Lab 2: Kho ngữ liệu

Nội dung

1. Khái niệm 22
2. Các Corpora có sẵn của NLTK 22
2.1. Gutenberg Corpus 22
2.2. Web và Chat Text Corpus 25
2.3. Brown Corpus 26
2.4. Reuters Corpus 28
2.5. Annotated Text Corpus 29
3. Bài tập 29

1. Khái niệm

Corpus/corpora là các dữ liệu văn bản, ngôn ngữ đã được số hoá (***kho ngữ liệu***). Corpora thường là các dữ liệu đã được xử lý, được sử dụng như đầu vào của các thuật toán trong NLP.

Ví dụ: NLTK Book là một Corpora được cung cấp sẵn.

2. Các Corpora có sẵn của NLTK

2.1. Gutenberg Corpus

Gutenberg là một dự án cung cấp 25.000 cuốn sách điện tử miễn phí (https://www.gutenberg.org/). NLTK đã lấy một phần nhỏ trong dự án này.

Để hiển thị các cuốn sách trong dự án Gutenberg mà NLTK cung cấp, chúng ta thực hiện như sau:

```
>>> import nltk

>>> nltk.corpus.gutenberg.fileids()

['austen-emma.txt',      'austen-persuasion.txt',      'austen-
sense.txt',   'bible-kjv.txt',   'blake-poems.txt',   'bryant-
```

```
stories.txt',  'burgess-busterbrown.txt',  'carroll-alice.txt',
'chesterton-ball.txt',   'chesterton-brown.txt',   'chesterton-
thursday.txt',       'edgeworth-parents.txt',       'melville-
moby_dick.txt',      'milton-paradise.txt',      'shakespeare-
caesar.txt',      'shakespeare-hamlet.txt',      'shakespeare-
macbeth.txt', 'whitman-leaves.txt']
```

Hay để ngắn gọn, ta có thể viết

```
>>> from nltk.corpus import gutenberg

>>> gutenberg.fileids()

['austen-emma.txt',      'austen-persuasion.txt',      'austen-
sense.txt', ...] >>> emma = gutenberg.words('austen-emma.txt')
```

Bây giờ hãy ta chọn một cuốn sách và đưa nội dung của nó vào một biến trong Python.

```
>>> emma = nltk.corpus.gutenberg.words('austen-emma.txt')

>>> len(emma)

192427
```

Bài tập: Với mỗi cuốn sách của Gutenberg mà NLTK cung cấp, hãy thống kê:

- Độ dài trung bình của mỗi từ.
- Độ dài trung bình của mỗi câu.
- Số lần xuất hiện trung bình của một từ.

Lời giải:

Bước 1: Ta lấy thông tin về tổng độ dài các từ, số từ đơn và số câu

Bước 2: Thực hiện tính toán

- Trung bình độ dài mỗi từ = Tổng số ký tự / Tổng số từ.
- Trung bình độ dài mỗi câu = Tổng số từ / Tổng số câu.
- Trung bình lần xuất hiện của mỗi từ = Tổng số từ đơn / Tổng số từ. Ta sẽ có

đoạn Code như sau:

```
from nltk.corpus import gutenberg

for fileid in gutenberg.fileids():

    num_chars = len(gutenberg.raw(fileid)) #Tính số lượng ký tự

    num_words = len(gutenberg.words(fileid)) #Tính số lượng từ đơn

    num_sents = len(gutenberg.sents(fileid)) #Tính số lượng câu.

    num_vocab = len(set([w.lower() for w in gutenberg.words(fileid)])) # Tính số lượng từ vựng trong văn bản

    print(int(num_chars / num_words), int(num_words / num_sents), int(num_words / num_vocab), fileid)

# Kết quả

4 24 26 austen-emma.txt

4 26 16 austen-persuasion.txt

4 28 22 austen-sense.txt

4 33 79 bible-kjv.txt

4 19 5 blake-poems.txt

4 19 14 bryant-stories.txt

4 17 12 burgess-busterbrown.txt

4 20 12 carroll-alice.txt

4 20 11 chesterton-ball.txt

4 22 11 chesterton-brown.txt
```

```
4 18 10 chesterton-thursday.txt

4 20 24 edgeworth-parents.txt

4 25 15 melville-moby_dick.txt

4 52 10 milton-paradise.txt

4 11 8 shakespeare-caesar.txt

4 12 7 shakespeare-hamlet.txt

4 12 6 shakespeare-macbeth.txt

4 36 12 whitman-leaves.txt
```

Lưu ý:

- **gutenberg.raw()** không trả lại list các Token mà trả lại tổng số lượng ký tự, có chứa cả các dấu cách.
- **gutenberg.sents()** sẽ trả lại một list các câu và mỗi câu là một list các từ.

2.2. Web và Chat Text Corpus

- Gutenberg cung cấp rất nhiều cuốn sách, đó là các văn bản mang tính trang trọng, các cuốn cách kinh điển.
- Ta muốn có thêm các văn bản, các đoạn hội thoại trên Web hay mạng xã hội. NLTK cũng cung cấp các văn bản này.

```
>>> from nltk.corpus import webtext

>>> for fileid in webtext.fileids():

... print fileid, webtext.raw(fileid)[:65], '...'

#Kết quả

firefox.txt Cookie Manager: "Don't allow sites that set removed
cookies to se

grail.txt SCENE 1: [wind] [clop clop clop]
```

```
KING ARTHUR: Whoa there!  [clop

overheard.txt White guy: So, do you have any plans for this evening?

Asian girl

pirates.txt PIRATES OF THE CARRIBEAN: DEAD MAN'S CHEST, by Ted Elliott & Terr

singles.txt 25 SEXY MALE, seeks attrac older single lady, for discreet encoun

wine.txt Lovely delicate, fragrant Rhone wine. Polished leather and strawb
```

Chat Text:

```
>>> from nltk.corpus import nps_chat

>>> chatroom = nps_chat.posts('10-19-20s_706posts.xml')

>>> chatroom[123]

['i', 'do', "n't", 'want', 'hot', 'pics', 'of', 'a', 'female', ',', 'I', 'can', 'look', 'in', 'a', 'mirror', '.']
```

2.3. Brown Corpus

Đây là Corpus điện tử đầu tiên bằng tiếng Anh (1961), chứa văn bản từ 500 nguồn và các nguồn thì được chia mục.

```
>>> from nltk.corpus import brown

>>> brown.categories()

['adventure',   'belles_lettres',   'editorial',   'fiction', 'government', 'hobbies', 'humor', 'learned', 'lore', 'mystery',
```

```
'news', 'religion', 'reviews', 'romance', 'science_fiction']
```

Lấy ra các từ thuộc chuyên mục có tên là "news":

```
>>> brown.words(categories='news')

['The', 'Fulton', 'County', 'Grand', 'Jury', 'said', ...]
```

Lấy ra các câu, từ những chuyên mục khác nhau:

```
>>> brown.sents(categories=['news', 'editorial', 'reviews'])

[['The', 'Fulton', 'County', 'Grand', 'Jury', 'said', 'Friday',
'an', 'investigation', 'of', "Atlanta's", 'recent', 'primary',
'election', 'produced', '``', 'no', 'evidence', "''", 'that',
'any', 'irregularities', 'took', 'place', '.'], ['The', 'jury',
'further', 'said', 'in', 'term-end', 'presentments', 'that',
'the', 'City', 'Executive', 'Committee', ',', 'which', 'had',
'over-all', 'charge', 'of', 'the', 'election', ',', '``',
'deserves', 'the', 'praise', 'and', 'thanks', 'of', 'the',
'City', 'of', 'Atlanta', "''", 'for', 'the', 'manner', 'in',
'which', 'the', 'election', 'was', 'conducted', '.'], ...]
```

Ta làm một thống kê về cách sử dụng các động từ khuyết thiếu (Modal verb) trong các thể loại văn bản mà Brown Corpus cung cấp.

```
>>> import nltk

>>> from nltk.corpus import brown

>>> cfd = nltk.ConditionalFreqDist((genre, word) for genre in
brown.categories() for word in brown.words(categories=genre))

>>> genres = ['news', 'religion', 'hobbies', 'science_fiction',
'romance', 'humor']
```

```
>>> modals = ['can', 'could', 'may', 'might', 'must', 'will']
>>> print(cfd.tabulate(conditions=genres, samples=modals))
# Kết quả
                  can could   may might  must  will
           news    93    86    66    38    50   389
       religion    82    59    78    12    54    71
        hobbies   268    58   131    22    83   264
science_fiction    16    49     4    12     8    16
        romance    74   193    11    51    45    43
          humor    16    30     8     8     9    13
```

2.4. Reuters Corpus

Corpus nay chứa hơn 10.000 tin tức và 1.3 triệu từ. Nó được chưa thành 90 chủ đề và được chưa làm hai tập là "training" và "test" (để phục vụ cho việc sử dụng Học Máy). Các tài liệu sẽ được đánh dấu dạng "training/1234" hay "test/1234".

```
>>> from nltk.corpus import reuters
>>> reuters.fileids()
```

Ta cũng có thể xem các chuyên mục mà Reutuers cung cấp:

```
>>> reuters.categories()
```

Tuy nhiên khi xem chuyên mục của một tài liệu, ví dụ

```
>>> reuters.categories('training/9865')
```

```
['barley', 'corn', 'grain', 'wheat']
```

Ta thấy một tài liệu có thể thuộc nhiều chuyên mục khác nhau. Và cũng tương tự như các Corpus khác, ta có thể truy cập và xử lý các từ và câu.

2.5. Annotated Text Corpus

Các Corpus được đánh dấu, NLTK cung cấp rất nhiều các Corpus đã được đánh dấu thể hiện như POP Tags, name entities, syntactic structures... (http://www.nltk.org/data)

3. Bài tập

Sử dụng PlaintextCorpusReader của NLTK để đưa một văn bản thành Corpus.