Narzędzie do rozpoznawania języka oparte na bibliotece fastText

Natalia Walczak, Kacper Pawlak, Krystian Pasiński

Abstrakt

W niniejszej prezentacji przedstawiono program do automatycznego rozpoznawania języka z wykorzystaniem biblioteki fasttext. Algorytm oparty jest o model uczenia maszynowego, który został wytrenowany na dużym zbiorze danych wejściowych, składających się z różnych języków. Po wytrenowaniu modelu, program jest w stanie skutecznie rozpoznawać język tekstu wejściowego z dokładnością sięgającą nawet 100%.

Wstęp

W celu utworzenia narzędzia do rozpoznawania języka w projekcie wykorzystano bibliotekę fasttext stworzoną przez korporację Meta. Fasttext jest biblioteką typu open-source, której głównym celem jest klasyfikacja oraz reprezentacja tekstu, przy jednoczesnym szybkim i dokładnym przetwarzaniu dużych zbiorów danych.

Rozwinięcie

Stworzyliśmy aplikację w dwóch wersjach:

- trenowaną na własnym zbiorze danych
  - 5 języków
  - ok. 30 000 zdań na język
  - rozmiar modelu ok. 50MB
- korzystającą z gotowego modelu
  - 176 języków
  - trenowane na danych z Wikipedia, Tatoeba i SETimes
  - rozmiar modelu ok. 120MB

Przykłady ze zbioru danych

__label__english A friend in need is a friend indeed.

__label__german Er ist sehr freundlich zu uns.

__label__spanish No acepto tu excusa.

__label__polish Poproszę o wyraźną odpowiedź.

__label__french La gare est loin d'ici.

...

Kluczowe elementy implementacji

fastText-language-recognizer - app.py

```python
# zaimportowanie biblioteki fasttext
import fasttext

# trening modelu
model = fasttext.train_supervised(input="data/dataset.txt")

# zapis modelu
model.save_model("model/model.bin")

# odczyt modelu
model = fasttext.load_model("model/model.bin")

# predykcja języka
predicted_language = model.predict(user_input)[0][0]
```

Wyniki działania (angielski)

Własny model

I just don't know what to say

The predicted language is: **english** Probability: **100.00%**

The predicted language is: **englishlish** Probability: **0.00%**

The predicted language is: **polish** Probability: **0.00%**

I just dont know what to say

The predicted language is: **english** Probability: **99.94%**

The predicted language is: **englishlish** Probability: **0.04%**

The predicted language is: **polish** Probability: **0.01%**

Gotowy model

I just don't know what to say

The predicted language is: **English** Probability: **99.73%**

The predicted language is: **Spanish; Castilian** Probability: **0.04%**

The predicted language is: **Russian** Probability: **0.03%**

I just dont know what to say

The predicted language is: **English** Probability: **89.51%**

The predicted language is: **French** Probability: **3.38%**

The predicted language is: **German** Probability: **2.28%**

Wyniki działania (polski)

Własny

Jak się masz?

The predicted language is: **polish** Probability: **100.00%**

The predicted language is: **polishicja** Probability: **0.00%**

The predicted language is: **englishlish**. Probability: **0.00%**

Jak sie masz?

The predicted language is: **polish** Probability: **100.00%**

The predicted language is: **german** Probability: **0.11%**

The predicted language is: **englishlish**. Probability: **0.00%**

Gotowy

Jak się masz?

The predicted language is: **Polish** Probability: **100.00%**

The predicted language is: **Esperanto** Probability: **0.00%**

The predicted language is: **Spanish; Castilian** Probability: **0.00%**

Jak sie masz?

The predicted language is: **German** Probability: **76.55%**

The predicted language is: **Polish** Probability: **17.63%**

The predicted language is: **Hungarian** Probability: **4.12%**

Wyniki działania (niemiecki)

Własny

ich habe nicht verstanden

The predicted language is: **german** Probability: **99.99%**

The predicted language is: **polish** Probability: **0.00%**

The predicted language is: **germantschland** Probability: **0.00%**

ich mochte tee

The predicted language is: **german** Probability: **98.33%**

The predicted language is: **polish** Probability: **1.49%**

The predicted language is: **englishlish**. Probability: **0.01%**

Gotowy

ich habe nicht verstanden

The predicted language is: **German** Probability: **99.92%**

The predicted language is: **English** Probability: **0.03%**

The predicted language is: **French** Probability: **0.02%**

ich mochte tee

The predicted language is: **German** Probability: **98.85%**

The predicted language is: **Finnish** Probability: **0.40%**

The predicted language is: **Czech** Probability: **0.14%**

Podsumowanie

Udało nam się stworzyć aplikację rozpoznającą język wprowadzanego tekstu z prawdopodobieństwem sięgającym nawet 100%. Dla przetestowanych zdań lepiej sprawdzał się model wytrenowany na naszych danych.

Bibliografia

1. https://fasttext.cc/docs/en/support.html
2. https://fasttext.cc/docs/en/supervised-tutorial.html
3. https://fasttext.cc/docs/en/language-identification.html
4. https://towardsdatascience.com/fasttext-for-text-classification-a4b38cbff27c