Czy w Unii Europejskiej mówiono po polsku?

J. MIRÓ, F. ROSSELLÓ, Hiszpania

Niejednego z odwiedzających nas polskich przyjaciół zaskakują graffiti z tekstami w rodzaju *Polacos, esto es España.* (*Polacy, tu jest Hiszpania.*) A kiedy wyjaśniamy im, że ludzie mówiący po katalońsku nierzadko spotykają się z ostrą reprymendą *¡Usted no me hable a mí en polaco!* (*Proszę do mnie nie mówić po polsku!*), przestają cokolwiek rozumieć.

Chodzi o to, że na ogromnym obszarze nad brzegiem Morza Śródziemnego, ogarniającym wschód Hiszpanii i południe Francji, mówimy nie tylko w języku hiszpańskim lub francuskim, ale także w innym języku romańskim, jakim jest kataloński. Często w historii było to powodem pogardliwych żartów ze strony mieszkańców centralnej Hiszpanii, a jednym z typowych żartów jest oskarżenie nas o mówienie... po polsku. Widać ktoś kiedyś uznał język kataloński, znacznie bogatszy w dźwięki od hiszpańskiego, za coś równie dziwnego i trudnego jak język polski.

Można się zastanawiać, czy istnieją podstawy naukowe do tego porównania. Czy kataloński jest rzeczywiście podobny do polskiego? Oczywiście nie, ale można pytać dalej, bardziej realistycznie: czy kataloński jest podobny do polskiego **bardziej** niż hiszpański? Pytanie to może okazać się interesujące dla Czytelników *Delty* nie tylko z powodu ich ojczystego języka, lecz także dlatego, że pozwala wyjaśnić pewne matematyczne narzędzie, bardzo modne w dzisiejszych czasach genomiki.

Poszukując na nie odpowiedzi, porównaliśmy zbiór polskich słów z ich odpowiednikami w językach katalońskim i hiszpańskim. A że mamy do czynienia ze słowami, użyliśmy pojęcia **odległości edycyjnej**, wprowadzonej przez V.I. Levenshteina w 1966 roku. Mając dwa słowa, staramy się przekształcić jedno w drugie za pomocą operacji edycyjnych: wstawienia lub usunięcia litery, lub zastąpienia jednej litery przez inną. Każdej takiej operacji przypisujemy pewien koszt i szukamy przekształcenia, dla którego koszt jest najmniejszy. Ów najmniejszy koszt jest odległością edycyjną danych dwóch słów.

Takie ciągi operacji edycyjnych reprezentuje się zazwyczaj poprzez uliniowienia. **Uliniowieniem** dwóch słów $X = x_1x_2\ldots x_n$ i $Y = y_1y_2\ldots y_m$ jest dwuwierszowa macierz, taka że w górnym wierszu występują kolejne litery X poprzetykane spacjami, w dolnym wierszu zaś litery słowa Y, także, być może, ze spacjami; żądamy jedynie, by żadna kolumna nie składała się wyłącznie ze spacji. Wystąpienie w jednej kolumnie litery x_i ze słowa X i litery y_j ze słowa Y oznacza zastąpienie x_i przez y_j; kolumna zawierająca w górnym wierszu x_i i w dolnym spację reprezentuje usunięcie x_i, natomiast kolumna z literą y_j w dolnym wierszu i spacją w górnym oznacza wstawienie litery y_j.

Levenshtein nie zaproponował żadnego algorytmu obliczającego odległość edycyjną, lecz lukę tę prędko wypełnili inni autorzy. Wszystkie opisane przez nich algorytmy są dość podobne. Jeden z pierwszych, służący do porównywania białek (patrz *Delta* 10/2002), podali S.B. Needleman i C.D. Wunsch w 1970 roku. Była to jedna z pierwszych prac z biologii obliczeniowej, teoretycznej gałęzi bioinformatyki. Do porównania słów użyliśmy uproszczonej wersji tego algorytmu, opracowanej w 1974 roku przez P.H. Sellersa dla szczególnego przypadku, gdy koszt wstawienia i usunięcia litery jest stały.

Rozpatrujemy słowa nad pewnym ustalonym alfabetem Σ.
Określmy **macierz kosztu**

$$(\sigma(a,b))_{a,b\in\Sigma},$$

w której $\sigma(a,b)$ reprezentuje koszt zamiany litery a na b. Zakładamy, że $\sigma(a,b) = \sigma(b,a)$ oraz $\sigma(a,b) = 0$, gdy $a = b$. Przyjmujemy ponadto stały koszt γ wstawienia lub usunięcia litery. Tak więc koszt uliniowienia dwóch słów

$X = x_1 \dots x_n$ i $Y = y_1 \dots y_m$ powstaje przez dodanie wszystkich wartości $\sigma(x_i, y_j)$ dla każdej kolumny uliniowienia, która łączy x_i z y_j, oraz wartości γ pomnożonej przez liczbę kolumn zawierających spację. Uliniowienie X i Y jest **optymalne**, gdy koszt jest najmniejszy (i to jest właśnie odległość edycyjna między X i Y).

Dla danych dwóch słów $X = x_1 \dots x_n$ i $Y = y_1 \dots y_m$ nad alfabetem Σ, algorytm Needlemana–Wunscha–Sellersa oblicza rekurencyjnie macierz

$$(F(i,j))_{\substack{i=0,\dots,n \\ j=0,\dots,m}}$$

taką że każdy wyraz $F(i,j)$ jest kosztem optymalnego uliniowienia przedrostków $x_1 \dots x_i$ oraz $y_1 \dots y_j$ słów X i Y, odpowiednio (gdy $i = 0$ lub $j = 0$, przedrostkami są *słowa puste*). W ten sposób wyraz $F(n,m)$ jest odległością edycyjną X i Y.

Koszt uliniowienia uzyskuje się przez dodawanie kolejnych wartości, zatem optymalne uliniowienie dwóch słów można zbudować na podstawie uliniowienia ich przedrostków. Wynika stąd, że jeśli optymalne uliniowienie przedrostków $x_1 \dots x_i$ i $y_1 \dots y_j$ łączy x_i z y_j, to indukuje ono optymalne uliniowienie dla $x_1 \dots x_{i-1}$ i $y_1 \dots y_{j-1}$ i wtedy

$$F(i,j) = F(i-1, j-1) + \sigma(x_i, y_j).$$

Podobnie, jeśli optymalne uliniowienie $x_1 \dots x_i$ i $y_1 \dots y_j$ łączy x_i ze spacją, to indukuje ono optymalne uliniowienie $x_1 \dots x_{i-1}$ i $y_1 \dots y_j$, a stąd

$$F(i,j) = F(i-1, j) + \gamma.$$

Symetrycznie, $F(i,j) = F(i, j-1) + \gamma$. Tak więc rzeczywista wartość $F(i,j)$ będzie najmniejszą z tych trzech możliwych wartości. Wyznacza to procedurę rekurencyjną, w której $F(i,j)$ oblicza się na podstawie znajomości wcześniej obliczonych wartości

$$F(i-1, j-1), \quad F(i-1, j) \quad \text{i} \quad F(i, j-1).$$

Z drugiej strony, koszty $F(i,0)$ i $F(0,i)$ w optymalnym uliniowieniu $x_1 \dots x_i$ oraz $y_1 \dots y_i$ ze słowem pustym, odpowiednio, muszą być równe $i\gamma$.

Powyższe rozważania wykazują w skrócie poprawność następującego algorytmu obliczającego odległość edycyjną dwóch słów:

`Input` $x_1 \dots x_n$, $y_1 \dots y_m$
$F(0,0) = 0$
`For` $i = 0, \dots, n$ $\quad F(i,0) = i\gamma$
`For` $j = 0, \dots, m$ $\quad F(0,j) = j\gamma$
`For` $i = 1, \dots, n$
  `For` $j = 0, \dots, m$
    $F(i,j) = \texttt{min}\{F(i-1, j-1) + \sigma(x_i,\ y_j), F(i-1, j) + \gamma,\ F(i, j-1) + \gamma\}$
`Output` $F(n,m)$

Jeśli ponadto zapamiętujemy, która z trzech możliwości została wykorzystana w obliczeniu każdego wyrazu $F(i,j)$ dla $i, j > 0$, możemy z łatwością uzyskać optymalne uliniowienie słów wejściowych.

Do naszych celów przyjęliśmy, że alfabet Σ składa się ze wszystkich liter używanych w języku polskim, katalońskim lub hiszpańskim, łącznie z tymi, które opatrzone są znakami diakrytycznymi. Taki alfabet składa się z 46 elementów. Dla uproszczenia zdefiniowaliśmy macierz kosztów w sposób następujący:

$$\sigma(a,b) = \begin{cases} 0 & \text{gdy } a = b \\ 1 & \text{gdy } a \text{ i } b \text{ są różnymi literami, lecz brzmią podobnie} \\ 2 & \text{w pozostałych przypadkach.} \end{cases}$$

Pełną wersję macierzy można znaleźć w sieci na stronie `bioinfo.uib.es/~joemiro/polcat`. Przyjęliśmy także $\gamma = 3$.

Jeśli zastosujemy algorytm do obliczenia odległości między polskim słowem „angielski” a katalońskim odpowiednikiem „anglès”, otrzymamy

następującą macierz (wszystkie litery w tym przykładzie są albo równe, albo różne z różnym brzmieniem):

			a	n	g	i	e	l	s	k	i
		0	1	2	3	4	5	6	7	8	9
	0	0	3	6	9	12	15	18	21	24	27
a	1	3	0	3	6	9	12	15	18	21	24
n	2	6	3	0	3	6	9	12	15	18	21
g	3	9	6	3	0	3	6	9	12	15	18
l	4	12	9	6	3	2	5	6	9	12	15
è	5	15	12	9	6	5	2	5	8	11	14
s	3	18	15	12	9	8	5	5	5	8	11

Widać stąd, że zgodnie z przyjętą macierzą kosztów, odległość między tymi słowami jest równa 11. Podkreśliliśmy w macierzy ścieżkę, która prowadzi do ostatecznego wyniku: 11 w dolnym prawym rogu bierze się z 8 w sąsiedniej kolumnie, ta wartość pochodzi z kolei od 5 z lewej strony itd. W ten sposób otrzymujemy optymalne uliniowienie (spacje oznaczamy symbolem „ _ ”):

a n g i e l s k i
a n g l è _ s _ _

Zastosowaliśmy ten algorytm do 427 polskich słów, losowo wybranych ze słownika polsko-angielskiego; szczegółowe obliczenia można zobaczyć na wymienionej wcześniej stronie web. W rezultacie okazało się, że średnia odległość między słowami polskimi a katalońskimi jest równa 12,43, podczas gdy średnia odległość między słowami polskimi a hiszpańskimi jest równa 12,31. Jak widać, wyniki dość zbliżone (co zapewne nie jest niespodzianką), ale słowa hiszpańskie są nieco bliższe polskim niż katalońskie. Niektórzy nasi hiszpanojęzyczni koledzy byliby tym mocno zdziwieni! Oczywiście próbka nie była duża ani reprezentatywna, a macierz kosztów można z pewnością ulepszyć. Zachęcamy do dalszych badań w tym kierunku.

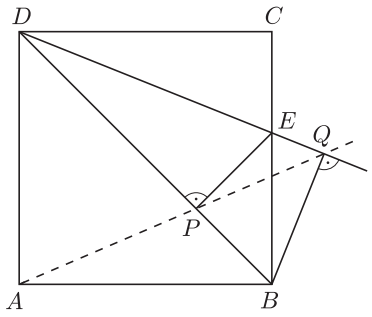
Musimy na koniec wspomnieć, że mimo wszystko nasze badania wspierał rząd Hiszpanii w ramach projektu DGES BFM2000-1113-C02-01 MOBIOCO.

Tłumaczył Wiktor BARTOL

Zadania

Redaguje Waldemar POMPE

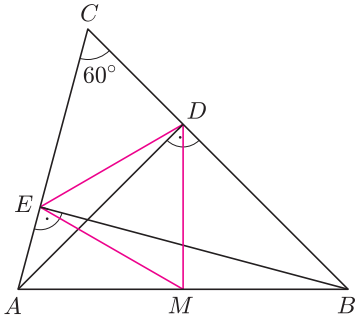
M 1060. Punkt E leży na boku BC kwadratu $ABCD$ (rys. 1). Punkty P i Q są rzutami prostokątnymi odpowiednio punktów E i B na proste BD i DE. Dowieść, że punkty A, P, Q leżą na jednej prostej.
Rozwiązanie na str. 3



Rys. 1

M 1061. Dane są liczby rzeczywiste $a \geqslant 1$, $b \geqslant 2$, $c \geqslant 3$. Wykazać, że $abc \geqslant a + b + c$.
Rozwiązanie na str. 16

M 1062. Dany jest trójkąt ostrokątny ABC, w którym $\sphericalangle ACB = 60^\circ$ (rys. 2). Punkty D i E są rzutami prostokątnymi odpowiednio punktów A i B na proste BC i AC. Punkt M jest środkiem boku AB. Wykazać, że trójkąt DEM jest równoboczny.
Rozwiązanie na str. 16



Rys. 2

Redaguje Mikołaj KORZYŃSKI

F 619. W okolicach Rowu Mariańskiego spuszczamy do wody obciążony gumowy balonik napełniony wodą oraz obciążoną, wypełnioną powietrzem stalową puszkę. Które z nich wytrzyma większe ciśnienie w głębi Oceanu?
Rozwiązanie na str. 2

F 620. Samolot lecąc poziomo osiąga maksymalną prędkość $v_p = 900$ km/h. Czy jest on w stanie przekroczyć prędkość dźwięku w powietrzu $m = 1200$ km/h pikując w dół, jeśli wiemy, że nie jest on w stanie startować pionowo do góry? Zakładamy, że opory ruchu są cały czas proporcjonalne do kwadratu prędkości.
Rozwiązanie na str. 12