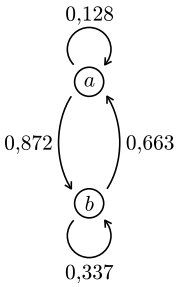
Twierdzenie ergodyczne

Rafał SZTENCEL

	a	b
a	0,128	0,872
b	0,663	0,337



Sturba wasza wlań chełbiasta!

Stanisław Ignacy Witkiewicz, *Szewcy*, cytat być może niedokładny.

Apentuła niewdziosek, te będy gruwaśne,
W koć turmiela weprzącłnie, kostrą bajtę spoczy,
Oproszędły znimęci, wyświrle uwzroczy,
A korśliwe porsacze dogremnie wyczkaśnie.

Stanisław Lem, *Cyberiada*.

Przez gwiezdne niebo jak przez durszlak
Noc sieje źdźbła mdłe blasków ostrych.
Gwóźdź Kacper mknie przez twardych dróg szlak
I idą za nim słupy wiorst pstre.

Julian Tuwim, *Jeszcze jeden wiersz poety Andrzeja Wiktora Butnego*.

W 1913 roku rosyjski matematyk Andriej Andriejewicz Markow (senior) dokonał analizy statystycznej ciągu 20 000 liter poematu Puszkina „Eugeniusz Oniegin”, badając następstwa samogłoska/spółgłoska [1], i uzyskał wyniki przedstawione obok.

Z pierwszego wiersza tabeli odczytujemy, że po samogłosce (a) w 87,2% przypadków następuje spółgłoska (b); pozostałe 12,8% przypadków to samogłoski.

Jest to prawdopodobnie pierwszy opublikowany przykład analizy statystycznej tekstu. Autor chciał sprawdzić, czy można traktować badany ciąg znaków – choćby w przybliżeniu – jako obiekt, zwany obecnie łańcuchem Markowa.

Wyobraźmy sobie najpierw, że litery są losowane niezależnie, a prawdopodobieństwa nie zależą od numeru losowania. Dobrze wiadomo, że częstości liter będą się stabilizować (jest to prawo wielkich liczb), będzie też zachodzić centralne twierdzenie graniczne: odchylenia częstości od prawdopodobieństwa w długiej serii mają rozkład zbliżony do normalnego.

Załóżmy teraz, że szanse wylosowania kolejnej litery zależą od tego, jaką wylosowano poprzednio. Można powiedzieć, że mamy do czynienia z grą planszową, gdzie ewolucja układu zależy od prawdopodobieństw przejścia z i-tego do j-tego stanu. Tak jak w grze planszowej, losowania są niezależne.

Markow bada zatem łańcuch o dwóch stanach: a, b i prawdopodobieństwach przejścia tworzących macierz 2×2. Okazuje się, że po długim czasie łańcuch niejako zapomina, skąd wystartował, i prawdopodobieństwa pobytu w stanach a i b zmierzają do dodatnich granic, które oznaczymy także literami a i b. Oczywiście, $a+b=1$. Liczby a i b wyznaczają tzw. rozkład stacjonarny.

Dokładniej, zbieżność do (jedynego) rozkładu stacjonarnego ma miejsce, jeśli stanów jest skończenie wiele, każde dwa stany komunikują się (tj. możliwe jest przejście między stanami, być może w wielu krokach), a ponadto stany są nieokresowe (największy wspólny dzielnik czasów powrotu równy 1). Jest to twierdzenie ergodyczne. Wynika z niego odpowiednik prawa wielkich liczb dla łańcuchów Markowa.

Jak wyznaczyć rozkład stacjonarny? Musi on być niezmienniczy ze względu na działanie macierzy przejścia (dlaczego?). Jak czytamy w [2], Markow rozwiązał układ równań:

$$(*)\qquad [a,b]\begin{bmatrix}0{,}128 & 0{,}872\\ 0{,}663 & 0{,}337\end{bmatrix}=[a,b],\quad a+b=1,$$

i okazało się, że $[a,b]=[0{,}432;0{,}568]$. Na tej podstawie powziął przypuszczenie, że w tekście „Eugeniusza Oniegina” powinno być 43,2% samogłosek i 56,8% spółgłosek. I rzeczywiście tak jest.

Analogiczny eksperyment przeprowadzony na X księdze „Pana Tadeusza” (Emigracja. Jacek; 31676 liter, 13087 spółgłosek, 18589 samogłosek), daje następujące wyniki (w nawiasach kwadratowych liczby przejść):

	a	b
a	0,152289 [1993]	0,847711 [11094]
b	0,596805 [11094]	0,403195 [7495]

Rozwiązanie układu równań $(*)$ daje $[a,b]=[0{,}413152;0{,}586848]$, co podejrzanie dobrze zgadza się z częstościami samogłosek i spółgłosek. Zabawne, że faktycznie zgodność jest idealna. Nietrudno zobaczyć, że dla konkretnego ciągu znaków tak musi być. Jest to prosta arytmetyka, a nie twierdzenie ergodyczne i prawo wielkich liczb, co zdaje się sugerować komentarz do wyników Markowa [2]. Na rozgrzewkę można wykazać, że liczba przejść z a do b jest taka sama, jak z b do a (jeśli uwzględnimy przejście z ostatniego znaku do pierwszego).

Macierze przejścia dla rosyjskiego i polskiego różnią się, można więc identyfikować za ich pomocą język, co więcej – również autora. Nie ulega wątpliwości, że próbki tekstów Witkiewicza i Lema to najczystsza polszczyzna. Tuwimowi udało się zredukować częstość samogłosek do 36%, a macierz przejścia sugerowałaby raczej ekstremalny język czeski (*strč prst skrz krk*).

A.A. Markow (1856–1922) brał czynny udział w rosyjskim ruchu liberalnym. W tymże 1913 roku, gdy uroczyście obchodzono w Petersburgu 300-lecie panowania Romanowów, Markow zorganizował konkurencyjne obchody 200-lecia odkrycia prawa wielkich liczb przez Bernoulliego. Fakty z życia Markowa podajemy za [2].

Literatura

[1] А. А. Марков, *Пример статистического исследования над текстом „Евгения Онегина” иллюстрирующий связь испытаний в цепь*, Изв. Акад. Наук, СПб. VI серия, т. 7, 1913, No. 3, стр. 153–162.

[1*] A. A. Markow, *An Example of Statistical Analysis of the Text of Eugene Onegin Illustrating the Association of Trials into a Chain*, Bulletin de l'Académie Impériale des Sciences de St. Petersbourg, Ser. 6, vol. 7 (1913), pp. 153–162.

[2] J. Laurie Snell, *Introduction to Probability*, Random House/Birkhäuser, New York 1988.