

考題 for Backend Engineer

Programming

Part 1

Counting

Given a list of urls, print out the top 3 frequent filenames.

ex.

Given

```
1  urls = [  
2      "http://www.google.com/a.txt",  
3      "http://www.google.com.tw/a.txt",  
4      "http://www.google.com/download/c.jpg",  
5      "http://www.google.co.jp/a.txt",  
6      "http://www.google.com/b.txt",  
7      "https://facebook.com/movie/b.txt",  
8      "http://yahoo.com/123/000/c.jpg",  
9      "http://gliacloud.com/haha.png",  
10 ]
```

The program should print out

a.txt 3

b.txt 2

c.jpg 2

Combination

$$C(5,2) = 5!/(3!*2!)$$

The above equation is one way to compute combination

but we don't want compute in this way.

We change to another way to compute combination.

the new way to compute combination is as follows:

$$C(n,r) = C(n-1, r) + C(n-1, r-1)$$

$$C(n,1) = n$$

$$C(n,n) = 1$$

While you refer the above equation to compute $C(990, 33)$, python will take too much time.

You **must** find some way to boost its performance.

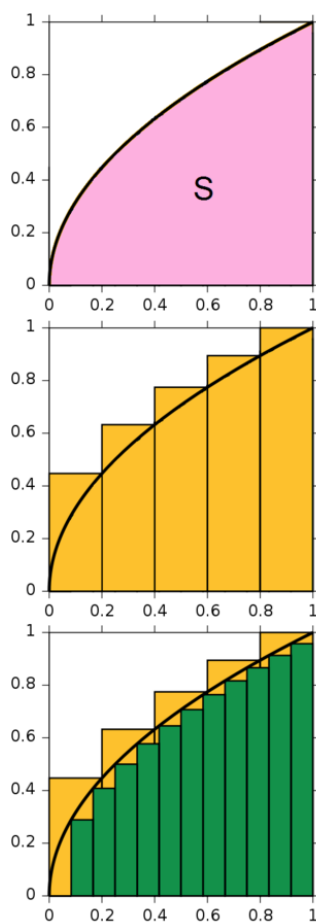
Integration

https://upload.wikimedia.org/wikipedia/commons/thumb/5/54/Integral_approximations-3-steps.png/320px-Integral_approximations-3-steps.png

Please try to add 1~3 line of code to finish the integration

```
def anonymous(x):  
    return x**2 + 1  
  
def integrate(fun, start, end):  
    step = 0.1  
    intercept = start  
    area = 0  
    while intercept < end:  
        intercept += step  
        ''' your work here '''  
    return area  
  
print(integrate(anonymous, 0, 10))
```

1 ■



Part 2

請用 Python 寫出一個可以爬 ptt 任意看板
(<https://www.ptt.cc>) 的爬蟲程式，可以使用任意
Python 套件

以下欄位為必要

- 日期
- 作者
- 標題
- 內文
- 看板名稱

Design Doc (三選一)

1. 我們的客戶很開心的透過 GliaStudio 把一堆影片做好了之後，他們希望可以把影片上架到各大平台 (Youtube、Facebook、優酷 ... 等)，你可以設計一個系統來讓客戶更方便的上架嗎？
2. 請設計一個咖啡廳分享、評分平台的 Backend，請考慮會有 APP 以及 Web 當作 client
3. 請設計一個新聞、論壇爬蟲系統，會定期抓取各大網站最新的資料，並且可以讓使用者檢索