Tongle Yao 4/28/2019
Scantist Software Engineer (Data) Coding Assignment
Language: Python 3.7
Dependencies:
  urllib.request
  urllib.error
  bs4
  time
  json

Part 1

Part 1 requests me to collecting all release version number from three repo. The first idea came up was using GitHub API. I wrote a demo to connect their API, but it failed. I did not realize I need authentic before using their API.

Then I turned to write a python crawler to catch information from their website. Learning how to write a crawler took about 3 hours. I used 'urllib' to open each repo's release tag URL and used 'bs4' to find all 'h4' tag with 'flex-auto min-width-0 pr-2 pb-1 commit-title' as class name. Release version number should be able to find in it's child component. After putting all version number into dict, program should find 'next' bottom and acquire next page's URL. Keep doing this until no more next page. One special thing here is, GitHub's server has connection limit. Program should sleep 0.5s before it enters the next page. If it still happens, program will sleep 1s and retry.

After acquiring all release version tags, write dict into a json file called 'release_list.json'

Part 2

Part 2 requests me to format the data from the previous part. I referenced 'Software versioning' on Wikipedia. Output should be like 'v1.0.0-rc1'.

The most difficult part is all three projects have version format chaos at their early stage. The program analysis every version number’s major number and type and changes them to the same format. For some special cases like apache/kafka has a release version called ‘show’, I keep as it is.