If it's on web it's yours!

Abdul Mateen Rajput

Life Science Informatics, Bonn University, Bonn, Germany

1 Introduction:

Text mining is an emerging field and there are many applications of this field since the rate of information production has increased many folds in recent past. Despite exponentially rate of data production we are still struggling for the answer of the question which can satisfy our needs as it has been said that we are drowning in sea of data while dying of thirst for knowledge. One important area which seeks answer from massive datasets is biomedical sciences, where text mining facilitates to add value and provides different procedures to analyze bulk data being produced either after each new experiment of microarray, fMRI etc or by scientific publications.

To explore the knowledge from data one needs to have access to it to get valuable information [datasets may vary in size and it depends upon the questions you are going to ask from it]. The availability of some datasets is usually restricted to the provider and user may sometime doesn't find the correct dataset he/she is interested in, though it may be browsable on the web but not available as repository to apply natural language processing and text mining tools and user finds difficulties to achieve what is required. There are many web crawlers (HTTrack[1], GRUB[2] etc) but the problem with these programs is they bring too much noise and uncleaned data. The cleaning of this data is also an issue and usually takes more time than downloading. In the current paper we discuss a smart approach to make clean dataset from any online website. The resultant dataset could be any file format you are interested in and the method will provide you different possibilities to extract from many layers of web pages. The methodology we are going to discuss is freely available and following programs are required for it:

- Mozilla Firefox[1]
- DownThemALL, Firefox Plugin[2]
- Notepad++[3]
- Linkgopher, Firefox plugin[4] /GREP (shareware) [5]

2 Methods:

The initial steps of the corpora creation requires to look for the pattern of the hyperlinks of the data you are interested in and if the links of data is available on one page

1 http://www.httrack.com/
2 http://www.gnu.org/software/grub/

then DownThemALL can automatically detects the links and you can start downloading instantly. If the actual data is under few layers of web pages then you can download the source pages and then actual data by combining all the source html pages and extracting links via LinkGopher or by using Grep program. The good feature of Grep is that it will also bring the data within the proximity of upto 5 lines from the actual search term.

3 Use Case:

The use case discusses the task we did with linkedCT.org [6], which is a RDF processed repository of clinicaltrials.gov [7]. We needed to download all the clinical trials associated with a particular disease and those clinical trials were stored under 4 different names (Multiple sclerosis Relapsing-Remitting, Relapsing-remitting Multiple Sclerosis, Relapse-Remitting Multiple Sclerosis, Relapsing Remitting Multiple Sclerosis). The actual data we were looking for was stored under 2 html pages where all the label of clinical trials associated with the disease state was mentioned (see figure 1). We stored the source html pages of actual clinical trials (4 pages associated with the disease titles) and then merge them together so we can have all the names of files on one html page. We found that the pattern of RDF storage and the page where it contains the link of it doesn't differ much and there is a similar pattern for each RDF file associated with the webpage link. Further we extracted all the links by using LinkGopher from the merged page and then looked at the patterns of RDF and html page. After finding out the pattern we simply replace the keywords with the one which was associated with RDF and then downloaded all the RDF files by simply using DownThemALL.

Fig. 1. The overall view of the dataset. We needed many different RDFs (in green box) stored under different pages, description page of clinical trial, label page of different clinical trial and on top the disease page.

4 Conclusion:

We have used this method with several different websites and collect a large repository for using different text analytics tools. However, the procedure also has some limitation (doesn't work with Java links) and you have to carefully find out the patterns of dataset etc. On the contrary the good thing is that it is freely available and very quick rather than clicking the links and saving it manually.

5 Reference:

1. *Firefox*. Available from: http://www.mozilla.org/en-US/firefox/new/.
2. *DownThemAll*. Available from: http://www.downthemall.net/.
3. *Notepad++*. Available from: http://notepad-plus-plus.org/.
4. *LinkGopher*. Available from: https://addons.mozilla.org/en-us/firefox/addon/link-gopher/.
5. *Windows Grep*. Available from: http://www.wingrep.com/.
6. *LinkedCT*. Available from: http://linkedct.org/.
7. *ClinicalTrials*. Available from: http://www.clinicaltrials.gov/.