**How to Process Barnes & Noble Text File into Reserve Reading Lists**

**Purpose:**

Using the “qprint.txt” file that Martha retrieves every semester from Barnes & Noble, of all citations that professors have submitted for their courses to Barnes & Noble, create lists that will help Reserves and Collections identify the citations for each course, separate titles we already have from titles we need to purchase,

**Set up environment**

* Install Python 2.7.x on your computer if it’s not already. If you have both 2.7 and 3 already, make sure you invoke 2.7 when you run commands, including pip install
* Make sure you have the necessary modules to run this script.
  + pip install –r requirements.txt
  + **check with Sarah or Kate to make sure the tk packages install OK the module name you call is not the same one you install**

**Get source data:**

* ask or wait to receive Barnes and Noble citations file. This is something they just pull off a webpage. It barely has any structure, but the scripts look for formatting that indicates its needed structure. This file is usually called “qprint.txt” Martha usually sends it to me several months before the beginning of each semester
* take this raw input file, and run parseBN.py on it
  + python parseBN.py
  + you will be asked for the input file mentioned above.
* Look at the output of this script. It will be called “Cleaned Barnes and Noble File <date>.xls” . Because the formatting can be irregular, there may be some lines that did not got processed correctly.
* To assure this, look through every line and make sure the columns line up.
* Then go back to the source data .txt file and perform a number of regex cleaning so you can be sure the number of citations in the output file matches what’s available in the input file.
* In Notepad ++
  + ^[ ]{3}(\S) replace with \t\t\1
  + ^[ ]{9}(?=\S) replace with \t\t\t
  + ^(\t{1}[^\t].+?$\r\n)|(\t{3,}[^\t].+?$\r\n) replace with blank
  + Manually take out the first 9 or so lines before actual data, including taking out headers
  + ^[ ]\d+\/.+?$\r\n replace with blank
  + ^\d\d\:.+?$\r\n replace with blank
  + \t\t[- \t]+$\r\n replace with blank
  + ^\t\tAuthor.+?$\r\n replace with blank
  + Take out last lines beginning with “Total”
* (**This should be scripted)**
* The line numbers of this list should match what’s in the output file noted above. If it doesn’t, manually adjustments to the output data, or adjustments to the script may be necessary. As a committer you will be able to save a new version of this under github.

**Find out what titles we Own:**

You will use the ISBNs from the Barnes and Noble file to create a Managed Set in Alma. The set will contain as members those items on the reading list we own

* Highlight columns “D” and “E” from the “Cleaned …” output file above, not including the headers, and paste into Notepad ++
* Transform the data into a list
  + \\*\\*[ ]E[ ]Book[ ]\\*\\* replace with blank
  + \;[ ] replace with \r\n
  + \r replace with \r\n
  + (\r\n){2,} replace with \r\n
* Create a new line at the top and enter “ISBN” all the way at the left.
* Save this file onto your desktop.
* Import into Excel and save as an .xlsx. **This is because Alma managed sets have had trouble importing text files directly, because of a problem with BOM**
* Import this Excel sheet as an input into a itemized managed set
* Export the managed set, including all fields

**Compare the list to what we have in Alma:**