**Named Entity Recognition**

**Perl Program** – automation.pl

**Input** – download.txt

If you want to change the input file, just change the $filename variable in automation.pl program.

**Output** text files contains affiliation and publication date along with all the locations.

Change $outputfolder variable in automation.pl to the required location and output text files will be created in that location.

This program automation.pl is calling java program which does Named entity recognition on the input text.

**Java Program** – prac.java

$javaArgs variable in automation.pl contains the location of all the jar files required for running the NER and java program.

To execute the perl program

1) open terminal and change directory to the location where automation.pl is present

2) Execute automation.pl script using the following command -> Perl automation.pl

**Summary:**

This project is designed to create space and time datasets (text files) on Pubmed central data which contains locations of the author, date of publication and all other locations present in the content of the paper. automation.pl is the final perl script which does the complete process of above said datasets.

Automation.pl script consists of main program and two sub routines. One sub routine is worker which was created in order to implement the multiple threads which parallelize the process. Here, we had 5 threads running simultaneously. The input is loaded in to a queue which contains the links of the files we need to download. This sub-routine worker downloads the file from the website by making a system call to linux and extract the tar.gz file. The extracted folder contains nxml file which is given as input to the next sub-routine perlscript. This sub-routine perlscript creates 2 output files, one of which contains the locations of author and publication date of the article. Another file contains the remaining text content of the xml tags in the inputted nxml file. This text content file is given as input to the java program which performs named entity recognition and collects all the locations present in to another output file.

The output file above which contains locations of text content file is appended to the file which contains locations of author and publication date of the article. All the tar.gz files and extracted folders are deleted once the above output files are generated.