

Narzędzie do rozpoznawania języka oparte na bibliotece FastText

Dawid Laska, Jakub Kurc, Jan Kołek



Cel projektu

- Projekt dotyczy stworzenia programu do automatycznego rozpoznawania języka na podstawie wprowadzonego tekstu.
- W tym celu wykorzystano gotowy model **FastText**, który potrafi identyfikować 176 języków.
- Drugi model, opracowany przez autorów projektu, zaprojektowany do rozpoznawania pięciu języków: angielskiego, polskiego, hiszpańskiego, francuskiego i niemieckiego.



Zastosowanie

- Tłumaczenia
- Analiza nastrojów
- Personalizacja treści
- Edukacja
- Obsługa klienta



Interfejs graficzny

- Zaimplementowany za pomocą **HTML**, **CSS** i frameworka **Flask**
- Użytkownik może wprowadzić dowolny tekst
- Po wysłaniu formularza metodą POST użytkownik otrzymuje odpowiedź w jakim języku jest wprowadzony tekst
- Możliwość przełączania się pomiędzy modelami FastText i autorskim



Language Identifier

Enter text:

☐ Use Fasttext model?

Identify Language

english



Trenowanie modelu

- Trenowanie odbywa się za pomocą funkcji **fasttext.train_supervised()**
- Posiada wiele parametrów odpowiedzialnych za trenowanie
- Parametry były wielokrotnie modyfikowane, aby osiągnąć jak najlepsze wyniki

Przykładowe dane

```
__label__english "Why do I have to wake up so early?" complained the student.  
__label__english "I don't know how I can get all of this done before the deadline!" wondered the office worker.  
__label__english "This cake tastes amazing!" exclaimed the chef.  
__label__english "Has anyone seen my keys?" asked the desperate driver.  
__label__english "I've always dreamed of going to Paris!" said the young girl.  
__label__french "Pourquoi dois-je me lever si tôt?" se plaignait l'étudiant.  
__label__french "Je ne sais pas comment je vais tout terminer avant la date limite!" se demanda l'employé.  
__label__french "Ce gâteau a un goût incroyable!" s'exclama le chef.  
__label__french "Quelqu'un a vu mes clés?" demanda le conducteur désespéré.  
__label__french "J'ai toujours rêvé d'aller à Paris!" dit la jeune fille.  
__label__german "Warum muss ich so früh aufstehen?" beschwerte sich der Student.  
__label__german "Ich weiß nicht, wie ich das alles vor der Deadline schaffen soll!" fragte sich die Büroangestellte.  
__label__german "Dieser Kuchen schmeckt fantastisch!" rief der Koch aus.  
__label__german "Hat jemand meine Schlüssel gesehen?" fragte der verzweifelte Fahrer.  
__label__german "Ich habe immer davon geträumt, nach Paris zu reisen!" sagte das junge Mädchen.  
__label__spanish "¿Por qué tengo que levantarme tan temprano?" se quejó el estudiante.  
__label__spanish "No sé cómo voy a terminar todo esto antes de la fecha límite!" se preguntó la trabajadora.  
__label__spanish "¡Este pastel sabe increíble!" exclamó el chef.  
__label__spanish "¿Alguien ha visto mis llaves?" preguntó el conductor desesperado.  
__label__spanish "¡Siempre he soñado con ir a París!" dijo la joven.
```



Parametry do trenowania

- **input**: ścieżka do pliku z danymi treningowymi.
- **label_prefix**: prefiks dodawany przed etykietami klas w pliku z danymi.
- **epoch**: liczba epok treningowych. Niska wartość prowadzi do niedouczenia modelu, natomiast zbyt duża może skutkować przeuczeniem, co oznacza, że model będzie zbyt dopasowany do danych treningowych i słabo generalizował na nowe dane, co może skutkować niedokładnymi wynikami w praktyce.
- **lr (learning rate)**: współczynnik uczenia. Zbyt mały sprawia, że proces uczenia przebiega wolno, natomiast zbyt duży może pogorszyć zdolność modelu do generalizacji, prowadząc do słabszych wyników na danych testowych.



- **wordNgrams**: długość n-gramów używanych do reprezentacji słów. Krótkie n-gramy mogą obniżyć skuteczność detekcji języka, natomiast zbyt długie mogą zwiększyć wymiarowość wektorów, co z kolei komplikuje proces trenowania i może prowadzić do przeuczenia.
- **bucket**: liczba kubeków wykorzystywanych do haszowania. Zbyt mała liczba zwiększa ryzyko kolizji haszy (różne słowa mogą być traktowane jako jedno), co obniża jakość modelu. Z kolei zbyt duża wartość niepotrzebnie wydłuża czas treningu i zwiększa zapotrzebowanie na pamięć.
- **dim**: liczba wymiarów wektorowej reprezentacji słów (tu 720). Niewystarczająca liczba wymiarów ogranicza zdolność modelu do uchwycenia złożonych informacji o słowach, co obniża jakość klasyfikacji. Zbyt duża wartość może prowadzić do przeuczenia i problemów z generalizacją na nowe dane.
- **thread**: liczba wątków wykorzystywanych do treningu.
- **ws (window size)** określa szerokość okna kontekstowego, czyli liczbę słów z sąsiedztwa, które są brane pod uwagę podczas tworzenia reprezentacji danego słowa.
- **loss** określa funkcję straty używaną podczas trenowania modelu..



Wyniki i podsumowanie

- Testy wykazały, że precyzja modelu wynosi około 96%, co jest znakomitą wynikiem, biorąc pod uwagę niewielki rozmiar zbioru danych użytego do trenowania modelu.
- W niektórych, szczególnych przypadkach model opracowany przez nasz zespół przewyższył dokładnością gotowy model dostarczany przez FastText.
- Biblioteka **FastText** to wszechstronne narzędzie, które poza rozpoznawaniem języka może być z powodzeniem wykorzystywane w wielu innych obszarach przetwarzania języka naturalnego.

Dziękujemy!

