实验四　PYTHON 数据分析与界面

一、目的和要求

1. 熟悉 Python 的面向对象定义；

2. 熟悉 Python 的字符串处理；

3. 掌握 Python 语言基本语法；

二、实验环境

1. Win 7 操作系统；

2. Python IDLE、PyCharm 等开发环境；

三、实验内容

（一）验证实验（每个同学完成）

1．运行调试第四章各小节例示代码及课后练习的程序设计题，检查运行结果是否正确，记录实验结果。

（二）设计实验（小组验收，代码提交，算法设计和测试写入实验报告）

1．编制系列单词处理函数，分别实现下述功能，并设计测试用例验证程序的正确性，请在实验报告中说明所使用的正则表达式。

（1）编写函数 rotateword，接收一个字符串 strsrc 以及一个整数 n 作为参数，返回新字符串 strdes，其各个字母是 strsrc 中对应位置各个字母在字母表中“轮转”n 字符后得到的编码。

（2）编写函数 avoids，接收一个单词和一个含有禁止字母的字符串，判断该单词是否含有禁止字母。

（3）编写函数 useonly，接收一个单词和一个含有允许字母的字符串，判断该单词是否仅仅由允许字母组成。

（4）编写函数 useall，接收一个单词和一个含有需要字母的字符串，判断该单词是否包含了所有需要字母至少一个，并输出 words.txt 中使用了所有元音字母 aeiou 的单词。

（5）编写函数 hasnoe，判断一个英语单词是否包含字母 e，并计算 words.txt

中不含 e 的单词在整个字母表中的百分比。

（6）编写函数 isabecedarian，判断一个英语单词中的字母是否符合字母表序，并且输出 words.txt 中所有这样的单词。

2. 为实验三设计实验 4 编写图形用户界面。

3．数据分析综合应用：

（1）以文本文件格式读入文件夹\dataanalysis\label\下的 MTL_*.dat, CMTL_*.dat, CEMTL_*.dat(*表示 White 或者 Male，选择其中一种处理即可)中数据，并且分别读入 numpy 数组 MTLLabel、CMTLLabel 或者 CEMTLLabel 中，对各个数组取绝对值后按照降序排序，并且记录数据元素排序前的下标号；

（2）以文本文件格式读入文件夹\dataanalysis\train\下的 MTL_*_train.dat(*表示 White 或者 Male，选择其中一种处理即可)中的数据，并且读入 numpy 矩阵 TrainSample 中，计算矩阵的行列数（该矩阵包含了 1000 个维数为 3304 的样本的观测值，第 1-500 个样本属于第一类，第 501-1000 个样本属于第二类，每类含 500 个样本顺序保存在文件中）。根据（1）中数组的排序（3 个数组分别实验），选择最大的 k 个值（k 取 200，400，600，800，1000，…3304 维）对应的维度，把 TrainSample 中的 1000 个样本降维为 k 维，并保存到新的矩阵中 TrainSub 中；

（3）对于\dataanalysis\test\下文件作和（2）相同的处理（其中数据矩阵包含了 800 个维数为 3304 的样本，第 1-400 个样本属于第一类，第 401-800 个样本属于第二类，每类含 400 个样本顺序保存在文件中）；

（4）阅读和学习\knnexample\下面关于最近邻分类算法 Knn 的实现，用（2）中数据训练分类模型，用（3）中数据测试分类结果，统计错误率。

4．文本分析综合应用：

（1）编写模块实现中文文本中给定字或词的频率统计功能；

（2）运用（1）中功能模块分析文件“dreamofredmaison.txt”中前 80 回和后 40 回中常见文言虚实词的词频，分析结果存入文本文件；

（3）采用 Matplotlib 可视化（2）中的分析结果；

（4）运用 GUI 编制用户界面，为用户提供选择文言虚实字词的交互界面，按照用户选择采用（1）中功能实现频率统计，并且把（3）中实现的分析结果动态呈现给用户。