Product offer classification

Time limit: 30 minutes
Memory limit: 1 Gb
Disk memory limit: 2Gb

Yandex is the biggest search engine in Russia and biggest web-portal but only for Russian speaking people so it is very interesting if somebody with no scene of Russian solves this problem.

Yandex.Market is a product choosing and buying service. If finds clients for sellers. For Yandex it is one of main services that earns essential part of total income.

Correct offer classification allows sellers to compete with each other on target set of offers, that increase average bid and as result profits. Also correct classification allows to propose accessories and tying products to clients.

There is learning set (archive learn.in.bz2 74Mb, about 1.9 million examples), where there is learning example with following format on each line:

Title\tDescription\tShopCategoryName\tPrice\tCategoryId

- Title – as example, Nokia n95
- Description – non formal description, it can be both characteristics (color, sizes) and commercial information (as example "this is the best cellphone in the world")
- ShopCategoryName has following :

  Host_name:CatnameLevel1\CatnameLevel2…

  Example: nokia-shop.ru:All staff\Phones\Nokia

- Price – price in roubles
- CategoryId – id from category tree (look below)

And category tree(file categories.txt) with following format:

CategoryId\tParentId\tName

Root vertex has ParentId = -1

You should write program that classify offers using learning set and category tree.

In learning set file "\\n" symbol sequence means line break and "\\" means \ symbol

Scoring principle:

For each testing example distance between output category and correct category is calculated (minimum edge number on category tree).

Score on particular testing example is:

$score = (\frac{1}{4})^{dist}$, where dist = 0 score = 1

Team score is average scores on all testing examples

Remarks:

1. There ARE errors in learning set. This set was done in semi-automatic way by people. Error count is less than 5%. It can be that some categories in learning where all examples are incorrect. But there are not many such categories and they will not effect much on resulting score.

2. Your program can use external data (files)

Input:

Several testing offers with following format:

Title\tDescription\tShopCategoryName\tPrice

It is about 100000 testing lines in input

Output:

For each testing offer output proper category id

Example:

Вход	Выход
Nokia n95 Хороший телефон shop-nokia.ru:Товары 14500	91491
Стул KUADRA 1271 Каркас-сатин,сиденье и спинка-пластик www.newbar.ru:Стулья\Стулья на металлокаркасе7560	90673