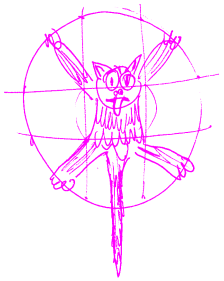
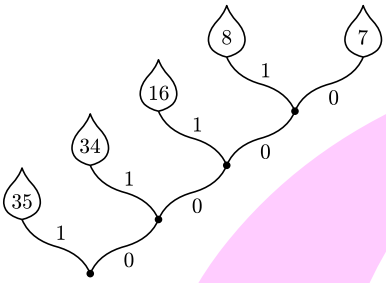
Kod Huffmana

Rafał SZTENCEL [*] (*scripsit et pinxit*)



Rozwiążemy problem z poprzedniego odcinka. Przypuśćmy, że w pewnym miasteczku występuje tylko pięć imion żeńskich: Agnieszka, Barbara, Celina, Dorota i Elżbieta, z częstościami odpowiednio 35%, 34%, 16%, 8% i 7%. Mamy ustalić imię losowo wybranej kobiety, zadając pytania, na które otrzymujemy odpowiedź „tak” lub „nie”. Jak pytać, by średnia liczba pytań była najmniejsza?



Imię	kod	częstość
Agnieszka	1	0,35
Barbara	01	0,34
Celina	001	0,16
Dorota	0001	0,08
Elżbieta	0000	0,07

Konstruujemy drzewko. Najpierw łączymy dwa „liście” o najmniejszej wadze (D i E), które będą w efekcie wyrastać ze wspólnego węzła. Otrzymujemy nową listę wag: 35, 34, 16, 15. Następnie łączymy dwie najmniejsze wagi i powtarzamy procedurę, dopóki się da. Rezultat widać na rysunku, drzewko sugeruje (mamy nadzieję) sposób zadawania pytań. Jeśli odpowiedź „tak” oznaczymy cyfrą 1, a odpowiedź „nie” cyfrą 0, to ciągi odpowiedzi (kody) identyfikujące wszystkie imiona będą wyglądać jak w tabeli na marginesie.

Do identyfikacji potrzebujemy średnio

$$q = 1 \cdot 0{,}35 + 2 \cdot 0{,}34 + 3 \cdot 0{,}16 + 4 \cdot 0{,}08 + 4 \cdot 0{,}07 = 2{,}11$$

pytania na osobę, podczas gdy entropia H odpowiedniego rozkładu prawdopodobieństwa jest równa 2,04. Przypominamy, że

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i,$$

gdzie w naszym przypadku $p_1 = 0{,}35, \ldots, p_5 = 0{,}07$. Widzimy, że spełniona jest nierówność z poprzedniego odcinka:

$$H \leqslant q < H + 1.$$

Przedstawiona metoda pochodzi od Huffmana. Można udowodnić, że średniej liczby pytań nie da się zmniejszyć.

Wyobraźmy sobie, że mamy tekst, w którym występują wyłącznie litery A, B, C, D i E z podanymi wcześniej częstościami. Gdybyśmy chcieli go zakodować za pomocą ciągu cyfr dwójkowych, czyli zer i jedynek, to, na przykład, słowo BABA przybrałoby postać 011011. Kod o stałej długości wymagałby trzech cyfr dwójkowych (bitów) na znak. Kod Huffmana wymaga tylko 2,11 bita na znak. Rzut oka do tabeli pozwala stwierdzić, że więcej oszczędzamy na często występujących symbolach A i B o krótkich kodach, niż tracimy na rzadkich symbolach D i E o długich kodach.

Kod Huffmana jest w pewnym sensie optymalny, ale w jakim? Jeśli kodujemy pojedyncze znaki, to liczby bitów na znak nie da się już zmniejszyć. Stąd oczywiste zastosowania w programach archiwizujących pliki komputerowe. Z drugiej strony kodowanie tekstu BA...BA (powtórzone milion razy) metodą Huffmana jest, delikatnie mówiąc, nieoptymalne. W języku naturalnym takie teksty zdarzają się rzadko, choć, na przykład, w zapisie cyfrowym fotografii analogiczne ciągi nie są niczym dziwnym. Skoro jednak mowa o języku naturalnym, to doskonale wiadomo, że żaden język nie przypomina ciągu symboli generowanych niezależnie z ustalonymi częstościami. Przeciwnie, znając część wyrazu można na ogół odtworzyć całość.

Nad oszacowaniem faktycznej entropii tekstu w bitach na znak zastanawiał się już twórca teorii informacji, Claude Shannon. Jeśli wziąć pod uwagę tylko częstości występowania liter w języku angielskim, to wynosi ona 2,14 bita na znak. Shannon zaproponował zadziwiająco prostą i pomysłową metodę szacowania faktycznej entropii tekstu: odczytywał kolejne litery, a druga osoba zgadywała dalszy ciąg. Okazało się, że entropia zawiera się w granicach 0,6–1,3 bita na znak. Jak wykorzystać ten fakt przy kodowaniu? Chyba będziemy musieli zastanowić się nad tym w przyszłości.

Czasem znając pierwszą literę tekstu, można odtworzyć całkiem pokaźny fragment, jak w przypadku artykułu z pierwszej strony „Trybuny Ludu” z lat siedemdziesiątych ubiegłego stulecia: „**P**ierwszy sekretarz KC PZPR Edward Gierek przebywał wczoraj z gospodarską wizytą w Zakładach Mięsnych im. Feliksa Dzierżyńskiego w Mławie...” – itd. w tym stylu. Przykłady współczesne tekstów o zawartości informacyjnej 0 bitów na znak każdy dośpiewa sobie sam.

Podziękowania. Kącik „Omega”ukazuje się już od roku. Docierają do mnie komentarze Czytelników, na ogół życzliwe, za które pragnąłbym w tym miejscu podziękować. Nie jest łatwo wyobrazić sobie modelowego czytelnika. Dlatego też większość tekstów przed publikacją przeczytała Agnieszka Strużyńska, studentka Wydziału Nauk Ekonomicznych UW. Co więcej, kilka tematów, w tym bieżący, ma bezpośredni związek z konsultacjami, jakich jej udzielałem. Ten przykład pożytków z działalności dydaktycznej zasługuje na wzmiankę, a osoba, która podjęła się krytycznej lektury – na podziękowanie.

[*] Instytut Matematyki, Uniwersytet Warszawski