Rozpoznawanie wiadomości dotyczących katastrof

Jakub Rymuza Karol Nowiński

Marzec 2022

Spis treści

1 Wstęp

Projekt polega na przetwarzaniu wiadomości z serwisu Twitter w celu ustalenia, które z nich mówią o katastrofach. Może to pomóc m.in. we wcześniejszym rozpoznawaniu sytuacji zagrażajacych życiu. Może to też pomóc służbom w ocenie skali zagrożenia zdarzeniem.

2 Podział pracy

W związku z tym, że grupa składa się z dwóch osób, zdecydowaliśmy, że nie będziemy dzielić pracy na poszczególne role. Praca wykonywana będzie w większości wspólnie, pracując wspólnie nad rozwiązaniem problemu.

3 Wykorzystywane narzędzia

Projekt opiera się na procesowaniu języka naturalnego za pomocą technik uczenia maszynowego. Projekt zostanie wykonany w języku Python. W szczególności wykorzystane zostaną biblioteki *NumPy* oraz *Pandas* w środowisku *VS Code*.

4 Wykorzystane dane

W projekcie wykorzystano dwa zbiory danych z [1]:

- zbiór treningowy *train.csv*, który zostanie wykorzystany do wytrenowania modelu do rozwiązywania problemu. Zawiera on 7613 rekordów. Rekordy zawierają następujące pola:
  - *id* - identyfikator rekordu,
  - *keyword* - słowo kluczowe uprzednio wyciągnięte z wiadomości - wszystkie słowa kluczowe dotyczą katastrof, są tą słowa takie jak na przykład "crash", "earthquake" i tym podobne. To pole jest opcjonalne, tzn. nie każdy rekord je zawiera,
  - *location* - lokalizacją z której wiadomość została wysłana. Podobnie jak *keyword*, jest to pole opcjonalne,
  - *text* - najważniejsza pole - zawiera treść wiadomości,

  - *target* - wartość logiczna stwierdzająca czy dana wiadomość mówi o katastrofie czy nie.

- zbiór testowy *test.csv* - zbiór na którym model będzie testowany. Zawiera on 3263 rekordów. Rekordy wyglądają tak samo, jak w przypadku zbioru treningowe, za wyjątkiem oczywiście braku pola *target*, którego obliczenie jest celem projektu.

5 Bibliografia

[1] https://www.kaggle.com/c/nlp-getting-started