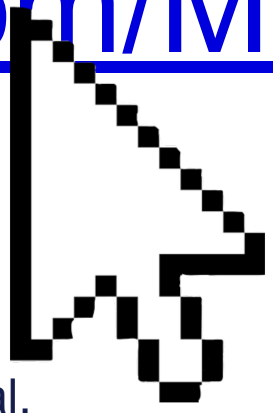
Second Term Project Assigment

BEIR CQADupStack and ARQMath Collection

https://github.com/MIR-MU/pv211-utils



MUNI FI

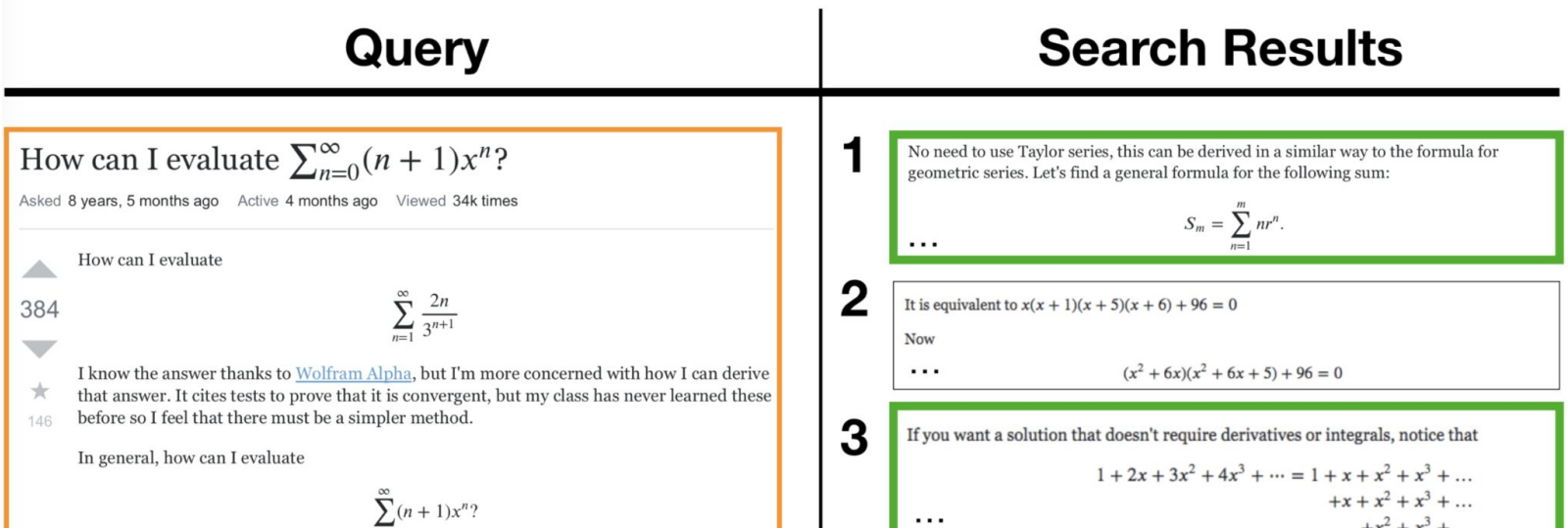
Vít Starý Novotný et al.

PV211: Introduction to Information Retrieval

Spring 2024

Introduction

- **CRANFIELD** (1950s) evaluates retrieval of scientific article metadata.
- **CQADupStack** (2015) evaluates the retrieval of text+math+code questions.
- **ARQMath 1–3** (2020–2) evaluates retrieval of text+math questions/answers.



Requirements

- Implement a *supervised* ranked retrieval system using relevance judgements.
  - Experiment with techniques such as weighted zone scoring and large language models.
  - NVIDIA GPUs are available at JupyterHub:
    - NVIDIA A10 (24G VRAM), A40 (48G VRAM), and A100 (80G VRAM)
- Document your code in accordance with PEP 8 and PEP 257.
- Reach at least 25% MAP for **CQADupStack** and 10% MAP for **ARQMath**.
- Submit .ipynb file by 2024-05-07 for either **CQADupStack** or **ARQMath**.

- You will be awarded ≤ 20pt for project, ≤ 6pt for explanation, ≤ 6 pt for review.
- You can get extra 20 / 10 / 9 / 8 /… / 1 points for 1st / 2nd / … / 11th final place in the leaderboard.

Retrieval Units

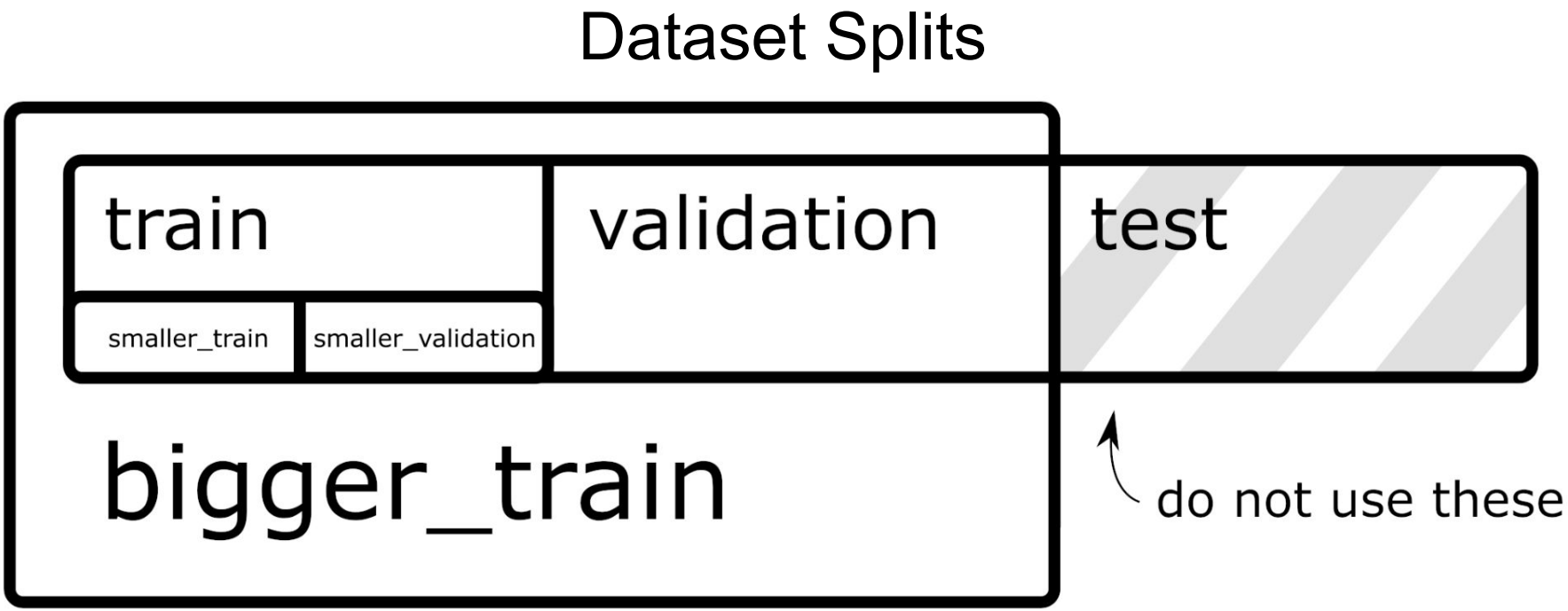
- **CRANFIELD** contains 225 queries over 1398 abstracts of journal articles.
  - **Task:** Finding journal articles using short queries.
- **CQADupStack** contains 300 queries over 457,188 questions from SE forums.
  - **Task:** Finding old duplicate questions to new questions.
- **ARQMath** contains 200 queries over 1,445,495 answers from MSE forums.
  - **Task:** Finding old answers to new math questions.
  - Each answer is connected to one of 1,020,585 questions, which can be used for enrichment.
  - Queries, questions, and answers are available with seven different *math representations*.

- **CQADupStack** and **ARQMath** don't fit in 12G RAM at Google Colab.
  Use JupyterHub, where we provide up to 64G RAM.

Relevance Judgements

- **CRANFIELD** contains 314,550 judgements with 1,837 (0.58%) relevant.
  - The relevance judgements are *exhaustive*: all 1398 abstracts × 225 queries have been judged
- **CQADupStack** contains 23,692 judgements, all of which (100%) are relevant.
- **ARQMath** contains 70,912 judgements with 3,765 (5.31%) relevant.
  - The relevance judgements are *graded* from 0 (non-relevant) to 3 (relevant).

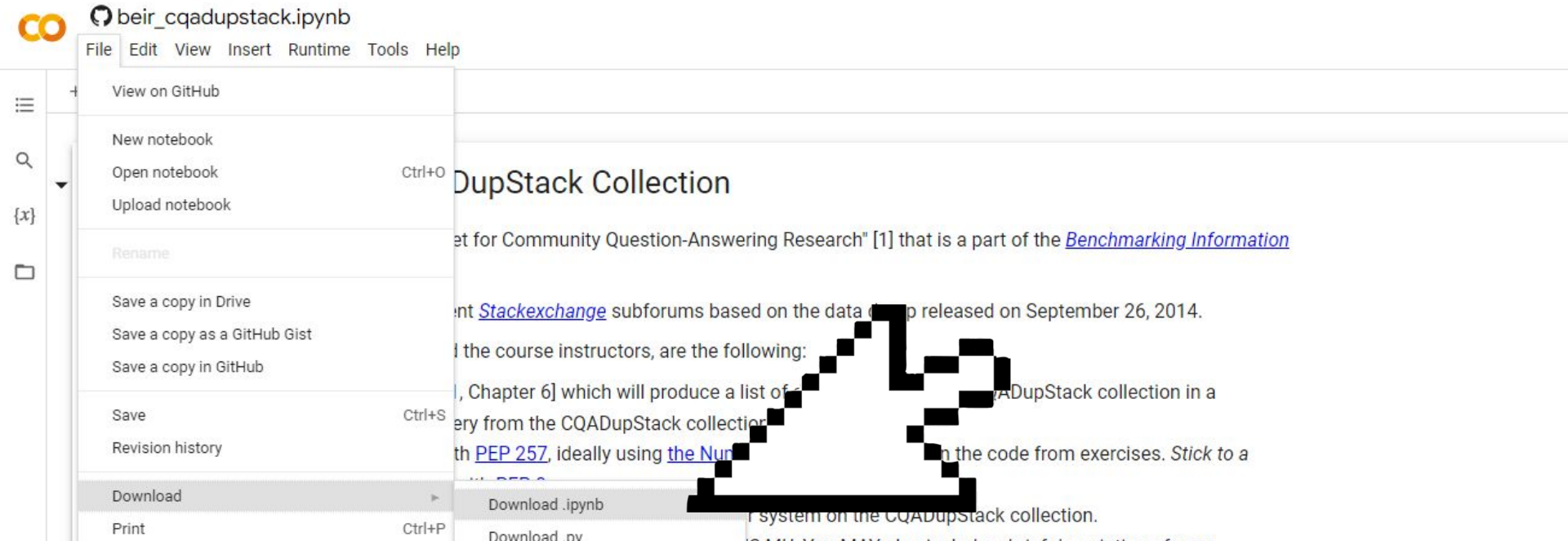
- For evaluation, we use *MAP*, which uses exhaustive non-graded judgements.
  - We made judgements for **CQADupStack** & **ARQMath** exhaustive by non-judged = non-relevant.
  - We made judgements for **ARQMath** non-graded by making 0–1 non-relevant and 2–3 relevant.



Dataset splits for supervised training on **CQADupStack** and **ARQMath**.

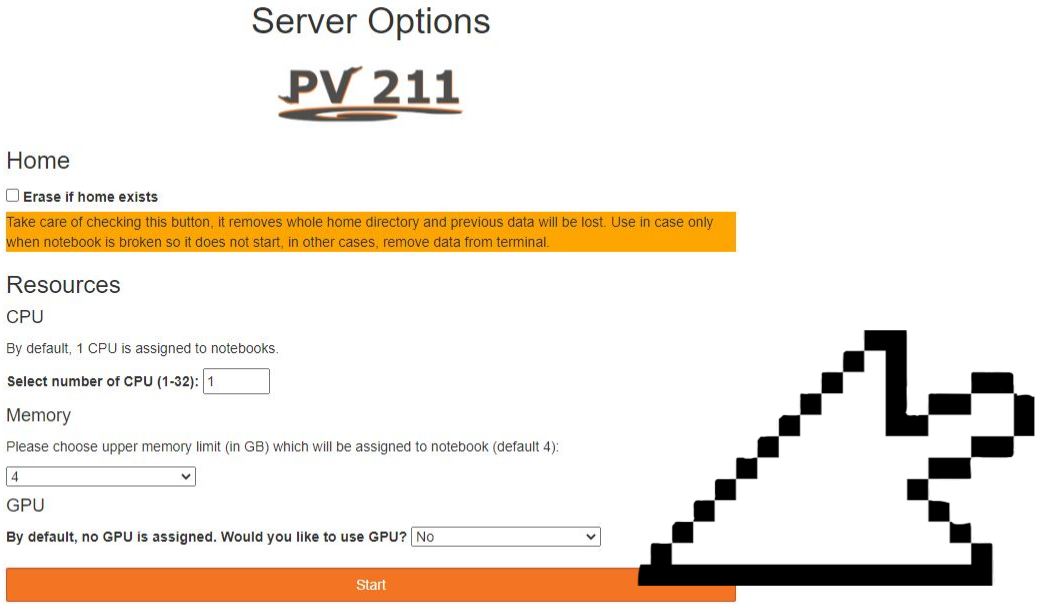
How to Use JupyterHub

1. Download notebook with example solution for [CQADupStack](#) / [ARQMath](#):



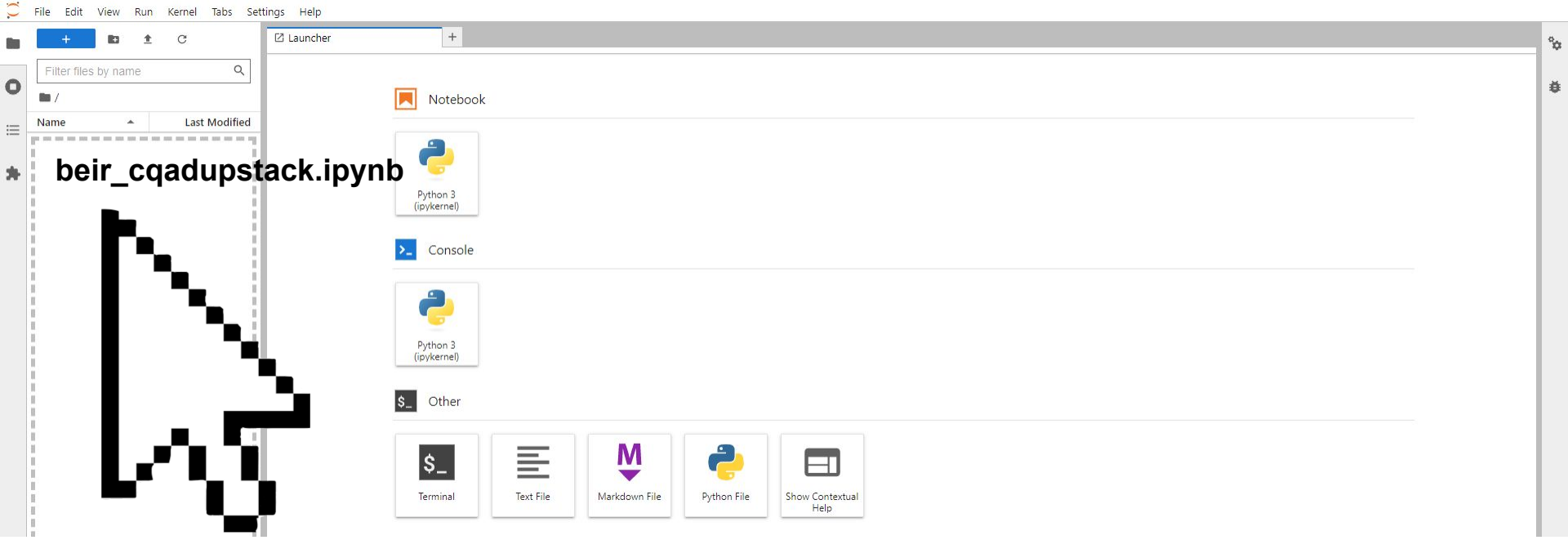
How to Use JupyterHub

2. Set up (CPU, GPU, RAM) and launch a server at [JupyterHub]():



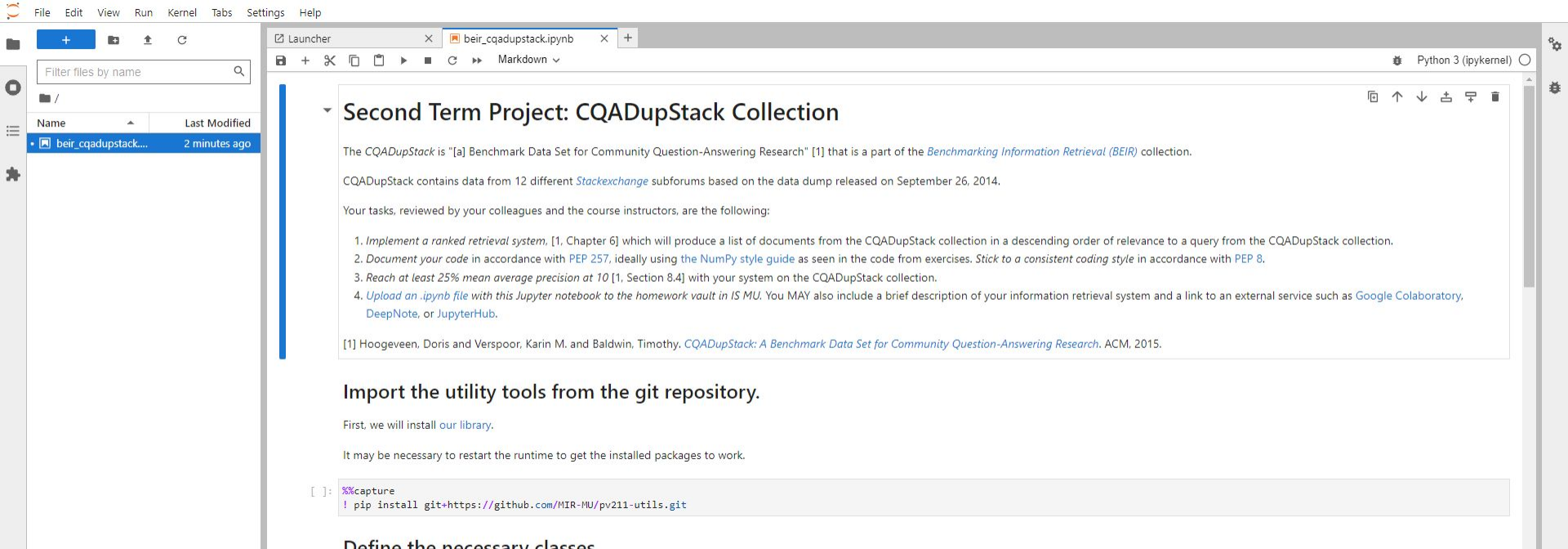
How to Use JupyterHub

3. Drag’n’drop notebook with example solution to JupyterHub**:**



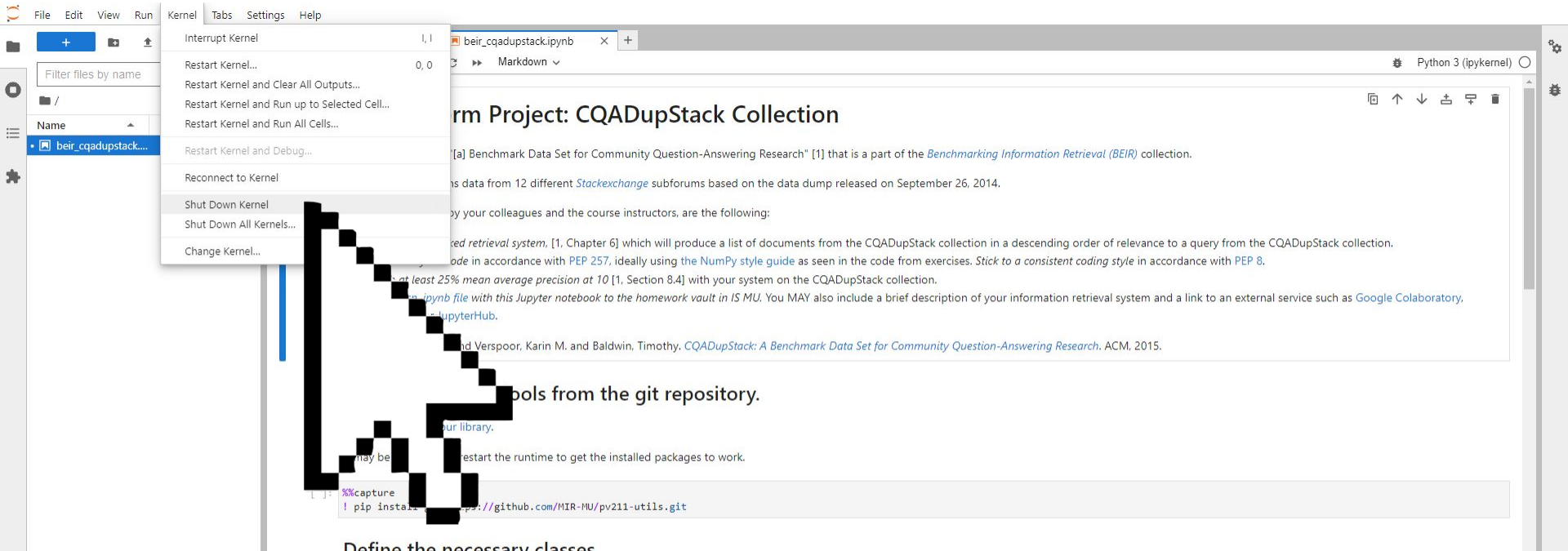
How to Use JupyterHub

4. Work on the assignment as you would at Google Colab or elsewhere.



How to Use JupyterHub

5. When you are done for the day, shut down Python, so the server can be reaped.



Peer Review

- **Same** Questions as in the **first peer review**

(1) Reproducible submission

Your colleague uploaded a Jupyter notebook file (in .ipynb format) to the homework vault in IS MU. After hitting "Run all cells", the notebook will run without issues all the way down to successful evaluation.

(2) Transparency

You are able to understand the methodology of the reviewed implementation, including the details, such as used constants, or data resources. Where needed, the code is documented and typed in accordance with PEP 257, ideally using the NumPy style guide as seen in the code from exercises.

(3) Correctness

The reviewed implementation is fair; It does not use the test data, or does not adjust the shared evaluation. The score reported in the leaderboard matches the evaluation score that you obtain from running the script. The submission passes minimum score threshold. In case you think the solution could be plagiarised or copy-pasted without a reference to the source, please give more details in the description below.

(4) Innovativeness

You personally think this solution is exciting, creative or in any way inspiring. This can include, for instance, innovativeness in the method design, or efficiency. Alternatively, you learned something new from this solution, including small things, like the impact of specific design decisions.

Peer Review

- **Same** Questions as in the **first peer review**
- **Extra (6pts)**: To help your reviewers to assess the quality of your solution, we'll ask you to submit your notebook with a **short Technical Report** (max 400 words):

  Report Objectives & Evaluation criteria:
  - To allow your reviewer to **understand your solution**, even if it uses technologies outside the scope of the course
  - To **present findings** from the development that you do not submit, but led you to final solution
    - comparison of scores with different choice of preprocessing, encoding, distance measure, …
- We will provide an optional Report template 1 week before the deadline