Это код для этого эксперимента.

import re,requests

from bs4 import BeautifulSoup

from urllib.request import urlopen

html = urlopen('http://books.toscrape.com')

bs = BeautifulSoup(html, "html.parser")

bookList = bs.findAll('h3')

word = []

Note = open('x.txt',mode = 'w')

# for i in range(10) # for (let i=0; i < 10; i++) {}

for name in bookList:

word = name.get\_text()

Note.writelines(word)

print(word)

На этот раз вызывается библиотека re, request.

Используйте библиотеку BeautifulSoup в качестве основного метода парсинга.

выберите сайт：

Мы выбрали "http://books.toscrape.com" для сканируемого веб-сайта, который является относительно простым веб-сайтом, подходящим для новичков, которые только начинают изучать сканирование. На этом сайте нет механизма защиты от сканирования, так что лучше смириться.

Укажите парсер как "html.parser"

заключается в том, чтобы указать синтаксический анализатор Beautiful как "html.parser"，При этом обработанный результат сохраняется в bs.

Найдите нужные данные：

Здесь мне нужно найти название книги на веб-сайте, я открываю браузер, чтобы просмотреть исходный код, и обнаруживаю, что все названия находятся в теге <h3>, поэтому я использую оператор find\_All() для их извлечения.После операции find\_All() я распечатал результат. Обнаружено, что данные очень грязные, и тогда оператор .get\_text используется для «очистки» полученных данных. На данный момент у нас есть «чистые» данные.

хранилище данных：

Сначала создайте пустой список, используйте Note = open('x.txt', mode = 'w') для открытия пустого текста, а затем сохраните «очищенные» данные в список один за другим через цикл for, а затем оператор use Note.writelines() записывает данные, хранящиеся в списке, в текст .txt