Homework 2: 文本处理

学号：16307130194 姓名：陈中钰

In [1]:
```
%matplotlib inline
```

导入nltk库和nltk.book语料：

In [2]:
```
import nltk
#nltk.download()
from nltk.book import *
```

```
*** Introductory Examples for the NLTK Book ***
Loading text1, ..., text9 and sent1, ..., sent9
Type the name of the text or sentence to view it.
Type: 'texts()' or 'sents()' to list the materials.
text1: Moby Dick by Herman Melville 1851
text2: Sense and Sensibility by Jane Austen 1811
text3: The Book of Genesis
text4: Inaugural Address Corpus
text5: Chat Corpus
text6: Monty Python and the Holy Grail
text7: Wall Street Journal
text8: Personals Corpus
text9: The Man Who Was Thursday by G . K . Chesterton 1908
```

导入正则表达式模块：

In [3]:
```
import re
```

问题1

说明以下的正则表达式匹配的字符串类： [a-zA-Z]+； [A-Z][a-z]*； p[aeiou]{,2}t； \d+(\.\d+)?；([^aeiou][aeiou][^aeiou])* ；\w+|[^\w\s]+。

[a-zA-Z]+

匹配的字符串是：1个或多个连续的字母（大写/小写都可以）。匹配的是连续英文字母组成的词。

[A-Z][a-z]*

匹配的字符串是：第1个字符是大写字母，后面连着0个或多个连续的小写字母。匹配的是titlecase的词。

p[aeiou]{,2}t

匹配的字符串是：以字母p开头，接上0~2个连续的a、e、i、o、u这5个字母中的任意字母，最后接上1个字母t。匹配的是以字母p开头、字母t结尾、中间有0~2个元音字母的词语。

\d+(\.\d+)?

匹配的字符串是：1个或多个连续的数字，后面可以接上这样的部分——字符'.'后连着1个或多个连续的数字，也可以没有这个部分。匹配的是数字，可以是整数，也可以是浮点数。

([^aeiou][aeiou][^aeiou])*

匹配的字符串是：0个或多个连续的这样的模块——由3个字符组成，第1、3个字符是除了a、e、i、o、u以外的任意字符，第2个字符是a、e、i、o、u中的任一字母。匹配的是0个或多个连续的上述模块组成的字符串。

\w+|[^\w\s]+

匹配的字符串是：1个或多个连续的字母、数字、下划线，或者是1个或多个连续的除了字母、数字、下划线、空白字符以外的字符。

问题2

创建一个文件，包含词汇和（任意指定）频率，其中每行包含一个词，一个空格和一个正整数，如：fuzzy 53。使用open(filename).readlines()将文件读入Python 链表。接下来，使用split()将每一行分成两个字段，并使用int()将其中的数字转换为一个整数。结果要求是链表形式：[['fuzzy', 53], ...]。

由于text8的长度最短，处理速度最快，接下来使用text8来制作题目要求的词频文件。

In [4]:
```
len(text8)
```

Out[4]: 4867

使用FreqDist()统计text8中的词语的词频，同时使用正则表达式选出只由英文字母组成的词语，可以过滤掉含有标点符号、只有标点符号的词。获得词频统计后，调用most_common()获得按照词频从大到小排序的词语及对应词频的列表。为了使生成文件的篇幅短，便于在报告中展示，只选择了词频大小前20的词语来生成文件。按照题目要求处理词语和词频的格式，并生成文件，其中文件的每行包含1个词、1个空格和1个正整数（词频）。生成的文件为'flist.txt'。

In [5]:
```
flist = FreqDist([w for w in text8 if re.search('^[A-Za-z]+$', w)]).most_common(20)
flist = ['{} {}\n'.format(w, freq) for (w, freq) in flist]
with open('flist.txt', 'w') as f:
    f.writelines(flist)
```

展示按照题目要求生成的flist.txt文件：

In [6]:
```
with open('flist.txt', 'r') as f:
    for line in f:
        print(line, end='')
```

```
for 99
and 74
to 74
lady 68
seeks 60
a 52
with 44
S 36
ship 33
relationship 29
fun 28
in 27
slim 27
build 27
o 26
s 24
y 23
smoker 23
non 22
I 22
```

使用open(filename).readlines()把整个文件读取进列表，把列表中的每行使用split()分成2个字段，并使用int()将其中的数字string转换为整数，生成最终列表结果：

In [7]:
```
lines = open('flist.txt').readlines()
lines = [line.split() for line in lines]
lines = [[w, int(freq)] for (w, freq) in lines]
print(lines)
```

```
[['for', 99], ['and', 74], ['to', 74], ['lady', 68], ['seeks', 60], ['a', 52], ['wi
th', 44], ['S', 36], ['ship', 33], ['relationship', 29], ['fun', 28], ['in', 27],
['slim', 27], ['build', 27], ['o', 26], ['s', 24], ['y', 23], ['smoker', 23], ['no
n', 22], ['I', 22]]
```

问题3

定义一个变量silly 包含字符串：'newly formed bland ideas are inexpressible in an infuriating way'。编写代码执行以下任务：分割silly 为一个字符串链表，每一个词一个字符串，使用Python 的split()操作，并保存到叫做bland 的变量中；提取silly 中每个词的第二个字母，将它们连接成一个字符串，得到'eoldrnnnna'；使用join()将bland 中的词组合成一个单独的字符串。确保结果字符串中的词以空格隔开。

定义silly字符串：

In [8]:
```
silly = 'newly formed bland ideas are inexpressible in an infuriating way'
silly
```

Out[8]:
```
'newly formed bland ideas are inexpressible in an infuriating way'
```

使用split()操作把silly字符串分割为字符列表，每个词一个string，并保存到bland中：

In [9]:
```
bland = silly.split()
print(bland)
```

```
['newly', 'formed', 'bland', 'ideas', 'are', 'inexpressible', 'in', 'an', 'infuriat
ing', 'way']
```

提取bland中每个词的第2个字母，将它们连接成一个字符串，得到'eoldrnnnna'：

In [10]:
```
sec = ""
for w in bland:
    if len(w) >= 2:
        sec += w[1]
sec
```

Out[10]: 'eoldrnnnna'

使用join()将bland中的词用空格组合成一个单独的字符串：

In [11]:
```
' '.join(bland)
```

Out[11]: 'newly formed bland ideas are inexpressible in an infuriating way'

In [ ]: