**Parsing mechanism of webpage:**

1. I used Jsoup Library to parse all the web page by providing the URL, and extracted all the publications present in particular html tag.
2. In each publication I searched for all 4 digit number and check whether it falls under 1950 to 2020 and if it is then I extracted it as year.

**Problems faced:**

1. Some of the authors put the month in combination like Sept/Oct, Oct/November, and July/Aug and so on. So first I had to generate all the combinations of 12 month.
2. Then I have to store generated combination in Array and sort on the basis of array index. That would reduce the efficiency I suppose.
3. I split each publication into tokens for extracting month, and some authors doesn’t even provide a space between month and year e.g. Aug2010 therefore I first had to put space between them (Aug 2010).
4. After extracting year and month from publications I had 3 separate array containing years, months and publications. And at that point I need to sort year’s array and want to change the months and publications array index in parallel according to year’s array.

**Feasibility in extracting author and topics:**

1. The first problem perhaps here is how the system understands that it is author name, publications or journal/conference?
2. I think it will require Natural Language Processing (NLP) concepts to do so.
3. If you consider, I can put the entire author’s name in an array and extract the author name from the publications by searching the author name in publications but it will reduce the efficiency.

Sir, some authors mentioned year alone and some mentioned year and month both in their publication. The authors that don’t specify months, their publications are sorted in alphabetically within same year.

If I will sort the publications for the same author listed together for same year than it will disturb the chronological order by month I suppose.

Do you want me do so?

**Scheme used to separate Author name, title and other information:**

1. First I extracted a text present in double quotes from publication as a Title of the paper and replace it with underscore (\_).
2. Then, I split the remaining part on the basis of that underscore (\_) and get the result in array.
3. Now, the first part most probably is an Author name and second part contains other information.

**Issues faced in above process:**

1. Some authors put the title first then Author’s name and then other information. In this case if we apply the above scheme after extracting title it considers the remaining part as other information i.e. author contains blank value.
2. Then to get the author name from other I split the other information part on the basis of (;) in stores in an array.
3. Now, the first part is author name and second part is refined other information.