

MPP 3

Zbierać z publicznych źródeł (Wikipedia, Projekt Gutenberg, WWW) teksty w 3-4 różnych językach. Jeden język – jeden podkatalog. Liczba języków nie jest z góry znana. Teksty przechowujemy w plikach tekstowych. Minimalna długość tekstu – 1000 znaków. Minimalna liczba plików w każdym języku – 10.

Dane dzielimy na zbiór treningowy i testowy w proporcji 50-50.

Wczytujemy pliki, usuwamy znaki „narodowe”, zliczamy częstości wystąpień znaków „a”-„z” (zakres od 0.0 do 1.0) i tworzymy dla każdego tekstu wektor częstości.

Następnie algorytmem delty trenujemy perceptrony, które rozróżniają poszczególne języki. Zbudujemy z nich jednowarstwową sieć perceptronową.

Testujemy ją potem na danych ze zbioru testowego. Wypisujemy macierz omyłek i miary jakości dla każdej klasy.

Program musi umożliwiać wielokrotne wprowadzenie tekstu przykładowego i podanie dla takiego tekstu jego wynik klasyfikacji.

Nie można używać żadnych bibliotek ML, wszystko ma być zaimplementowane od zera w pętlach, if-ach, wyjście perceptronu trzeba liczyć za pomocą działań arytmetycznych, etc.