Python程序设计

案例：词频统计程序

张 华

WHU

词频统计程序

需求

- 一篇文章，出现了哪些词？
- 哪些词出现得最多？
- 英文文本
- 中文文本

词频统计程序

设计

- 文本是保存在文件中的一个长字符串
- 从文本中分离出单词
- 统计单词出现的次数
- 用字典来保存单词及其出现次数

词频统计程序

实现1

- 统计《哈姆雷特》中出现次数最多的10个单词

```
f = open('hamlet.txt', 'r')
txt = f.read()
txt = txt.lower()
for c in '\'\"\\-!?$%()<>=_|{}~,;./*&@^':
    txt = txt.replace(c, ' ')

wordList = txt.split()
wordDict = {}
for word in wordList:
    wordDict[word] = wordDict.get(word, 0)+1

items = list(wordDict.items())
items.sort(key=lambda x:x[1], reverse=True)

for i in range(10):
    word, count = items[i]
    print("{0:<20}({1:>5})".format(word, count))
```

词频统计程序

实现2

- 统计《三国演义》中出场次数最多的10个人物

```
import jieba
f = open('threekingdoms.txt', 'r', encoding='utf-8')
txt = f.read()

wordList = jieba.cut(txt)
wordDict = {}
for word in wordList:
    if len(word)!=1:
        wordDict[word] = wordDict.get(word, 0)+1

items = list(wordDict.items())
items.sort(key=lambda x:x[1], reverse=True)

for i in range(10):
    word, count = items[i]
    print("{0:<10}({1:>5})".format(word, count))
```

要改进

```
曹操        (  953)
孔明        (  836)
将军        (  772)
却说        (  656)
玄德        (  585)
关公        (  510)
丞相        (  491)
二人        (  469)
不可        (  440)
荆州        (  425)
```