This is a test project on “data parsing”.

Task Description:

The attached file contains a raw data-set of a list of companies. The data is a little unstandardized. It contains company names like:

* Microsoft-corp
* Microsoft-corporation
* Microsoft-india-inc
* Microsoft-co-uk
* Microsoft-gmbh

These names need to be standardised and changed to “Microsoft”. This is an example to explain what is to be done. All companies in the list are to be standardised. In short, one needs to remove all country names (India, USA, Germany, Australia, etc) that appear in the list and also, remove the type of company (co, llc, corp, corporation, bv, nv, etc)

Steps to follow:

* Make a word list of all unique country names and variations in company type (could be in any language).
* Write an algorithm to match and remove the above word-lists from the results.
* Code.
* Quality check to ensure the final output is correct. If and only if need be, one can change the remaining companies manually.
* Record a counter of the number of company names changed 1) using the algorithm, and 2) manually. Send this along with the final output file. Also, send any algorithms and codes used.

You can use Java programming language for this project.

Task deadline: September 19, 2013, 23:59 hours

Shortlisting criteria:

* Quality of the output. The percentage of accuracy will determine one’s selection.
* Approach: a thorough documentation of the approach and the algorithm used.
* If there is any copying of the source codes or algorithm, we will cancel all the matching results, irrespective of who copied from whom.
* Code: a short description of the code, the list of variables used and its requirement (alternatively, you can add comments in the code, so we can interpret them).
* Efficiency of the code: The same code which completes the query in lesser time will be rated higher.
* Speed of completion: needless to say, the one who finishes the test project first, would be given higher preference.