**Flowchart for Python Code for metadata extraction: (Change year value on line 65 of the code before you execute the code.)**

**Steps:**

1. Import packages and modules
2. Define pdf to html conversion function and string clean function
3. Convert all files from pdf to html for a particular year and store in stxt array variable. Step 3 calls function in Step2.
4. Define arrays for holding metadata like title, abstract, keywords, etc.
5. VERY Big FOR LOOP for metadata extraction. (80% of code lies in this for loop). Here, each iteration of big for loop accesses each element of stxt array. Remember, each stxt element contains one full article. Together, it contains all articles for one year.
   1. This loop accesses stxt array in step 3 one by one and points to each article for a year and all quarters
   2. Every metadata is stored in its own array.

For example: title array contains titles of all articles for a particular year. Length of each metadata array would be same: title, abstract, etc. If any metadata is missing in article then I put blank in that position. Like some papers didn’t contain conclusion. This ensures that each array is of same length.)

* 1. Extracts metadata in the following order:

1. Actual file name
2. Page count
3. Category: Special issue, theory and review, etc.
4. Title
5. Actual Keywords
6. Count of keywords
7. Abstract
8. Count of figures in article
9. Count of tables in article
10. Name of senior editor
11. Publication year, month, paper id
12. Composite id
13. Introduction
14. Conclusion
15. Count of authors in article
16. Author names, email id country
17. Compile data and export in csv