Natural Language Processing Summer 2021 #0

Chi-Jen Wu

Agenda

- About Me Chi-Jen Wu (吳齊人, CJ)
- Syllabus
  - General Information
  - Objectives
  - Topics
- Conclusion
- Q&A
- https://cjwu.github.io/courses/nlp.html
- cjwu@mail.cgu.edu.tw

About Chi-Jen Wu (CJ)

- Assistant professor, Feb 2021
- EECS Ph.D., NTU 2012
- Academia Sinica Postdoc Scholar
  - 2012-2013
- 7 years startup
  - 2013-2020
  - CEO/CTO
  - Director of Board

General Information

- Time
  - Thu 09:10 - 12:00 and Thu 14:10 - 17:00
- Course Assistants
  - 陳秉宏, 沈育賢, 黃教華
  - m0829014@cgu.edu.tw, b0629056@cgu.edu.tw, b0629048@cgu.edu.tw
- Classroom:
  - Google Meet
  - https://csiecgu.slack.com
    - nlp
- Office Hour:
  - 1, 2, 3

Objectives

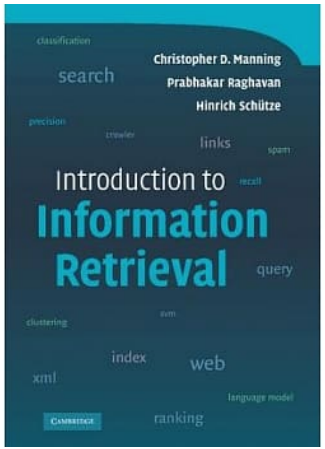
- It is about a variety of ways to **represent human languages** for Computer Science (CS) undergraduate students.
- to exploit languages representations to write programs based on the modern data-driven techniques
  - Machine Learning
  - Deep Learning
  - Rapid prototyping

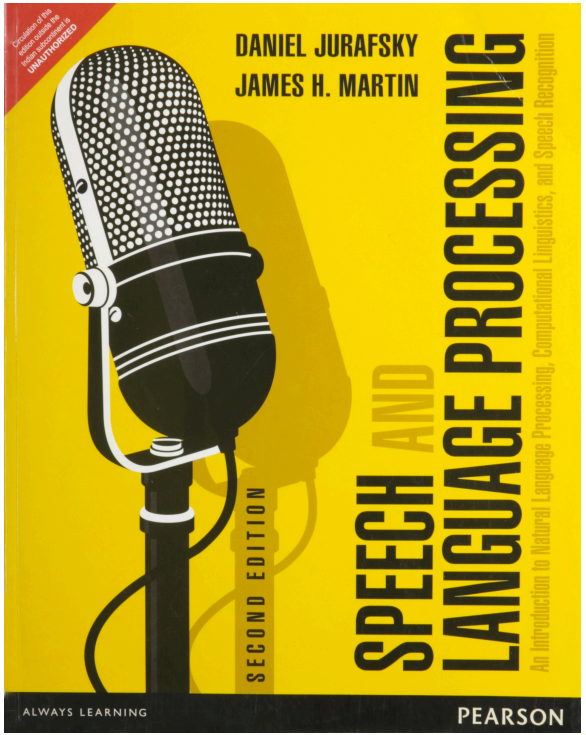
Topics

- Web crawling and indexes
- Language modeling
- Representation learning
- Word Embeddings
- Text classification
- Sequence modeling
- Machine learning models
- Deep learning models

Textbook

- here is no required textbook





References

- 1. Jurafsky and Martin,. Speech and Language Processing. 3 edition
  - https://web.stanford.edu/~jurafsky/slp3/ed3book_dec302020.pdf
- 2. Raghavan, and Schutze. 2008. Introduction to Information Retrieval. Cambridge University Press.
  - https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf
- 3. Natural Language Processing With Python's NLTK Package.
  - https://realpython.com/nltk-nlp-python/
- 4. Tencent AI LAB (中文)
  - https://ai.tencent.com/ailab/zh/paper/?page=1
- 5. CKIP Lab (中文)
  - https://ckip.iis.sinica.edu.tw/
- 6. Kaggle data
  - https://www.kaggle.com/

Grading

- 35% Homework / Exercises
- 40% Final project
- 25% Participation/Quizzes

About HW & Quizzes

- Using github, only
- 每一個作業或是隨堂練習都是一個repo
- 作業的commit 都要大於等於**五**次
- 第一個commit 和 最後一個commit 時間差要大於**30**分鐘
- **一個function 不能超過50行**
- 例如
  - A的作業全對 100分
    - 但只有一個commit, 100 - (5-1)*10 - 10 = 50分
    - 有n個commits
      - < 30分鐘, 100 - (5-n)*10 - 10 = 60分
      - >= 30分鐘, 100 - (5-n)*10 = 70分

About final project

- Using github, only
- 兩種 個人 或兩人組隊
  - 個人
    - 選取 Kaggle 上的data set
    - 分析訓練模型並上傳至kaggle 上 評分
  - 兩人組隊
    - 自行爬網頁資料 標註
    - 分析訓練模型得出精準度

Course Progress Outline

項次 No.	上課日期/星期 Date / Weekday	開始/結束 Begin/End	時數 Hours	授課教師 Instructor	教學進度 Outline	訊息 Note
1	2021-07-15 四 (Thu)	2 (09:10) ~ 4 (12:00)	3	資工系 吳齊人	Introduction to NLP	
2	2021-07-15 四 (Thu)	6 (14:10) ~ 8 (17:00)	3	資工系 吳齊人	Language modeling & Representation learning	
3	2021-07-22 四 (Thu)	2 (09:10) ~ 4 (12:00)	3	業師 張維元	Web crawling and indexes	
4	2021-07-22 四 (Thu)	6 (14:10) ~ 8 (17:00)	3	業師 張維元	Web crawling Exercises	
5	2021-07-29 四 (Thu)	2 (09:10) ~ 4 (12:00)	3	業師 張維元	Web crawling and indexes	
6	2021-07-29 四 (Thu)	6 (14:10) ~ 8 (17:00)	3	業師 張維元	Web crawling Exercises	

Course Progress Outline

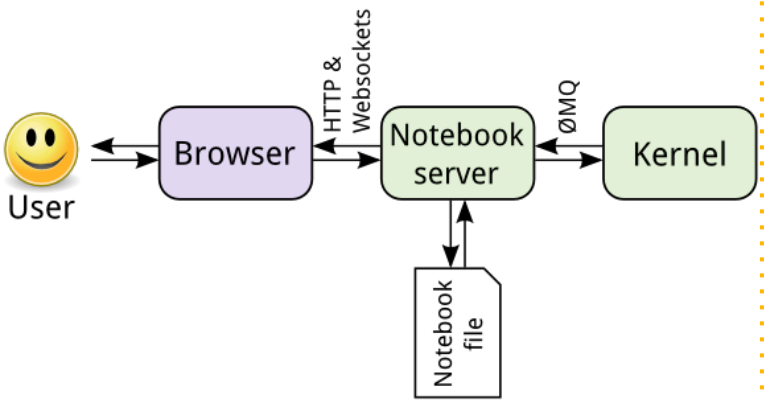
7	2021-08-05 四 (Thu)	2 (09:10) ~ 4 (12:00)	3	資工系 吳齊人	Word Embeddings Text classification	
8	2021-08-5 四 (Thu)	6 (14:10) ~ 8 (17:00)	3	資工系 吳齊人	Exercises	
9	2021-08-12 四 (Thu)	2 (09:10) ~ 4 (12:00)	3	資工系 吳齊人	Machine learning and Deep learning models for NLP	
10	2021-08-12 四 (Thu)	6 (14:10) ~ 8 (17:00)	3	資工系 吳齊人	Exercises	
11	2021-08-19 四 (Thu)	2 (09:10) ~ 4 (12:00)	3	資工系 吳齊人	Final project demo	
12	2021-08-19 四 (Thu)	6 (14:10) ~ 8 (17:00)	3	資工系 吳齊人	Final project demo	

作業環境

- python 3.x
- Jupyter
- Or Google colab
- git/github

Jupyter / colab

- 互動式計算
- 直接在上面跑python
- 多種ML/DL libs
- 直接在browser上互動
- 下午課程實習請自行安裝練習
  - 實作HW#1



Git

- Git Commands
- Git Flow
- GitHub

Git Commands

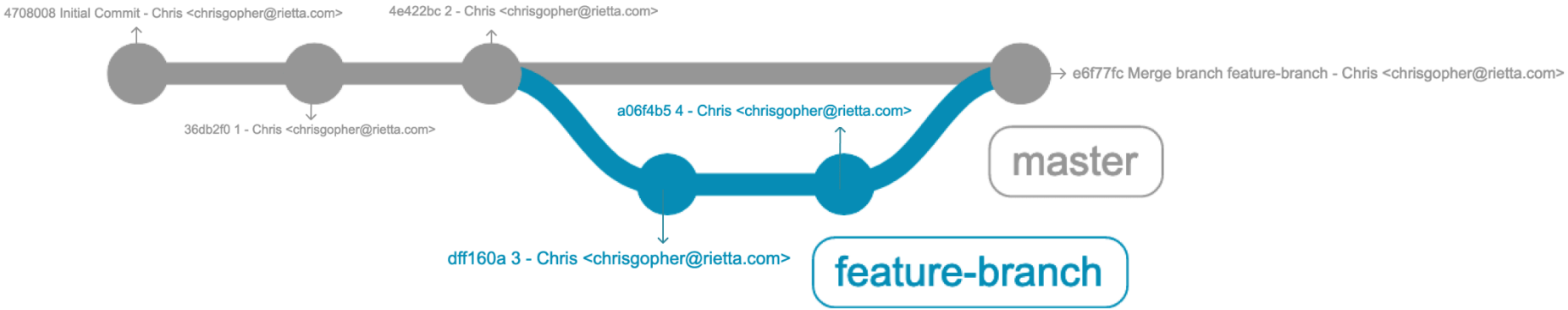
- Git init
- Git clone
- Git checkout
- Git add
- Git branch
- Git commit
- Git push
- Git pull
- Git status
- Git diff
- Git log
- Git fetch
- Git rm
- Git stash
- Git merge
- Git revert & reset
- Git rebase
- Git tag

```
22:04:23  py3.7  ~/projects/@homepage  master ✘ ★
$ git status
On branch master
Your branch is up to date with 'origin/master'.

Untracked files:
  (use "git add <file>..." to include in what will be committed)
        cgulogo 拷貝.png
        cjwu.jpg
        cjwu3.jpg
        courses/webapp/cgu-talk.pdf
        courses/webapp/web_0.key
        courses/webapp/web_0.pdf
        courses/webapp/web_1.key
        courses/webapp/web_1.pdf
        courses/webapp/web_2.key
        courses/webapp/web_2.pdf
        courses/webapp/web_3.key
        courses/webapp/web_3.pdf
        courses/webapp/web_4.key
        courses/webapp/web_4.pdf
        courses/webapp/web_5.key
        courses/webapp/web_5.pdf
        favicon.ico
        index 拷貝.html

nothing added to commit but untracked files present (use "git add" to track)
(py3.7)
22:04:26  py3.7  ~/projects/@homepage  master ✘ ★
$
```

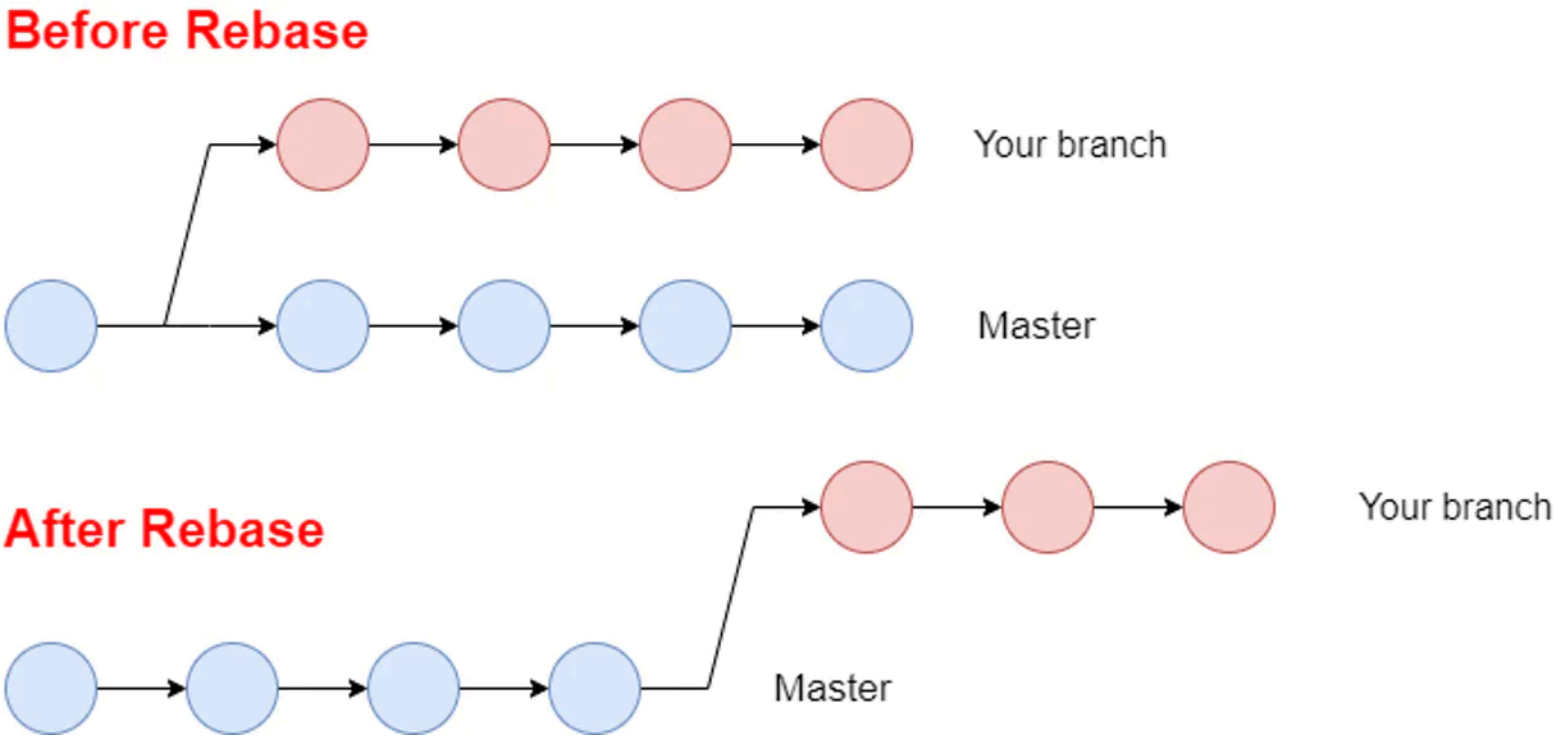
```
$ git diff
diff --git a/a.js b/a.js
index 01aafe1..2f6fc14 100644
--- a/a.js
+++ b/a.js
@@ -1,10 +1,4 @@
    this.retrieveUnitTypes = function(req, res, next) {

-     if (foo) {
-       return '1';
-     } else if (bar) {
-       return '2';
-     }
-
      return 3;
    };
```

Git branch

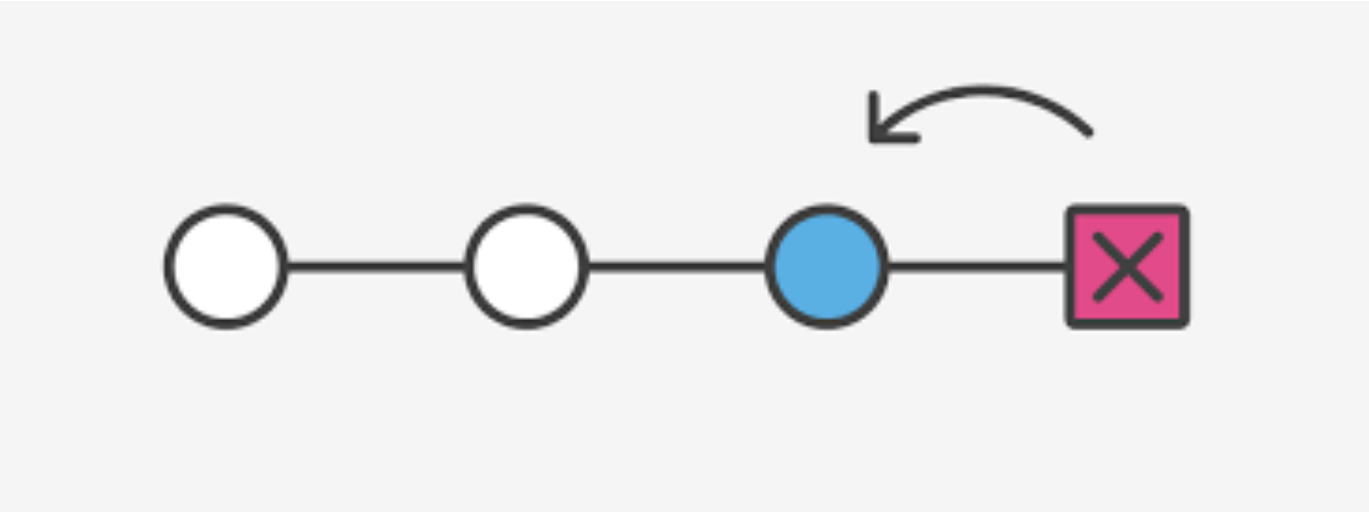
Git merge


4708008 Initial Commit - Chris <chrisgopher@rietta.com>
4e422bc 2 - Chris <chrisgopher@rietta.com>
e6f77fc Merge branch feature-branch - Chris <chrisgopher@rietta.com>
36db2f0 1 - Chris <chrisgopher@rietta.com>
a06f4b5 4 - Chris <chrisgopher@rietta.com>
master
dff160a 3 - Chris <chrisgopher@rietta.com>
feature-branch

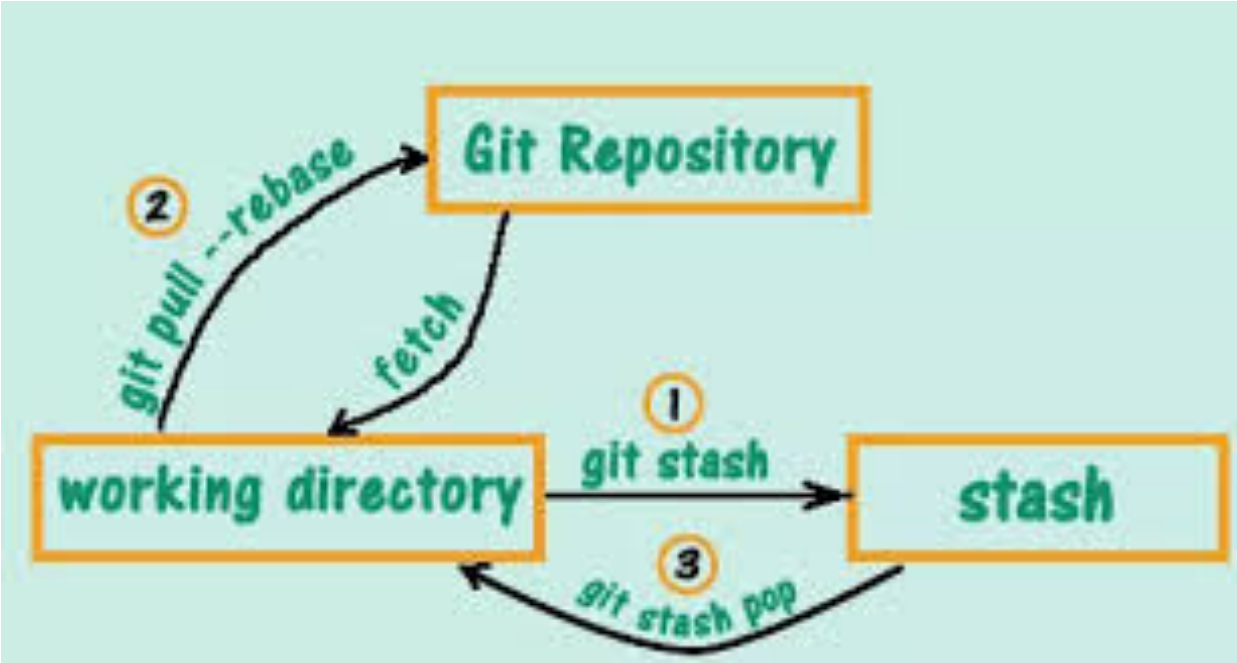
Git rebase

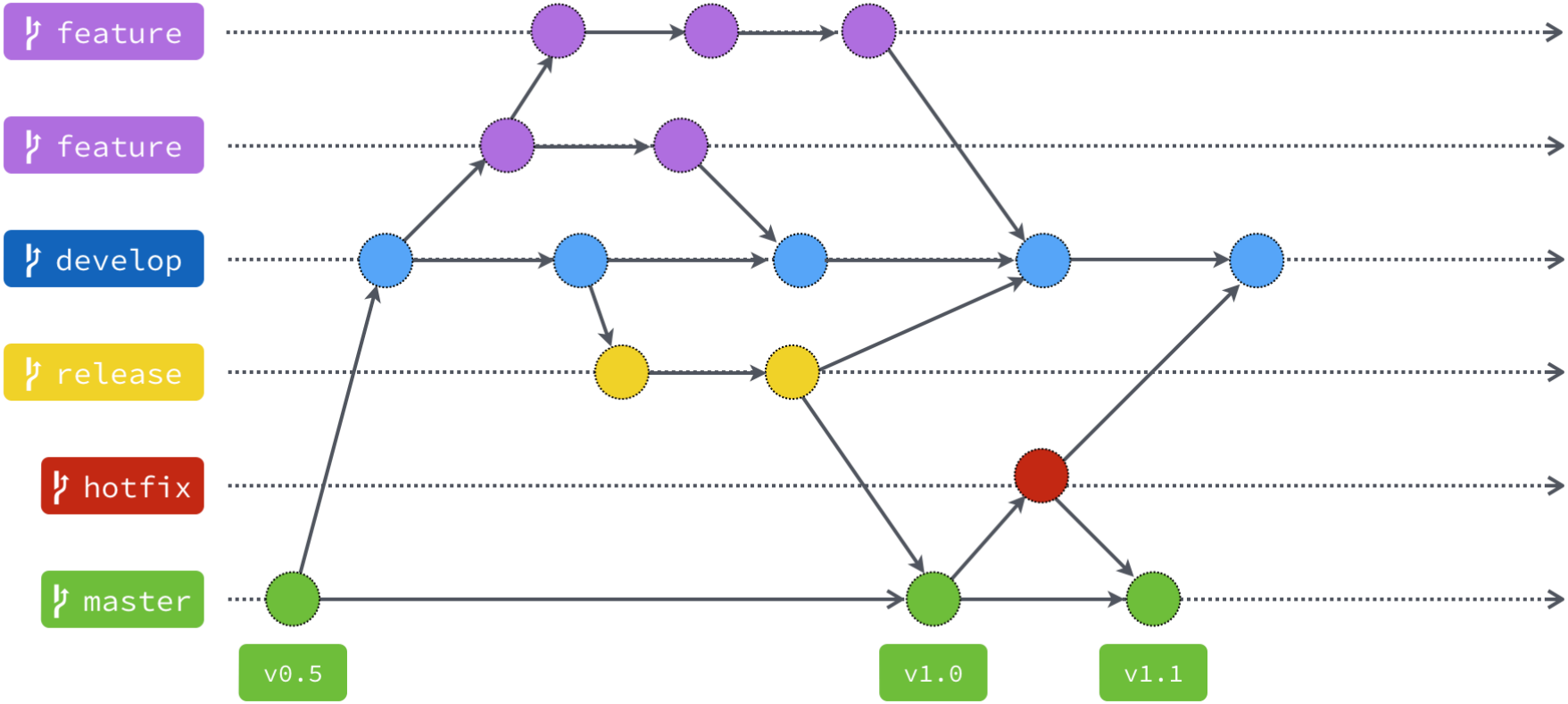

Before Rebase
Your branch
Master
After Rebase
Your branch
Master

Git revert

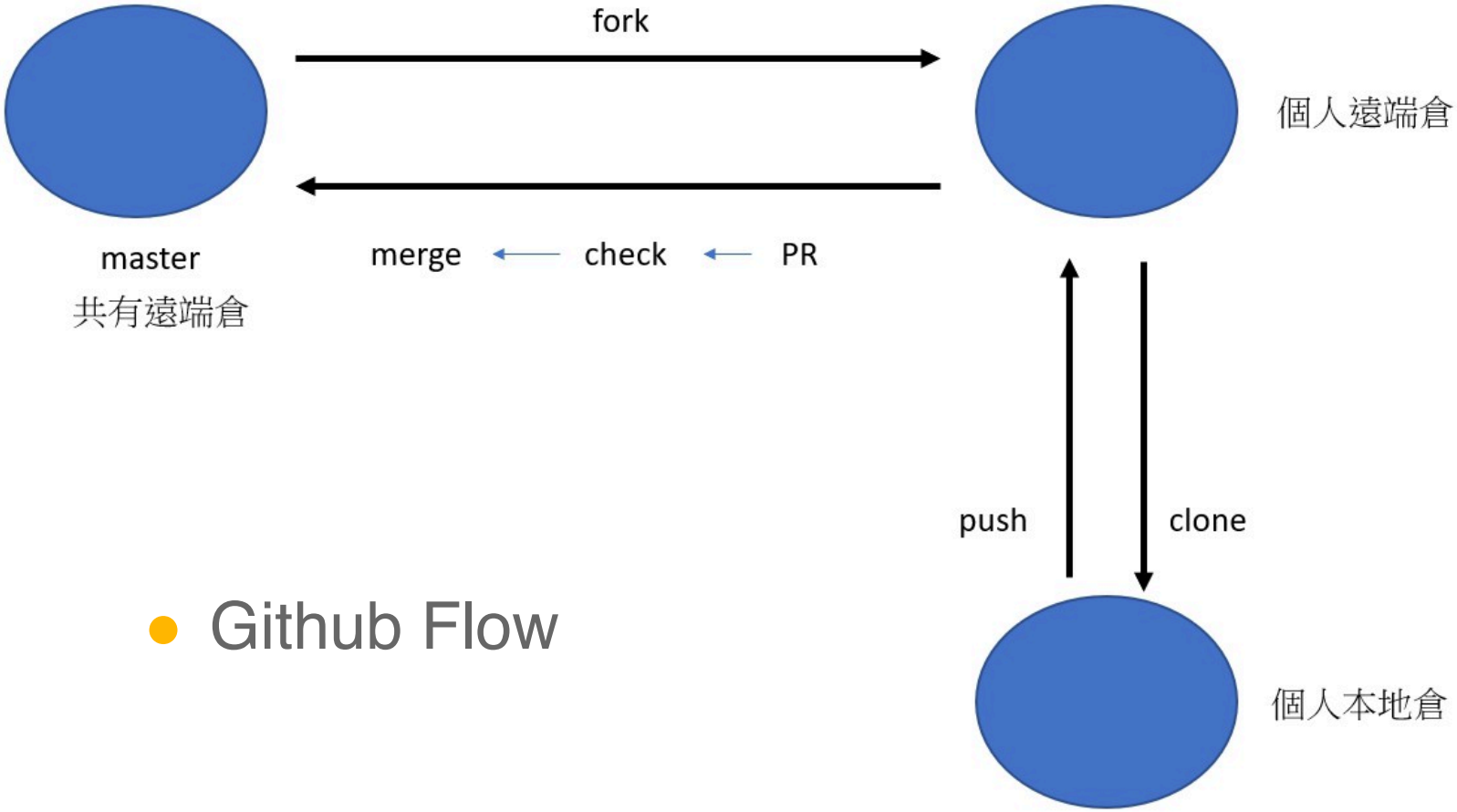


Git stash

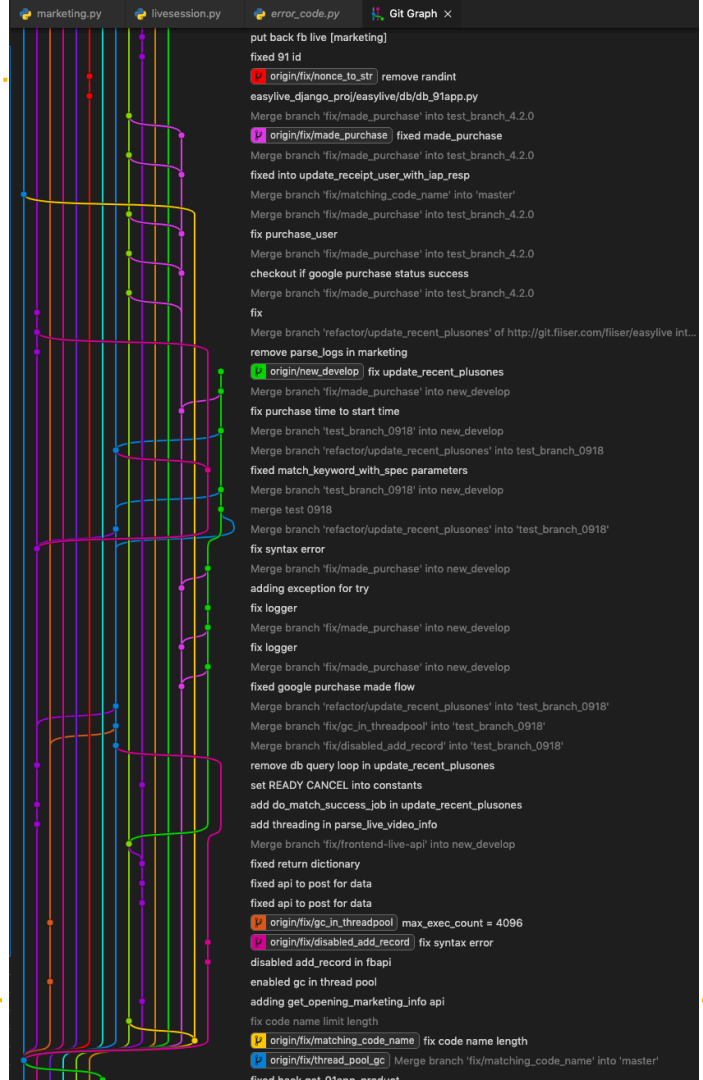

Git Repository
2
git pull --rebase
fetch
1
git stash
working directory
stash
3
git stash pop

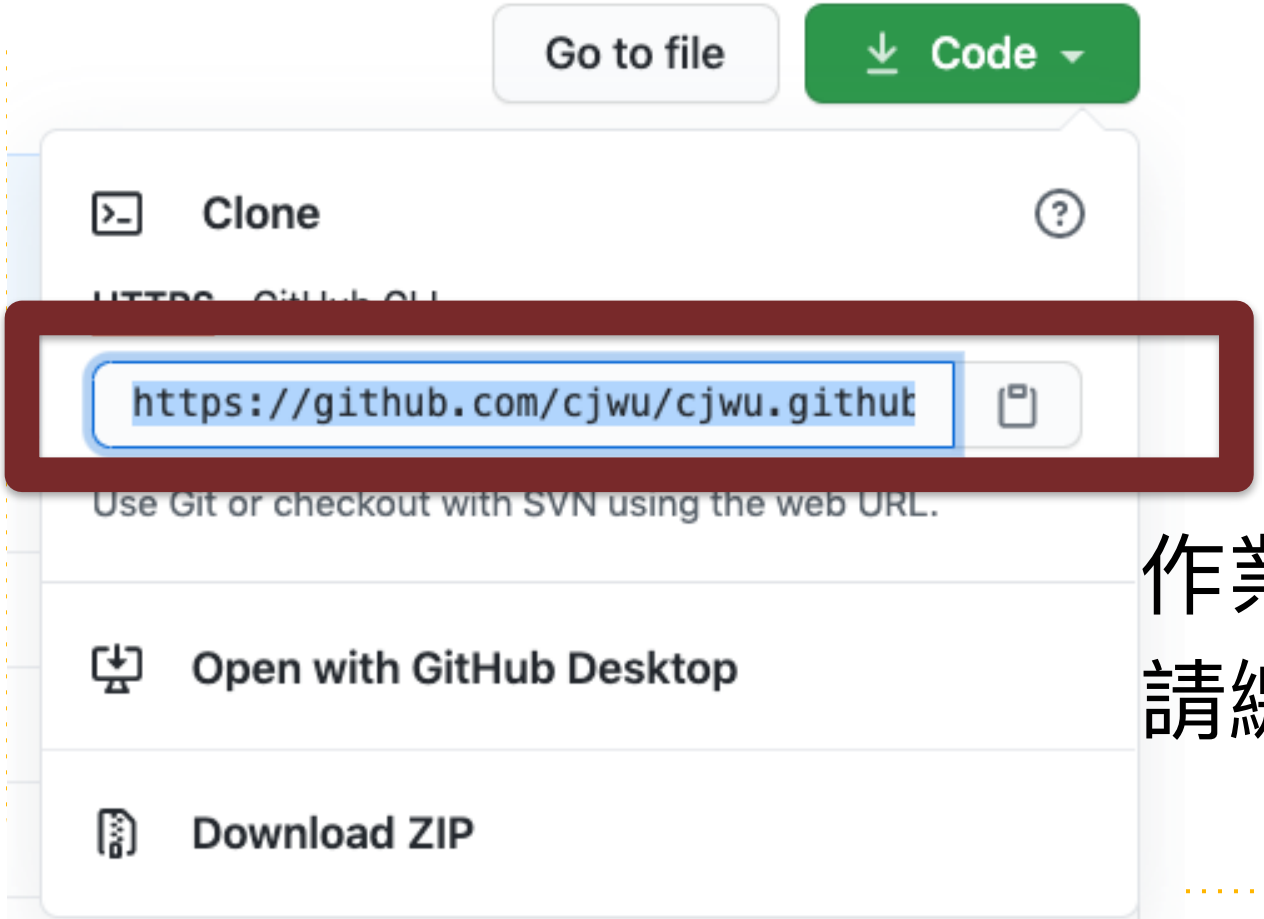


Git Flow



- Github Flow


marketing.py
livesession.py
error_code.py
Git Graph
put back fb live [marketing]
fixed 91 id
origin/fix/nonce_to_str remove randint
easylive_django_proj/easylive/db/db_91app.py
Merge branch 'fix/made_purchase' into test_branch_4.2.0
origin/fix/made_purchase fixed made_purchase
Merge branch 'fix/made_purchase' into test_branch_4.2.0
fixed into update_receipt_user_with_iap_resp
Merge branch 'fix/matching_code_name' into 'master'
Merge branch 'fix/made_purchase' into test_branch_4.2.0
fix purchase_user
Merge branch 'fix/made_purchase' into test_branch_4.2.0
checkout if google purchase status success
Merge branch 'fix/made_purchase' into test_branch_4.2.0
fix
Merge branch 'refactor/update_recent_plusones' of http://git.fiiser.com/fiiser/easylive int...
remove parse_logs in marketing
origin/new_develop fix update_recent_plusones
Merge branch 'fix/made_purchase' into new_develop
fix purchase time to start time
Merge branch 'test_branch_0918' into new_develop
Merge branch 'refactor/update_recent_plusones' into test_branch_0918
fixed match_keyword_with_spec parameters
Merge branch 'test_branch_0918' into new_develop
merge test 0918
Merge branch 'refactor/update_recent_plusones' into 'test_branch_0918'
fix syntax error
Merge branch 'fix/made_purchase' into new_develop
adding exception for try
fix logger
Merge branch 'fix/made_purchase' into new_develop
fix logger
Merge branch 'fix/made_purchase' into new_develop
fixed google purchase made flow
Merge branch 'refactor/update_recent_plusones' into 'test_branch_0918'
Merge branch 'fix/gc_in_threadpool' into 'test_branch_0918'
Merge branch 'fix/disabled_add_record' into 'test_branch_0918'
remove db query loop in update_recent_plusones
set READY CANCEL into constants
add do_match_success_job in update_recent_plusones
add threading in parse_live_video_info
Merge branch 'fix/frontend-live-api' into new_develop
fixed return dictionary
fixed api to post for data
fixed api to post for data
origin/fix/gc_in_threadpool max_exec_count = 4096
origin/fix/disabled_add_record fix syntax error
disabled add_record in fbapi
enabled gc in thread pool
adding get_opening_marketing_info api
fix code name limit length
origin/fix/matching_code_name fix code name length
origin/fix/thread_pool_gc Merge branch 'fix/matching_code_name' into 'master'



作業/隨堂練習
請繳交這個Link

課堂練習#0

1. 申請github帳號
2. Git 安裝 & try it
3. Commit a readme file to github

https://blog.techbridge.cc/2018/01/17/learning-programming-and-coding-with-python-git-and-github-tutorial/

Conclusion

- Make things and have fun

- Slack 私訊我或是助教

Thanks!

Open for any questions

CJ Wu

cjwu@mail.cgu.edu.tw