Thực hành xử lý văn bản

Nguyễn Mạnh Hiển
hiennm@tlu.edu.vn

Cài đặt NLTK

- NLTK là thư viện xử lý ngôn ngữ tự nhiên.
- Kiểm tra xem máy tính của bạn đã cài NLTK chưa:
  - Gõ lệnh **`import nltk`** ở dấu nhắc lệnh Python.
  - Không thấy báo lỗi gì nghĩa là đã cài NLTK.
- Nếu máy tính của bạn chưa cài NLTK, gõ lệnh **`pip install nltk`** ở cửa sổ lệnh Windows (không gõ ở dấu nhắc lệnh Python).
- Ở dấu nhắc lệnh Python, gõ các lệnh sau đây để cài dữ liệu đi kèm của NLTK:

```
import nltk
nltk.download('popular')
```

Tải một văn bản trên web xuống

```
>>> from urllib import request
>>> url = 'http://www.gutenberg.org/files/2554/2554-0.txt'
>>> response = request.urlopen(url)
>>> raw = response.read().decode('utf8')
>>> type(raw)
<class 'str'>
>>> len(raw)
1176893
>>> raw[1:74]
'The Project Gutenberg EBook of Crime and Punishment, by
Fyodor Dostoevsky'
```

Tách từ

(Gõ tiếp từ slide trước...)

```
>>> tokens = nltk.word_tokenize(raw)
>>> type(tokens)
<class 'list'>
>>> len(tokens)
257727
>>> tokens[1:10]
['The', 'Project', 'Gutenberg', 'EBook', 'of', 'Crime',
'and', 'Punishment', ',', 'by']
```

Đọc văn bản từ HTML

- Kiểm tra thư viện Beautiful Soup đã cài chưa, bằng cách thử lệnh **`from bs4 import BeautifulSoup`** ở dấu nhắc Python.
- Nếu chưa cài Beautiful Soup thì cài bằng lệnh **`pip install beautifulsoup4`** ở dấu nhắc lệnh Windows.

```
>>> url = 'http://news.bbc.co.uk/2/hi/health/2284783.stm'
>>> html = request.urlopen(url).read().decode('utf8')
>>> html[:60]
'<!doctype html public "-//W3C//DTD HTML 4.0 Transitional//EN'
>>> from bs4 import BeautifulSoup
>>> raw = BeautifulSoup(html, 'html.parser').get_text()
>>> tokens = nltk.word_tokenize(raw)
>>> tokens[:10]
['BBC', 'NEWS', '|', 'Health', '|', 'Blondes', "'to", 'die',
'out', 'in']
```

Tách gốc từ dùng thuật toán Porter

```
>>> raw = 'DENNIS: Listen, strange women lying in ponds
distributing swords is no basis for a system of
government.  Supreme executive power derives from a
mandate from the masses, not from some farcical aquatic
ceremony.'
>>> tokens = nltk.word_tokenize(raw)
>>> porter = nltk.PorterStemmer()
>>> [porter.stem(t) for t in tokens]
['denni', ':', 'listen', ',', 'strang', 'women', 'lie',
'in', 'pond', 'distribut', 'sword', 'is', 'no', 'basi',
'for', 'a', 'system', 'of', 'govern', '.', 'suprem',
'execut', 'power', 'deriv', 'from', 'a', 'mandat',
'from', 'the', 'mass', ',', 'not', 'from', 'some',
'farcic', 'aquat', 'ceremoni', '.']
```

Bài tập

Bài 1: Cho một văn bản. Viết các câu lệnh Python để tách ra các từ riêng biệt và đếm xem mỗi từ đó xuất hiện bao nhiêu lần trong văn bản. *Gợi ý: Dùng kiểu dữ liệu từ điển (dictionary) đã học trong buổi thực hành trước*.

Bài 2: Cho một ***tập*** văn bản. Yêu cầu giống như bài 1. Yêu cầu bổ sung là đếm xem mỗi từ đó xuất hiện trong bao nhiêu văn bản khác nhau.

Tự tìm hiểu thêm

- Đọc ở đây: https://www.nltk.org/book/ch03.html
- Phần 3.1: Đọc phần “Reading Local Files” về cách mở file văn bản đang có trên ổ cứng.
- Phần 3.2: Các hàm xử lý xâu ký tự trong Python. (Xem lướt qua)
- Phần 3.4 và 3.7: Biểu thức chính quy (regular expression) cho phép tách ra các dãy ký tự hoặc các từ thỏa mãn tiêu chuẩn nào đó.