

02.06.2024, Kraków

Narzędzie do rozpoznawania języka w oparciu o bibliotekę fasttext

Szymon Twardowski

Mateusz Moneta

Cel projektu

Celem projektu było przygotowanie narzędzia, które będzie rozpoznawać język, tekst w oparciu o bibliotekę fasttext. Kluczowa technika w wielu zastosowaniach przetwarzania języka naturalnego, takich jak analiza sentymentów, rozpoznawanie mowy, tłumaczenie maszynowe.

Przykład zastosowania

W erze, gdy informacje są generowane i dostępne w olbrzymich ilościach, klasyfikacja pomaga uporządkować te dane i uczynić je łatwiejszymi do zrozumienia i analizy. Przykładowo, klasyfikacja e-maila pod kątem tego, czy jest to SPAM, pomaga w efektywnym zarządzaniu pocztą.

Czym jest biblioteka fasttext?

Biblioteka fasttext jest biblioteką typu open-source stworzoną przez zespół Facebook AI Research (FAIR), która pozwala na klasyfikację tekstu. W swojej początkowej wersji biblioteka ta wykorzystywała technikę trenowania modelu o nazwie skip-gram, ale aktualnie wspiera również metodę Continuous Bag of Words (CBOW). Fasttext pozwala na trenowanie około 1 miliarda słów w czasie mniejszym niż 10 minut.

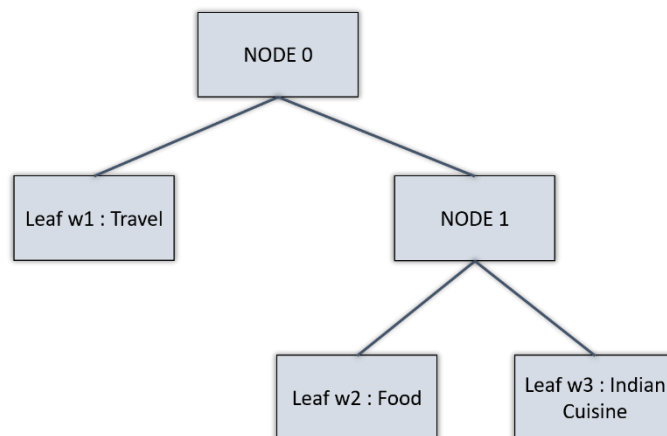
Klasyfikacja tekstu

Celem klasyfikacji tekstu, jest przypisanie dokumentów do jednej lub wielu kategorii. Takimi kategoriami mogą być język danego tekstu, oceny recenzji czy również jedzenie. Obecnie jednym z dominujących podejść do budowy tego typu klasyfikatorów jest wykorzystywanie do tego uczenia maszynowego, czyli przekazanie sztucznej inteligencji reguł klasyfikacji na podstawie przykładów. Do budowy takich klasyfikatorów potrzebne są dane pogrupowane, które składają się z dokumentów oraz odpowiadających im kategorii.

Rodzaje klasyfikatorów

Klasyfikator liniowy polega na tym, że tekst oraz etykiety są reprezentowane jako wektory. Znajdujemy takie reprezentacje wektorowe, których tekst i związane z nim etykiety mają zbliżone wektory. Tłumacząc dokładniej to wektor, który odpowiada danemu tekstowi jest bliższy odpowiadającej jej etykietce.

Klasyfikator hierarchiczny polega na tym, że reprezentowane etykiety są przedstawione w formie drzewa binarnego. Każda gałąź w drzewie binarnym reprezentuje prawdopodobieństwo. Etykieta natomiast jest reprezentowana przez prawdopodobieństwo wzdłuż ścieżki do danej etykiety. Oznacza to, że węzły liści reprezentują etykiety. Biblioteka fasttext używa w tym celu algorytmu Huffmana do budowy drzew, aby w pełni wykorzystać możliwość, że klasy mogą być niebalansowane, niewyważone.



Klasyfikator hierarchiczny używany przez FastText

Polecenia biblioteki

supervised	train a supervised classifier
quantize	quantize a model to reduce the memory usage
test	evaluate a supervised classifier
test-label	print labels with precision and recall scores
predict	predict most likely labels
predict-prob	predict most likely labels with probabilities
skipgram	train a skipgram model
cbow	train a cbow model
print-word-vectors	print word vectors given a trained model
print-sentence-vectors	print sentence vectors given a trained model
print-ngrams	print ngrams given a trained model and word
nn	query for nearest neighbors
analogies	query for analogies
dump	dump arguments, dictionary, input / output vectors

Trenowanie modelu

```
fasttext skipgram -input amazon_reviews.txt -output model_trained
```

Tutaj plikiem wejściowym jest *amazon_reviews.txt*. Upewnij się, że podałeś pełną ścieżkę do pliku, jeśli nie znajduje się on w katalogu dane. *model_trained* to nazwa nadana plikowi wyjściowemu.

Najpierw rozpoczyna czytanie słów znajdujących się w dokumencie wejściowym. Dokument składał się z 32 milionów słów, a jego szacowany czas dotarcia wynosił około 15 minut.

```
(base) C:\Users\hp\Desktop\v0.9.2\fastText-0.9.2>fasttext skipgram -input amazon_reviews.txt -output model_trained
Read 32M words
Number of words: 142106
Number of labels: 2
Progress: 0.8% words/sec/thread: 19711 lr: 0.049585 avg.loss: 2.199082 ETA: 0h15m14s
```

Podaje szczegółowe statystyki szybkości uczenia się sieci neuronowej, ile słów jest przetwarzanych w każdej sekundzie w każdym wątku. Pokazuje również wartość straty, która maleje w miarę uczenia modelu.

Po przeszkoleniu modelu generowane są dwa pliki, tj. *model_trained.bin* i *model_trained.vec*.

- Plik *.bin* zawiera parametry modelu wraz ze słownikiem. To jest plik, który używa *fasttext*.
- Plik *.vec* to plik tekstowy zawierający wektory słów. To jest plik, który będziesz używać w swoich aplikacjach.

Konfiguracja

1. Aby móc korzystać z programu należy zainstalować bibliotekę fasttext

pip3 install fasttext

```
PS C:\Users\asus> pip install fasttext
Collecting fasttext
  Using cached fasttext-0.9.2.tar.gz (68 kB)
  Preparing metadata (setup.py) ... done
Requirement already satisfied: pybind11<=2.2 in c:\users\asus\appdata\local\programs\python\python312\lib\site-packages (from fasttext) (2.12.0)
Requirement already satisfied: setuptools<=0.7.0 in c:\users\asus\appdata\local\programs\python\python312\lib\site-packages (from fasttext) (70.0.0)
Requirement already satisfied: numpy in c:\users\asus\appdata\local\programs\python\python312\lib\site-packages (from fasttext) (1.26.4)
Building wheels for collected packages: fasttext
  Building wheel for fasttext (setup.py) ... error
error: subprocess-exited-with-error

  × python setup.py bdist_wheel did not run successfully.
  | exit code: 1
  |
  | [33 lines of output]
  | C:\Users\asus\AppData\Local\Programs\Python\Python312\Lib\site-packages\setuptools\dist.py:476: SetuptoolsDeprecationWarning: Invalid dash-separated options
  | !!
  |
  | *****
  | Usage of dash-separated 'description-file' will not be supported in future
  | versions. Please use the underscore name 'description_file' instead.
  |
  | By 2024-Sep-26, you need to update your project and remove deprecated calls
  | or your builds will no longer be supported.
  |
  | See https://setuptools.pypa.io/en/latest/userguide/declarative_config.html for details.
  | *****
  |
  | !!
  |     opt = self.warn_dash_deprecation(opt, section)
  |     running bdist_wheel
  |     running build
  |     running build_py
  |     creating build
  |     creating build\lib.win-amd64-cpython-312
  |     creating build\lib.win-amd64-cpython-312\fasttext
  |     copying python\fasttext_module\fasttext\FastText.py -> build\lib.win-amd64-cpython-312\fasttext
  |     copying python\fasttext_module\fasttext\__init__.py -> build\lib.win-amd64-cpython-312\fasttext
  |     creating build\lib.win-amd64-cpython-312\fasttext\util
  |     copying python\fasttext_module\fasttext\util\util.py -> build\lib.win-amd64-cpython-312\fasttext\util
  |     copying python\fasttext_module\fasttext\util\__init__.py -> build\lib.win-amd64-cpython-312\fasttext\util
  |     creating build\lib.win-amd64-cpython-312\fasttext\tests
  |     copying python\fasttext_module\fasttext\tests\test_configurations.py -> build\lib.win-amd64-cpython-312\fasttext\tests
  |     copying python\fasttext_module\fasttext\tests\test_script.py -> build\lib.win-amd64-cpython-312\fasttext\tests
  |     copying python\fasttext_module\fasttext\tests\__init__.py -> build\lib.win-amd64-cpython-312\fasttext\tests
  |     running build_ext
  |     building 'fasttext_pybind' extension
  |     error: Microsoft Visual C++ 14.0 or greater is required. Get it with "Microsoft C++ Build Tools": https://visualstudio.microsoft.com/visual-cpp-build-tools/
  | [end of output]
  |
  | note: This error originates from a subprocess, and is likely not a problem with pip.
  | ERROR: Failed building wheel for fasttext
  | Running setup.py clean for fasttext
  | Failed to build fasttext
  | ERROR: Could not build wheels for fasttext, which is required to install pyproject.toml-based projects
```

Błędy podczas instalacji pakietu na systemie Windows, które wynikają z brakujących bibliotek „Visual Build Tools”. W przypadku naszego zespołu problem występował nawet po ich zainstalowaniu. W związku z tym środowisko zostało skonfigurowane na MacOS.

2. Należy również pobrać plik lid.176.bin poprzez poleceni wnet

wget <https://dl.fbaipublicfiles.com/fasttext/supervised-models/lid.176.bin>

3. Importujemy bibliotekę na potrzeby naszego programu

```
from fasttext import load_model
```

- load_model - jest to funkcja z biblioteki Fasttext, która ładuje wcześniej trenowany model do pamięci

Podsumowanie

Wektory słów są technologią, która umożliwi znaczący przełom w zastosowaniach i badaniach „Przetwarzania języka naturalnego”. Podkreślają one moc wyuczonych reprezentacji danych wejściowych w warstwach ukrytych. Budowanie lepszych aplikacji wymaga dobrego zrozumienia wektorów słów, a biblioteka fasttext w znacznym stopniu wesprze w rozwoju tego zrozumienia.

Bibliografia

- <https://www.promptopedia.pl/techniki-podstawowe/klasyfikacja-tekstu>
- <https://fasttext.cc>
- <https://www.geeksforgeeks.org/fasttext-working-and-implementation/>
- <https://pythonwife.com/fasttext-in-nlp/>