ItC Python & GitHub Homework

National Taiwan University
Introduction to Computer
2019 Fall - Python and GitHub Homework

> Due date
> — Mon, Dec 30, 2019 11:59 PM

TAs

> — r08922053 Yu-Kai Huang
>
> — r08922042 Kuang-Yu Jeng

- Instructor

> — Winston Hsu

Requirements

- Crawl the announcement page (https://www.csie.ntu.edu.tw/news/news.php?class=101) of CSIE website within specified range of dates. (Please use the request headers in TA sample codes) The results should contain but not limited to the following fields:
  - Post date
    - e.g. `2019-05-14`
  - Title
    - e.g. `107學年度資訊學群畢業典禮重要公告(典禮前請務必詳閱) - 5/31更新`
  - Content
    - recursively find all the *text* in `<div class="editor content">`
- Please save the results to a CSV file which **can be opened by Excel** using utf-8. Please note that:
  - User should be able to specify the path to write the CSV file with `--output` argument.
  - Formats
    - Each record in one line.
    - Fields of a record are seperated by a comma “,” with no space or new line

between.

    - Strings in the CSV file are enclosed by a pair of double quotation mark (e.g. "I'm string "). And any double quote within a string should be replace by 2 double quotation mark. For instance, the string: "Prof. Yuguang "Michael" Fang, University of Florida" should be replaced by "Prof. Yuguang ""Michael"" Fang, University of Florida"

What you should do

- Create programming environment (*Linux environment (https://docs.google.com/presentation/d/1O43qZ5th7l5kpojirpqSCzVXqZtvL_7WZ7Z05wSCWig/edit?usp=sharing)*)
- Init your git repository with a README
    - If not, there will have no master branch. Reference this (https://docs.google.com/presentation/d/123JcZ-YwsCXcY6PYHk31_1wCss0ukQFGQ05SrOa6ZIg/edit#slide=id.g6c70dc8c07_0_0).
- Clone the repository to local
- (optional) Copy *TA sample codes (https://github.com/kaikai4n/ItC-python-hw-sample-code)* to your local repo, push to origin, and star TA repository
- Start programming (*Python Toturial (https://docs.google.com/presentation/d/14pCla_krES-uVRrrv-aW1XtZNFV0ArhNn89WVedeV3Y/edit?usp=sharing)*)
- After, finishing the crawler, remember to write
    - team members' names
    - school ids
    - Brief introduction to what the project does
    - Environment
        - e.g. CSIE Workstation, Python 3.6.2, lxml==4.4.2, tqdm==4.28.1, ...
    - collaboration contribution (which programming parts you are responsible for)
- Put your git url in a file and upload to ceiba, only one person in the team should upload.

What TAs will run

```
python3 main.py --start-date [start date] --end-date [end date] --output [out filena
```

- `--start-date` and `--end-date` will be in the format of `[Year]-[month]-[day]`. For

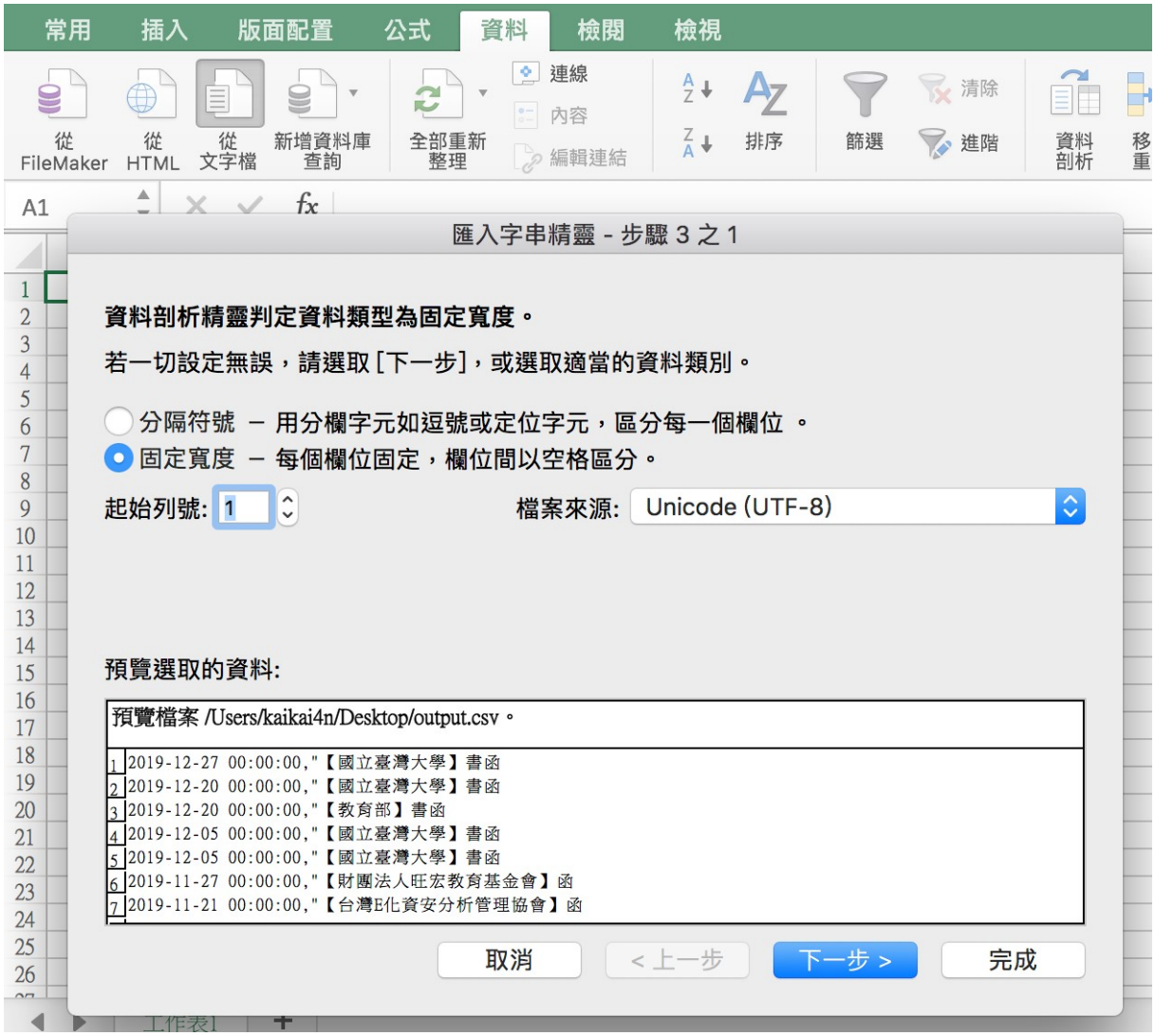
instance, `2019-12-09`.

- `--output` is the csv filename to save. For instance, `output.csv`.

Score

Hope you get 💯

(60 pts) Python

- (10 pts) Run without error
- (5 pts) Correctly parse arguments
- (10 pts) Output files to correct place and can be opened by Excel and pandas.read_csv
  - 

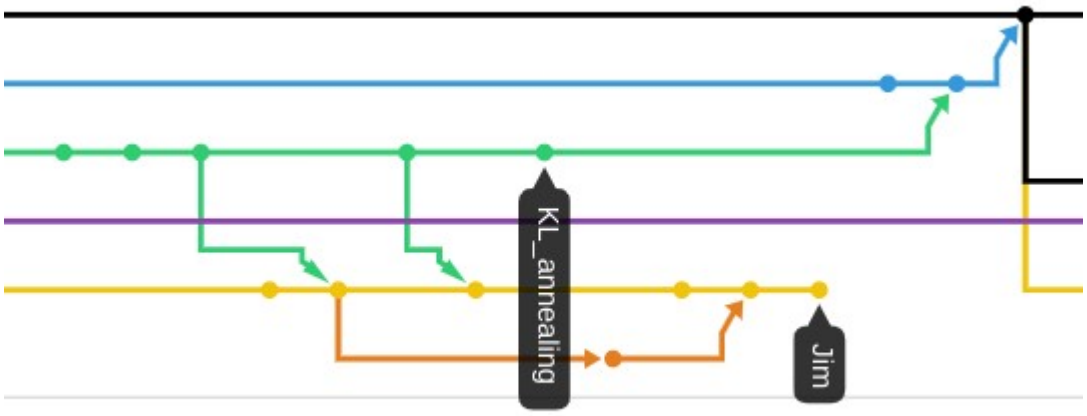
- (5 pts) Sort by post date (current to before)

- (-20 pts) Sleep 0.1 seconds before every request. This rule is **required**. You will lose points if you violate the rule.

- (30 pts) Contents are correct

- If your python contains malicious codes, all the team members will fail .

(40 pts) GitHub

- (10 pts) Protect master branch
- (10 pts) Pull Request
  - Always have pull request when merging to master branch
  - Peer review with comment
- (10 pts) Collaboration: ≥ 2 people in the team commit codes
- (5 pts) Neat and tidy network: Rebase is **required** when merging codes with conflict. Abnormal branch networks would be thought of as **not** neat and tidy.
  - 

The above is considered as not neat and tidy.
- (5 pts) Branch
  - Merge or delete unneccessary branches after finishing the homework
  - Branch name should have meaning
    - For example, branch name `kai` is meaningless

Others

- (-10 pts) Readme contains no team members' names or school ids or descriptions or environment
- If no collaboration contribution is specified, TAs would think team members equally

contributed to the homework.

- README is generally written in markdown format, but it is optional to use the format. If you are interested in how to use markdown, you can reference *markdown tutorial (https://hackmd.io/s/features-tw)*

Related Links

- *GitHub Tutorial (https://docs.google.com/presentation/d/123JcZ-YwsCXcY6PYHk31_1wCss0ukQFGQ05SrOa6ZIg/edit?usp=sharing)*
- *TA Sample code (https://github.com/kaikai4n/ItC-python-hw-sample-code)*
- *Python Introduction (https://docs.google.com/presentation/d/14pCla_krES-uVRrrv-aW1XtZNFV0ArhNn89WVedeV3Y/edit?usp=sharing)*
- *Environment Setting (https://docs.google.com/presentation/d/1O43qZ5th7l5kpojirpqSCzVXqZtvL_7WZ7Z05wSCWig/edit?usp=sharing)*