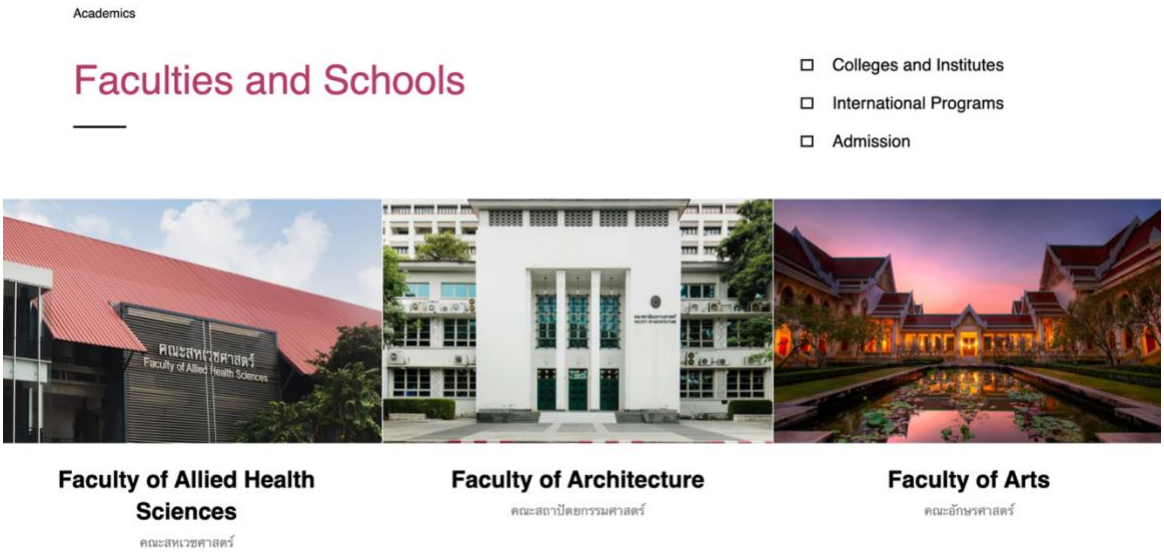
Assignment 8: Web scraping

การบ้านนี้จะให้นิสิตได้ทดลองเขียนโปรแกรมเพื่อดึงข้อมูลประเภทต่าง ๆ จากเว็บไซต์ โดยให้นิสิตทดลองดึงข้อมูลจาก https://waiiinta.github.io/ ซึ่งคัดลอกมาจากเว็บไซต์ของจุฬา



ในการบ้านนี้นิสิตจะต้องดึงข้อมูลจากเว็บไซต์ดังกล่าว ทั้งหมดสามประเภท

1. ชื่อคณะ get_faculty_names() คืน list ที่มีชื่อคณะภาษาอังกฤษทั้งหมด **คณะจะต้องขึ้นต้นด้วยคำว่า Faculty of เท่านั้น**
2. รูปของคณะ download_faculty_images() ดาวโหลดรูปจากคณะต่างๆมาเซฟลงในเครื่องของนิสิต
3. เบอร์โทรของคณะ print_faculty_numbers() แสดงชื่อคณะและเบอร์โทรในรูปแบบที่กำหนด

เว็บไซต์จะถูกกำหนดรูปแบบไว้ด้วยภาษา html สามารถเรียนรู้เรื่องภาษา html เบื้องต้นได้จาก คลิปวิดีโอ https://youtu.be/ZfxA3AxGbHE

ทั้งนี้ นิสิตจะต้อง import library urllib เพิ่มเติมเพื่อดึงข้อมูลผ่านเวบไซต์ด้วยคำสั่งเหล่านี้

```
import urllib
import urllib.request as urq
html = str(urq.urlopen(url).read().decode('utf-8'))
```

คำอธิบายเพิ่มเติม

urq.urlopen(url).read() จะอ่านข้อมูลจากurlออกมา เราเปลี่ยนรูปแบบข้อมูลให้เป็น utf-8 ด้วย .decode('utf-8') เพื่อให้อ่านข้อความภาษาไทยได้ สุดท้ายเปลี่ยนให้กลายเป็น str ด้วย str()

ถ้าอ่าน url อื่นที่ไม่ใช่ .html แต่เป็น binary file เช่น รูปภาพ ไม่ต้องเติม .decode('utf-8')

รูปแบบการคืนค่าของฟังก์ชัน

get_faculty_names()

- list ของชื่อคณะโดยต้องมีการใช้ตัวพิมพ์เล็กและพิมพ์ใหญ่ถูกต้องตามในหน้าเวบ
- ไม่มี space หรือ tab ใดๆก่อนหรือหลังชื่อ
- ตัวอย่าง ['Faculty of Allied Health Sciences', 'Faculty of Architecture']

save_image()

- อ่านรูปตาม url ที่ให้และเซฟลงเครื่องตามชื่อที่กำหนด

download_faculty_images()

- เรียกใช้ save_image()
- save ไฟล์รูปภาพของแต่ละคณะ คณะละหนึ่งรูป ลงในโฟลเดอร์ปัจจุบันบนเครื่อง
- ในเวบจะมีรูปคณะหลายขนาด ให้เซฟเฉพาะรูปขนาด 300x188 มาเท่านั้น
- ใช้ชื่อไฟล์เดิมของรูปภาพนั้นๆ

print_faculty_numbers()

- เลขโทรศัพท์สามารถค้นหาได้โดยตามลิงค์ไปที่หน้าเวบของแต่ละคณะ (มีลิงค์อยู่กับรูปแต่ละคณะ)
- แสดงผลทางหน้าจอ โดยพิมพ์ชื่อและเบอร์โทรศัพท์ตามตัวอย่าง
- ใช้ 0 แทน +66
- แสดงเบอร์หลังส่วน Contact
- ในกรณีที่มีหลายเบอร์ ให้แสดงแค่เบอร์แรกเท่านั้น ไม่ต้องมีส่วน ext.
- ไม่มีตัวเลขแปลกปลอมหรือการเว้นวรรคอื่นใด
- การแสดงผลควรเหมือนกับในไฟล์ตัวอย่าง tel.txt ที่มีให้

หน้าเวบจะถูกเปลี่ยนเมื่อถึงเวลาให้คะแนน นิสิตควรเขียนโค้ดที่รองรับความเปลี่ยนแปลงของหน้าเวบได้ เช่น ถ้ามีการเพิ่มชื่อคณะใหม่ หรือ มีการเปลี่ยนเบอร์โทร โค้ดก็ยังสามารถทำงานได้อย่างถูกต้อง ทั้งนี้ รูปแบบของเวบจะยังคงเดิม

งานของคุณ

จากโปรแกรมต้นฉบับข้างล่างนี้จงเขียนชุดคำสั่งให้กับฟังก์ชัน 3 ฟังก์ชันในบริเวณสีเขียว (20 คะแนน ไม่มีโบนัส)

```
# Prog-08: Web Scraping
# Fill in your ID & Name
# ...
# Declare that you do this by yourself
import urllib
import urllib.request as urq

def load_html(url):
  return str(urq.urlopen(page_url).read().decode('utf-8'))
#------------------------------------------------
def get_faculty_names(url):
  return ['Faculty of Allied Health Sciences', 'Faculty of Architecture']

def save_image(img_url, filename):
  pass

def download_faculty_images(url):
  pass

def print_faculty_numbers(url):
  print("faculty-of-allied-health-sciences")
  print("0 2218 1100")
#------------------------------------------------
def main():
  pageurl = "https://waiiinta.github.io/"

  print(get_faculty_names(pageurl))

  download_faculty_images(pageurl)

  print_faculty_numbers(pageurl)
#------------------------------------------------
main()
```

ห้ามเปลี่ยนโค้ดส่วนสีแดง

load_html() จะเปิดเวบและคืนหน้าเวบเป็น str

เขียนชุดคำสั่งบริเวณนี้เท่านั้น
จะเขียนฟังก์ชันอื่นเพิ่มเติมในบริเวณนี้ก็ได้
ไม่ใช้ตัวแปรใด ๆ ที่อยู่นอกฟังก์ชัน

สามารถดาวน์โหลดโค้ดและไฟล์ต่างๆได้ที่ คลิกเพื่อดาวโหลด

ปฏิบัติตามอย่างเคร่งครัด

- ห้ามเขียนคำสั่งในฟังก์ชันที่ไปใช้ตัวแปรที่อยู่นอกฟังก์ชัน (หรือที่เรียกว่าตัวแปร global)
- อนุญาตให้เพิ่มคำสั่งในบริเวณพื้นสีเขียวเท่านั้น จะเขียนฟังก์ชันเพิ่มเติมก็ได้
- ตั้งแต่ชื่อแฟ้ม ให้ถูกต้องตามที่เขียนใน CourseVille
- เพิ่มเติมข้อมูลใน comment ต้นโปรแกรมให้ตรงตามความจริง
- ก่อนส่งโปรแกรม ควรสั่งทำงานโปรแกรมที่ส่งอีกครั้ง ว่าทำงานได้ตามที่ต้องการ