TI – zad 2

Kamil Kowalczyk, 136742, L1, czwartek 13:30

Do rozwiązania wszystkich zadań, na wejściu podano korpus Wikipedii.

Zad 1

Poniżej 10 najpopularniejszych słów w tekście Wikipedii, wraz z ich liczebnością i informacją jaki procent wszystkich słów stanowią:

the	118991	6.47%
of	59073	3.21%
and	48804	2.65%
in	47667	2.59%
a	36762	2.00%
to	33997	1.85%
was	19579	1.06%
is	16649	0.90%
for	14178	0.77%
on	13896	0.76%

Według wyliczeń, przeciętny Polak byłby w stanie przekazać zaledwie **3.36%** wiedzy z korpusu Wikipedii.

Zbiór 30000 najpopularniejszych słów stanowi **94.72%** wszystkich słów. Natomiast zbiór 6000 najpopularniejszych słów to **82.25%** wszystkich słów.

Zad 2

```
and the been 1996 garage during is estonian by pilots reviews is the cash
or greene employees two on in yoldi listed
```

Tak jak można było się spodziewać, najczęściej występujące słowa w przybliżeniu pierwszego rzędu to m.in.: „the”, „and”, „is”, „in”.

Zad 3

Przybliżenie źródła Markova pierwszego rzędu:

```
was released a bit bigger than synthetic organic chemistry it would take
the black talon made at the
```

Przybliżenie źródła Markova drugiego rzędu:

```
1 may 1920 and der falsche dimitry the highlight of the collegiate level
at duke university florida state university
```

W przypadku przybliżenia drugiego rzędu, ciąg wyliczany był na podstawie początkowych słów „probability of". Dla pierwszego rzędu było to samo „probability". W obu przykładach można zauważyć, że połączenia kolejnych wyrazów nie są chaotyczne i dobrane w sposób losowy. Dzięki prawdopodobieństwu warunkowemu dwójki wyrazów z pierwszego przykładu są zazwyczaj znanymi połączeniami dwóch słów z języka angielskiego. To samo się tyczy przybliżenia drugiego rzędu, w którym dodatkowo można zauważyć mnóstwo trójek, które zazwyczaj tworzą sensowne połączenia słów.
Niestety, nie wiem dlaczego metoda Shannona mogłaby generować odmienny rozkład do rozkładu wygenerowanego z użyciem łańcucha Markova. Po dłuższym zastanowieniu, wydaje mi się, że rozkłady te powinny być podobne.