# 广州理工学院实验报告

**学院：计算机科学与工程学院 专业： 数据科学与大数据技术 年级： 2022**

**姓名： 吴恩威 学号：20220409430251 组 实验时间：2023.9.17**

**指导教师签字： 成绩：**

|  |
| --- |
| **实验项目名称：**  实验3字符串与正则表达式应用   * **实验目的和要求：**   1、熟练掌握字符串的概念；  2、使用内置函数实现对字符串的处理。  3、掌握正则表达模式匹配字符串，实现复杂的查找、替换等处理；  4、熟悉使用字符串及正则表达式在程序中的应用。 |
| * **主要仪器设备：**   PC机一台；Windows操作系统；Python运行环境；开发工具（PyCharm)。 |
| * **实验内容和原理：**   1、字符串表示及内置函数的使用；  2、正则表达式的匹配；  3、字符串与正则表达式在程序设计中的应用。 |
| * **操作方法和实验步骤（绘图）：**   1、编写程序：输入一个字符串，找出其中最大的字符，输出该字符及其在字符串中的索引位置。  操作方法与步骤：先定义str变量来存储输入的字符串，然后使用input函数实现数据输入，然后定义ch变量来存储str中的最大值，并使用print函数打印输出，最后调用字符串的函数来实现str中ch的索引位置，如下图所示：    2、请问如何拆一个包含有多种分隔符的字符串？  比如：现有一字符串：s = "ab;cd|ef\ng|hi,jkl|m.n\topq;rst,u vw\txyz"。  其中：”;,.空格|\t\n “都是分隔符号，如何处理？  操作方法与步骤：导入re模块，使用str1存储规则，将‘空格’换成‘ ’并使用[]包起来，然后使用re模块下的compile()函数实现规则的制定,再使用re下的split进行分割,如下图所示：    3、编程程序：匹配合法的身份证号码（15或18位）。从键盘上输入一串字符（身份证号），然后进行匹配，根据匹配结果输出是否是合法的身份证号。  操作方法与步骤：先定义id存储从键盘输入的身份证号码，然后定义str1匹配身份证号码的存储规则，然后使用re模块下的match函数进行匹配    4、编程程序，制作文本进度条。效果如下图所示，形成一个动态的进度完成状态显示效果。        操作方法与步骤：先使用print()函数打印开始下载的语句，然后通过for循环控制数字变化次数，星星个数和圆点个数，定义一个num存储变化的数字，我这里设置的是+2的变化，定义star表示星星个数，定义point表示圆点个数，然后使用time模块下的sleep函数实现慢的动态效果，最后打印下载完成语句    5、敏感词过滤。敏感词是指带有敏感政治倾向、暴力倾向、不健康色彩的词或不文明的词语。请使用字符串中的replace、sub等方法模拟敏感词过滤，将含有敏感词的字语句使用对应长度的“\*”符号替换。  操作方法与步骤：定义一个列表words来存储敏感词汇，使用input实现语句输入，通过for循环遍历列表words，将content中的敏感词汇替换成\*，最后print实现输出    6、附加题：统计每个英文单词的个数（频次）。  思路如下：  (1)、从键盘上输入一段英文字符串，并将该字符串赋值给某个变量。  (2)、使用正则表达式，以非英文字符为间隔符，切片（调用split函数完成），此操作会生成一个以单词为基本元素的list列表。  (3)、由于原先非字母可能连续，故生成的list可能存在空单词，需要去除空字符串。  (4)、全部转成小写，并排序。  (5)、将单词顺次存入dict，如果单词存在，则个数加一，如果不存在，则将此单词存入dict，个数设置为1。  (6)、输出。  操作方法与步骤：定义str存储input输入的数据，str1为制定的规则，然后使用re模块下的split对str进行切割，并使用for循环进行去空处理，然后变成小写存入list中，for循环结束后使用sorted对list进行排序。创造一个空字典dict，遍历list，通过if-else语句将list的元素放入dict中或者对dict中已有的元素的value进行更新 |
| * **实验结果与分析**  1. 实验结果如下：     分析：使用max函数得到字符串str中的最大字符，然后使用字符串的index()函数或者find()函数找到ch在字符串str中的索引位置   1. 实验结果如下：     分析： 使用字符串的split函数进行分割需要进行多次分割，而使用re模块下的split函数则可以一次完成全部分割，需要将 ”;,.空格|\t\n “分割符变为”;,. |\t\n”,+表示有一个或多个以上字符   1. 实验结果如下：     分析：题目要求匹配的身份证号码是15位或18位，所以在指定规则时需注意同时制定15和18位的规则，\d表示0-9的数组，{17}表示17位,[]表示范围   1. 实验结果如下：       分析：使用range()函数得到0-50的数字，num=i\*2就是下载进度，由于星星与圆点加起来的个数是50，所以star=1，point=50-star，然后在print中使用’\r’配合sleep实现动态的效果，format实现数据的格式化，最后使用print()实现换行。sleep()表示睡眠多少毫秒   1. 实验结果如下：     分析：需要使用到字符串的split函数对敏感词汇进行切割成列表，然后使用replace方法和len方法将敏感词汇替换成同等长度的\*，需要注意的是：字符串的replace方法是有返回值的(字符串)，替换后需要赋值给原来的content，如果没有此操作，最后输出的content还是你输入的语句   1. 实验结果如下：     分析：制定str1的规则是以非英文字符为切割，我使用了[^0-9a-zA-Z],^代表着反，就是除了[]里面的字符都进行切割。去空使用了字符串的strip方法，并将字符串变小写后排序，需要注意的是使用sorted()后会返回一个新的列表，需要赋值给原来的list。然后就是遍历list对字典dict进行更新，更新也可使用update()方法 |

**批阅评语：**