



### Rozwiązanie zadania M 1298.

Liczba  $a^2 + b^2 + c^2$  jest nieparzysta, a więc postaci  $2p + 1$ . Wybierzmy  $d = p$ . Wtedy  $a^2 + b^2 + c^2 + d^2 = (p + 1)^2$ . Pozostaje więc wykazać, że liczba  $p$  jest nieparzysta.

Liczby  $a, b, c$  są nieparzyste, a więc liczby  $a^2, b^2, c^2$  dają z dzielenia przez 4 resztę 1. Wobec tego  $2p + 1 = a^2 + b^2 + c^2 \equiv 3 \pmod{4}$ , skąd wynika, że liczba  $p$  jest nieparzysta.

odnajdujemy tę prawidłowość. Możemy się zatem pokusić o pytanie, czy rozkład Benforda działa dla każdego zebranych danych liczbowych? Odpowiedź, oczywiście, brzmi: nie!

W eksperymencie III posłużyliśmy się danymi geograficznymi: powierzchnią w  $\text{km}^2$  wszystkich państw świata. Badamy zatem dane, na które ma wpływ wiele czynników. Powierzchnia poszczególnych państw jest bardzo zróżnicowana – od Rosji o powierzchni 17 075 400  $\text{km}^2$  po Watykan – 0,44  $\text{km}^2$ .

Jeżeli za bardzo zawężymy zakres danych, okaże się, że prawo Benforda nie ma dla nich zastosowania. Na przykład, badając długości samochodów osobowych lub wysokość dorosłej żyrafy stwierdzimy, że niewiele z nich zaczyna się od cyfry 1. Wynika to z faktu, iż wartości tych danych są silnie ograniczone innymi czynnikami. Mało która żyrafa, zwłaszcza dorosła, mierzy poniżej 2 metrów.

Może warto zatem pamiętać o prawie Benforda, rzucając sześcienną kostką do gry? Niestety, także nie. Każda liczba oczek ma takie samo prawdopodobieństwo wylosowania. Powtarzając wielokrotnie losowanie, uzyskamy rozkład prawdopodobieństwa zbliżony do równomiernego.

W 1995 roku amerykański profesor matematyki z Georgia Institute of Technology, Theodore P. Hill, przedstawił dowód prawa Benforda na łamach magazynu *Statistical Science* w tekście *A statistical derivation of the significant-digit law* [3].

### Tylko ciekawostka?

Prawo Benforda jest samo w sobie bardzo ciekawym zjawiskiem, a w niektórych dziedzinach ma zastosowanie praktyczne. Służy jako narzędzie do sprawdzania poprawności obliczeń, prawdziwości danych statystycznych czy wykrywania oszustw w zeznaniach podatkowych i rozliczeniach finansowych.

Za pomocą prawa Benforda sprawdza się dokładność działania modeli matematycznych opisujących ewolucję danych z różnych dziedzin, na przykład modeli zmian populacji. Dla danych wejściowych spełniających prawo Benforda powinniśmy otrzymać dane wyjściowe, które również tę zależność spełniają. Jeżeli tak nie jest, oznacza to, że zastosowany model (algorytm) zakłócił „naturalny” rozkład danych.

Najpopularniejszym zastosowaniem prawa Benforda jest sprawdzanie poprawności zeznań podatkowych i rozliczeń. Okazuje się, że fałszerze bardzo często wybierają liczby rozpoczynające się od 4, 5 i 6 zamiast od 1, 2 i 3! Stąd, jeśli rozkład częstości występowania cyfr na pierwszych pozycjach nie jest zbliżony do rozkładu Benforda, to sprawdzający powinien zwrócić na to rozliczenie większą uwagę. Z całą pewnością o prawie Benforda nie wiedział skarbnik stanu Arizona, James Nelson, którego fałszerstwa na kwotę bliską 2 mln dolarów zostały wykryte przy zastosowaniu prawa pierwszych cyfr.

### Literatura

- [1] F. Benford, *The law of anomalous numbers*, Proc. Amer. Philos. Soc. 78 (1938), 551–572.
- [2] T.P. Hill, *The first digit phenomenon*, Amer. Scientist 86 (1998), 358–363.
- [3] T.P. Hill, *A statistical derivation of the significant-digit law*, Statist. Sci. 10 (1995), 354–363.
- [4] S. Newcomb, *Note on the frequency of use of the different digits in natural numbers*, Amer. J. Math. 4 (1881), 39–40.
- [5] R.S. Pinkham, *On the distribution of first significant digits*, Ann. Math. Statist. 32 (1961), 1223–1230.

## Słowa pierwsze

Jakub RADOSZEWSKI

W numerze 10/2010 *Delty* pojawił się artykuł Wojciecha Plandowskiego, w którym autor po ciężkich bojach pokazuje rozwiązanie pewnego konkretnego typu równania na słowach. Mogłoby się wydawać: udało się, sprawa skończona. Tymczasem przy okazji w artykule pojawia się definicja i kilka ważnych własności słów pierwotnych, a stąd już tylko mały krok do innej ciekawej rodziny słów, mianowicie do słów pierwszych. To dobry pretekst, by coś o nich opowiedzieć.

Przypomnijmy, że słowo *pierwotne* to takie, które nie jest potęgą ( $u^k$  dla  $k \geq 2$ ) żadnego niepustego słowa. Znamy już kilka własności takich słów, w szczególności to, że każde słowo  $w$  przedstawia się jednoznacznie w postaci  $w = u^k$ , gdzie  $u$  jest pierwotne; w tym artykule będzie nam wygodnie nazwać  $u$  *pierwiastkiem pierwotnym* słowa  $w$ . Dalej, wiemy, że każdy obrót cykliczny słowa pierwotnego jest pierwotny, a dodatkowo wszystkie takie obroty stanowią różne słowa. Wśród tych obrotów jedno słowo jest, w szczególności, najmniejsze *leksykograficznie*,



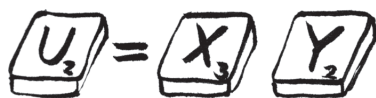
Obrót cykliczny słowa polega na przetrzuceniu dowolnej (w tym zerowej) liczby liter z początku słowa na koniec, np. wszystkimi obrotami cyklicznymi słowa *aba* są: *aba*, *baa* oraz *aab*.

Mówimy, że słowo  $u$  jest nie większe leksykograficznie niż słowo  $v$  (co zapisujemy po prostu jako  $u \leq v$ ), jeśli albo  $u$  jest prefiksem (czyli początkowym fragmentem)  $v$ , albo na pierwszej pozycji, na której  $u$  i  $v$  różnią się, w  $u$  występuje litera mniejsza niż w  $v$ . Oczywiście  $u < v$ , jeśli  $u \leq v$  i  $u \neq v$ . Przykładowo:  $aba < abaab$ ,  $aabab < abaab$ .

1 oznacza słowo puste (czyli słowo zeroliterowe).

Miłośnicy algorytmów tekstowych (takich jak algorytm wyszukiwania wzorca KMP) rozpoznają zapewne w tym równaniu i jego rozwiązaniu związek między prefikso-sufiksami słowa a jego okresami.

Otrzymana sprzeczność dowodzi, że słowa pierwsze są bezokresowe (nie mają żadnego nietrywialnego okresu).



$|u|$  oznacza długość słowa  $u$ , czyli liczbę liter w tym słowie.

Polecamy Czytelnikowi sprawdzenie Twierdzenia 3 na jakimś przykładzie, np. dla słowa  $abaababaabaab$ .

tj. najmniejsze w porządku słownikowym. Każde takie słowo pierwotne najmniejsze w klasie swoich obrotów cyklicznych nazwiemy właśnie słowem *pierwszym* (inna nazwa: słowo Lyndona). Kilka przykładów słów pierwszych:  $b$ ,  $aabab$ ,  $aaaab$  oraz niepierwszych:  $aabaab$ ,  $abaab$ . Zanim wyjaśnimy nieco tajemniczą nazwę rozważanej rodziny słów, przyjrzyjmy się następującej, równoważnej definicji słów pierwszych:

**Twierdzenie 1.** *Niepuste słowo  $u$  jest pierwsze wtedy i tylko wtedy, gdy każdy właściwy sufiks  $u$  (czyli końcowy fragment różny od  $u$  i od słowa pustego 1) jest leksykograficznie większy niż  $u$ .*

**Dowód.** ( $\Leftarrow$ ) Załóżmy, że słowo  $u$  jest mniejsze leksykograficznie od wszystkich swoich właściwych sufiksów. Słowo  $u$  musi być pierwotne, gdyż w przeciwnym razie mielibyśmy  $u = w^k$  dla  $k \geq 2$  i sufiks  $u$  postaci  $w^{k-1}$  byłby mniejszy leksykograficznie niż  $u$ . Dalej, jeśli mamy  $u = xy$  i  $x, y \neq 1$ , to  $u < y$ , a skoro  $y$  jest krótsze niż  $u$ , to  $u$  nie może być prefiksem  $y$  i mamy  $u < yx$ . To pokazuje, że dowolny obrót cykliczny  $u$  jest większy leksykograficznie niż  $u$ , czyli rzeczywiście  $u$  jest pierwsze.

( $\Rightarrow$ ) Załóżmy, że  $u$  jest pierwsze, i niech  $u = xy$  dla  $x, y \neq 1$ ; chcemy wykazać, że  $u < y$ . Dowód przeprowadzimy nie wprost. Gdyby zachodziło  $y \leq u$  i  $y$  nie byłoby prefiksem  $u$ , to mielibyśmy także  $yx < u$  i  $u$  nie byłoby pierwsze. Skoro tak, to  $y$  jest prefiksem  $u$ , ale przecież także i jego sufiksem, czyli  $u = xy = yz$  dla pewnego  $z$ . Czytelnikowi zaznajomionemu z artykułem Wojciecha Plandowskiego taka równość może coś przypominać: otóż jest to dokładnie „główne” równanie, jakie zostało w nim rozwiązane! Korzystając z gotowego wyniku, mamy, że  $x = (pq)^i$ ,  $z = (qp)^i$  oraz  $y = (pq)^j p$  dla pewnych słów  $p$  oraz  $q \neq 1$ , czyli  $u = (pq)^k p$  dla  $k = i + j \geq 1$ . Dodajmy, że słowo  $p$  także musi być niepuste, gdyż w przeciwnym przypadku  $j \geq 1$  (jako że  $y \neq 1$ ) i  $u = q^k$  dla  $k \geq 2$ , czyli  $u$  nie byłoby pierwotne ani, tym bardziej, pierwsze.

No to teraz pójdzie już z górki. Słowo  $u$  jest pierwsze, więc rozważając jego dwa wybrane obroty cykliczne, otrzymujemy:  $u = (pq)^k p = p(qp)^k < (qp)^k p$  oraz  $u = (pq)^k p < p(pq)^k$ . Na mocy pierwszej nierówności mamy  $pq < qp$ , natomiast z drugiej, po usunięciu początkowego  $p$  otrzymujemy  $qp < pq$ . Te dwie nierówności dają nam oczekiwaną sprzeczność.  $\Delta$

Nie koniec na tym – istnieje jeszcze inna, równoważna, choć nieco bardziej egzotyczna definicja słów pierwszych. Jej uzasadnienie pozostawiamy Czytelnikowi, nam i tak będzie ona potrzebna jedynie jako własność tej rodziny słów.

**Twierdzenie 2.** *Słowo  $u$  jest słowem pierwszym wtedy i tylko wtedy, gdy jest jednoliterowe albo jest postaci  $u = xy$ , przy czym  $x < y$  i  $x, y$  są słowami pierwszymi.*

Po tym wstępie możemy już wreszcie zdradzić, skąd wzięła się nazwa rozważanej rodziny słów. Chodzi mianowicie o następującą analogię do liczb pierwszych: każdą liczbę naturalną można przedstawić jednoznacznie (oczywiście z dokładnością do kolejności) jako iloczyn liczb pierwszych, a każde słowo w pewnym sensie jednoznacznie jako sklejenie słów pierwszych.

**Twierdzenie 3.** *Dowolne słowo  $u$  można przedstawić jednoznacznie jako sklejenie pewnej liczby słów pierwszych  $u = l_1 l_2 \dots l_k$  dla  $l_1 \geq l_2 \geq \dots \geq l_k$ .*

**Dowód.** Czy każde słowo można w ogóle przedstawić jako sklejenie (jakichkolwiek) słów pierwszych? Można, wystarczy podzielić je na pojedyncze litery, które, jak by nie patrzeć, są słowami pierwszymi. Teraz, dopóki w bieżącym rozkładzie występują dwa kolejne słowa pierwsze  $l_i, l_{i+1}$ , takie że  $l_i < l_{i+1}$ , dopóty możemy takie słowa sklejać w jedno słowo pierwsze  $l_i l_{i+1}$  na mocy Twierdzenia 2. Powtarzając tę operację do skutku, znajdziemy jakieś przedstawienie  $u$  w postaci sklejenia nierosnącego ciągu słów pierwszych.

No dobrze, a czemu takie sklejenie jest tylko jedno? Załóżmy, że byłyby dwa różne,  $u = l_1 l_2 \dots l_k = l'_1 l'_2 \dots l'_m$ . Możemy założyć, że  $l_1 \neq l'_1$ , w przeciwnym przypadku możemy tę parę słów wykreślić z rozkładu. Niech więc, bez straty ogólności, będzie  $|l_1| > |l'_1|$ . Wówczas musi istnieć rozkład:  $l_1 = l'_1 l'_2 \dots l'_i \alpha$ , przy czym  $\alpha$  jest *niepustym* prefiksem  $l'_{i+1}$ . No i teraz możemy zapisać ciąg nierówności, który prowadzi do sprzeczności, a którego uzasadnienie pozostawiamy Czytelnikowi (jedynie nietrywialne miejsce polega na zastosowaniu Twierdzenia 1):

$$l'_1 < l_1 < \alpha \leq l'_{i+1} \leq l'_1. \quad \Delta$$

To w takim razie już „wszystko” jasne, poznaliśmy jakieś mniej lub bardziej interesujące własności słów pierwszych, w tym wyjaśniliśmy ich nazwę, można by właściwie zakończyć niniejszy artykuł. Niewykluczone jednak, że Czytelnikowi, który dobrnął do tego miejsca, pozostał po lekturze pewien niedosyt. Chciałoby się, żeby te słowa pierwsze miały jakieś *naprawdę* fajne własności (tak jak liczby pierwsze), a może nawet żeby miały jakiś

związek z teorią liczb... Autor niniejszego artykułu przygotował coś specjalnego, czym ma nadzieję uraczyć wciąż nienasyconych Czytelników.

Słowo cykliczne to ciąg liter bez wyróżnionego początku i końca – aby przekształcić zwykle słowo w cykliczne, wystarczy je sobie wyobrazić jako zapisane wzdłuż okręgu.

W *Delcie* 3/2009 pojawił się artykuł o ciągach de Bruijna. Przypomnijmy ich definicję w wersji ogólniejszej: *ciągami de Bruijna rzędu  $n$*  nazywamy słowo cykliczne długości  $k^n$  złożone z „liter” należących do zbioru  $\{0, 1, \dots, k-1\}$ , w którym każde  $n$ -literowe słowo występuje jako podsłowo (tzn. spójny fragment) dokładnie raz. Najłatwiej wyobrazić sobie ciągi de Bruijna dla  $k=2$ , przykładem takiego ciągu rzędu 4 jest 0000100110101111. Jak nietrudno się już teraz domyślić, asem w rękawie autora artykułu jest pewien związek między ciągami de Bruijna a słowami pierwszymi. A jest to związek zaiste niespodziewany – otóż, *sklejenie wszystkich słów pierwszych o długościach dzielących  $n$  w kolejności leksykograficznej daje ciąg de Bruijna rzędu  $n$ , a do tego najmniejszy leksykograficznie dla danych parametrów  $n, k$* . Kilka przykładów:

$$n=4, k=2: 0 \cdot 0001 \cdot 0011 \cdot 01 \cdot 0111 \cdot 1,$$

$$n=3, k=3: 0 \cdot 001 \cdot 002 \cdot 011 \cdot 012 \cdot 021 \cdot 022 \cdot 1 \cdot 112 \cdot 122 \cdot 2.$$

Po napatrzeniu się na te i inne przykłady Czytelnik zada sobie zapewne pytanie, skąd się ten związek bierze. No i tu kolejna niespodzianka: autorowi nie jest znany żaden naprawdę prosty dowód faktu, że opisana konstrukcja faktycznie daje ciąg de Bruijna. Podobnie (a właściwie jeszcze gorzej) jest z uzasadnieniem faktu, że jest to zawsze najmniejszy leksykograficznie ciąg de Bruijna dla zadanych parametrów  $n$  oraz  $k$ . Być może któremuś z Czytelników uda się wymyślić jakiś elegancki dowód tych faktów, a tymczasem spróbujmy odpowiedzieć choćby na najprostsze z tego typu pytań: czy w ogóle *długość* sklejonego ciągu słów pierwszych zgadza się z długością ciągu de Bruijna?

Oznaczmy przez  $L_d(k)$  liczbę słów pierwszych długości  $d$ , dla ustalonego  $k$ . Wówczas długość sklejonego ciągu dla danego  $n$  to  $S(n, k) = \sum_{d|n} d \cdot L_d(k)$ . Zauważmy, że  $d \cdot L_d(k)$  to tak naprawdę liczba słów *pierwotnych* długości  $d$ . Faktycznie, wszystkie  $d$  obrotów cyklicznych danego  $d$ -literowego słowa pierwszego stanowi różne słowa, z czego wszystkie pierwotne, obroty cykliczne różnych słów pierwszych dają, oczywiście, różne słowa pierwotne, a każde słowo pierwotne można obrócić cyklicznie tak, by stało się pierwsze. To oznacza, że  $S(n, k)$  jest równe łącznej liczbie słów pierwotnych o długościach dzielących  $n$ . A ile jest takich słów? Otóż jest ich dokładnie tyle, ile wszystkich  $n$ -literowych słów nad alfabetem  $k$ -literowym, czyli dokładnie  $k^n$ . A to dlatego, że dowolnemu słowu  $n$ -literowemu możemy przyporządkować jednoznacznie jego pierwiastek pierwotny, który jest słowem pierwotnym o długości dzielącej  $n$ . Z drugiej strony, dowolnemu słowu pierwotnemu o długości  $d$ , będącej dzielnikiem  $n$ , możemy przyporządkować (znow jednoznacznie) jego potęgę o wykładniku  $n/d$ , która jest słowem długości  $n$ . Podsumowując,  $S(n, k) = k^n$ , czyli długości rzeczywiście się zgadzają.

To jednak nie koniec, gdyż z uzyskanej równości możemy wyciągnąć pewne interesujące wnioski. Jednym z nich jest wyznaczenie wzoru na liczbę słów pierwszych długości  $n$ , czyli  $L_n(k)$ . Do tego celu możemy się posłużyć tzw. *wzorem inwersyjnym Möbiusa*, który dla dwóch funkcji  $f, g$ , przyjmujących argumenty całkowite dodatnie, wygląda następująco:

$$f(n) = \sum_{d|n} g(d) \Leftrightarrow g(n) = \sum_{d|n} \mu(d) f(n/d),$$

gdzie  $\mu$  to funkcja Möbiusa zdefiniowana dla liczb naturalnych rekurencyjnie jako:

$$\mu(1) = 1, \quad \sum_{d|n} \mu(d) = 0 \quad \text{dla } n > 1.$$

Stosując wzór inwersyjny do  $f(n) = S(n, k) = k^n$  i  $g(n) = n \cdot L_n(k)$ , otrzymujemy:

$$L_n(k) = \frac{1}{n} \sum_{d|n} \mu(d) k^{n/d}.$$

Na koniec przyjrzyjmy się, jak wygląda wzór  $S(p, k) = k^n$  dla  $p$  będącego liczbą pierwszą. Wówczas suma  $S(p, k)$  ma tylko dwa składniki,  $L_1(k)$  oraz  $p \cdot L_p(k)$ . Jeśli zauważymy, że pierwszy z tych składników jest równy po prostu  $k$ , możemy nasz wzór przekształcić do takiej postaci:

$$L_p(k) = \frac{1}{p}(k^p - k).$$

Ponieważ liczba słów pierwszych nie może być niecałkowita, więc otrzymujemy, że dla dowolnej liczby pierwszej  $p$  i liczby całkowitej dodatniej  $k$  zachodzi  $p | k^p - k$ , który to fakt znany jest powszechnie jako *Małe Twierdzenie Fermata*. Możemy więc stwierdzić, że przy okazji udało nam się udowodnić to twierdzenie, używając do tego jedynie faktów związanych ze słowami.

Poprawność tego wzoru można całkiem nietrudno sprawdzić, podstawiając lewą stronę w miejsce  $f(n/d)$  i, odwrotnie, prawą w miejsce  $g(d)$ .

Kilka pierwszych wartości:  $\mu(1) = 1$ ,  $\mu(2) = -1$ ,  $\mu(3) = -1$ ,  $\mu(4) = 0$ ,  $\mu(5) = -1$ ,  $\mu(6) = 1$ . W ogólności można udowodnić, że  $\mu(n) = 0$ , jeśli  $n$  jest podzielne przez kwadrat liczby pierwszej, a w przeciwnym przypadku  $\mu(n) = (-1)^k$ , przy czym  $k$  to liczba różnych dzielników pierwszych  $n$ .