DATA 601 : Project 2

SPRING 2023

Introduction

- Arxiv (pronounced archive) is a website that hosts a lot of pre-published research papers
- For this project you will need to access Arxiv metadata on papers from 2018 to 2022 for certain categories (ie. First uploaded)
- Using this data, you will do the rest of the tasks
- Submit your project folder as a .zip or .7z.
- Name convention for folder “<Lastname>_Pr2”
- A total of 100 pts + 15 bonus

Task 1 (25 pts)

- Access Arxiv and get metadata (title, authors, summary etc) of all the papers in the **primary** categories of:
  - Databases
  - Graphics
  - Robotics
  - Emerging Technologies
- Do the above for the years from start of 2018 to the end of 2022
- Store data in your choice of file (json,csv etc) or multiple files.
- Create notebook task1.ipynb inside your project folder. This should have the code you used for extracting and storing the data

Task 2 (25 pts)

- Create notebook task2.ipynb for the code and results of this task.
- Using the stored data from the last task, create a dataframe for each **primary** category (Databases,Graphics,Robotics,Emerging Technologies), the various fields of the metadata will become columns (title, authors, summary, etc), rows will be the papers from 2018 to 2022
- Show first 5 lines of each primary category

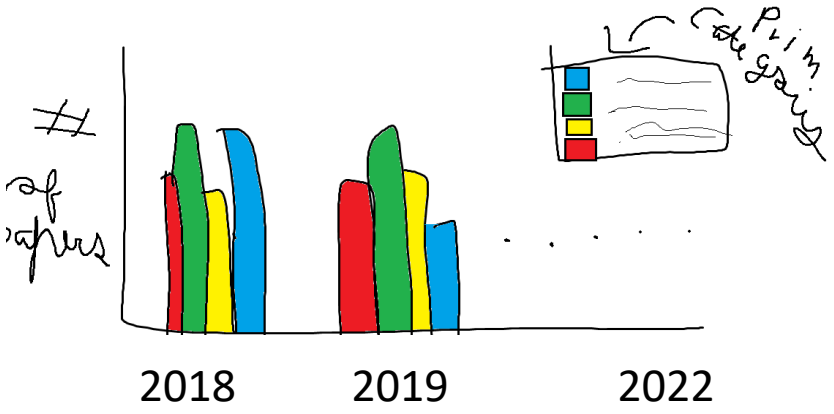
Task 3 (25 pts)

- Create notebook task3.ipynb for the code and results of this task.
- For each of the four primary categories considered (2018-2022) , draw a pie chart with slices (%age) for
  - single author papers
  - two authors papers
  - 3-4 authors papers
  - More than four authors.

(NOTE : Two authors means ONLY two authors)

Task 4 (25 pts)

- Create task4.ipynb
- Make a bar graph
- On the X axis, put the years
- On the Y axis, put the number of papers uploaded
- Bars should be grouped by category (see right for an example diagram)



Just a illustrative sketch, please don't reproduce this graph literally.

BONUS Task (15pts)

- Find the top 3 authors in cs.DB (Databases), cs.RO (Robotics) and cs.GR (Graphics) by number of papers on arxiv (between 2018 and 2022)
  - The criteria is the number of times their name appears as one of the author (i.e can be solo, or with other authors) in the papers.

Other Instructions

- Write comments in code
- Document what is being done for each task using Markdown cells (so that I understand what you are doing).

How to get the metadata of papers?

- Use the Arxiv API (Application Programming Interface) to get the desired metadata (title, authors, summary,…etc)
- Do this directly using web API
- Or with the Pypi arxiv package
- Do not use the arxivscraper, arxivabscraper or arxiv-miner packages (inaccurate results)

ARXIV ID

Year
18 = 2018

Month
03 = March

Number of paper for that month
Starts from 00001
Can go upto 99999

For example code…

Take a look at access_arxiv_paper.ipynb

(It also tells you about all of the information fields for each paper)

And a file called arxiv_helper.ipynb

Category IDs in ARXIV

Categories	Arxiv Category ID
Databases	cs.DB
Graphics	cs.GR
Robotics	cs.RO
Emerging Technologies	cs.ET

Credit: https://arxiv.org/category_taxonomy