

Programming Assignment 5

Wikipedia API

จุดประสงค์ของ PA นี้

1. ฝึกการออกแบบ class และ subclass เพื่อประหยัด code และลดโอกาสการผิดพลาด
2. ฝึกการอ่าน documentation ของ API และการใช้ API เพื่อดึงข้อมูลที่ต้องการ

Wikipedia article

บทความของ wikipedia นั้นจะมี title (หัวข้อของบทความ) และ pageid (หมายเลขประจำตัวของบทความ) ซึ่งทั้งสองอย่างสามารถใช้ชี้ตัวบทความได้ (identification) เช่น บทความของมหาวิทยาลัยเป็นภาษาอังกฤษ

Chulalongkorn University

URL คือ https://en.wikipedia.org/wiki/Chulalongkorn_University

Title คือ Chulalongkorn_University

Page id ต้องใช้ API ในการเปิดหา แต่เราสามารถ Title เพื่อระบุบทความที่ต้องการได้ กล่าวคือไม่มีบทความอื่นที่ title เป็น Chulalongkorn_University

จงเขียน module ที่มี class ที่ไว้สำหรับดึงข้อมูลดังต่อไปนี้

Method 1: English Search

Search หา summary ของบทความ wikipedia ที่เกี่ยวข้องกับคำที่ต้องการหา (query) มากที่สุด summary คือประโยคแรกของ wikipedia article และจะต้องได้ออกมาเป็น tuple ของภาษาอังกฤษและภาษาไทย ตัวอย่างเช่น

```
>>> query = 'Siam'
```

```
>>> wiki.search(query, otherthings, otherthings2)
```

```
("Thailand ( TY-land), officially the Kingdom of Thailand and formerly known as Siam, is a country at the center of the Southeast Asian Indochinese peninsula composed of 76 provinces.",
```

```
ประเทศไทย มีชื่ออย่างเป็นทางการว่า ราชอาณาจักรไทย เป็นราชชาติอันตั้งอยู่ในภูมิภาคเอเชียตะวันออกเฉียงใต้)
```

นิสิตจะต้อง call API สามครั้งเพื่อให้ได้ข้อมูลครบ

1. Search API เพื่อหา title (เช่น *Thailand*) ที่ใกล้เคียงที่สุดกับ query (เช่น *Siam*)
2. Parsing API เพื่อหา langlinks ของบทความเดียวกันแต่เป็นภาษาไทย (เช่น <https://th.wikipedia.org/wiki/%E0%B8%9B%E0%B8%A3%E0%B8%B0%E0%B9%80%E0%B8%97%E0%B8%A8%E0%B9%84%E0%B8%97%E0%B8%A2>) ซึ่งมี title ของภาษาไทย (เช่น ประเทศไทย)
ตัวพวกเปอร์เซนต์ยี่สิบคือมันเป็นภาษาไทยใน URL ซึ่งสามารถ convert เป็น unicode string ที่เราสามารถอ่านออกได้

```
>>> import urllib
```

```
>>> url =
```

```
'https://th.wikipedia.org/wiki/%E0%B8%9B%E0%B8%A3%E0%B8%B0%E0%B9%80%E0%B8%97%E0%B8%A8%E0%B9%84%E0%B8%97%E0%B8%A2'
```

```
>>> urllib.parse.unquote(url)
```

```
'https://th.wikipedia.org/wiki/ประเทศไทย'
```

3. ใช้ <https://www.mediawiki.org/wiki/Extension:TextExtracts#API> ในการดึงประโยคแรกของบทความที่เป็นภาษาอังกฤษออกมา

Thailand (TY-land), officially the Kingdom of Thailand and formerly known as Siam, is a country at the center of the Southeast Asian Indochinese peninsula composed of 76 provinces.

4. เหมือนกับ 3 แต่ว่าดึงของภาษาไทยออกมา ภาษาไทยมันแบ่งประโยคไม่เป็น นิสิตสามารถกะๆ เอาโดยใช้กฎง่ายๆ ในการหาว่าประโยคแรกจบเมื่อไร กฎนั้นไม่ต้องถูกเป๊ะก็ได้ หยวนๆ เช่น หยิบมาจนถึงการเว้นวรรคครั้งที่สี่ เป็นต้น

ประเทศไทย มีชื่ออย่างเป็นทางการว่า ราชอาณาจักรไทย เป็นราชอาณาจักรอันตั้งอยู่ในภูมิภาคเอเชียตะวันออกเฉียงใต้

Method 2: Thai Search

ผลลัพธ์ออกมาเหมือนกับ English search แต่ว่า query ที่ใช้ search นั้นเป็นภาษาไทย

ให้สังเกตว่ากระบวนการมันคล้ายๆ กับ English search มากเราควรจะต้องพยายาม reuse code ให้ได้มากที่สุด

Method 3: What is X? (นิสิตป.โทเท่านั้น)

Wikipedia article ทุกบทความมี categories ซึ่งสามารถใช้ตอบคำถาม What is X ได้ เช่น

```
>>> english_wiki.what_is('Thailand', otherparams1, otherparams2)
```

```
['Thailand', 'Countries in Asia', 'Kingdoms,' 'Member states of the Association of Southeast Asian Nations', 'Member state of the United Nations', 'Southeast Asian countries']
```

สำหรับ Method นี้เราต้องใช้ Search API เพื่อหา title จากนั้น Parse API เพื่อหา categories จากนั้นให้กรอง categories ที่ hidden ออกไป

Method 4: X คืออะไร (นิสิตป.โทเท่านั้น)

เหมือนกับ Method 3 แต่ว่า search query เป็นภาษาไทย และคำตอบต้องเป็นภาษาไทย

```
>>> thai_wiki.what_is('ฝรั่งเศส', otherparams1, otherparams2)
```

```
['ประเทศในระบบซีวิลลอว์', 'สาธารณรัฐ']
```

การส่งงาน

ให้ save เก็บไว้ในไฟล์ firstname_PA5.py และทุกคลาส ทุก method จะต้อง มี comment ครบถ้วน เพราะฉะนั้นนิสิตเป็นผู้ออกแบบคลาสเองทั้งหมด