TABA Batch 8 - Individual Homework 2 (NLP, NER, Chunking)

Task:

Step 1 – Select one well-known firm from the list of the Fortune 500 firms.

Step 2 – For the selected firm, scrape it's Wikipedia page.

Step 3 – Using openNLP, find all the locations and persons mentioned in the Wikipedia page. Its good practice to set timers and report runtimes for heavy functions.

Note: You can use either openNLP's NER functionality, or, alternately, use the noun-phrase home-brewed chunker we covered in class. If using the latter, manually separate persons and locations of interest.

Step 4 – Plot all the extracted locations from the Wikipedia page on a map. You may want to see 'NLP location extract and plot.R' file on LMS for this.

Step 5 – Extract all references to numbers (dollar amounts, number of employees etc) using Regex.

Deliverable: R markdown submitted as HTML page. Ensure that the lists of people, persons and amounts mentioned are clearly visible. Also that the map of places is visible too. Use the text part of the markdown to record your observations.

Deadline: Midnight of 6-May Saturday in the LMS Dropbox.

Instructions: Don't over-think this one. I'd say no more than a few hours of work, at most. Feel free to discuss these assignments with peers, help out and take help where necessary etc.

However, at the end of the day, your submission must be individual and driven primarily by your own effort.

Any Qs etc, contact me, Preethi or Aashish.

Sudhir