Zadanie kwalifikacyjne

Zadanie będzie składało się z dwóch podzadań. Pierwsze z nich będzie polegało na napisaniu generatora losowych słów w danym języku. W drugim natomiast trzeba będzie zaklasyfikować słowa do języka, z którego pochodzą. Oba podzadania trzeba będzie zrobić na platformie Google Colab.

Wymagania techniczne

Należy skopiować Google Colab dostępny pod adresem

`https://colab.research.google.com/drive/1nn1Ga2R7DyQxcM0TT4gL-JeKV8SfLEM6`

na własny Dysk Google, a następnie zmodyfikować kod, rozwiązując podzadania. Jako rozwiązanie należy załączyć plik dostępny po skorzystaniu z opcji *Plik/Pobier/Pobierz Plik Ipynb*.

Rozwiązanie powinno być dobrze opisane.

Kryteria oceny

Za zadanie będzie można uzyskać maksymalnie $\mathbf{50 = 20 + 30}$ punktów, przy czym **20** punktów jest za pierwsze podzadanie(generator), a **30** za drugie podzadanie(klasyfikator).

Ocenie będzie podlegał pomysł, jakość generowanych słów oraz sprawdzona automatycznie jakość klasyfikacji.

Podpowiedź

Zdecydowanie nie chodzi mi o to, aby implementować jakiekolwiek sieci neuronowe. Nawet jest to zakazane. Zasugeruje prostsze rozwiązanie, oparte na prawdopodobieństwie, które daje satysfakcjonujące rezultaty.

Analizując dane słowa możemy obliczyć jak często po literce 'a' występuje litera 'b', jak często litera 'c', ... itd. Jeśli chcemy wygenerować słowo, to losujemy pierwszą literę. Następnie patrzymy na rozkład prawdopodobieństwa liter, które występują po danej literze. Niektóre będą występować częściej – znacznie bardziej powszechne jest, aby po literze 'c' była litera 'z' niż litera 'f'. Losujemy tę literę korzystając z rozkładu otrzymanego za pomocą listy słów. Postępujemy tak dalej.

Ta idea nosi nazwę *Markov Word Generator*, można coś – choć niewiele – znaleźć na ten temat w internecie.

Do klasyfikatora podpowiedzi nie ma, ale stosunkowo łatwo jest wpaść na pewne heurystyki.

Kontakt

Pytania proszę kierować pod adres *pf.gadzinski@student.uw.edu.pl*.