Przetwarzanie języka naturalnego
Pracownia 4b[1]

Wszystkie zadania są ważne do końca semestru, wszystkie zadania mają oznaczenie E.

Lista nawiązuje do konkursu Zero Speech Challenge 2021 (ale w mocno uproszczonej postaci), w którym dysponując **jedynie** zbiorem nagrań mamy odpowiadać na różne pytania dotyczące języka. My skoncentrujemy się na jednym pytaniu: widząc dwa wyrazy zdecyduj, który z nich jest poprawnym słowem danego języka. Dodatkowo zrezygnujemy z aspektu związanego z dźwiękiem, czyli danymi będą (zniekształcone) teksty polskie. Plik definiujący zadanie składa się z par wyrazów (dla uproszczenia pierwszy wyraz zawsze jest poprawny, a drugi sztucznie utworzony, oczywiście z tej właściwości korzystamy tylko przy testowaniu). Rozwiązanie każdego z zadań powinno:

a) zawierać funkcję, która dla potencjalnego słowa podaje wartość punktową odpowiadającą przekonaniu programu, że jest to prawdziwe słowo (nie ma żadnych warunków na zakres tej wartości)

b) wykonać dostarczone testy, przypisując każdemu przypadkowi testowemu wartość 1, jeżeli poprawne słowo dostanie więcej punktów od niepoprawnego, 0.5 – jeżeli dostaną tyle samo, 0 – w pozostałym przypadku.

c) Przedstawić uśredniony wynik testów[2]

Można korzystać z faktu, że średnia długość słowa to około 6.2 znaku. Każda dodatkowa wiedza o języku może pochodzić jedynie z danych dostarczonych wraz z zadaniem (oczywiście nie wolno wykorzystywać danych z zadania 1 do zadania 2 i odwrotnie).

Zadanie 1. (5p+) W tym zadaniu zniekształcenie polega na usunięciu spacji z każdego zdania (ale przy zachowaniu podziału na zdania, które ciągle są w osobnych linijkach). Zatem dane wyglądają tak:

```
poiiwojnieświatowejniebyłojużtejseriinapkp
poparuminutachtocośbyłojużbardzobliskoekranu
jestopartynazłymwilkuzczerwonegokapturkaitrzechmałychświnek
```

Pary słów-niesłów do tego zadania charakteryzują się tym, że:

a) słowa poprawne występują w oryginalnym tekście (ze spacjami)

b) słowa niepoprawne są ciągami liter z tekstu bez spacji (mamy dodatkowo gwarancję, że nie występowały one jako słowa w oryginalnym tekście, choć oczywiście mogą być przypadkowo poprawnymi słowami polskimi)

c) słowa niepoprawne dodatkowo zostały zaakceptowane przez prosty n-gramowy literkowy model językowy (przykładowo `śbyłojużb` zostałoby zapewne odrzucone)

Przykładowe testy:

```
zastanawiam taktymiędzy
drugorzędnie czegospowodo
przechowywania stanieprzekaza
doprowadził jarozwiązan
parkowanie zdzierasza
panowie wypełni
publicznych icznościami
ograniczyć ropejskiej
```

[1] Jeszcze będzie 4c

[2] Dla ewaluacyjnych par słów, które zostaną utworzone tym samym programem co pary uczące

Do zaliczenia zadania trzeba osiągnąć poprawność 0.7 (aczkolwiek możliwe są częściowe punkty za mniejsze wyniki, uwzględniające nakład pracy studenta, wg uznania prowadzącego pracownię). Każde kolejne 0.04 to 1 punkt. W przypadku wyraźnego przekroczenia wartości 0.9 możliwe są dalsze premie egzaminacyjne, przydzielane podczas konsultacji z wykładowcą.

Zadanie 2. (5p+) W tym zadaniu zniekształcenie polega na zamianie średnio co piątej litery na losową literę, oraz skasowanie 2/3 spacji. Zatem dane wyglądają tak:

```
poiswajniefwidtowejziebołojyżtip yeiiixa pkp
poparu minutzch to cośłyłojużbardzkblesko edranu
jestooarty nazłymwilbu zczexwłnegokapturka h tróechmałych ywinzk
dalsłapodróżpa pcłnozmoqliwazestjeqynieóamofhodami terćncohmi
```

Pary słów-niesłów do tego zadania charakteryzują się tym, że:

a) słowa poprawne występują w oryginalnym tekście (ze spacjami, bez zniekształceń)

b) słowa niepoprawne zostały wygenerowane przez zmodyfikowany na potrzeby języka polskiego i uproszczony algorytm Wuggy

c) „prawe" słowo ma podobną strukturę sylabową do słowa „lewego"

d) słowa niepoprawne dodatkowo zostały zaakceptowane przez prosty n-gramowy literkowy model językowy

Przykładowe testy:

```
dekarskiej dokarskiej
proszę prosza
oprawa oprowa
do żo
powiedzieć powiedziem
pracowała pracowana
```

Do zaliczenia zadania trzeba osiągnąć poprawność 0.7 (aczkolwiek możliwe są częściowe punkty za mniejsze wyniki, uwzględniające nakład pracy studenta, wg uznania prowadzącego pracownię). Każde kolejne **0.05** to 1 punkt. W przypadku wyraźnego przekroczenia wartości 0.95 możliwe są dalsze premie egzaminacyjne, przydzielane podczas konsultacji z wykładowcą.