from \_\_future\_\_ import division

import nltk, re, pprint

from urllib import urlopen

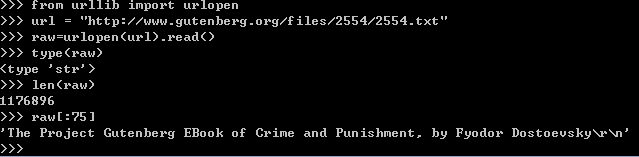
url = "http://www.gutenberg.org/files/2554/2554.txt"

raw=urlopen(url).read()

type(raw)

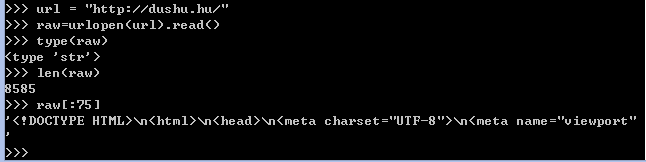
len(raw)

raw[:75]

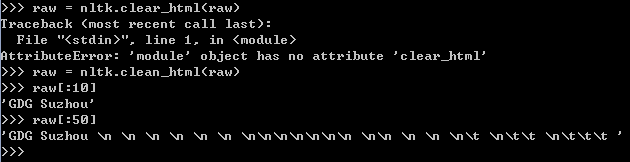


这样我们就加载了罪与罚的英文翻译。

试着加载一下dushu.hu，



加载出来的是html。。。。



使用clean\_html函数去掉html标记，但是还是有很多冗余。

关于如何分析html，可以使用beautifulSoup来处理，[http://www.crummy.com/software/BeautifulSoup/](http://www.crummy.com/software/BeautifulSoup/下载)  下载<http://www.crummy.com/software/BeautifulSoup/bs4/download/>

解压到D:\software\programming\Python27\Lib\site-packages文件夹

运行cmd，切换到安装目录下（根据自己解压缩后的目录和下载的版本号修改），

运行命令：

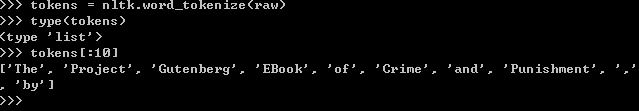
setup.py build

setup.py install

在IDE下from bs4 import BeautifulSoup，没有报错说明安装成功。

注意安装后文件在bs4文件夹中

对于自然语言处理，将字符串分解为词和标点成为分词。



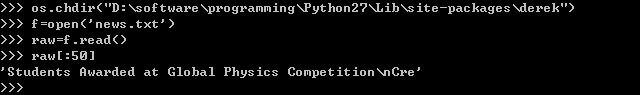
处理搜索引擎的结果

处理RSS订阅的结果

以上两部分跳过

读取本地文件

切换到文件目录然后读取



open(‘document.txt’,’rU’)r表示只读方式打开，U表示通用，忽略换行符公约

strip方法可以删除输入行尾的换行符。

捕获用户输入

