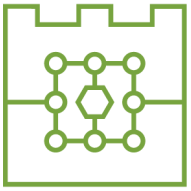


Politechnika Krakowska
Wydział Informatyki
i Telekomunikacji

Wykrywanie języka przy pomocy biblioteki FastText

Projekt z przedmiotu: Przetwarzanie Języka Naturalnego

Wykonali:
Dawid Banach
Adrian Dochnal
Mateusz Ciepał

Spis treści

1. Wstęp

Opis problemu

Rozpoznawanie języka wpisanego tekstu jest ważnym problemem w informatyce, ponieważ wiele aplikacji i systemów wymaga znajomości języka tekstu, aby móc go odpowiednio przetwarzać i interpretować. Na przykład, wyszukiwarka internetowa musi wiedzieć, w jakim języku jest napisany tekst, aby móc zwrócić odpowiednie wyniki wyszukiwania. Chatboty muszą rozpoznawać język, aby móc poprawnie odpowiadać na pytania użytkowników.

Rozpoznawanie języka tekstu jest trudnym zagadnieniem, ponieważ istnieje wiele różnych języków, a każdy z nich ma swoje własne cechy i szczególności. Wiele języków posiada różne alfabety, co dodatkowo utrudnia rozpoznawanie. Do tego dochodzi jeszcze kwestia błędów ortograficznych, które mogą być obecne w tekście i utrudniać jego interpretację.

Aby rozwiązać problem rozpoznawania języka tekstu, stosuje się różne metody i algorytmy, takie jak klasyfikacja tekstu oraz inne. W tym celu często wykorzystuje się również modele językowe, które są trenowane na dużych zbiorach danych i potrafią rozpoznawać język tekstu. Mimo to, rozpoznawanie języka nadal pozostaje trudnym problemem, szczególnie w przypadku krótkich lub niepoprawnie napisanych tekstów.

Opis biblioteki FastText

Biblioteka FastText to biblioteka opracowana przez Facebooka do szybkiego uczenia modeli oraz klasyfikacji tekstu. Służy do przetwarzania tekstu oraz do budowania modeli, które potrafią rozpoznawać i klasyfikować tekst na podstawie jego zawartości. Może być wykorzystywana do różnych zastosowań, takich jak klasyfikacja spamu, detekcja języka, klasyfikacja sentymentu czy klasyfikacja tematów.

Biblioteka FastText umożliwia szybkie trenowanie modeli na dużych zbiorach danych oraz dokonywanie predykcji na nowych danych. Pozwala również na

przetwarzanie tekstu bez potrzeby jego tokenizacji oraz umożliwia pracę z tekstem w językach, które nie posiadają spacji między słowami.

Biblioteka FastText jest wykorzystywana w różnych dziedzinach, takich jak przetwarzanie języka naturalnego, analityka internetowa, reklama czy edukacja. Może być również używana do tworzenia chatbotów oraz do automatyzacji procesów biznesowych. Biblioteka ta jest dostępna dla różnych języków programowania, w tym dla Pythona.

2. Implementacja rozwiązania

Fragmenty kodu programu wraz z opisem

Backend:

```
import fasttext
import fasttext.util
from iso_language_codes import language_name

class LanguageIdentification:
  def __init__(self):
    self.model = fasttext.load_model('lid.176 (1).bin')

  def predict_lang(self, text):
    predictions = self.model.predict(text, k=1)
    tmp = predictions[0][0]
    return language_name(tmp[9:])
```

Powyższa klasa odpowiada za identyfikację języka na podstawie otrzymanego tekstu. W konstruktorze ładowany jest model dostępny na stronie biblioteki FastText odpowiedzialny za rozpoznawanie języków.

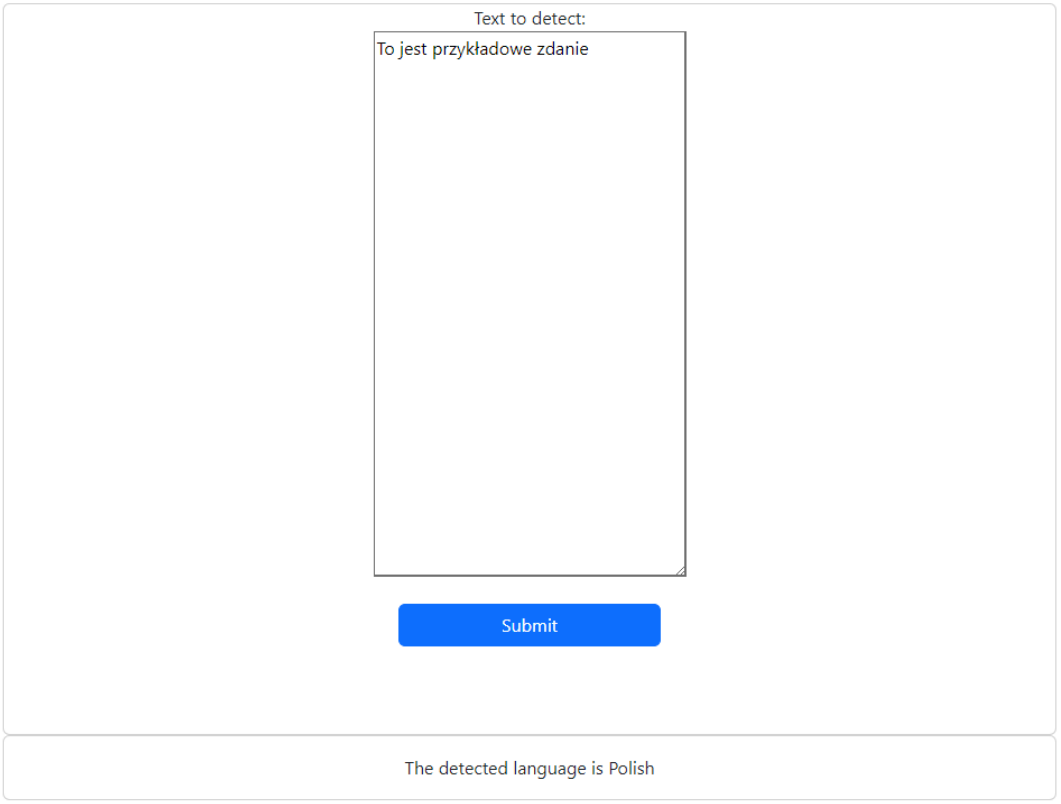
```
from wtforms import Form, StringField
from wtforms.widgets import TextArea
class TextForm(Form):
  text = StringField('Text to detect', widget=TextArea())
```

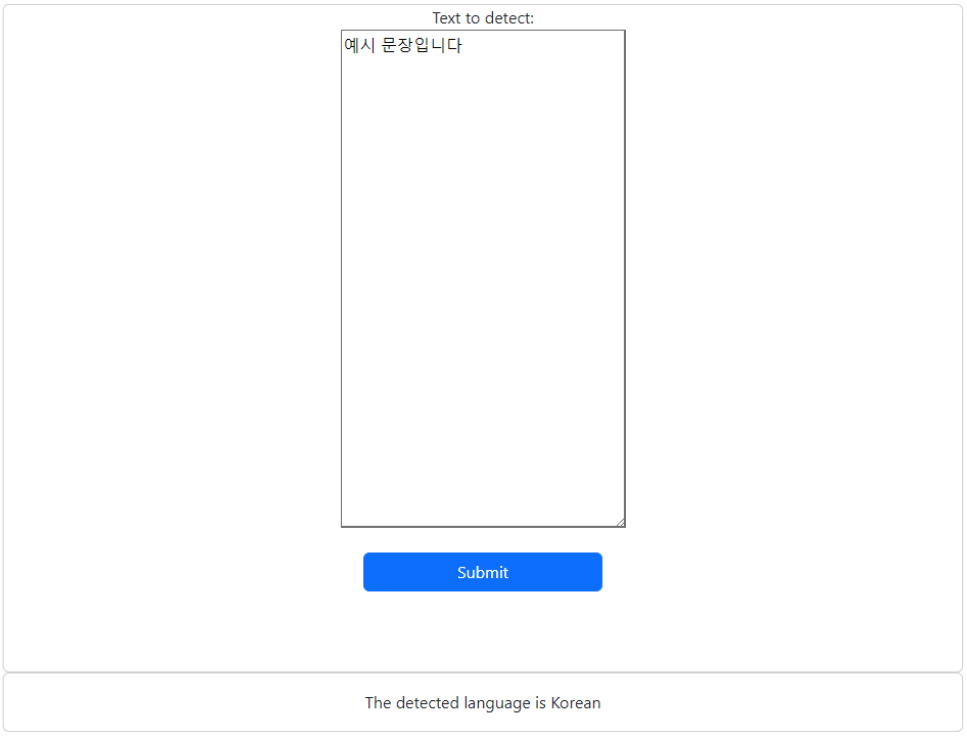
```
from flask.templating import render_template
from flask import Flask, request
from google.colab.output import eval_js
from flask_bootstrap import Bootstrap5

app = Flask(__name__, template_folder="templates")

languageIdentification = LanguageIdentification()
bootstrap = Bootstrap5(app)

print(eval_js("google.colab.kernel.proxyPort(5000)"))

@app.route('/detect', methods=["POST", "GET"])
def hello_world():
  if request.method == "POST":
    form = TextForm(request.form)
    text = form.text.data
    language = languageIdentification.predict_lang(text)
    return render_template('test.html',form=form, language=f"The
detected language is {language}")
  if request.method == "GET":
    form = TextForm()
    return render_template('test.html', form=form, language="The
detected language is")


if __name__ == '__main__':
    app.run()
```

Frontend:

```
<head>
  {{ bootstrap.load_css() }}
</head>

<div style="height: 70%;margin-top: 10vh;" class="card col-md-6
pagination-centered text-center mx-auto">
  <form method="POST" action="/detect" class="col-md-6
pagination-centered text-center mx-auto">
    <div class="">
      {{ form.text.label }}:
      <br/>
      <div class="">
        {{ form.text(cols="35", rows="20") }}
      </div>
    </div>
    <br/>
    <button type=submit class="btn btn-primary col-md-6
pagination-centered text-center mx-auto">Submit</button>
  </form>
</div>

<div class="card col-md-6 pagination-centered text-center mx-auto">
  <div class="card-body">
    {{ language }}
</div>


<div>
  {{ bootstrap.load_js() }}
</div>
```

Przykłady użycia





3. Podsumowanie

Podczas realizacji projektu zrealizowano program służący do rozpoznawania języka tekstu, oparty o bibliotekę FastText. Projekt ten polegał na opracowaniu algorytmu klasyfikacji tekstu oraz na wykorzystaniu modeli językowych do rozpoznawania języka tekstu.

Dzięki zastosowaniu biblioteki FastText udało się osiągnąć bardzo dobre rezultaty w rozpoznawaniu języka tekstu. Algorytm klasyfikacji tekstu oparty o tę bibliotekę charakteryzował się dużą dokładnością oraz szybkością działania. Modele językowe trenowane z wykorzystaniem biblioteki FastText również cechowały się dobrą skutecznością w rozpoznawaniu języka tekstu.

W trakcie realizacji projektu udało się rozwiązać problem rozpoznawania języka tekstu za pomocą biblioteki FastText. Algorytm klasyfikacji tekstu oraz modele językowe oparte o tę bibliotekę okazały się skuteczne w rozpoznawaniu języka tekstu w wielu różnych językach. Opracowane rozwiązanie może być wykorzystywane w różnych aplikacjach i systemach, które wymagają znajomości języka tekstu, np. w wyszukiwarkach internetowych

4. Bibliografia

Dokumentacja biblioteki FastText: https://fasttext.cc/docs/en/support.html
Dokumentacja Bootstrapa: https://getbootstrap.com/docs/5.2/getting-started/introduction/
Dokumentacja Pythona: https://docs.python.org/3.11/