000**
- repozytorium pod https://github.com/teh42eem00/PJN_Language_Recognition
- cel- stworzenie aplikacji, która automatycznie identyfikuje język tekstu
- problemy z instalacją pod Windows - wybór Linuxa i konteneryzacji
- aplikacja webowa napisana w języku Python z wykorzystaniem biblioteki Flask
- umożliwia użytkownikom przesyłanie tekstu i określenie języka w jakim jest napisany
- korzysta z dwóch modeli do detekcji języka - modelu wytrenowanego na własnym zbiorze danych oraz pretrenowanego modelu FastText zawierającym 176 języków

Technologie

- Python 3.11
- Flask 2.3.2
- FastText 0.9.2
- Docker
- Docker 3.11-slim-bullseye Debian image
- HTML
- CSS - Materialize Framework

Uruchomienie ogranicza się do czterech komend:

- git clone https://github.com/teh42eem00/PJN_Language_Recognition
- cd PJN_Language_Recognition
- docker build -t pjn_language_recognition .
- docker run -p 5000:5000 --name pjn_language_recognition -d pjn_language_recognition

Techniki i proces

- uczenie maszynowe do rozpoznawania języka tekstu
- przygotowanie danych treningowych
- trenowanie modelu
- testowanie modelu
- detekcja języka

Przygotowanie danych treningowych

- dostosowanie danych treningowych do formatu oczekiwanego przez fasttext
- __label__etykieta Zdanie do nauczenia które zostanie przypisane do danej etykiety

```
1276    eng Let's try something.
1277    eng I have to go to sleep.
1280    eng Today is June 18th and it is Muiriel's birthday!
1282    eng Muiriel is 20 now.
1283    eng The password is "Muiriel".
1284    eng I will be back soon.
1286    eng I'm at a loss for words.
```

```
for line in lines:
    line = line.strip()
    line = re.sub(r'^\d+\t\w+\t', f'__label__{language_code}\t', line)
```

```
__label__en Let's try something.
__label__en I have to go to sleep.
__label__en Today is June 18th and it is Muiriel's birthday!
__label__en Muiriel is 20 now.
__label__en The password is "Muiriel".
__label__en I will be back soon.
__label__en I'm at a loss for words.
```

Trening danych

- zbiór danych z portalu Tatoeba (zbiory zdań i tłumaczeń) w trzech językach: angielskim, polskim i niemieckim
- do trenowania modelu została wykorzystana funkcja train_supervised z biblioteki FastText
- model tworzy reprezentacje wektorowe słów, które pojawiają się w tekście
- reprezentacje te są przetwarzane przez sieć neuronową
- model stara się nauczyć, jakie cechy są charakterystyczne dla każdego języka, takie jak na przykład kolokacje słów czy częstość występowania pewnych wyrazów

```
def train_own_model():
    return fasttext.train_supervised(input='static/combined.txt', label_prefix="__label__", epoch=25, lr=0.1,
                                     wordNgrams=1, bucket=2000000, dim=300, thread=4)
```

```
Read 1M words
Number of words:  114243
Number of labels: 3
Progress:  24.2% words/sec/thread:  204726 lr:  0.075808 avg.loss:  0.007728 ETA:   0h 0m30s
```

Parametry train_supervised

- input: ścieżka do pliku z danymi treningowymi,
- label_prefix: prefiks używany przed etykietami klas w pliku z danymi,
- epoch: liczba epok treningowych - wartość 25 - zbyt mała wartość - doprowadza do niedouczenia modelu, natomiast zbyt duża wartość może prowadzić do przeuczenia modelu, czyli sytuacji, w której model nauczy się zbytnio dostosowywać do danych treningowych i będzie słabo generalizował na nowych danych, co może prowadzić do niedokładnych predykcji na danych testowych lub w praktyce,
- lr: learning rate - wartość 0.1 - zbyt mała wartość może spowodować, że model będzie uczył się zbyt wolno, co wydłuży czas treningu, natomiast zbyt duża wartość może spowodować, że model będzie miał problem z generalizacją i osiągnie gorsze wyniki na danych testowych,
- wordNgrams: długość n-gramów używanych do reprezentacji słów - wartość 1 - zbyt mała wartość może prowadzić do obniżenia jakości detekcji języka, natomiast zbyt duża wartość może prowadzić do zwiększenia wymiarowości wektorów słów, co zwiększa ilość parametrów do nauki i może prowadzić do przetrenowania modelu,

- bucket: liczba kubełków używanych do haszowania - wartość 2000000 - zbyt mała wartość może prowadzić do kolizji haszy, co oznacza, że dwa różne słowa zostaną przypisane do tego samego kubełka i będą traktowane jako jedno słowo, to może prowadzić do pogorszenia jakości modelu i wynikającej z niego analizy. Zbyt duża wartość może prowadzić do niepotrzebnego zwiększenia czasu trenowania i pamięci potrzebnej do przechowywania modelu,
- dim: liczba wymiarów reprezentacji wektorowej - wartość 300 - zbyt mała wartość może spowodować, że model nie będzie w stanie zachować odpowiedniej ilości informacji o słowach, co wpłynie na jakość klasyfikacji. Zbyt duża wartość może prowadzić do overfittingu, tj. model będzie zbyt dobrze dopasowany do zbioru treningowego, a nie będzie w stanie dobrze generalizować na nowych danych,
- thread: liczba wątków używanych do treningu - wartość 4.

Testowanie modelu

- model klasyfikuje teksty na podstawie występowania słów w tekście (reprezentacja n-gramów)
- szacuje etykietę języka na podstawie prawdopodobieństwa, że dany tekst należy do danego języka
- przetestowany na zbiorze danych testowych zawierających ostatnie 20 tysięcy zdań pochodzących z każdego z trzech zbiorów języków: angielskiego, polskiego i niemieckiego
- testy zostały przeprowadzone w celu oceny precyzji i skuteczności klasyfikacji

```
own_model_test_results = own_model.test('static/test_sentences.txt')
pretrained_model_test_results = model_176.test('static/test_sentences.txt')
```

Model Info:

No. of validated records: 59998

Precision: 0.9848828294276476

Recall: 0.9848828294276476

Detekcja języka

- użytkownik wprowadza tekst, który chce rozpoznać
- model dokonuje klasyfikacji tekstu na podstawie reprezentacji wektorowej, której nauczył się podczas treningu
- ocenia prawdopodobieństwo dla każdego języka i zwraca wyniki rozpoznania do języka, którego reprezentacja wektorowa daje najwyższy wynik prawdopodobieństwa klasyfikacji
- zwracana lista dwóch języków wraz z prawdopodobieństwem dla każdego języka, co pozwala nam dowiedzieć się, jaki język został rozpoznany na podstawie analizy tekstu

```python
def detect_language(text, own_model, pretrained_model):
    own_predictions = own_model.predict(text, k=2)
    own_results = [
        {'language': label.replace('__label__', ''), 'probability': probability}
        for label, probability in zip(own_predictions[0], own_predictions[1])
    ]

    pretrained_predictions = pretrained_model.predict(text, k=2)
    pretrained_results = [
        {'language': label.replace('__label__', ''), 'probability': probability}
        for label, probability in zip(pretrained_predictions[0], pretrained_predictions[1])
    ]

    return {'own_model': own_results, 'pretrained_model': pretrained_results}
```

Fasttext Language Recognition

Enter text

RECOGNIZE LANGUAGE

Input Text:

trochę dobry und deutschland

Own Model

Language: de
Probability: 0.6429899931

Language: pl
Probability: 0.3570289612

Model Info:

No. of validated records: 59998

Precision: 0.9843161438714624

Recall: 0.9843161438714624

Fasttext Model

Language: de
Probability: 0.8032852411

Language: pl
Probability: 0.1479703933

Model Info:

No. of validated records: 59998

Precision: 0.9846161538717957

Recall: 0.9846161538717957