Checkpoint 2

Warsztaty z technik uczenia maszynowego

Jakub Rymuza Karol Nowiński

21 kwietnia 2022

1 Przygotowanie danych

W celu poprawienia treningu modelu przeprowadzono następujące operacji czyszczące i przygotowujący dane:

- Przekształcenie całego tekstu do małych liter. Dzięki temu słowa *fire* i *Fire* będą traktowane jako to samo słowo, a nie dwa różne.
- Usunięcie linków z danych - nasz model nie będzie w stanie wchodzić w linki i ich analizować, więc tylko utrudniałyby one niepotrzebnie proces uczenia.
- Poprawienie błędów konwersji w danych - w używanych danych w niektórych miejscach można znaleźć ciągi *%20* (kod ASCII spacji) zamiast znaku spacji. Może to spowodować, że na przykład fraza *building%20on%20fire* będzie traktowana jako jedno słowo i nie rozpozna że chodzi tam o coś związanego z ogniem.
- Rozwinięcie skrótów - na przykład zamiana *asap* na *as soon as possible*.
- Usunięcie tzw. *stopwords*. Są to słowa typu *the* czy *is*, które wnoszą niewiele informacji do treści i mogą zostać pominięte przy treningu modelu.

Ponadto na tekście zostały wykonano operacje tokenizacji i wektoryzacji.

2 Wstępne modele

Na danych wstępnie użyto trzy proste modele testowe - *ridge regression classifier*, *logistic regression classifier* i *naive Bayes classifier*. Zgodnie z przypuszczeniami te klasyfikatory nie uzyskały szczególnie dobrych wyników - są one w granicach 55% do 70% skuteczności.