DATA 601 : Project 2

FALL 2022

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 2017 to 2021 for certain categories
- 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
- **No bonus pts in this Project**

Task 1 (25 pts)

- Access Arxiv and get metadata (title, authors, summary etc) of all the papers in the **primary** categories of:
  - AI
  - ML
  - Computational complexity
  - Hardware architecture
- Do the above for the years from start of 2017 to the end of 2021
- 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 (ML, AI, computational complexity, Hardware architecture), the various fields of the metadata will become columns (title, authors, summary, etc)
- Show first 5 lines of each primary category

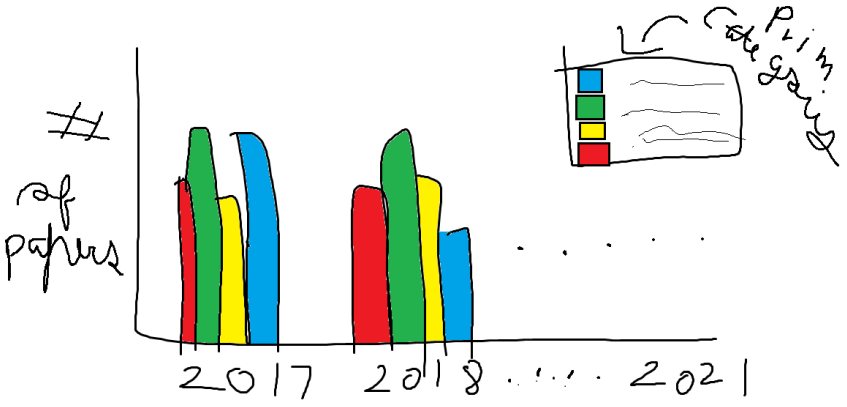
Task 3 (25 pts)

- Create notebook task3.ipynb for the code and results of this task.
- For each of the four primary categories considered, 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.

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).

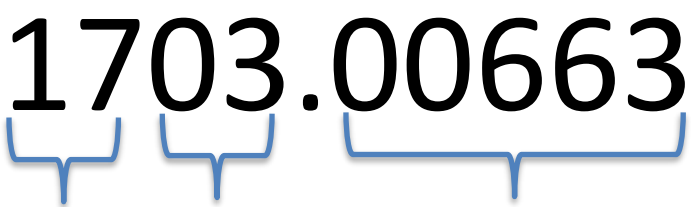
How to get the metadata of papers?

- 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)

Using the API : one way to do it

- **Problem :** You can't search by date using the arxiv package or API (directly). Also, package has a max return of 300,000.
- **Possible Solution** : We can search using the **arxiv ID**: each paper has a unique **arxiv ID** that is related to which year and month the paper was upload.
- If you have other ideas, you are free to do it.

ARXIV ID



Year
17 = 2017

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)

Category IDs in ARXIV

Categories	Arxiv Category ID
Artificial Intelligence	cs.AI
Machine Learning	cs.LG
Computational Complexity	cs.CC
Hardware Architecture	cs.AR

Credit: https://arxiv.org/category_taxonomy

General Strategy with the Pypi arxiv package

Solution :

1. Using arxiv id (`id_list`) access every file from Jan 1 2017 to Dec 31 2021.
2. Filter out (discard results) that do not have the primary_category as cs.AI, cs.LG, cs.CC or cs.AR

To access every paper using arxiv id

- You can start with 1701.00001 and go up
- Digits after .
  - Increment from 00001 up to the number where no paper exists (`result = []`)
- Digits before .
  - For each year, increment month from 01 to 12 to cover entire year
  - Repeat for years from 17 to 21 (include)

Or do it any other way you like

Within the bounds of the description of this project (see the packages that are forbidden)