Задача 3.

Съпоставителните изследвания в лингвистиката често се основават на *паралелни корпуси* — версии на едни и същи текстове на два (или повече) езика, които са *подравнени*, т. е. определено е и е отбелязано кои части отговарят една на друга. Подравняването обикновено се прави с помощта на програми, които обработват бързо големи количества текст, но нерядко допускат грешки, защото се ориентират само по формални критерии.

Ще разгледаме една такава програма, която приема два текста, разделени на редове, и разполага редовете един до друг, като дава на всяка двойка оценка, която е толкова по-висока, колкото по-уверена е програмата в съответствието. При това може някой ред от единия текст да обединява два или повече от другия или да не отговаря на нищо; не може обаче да се пренареждат редове. Програмата търси такова съотнасяне на редовете, при което средната оценка за целия текст да е колкото може по-висока.

Обикновено на програмата се дава абзац или по-дълъг текст, разделен на изречения, за да реши тя кои изречения отговарят едно на друго, но ние ще ѝ даваме по една дума на ред. Например да подравним българските изречения *В заведението влезе Марта с пазарска чанта в ръка* (П. Вежинов, „Везни“) и *Всичката тази плячка Авакум мигновено напъха във вътрешните джобове на сакото си* (А. Гуляшки, „История с кучета“) с украинските им преводи — съответно *До корчми зайшла Марта з господарською сумкою в руці* и *Всю цю здобич Авакум миттю порозпихав по внутрішніх кишенях піджака*:

В	*До*	0.1
заведението	*корчми*	0.0692308
влезе	*зайшла*	0.245455
Марта	*Марта*	1.8
с	*з*	0.3
пазарска	*господарською*	0.123529
чанта	*сумкою*	0.245455
в	*в*	1.8
ръка	*руці*	0.3

Всичката	*Всю*	-0.128571
тази	*цю*	0.06
плячка	*здобич*	0.3
Авакум	*Авакум*	1.8
мигновено	*миттю*	0.0818182
напъха	*порозпихав*	0.115385
във	*по*	0.18
вътрешните	*внутрішніх*	0.3
джобове	*кишенях*	0.3
на ~~~ сакото	*піджака*	-0.0230769
си		-0.3

Както виждаме, програмата е разбрала, че двете български думи *на сакото* отговарят на една украинска *пиджака* (знакът ~~~ показва, че са обединени два реда от текста), а местоимението *си* няма съответствие.

а) В подравняването на изречението *И през това време Лоренцо наглеждаше* (А. Гуляшки, „Откраднато на „Даная“) и превода му *І в цей час Лоренцо пильнував* са пропуснати оценките на съответствията:

И	*І*	
през	*в*	
това	*цей*	
време	*час*	
Лоренцо	*Лоренцо*	
наглеждаше	*пильнував*	

Във възходящ ред те са: {-0.3, 0.128571, 0.214286, 0.268421, 0.3, 1.8}.

Разположете ги на правилните места.

Решение:

Най-висока оценка (1.8) получава двойката, в която двете думи са еднакви. Оценките на останалите двойки зависят само от близостта на дължините на думите. От тях най-висока оценка (0.3) получава двойката, в която дължините са равни. От данните се вижда, че оценката е по-ниска, ако разликата в брой букви е по-голяма (двойката *това*:*цей* е по-добра както от *време*:*час*, така и от *през*:*в*), а при еднаква разлика тя тежи по-малко, ако думите са по-дълги (затова *наглеждаше*:*пильнував* е по-добра от *това*:*цей*). Остава да разпределим двете най-ниски оценки между *време*:*час* и *през*:*в*, а това можем да направим, като забележим, че *тази*:*цю* има оценка 0.06, а тази двойка е по-добра от *през*:*в*.

И	*І*	0.3
през	*в*	-0.3
това	*цей*	0.214286
време	*час*	0.128571
Лоренцо	*Лоренцо*	1.8
наглеждаше	*пильнував*	0.268421

б) Всъщност нищо не пречи вместо изречения, разделени на думи, да дадем на програмата просто списъци от думи. Разбира се, това няма да е съвсем коректно, ако знаем, че те отговарят една на друга. Но нека да видим какво става.

Ако двата „текста“ съдържат имената на месеците от годината съответно на български (1) и на украински (2), програмата предлага като най-добро следното подравняване (3):

(1) (2) (3)

януари	*січень*	*януари*	*січень*	0.3
февруари	*лютий*		*лютий*	-0.3
март	*березень*	*февруари*	*березень*	0.3
април	*квітень*	*март*	*квітень*	0.1
май	*травень*	*април*	*травень*	0.190909
юни	*червень*	*май ~~~ юни*	*червень*	0
юли	*липень*	*юли*	*липень*	0.0428571
август	*серпень*	*август*	*серпень*	0.253846
септември	*вересень*	*септември*	*вересень*	0.264706
октомври	*жовтень*	*октомври*	*жовтень*	0.26
ноември	*листопад*	*ноември*	*листопад*	0.26
декември	*грудень*	*декември*	*грудень*	0.26

Както виждате, програмата е сметнала за вероятно, че украинското *лютий* няма българско съответствие, а *червень* отговаря на *май ~~~ юни*. Обединяването на два реда е намалило оценката за реда (в случая до 0), а полето без съответствие получава даже отрицателна оценка, но сумата за петте реда (0.290909) все пак е по-висока, отколкото ако не беше станало това изместване (тогава щеше да е 0.2748921):

февруари	*лютий*	0.136364
март	*березень*	0.0333333
април	*квітень*	0.190909
май	*травень*	-0.0428571
юни	*червень*	-0.0428571

Ако решим годината да започва от април, програмата предлага друго подравняване, като оставя този път без българско съответствие украинското *березень*:

(1) (2) (3)

април
май
юни
юли
август
септември
октомври
ноември
декември
януари
февруари
март

квітень
травень
червень
липень
серпень
вересень
жовтень
листопад
грудень
січень
лютий
березень

април	*квітень*	0.190909
май ~~~ юни	*травень*	0
юли	*червень*	-0.0428571
август	*липень*	0.3
септември	*серпень*	0.22
октомври	*вересень*	0.3
ноември	*жовтень*	0.3
декември	*листопад*	0.3
януари	*грудень*	0.253846
февруари	*січень*	0.207692
март	*лютий*	0.233333
	березень	-0.3

И има само един месец, с който, като започне годината, програмата дава правилното (от наше гледище) подравняване на месеците. Кой е той?

- ☐ януари
- ☐ февруари
- ☐ март
- ☐ април
- ☐ май
- ☐ юни
- ☐ юли
- ☐ август
- ☐ септември
- ☐ октомври
- ☐ ноември
- ☐ декември

Обяснете накратко отговора си.

Решение:

☑ март

Разполагаме с две подравнявания, за които знаем, че програмата смята за по-добри от каноничното:

януари	*січень*	*април*	*квітень*
	лютий	*май ~~~ юни*	*травень*
февруари	*березень*	*юли*	*червень*
март	*квітень*	*август*	*липень*
април	*травень*	*септември*	*серпень*
май ~~~ юни	*червень*	*октомври*	*вересень*
юли	*липень*	*ноември*	*жовтень*
август	*серпень*	*декември*	*листопад*
септември	*вересень*	*януари*	*грудень*
октомври	*жовтень*	*февруари*	*січень*
ноември	*листопад*	*март*	*лютий*
декември	*грудень*		*березень*

Следователно верният отговор по условие не може да е *януари*:*січень* или *април*:*квітень*.

Ако предложим годината да започва с *февруари*:*лютий* или с който и да е месец от втората половина (VII – XII), програмата ще може да построи също такова подравняване като първото, само че с циклично изместване на месеците, което не променя общата оценка.

Ако поискаме годината да започва с *май*:*травень*, програмата ще може да отговори с второто подравняване с *април*:*квітень* преместени в края.

Ако годината започва с *юни*:*червень*, пак ще може да се получи по-добро според програмата подравняване, основано на второто: *април*:*квітень* и *май*:*травень* ще са в края,

вместо *май* и *юни* (подравнени с *травень*) ще се обединят *юни* и *юли* (подравнени с *червень*), а на оценката това няма да се отрази, защото в *травень* и *червень* има еднакъв брой букви.

Така остава само *март*:*березень*. Не е лесно да се убедим, че при такова начало на годината няма възможно обединяване или пропускане на месеци, което да повиши оценката, но не е и нужно: достатъчно е (но това пък беше необходимо), че изключихме останалите 11 месеца.