拼音输入法编程作业

一、作业内容

拼音输入法可以按注音符号与汉语拼音两种汉字拼音方案分成两大类。汉语拼音输入法的编码是依据汉语拼音方案（汉字的读音）进行输入的一类中文输入法。早期只有全拼这种方式，即完全依照汉字的整个音节来输入。随着技术的发展，拼音输入法不仅可以简拼还出现了一种只需两键就能输入整个音节的双拼方案。

在本次作业中，我们要求同学们自己编程实现一个简单的汉语拼音输入法，即实现从拼音（全拼）到汉字（字串）内容的转换。

二、 实验内容

1. 输入

a) 多个拼音串储存在指定的文本文件中（input.txt）；

b) 每个音之间用空格隔开，不含标点符号、阿拉伯数字和英文等非汉字内容；

c) 每行为每句（或短语）的拼音串，末尾没有标点符号。

```
qing hua da xue ji suan ji xi
ren gong zhi neng
ji qi xue xi
shu ju wa jue
```

Figure 1 **输入文件格式示例**

2. 输出

a) 转换后的汉字串，储存在指定的文本文件中（output.txt）；

b) 汉字间没有空格，每行为对应的汉字串。

```
清华大学计算机系
人工智能
机器学习
数据挖掘
```

Figure 2 **输出文件格式示例**

三、训练语料

1. 转换的汉字范围为国标一二级汉字，共 6763 个，以文本文件的形式提供；
2. 提供基于 sina news 的汉语语料库进行模型训练，也可以自己寻找其他资源。

四、基本要求

1. 使用基于字的二元模型，实现一个拼音到汉字的转换程序；
2. 完成实验报告，主要内容如下：
   a) 介绍算法的基本思路和实现过程；
   b) 展示实验效果，选取效果好和差的例子进行分析；
   c) 对比参数选择，进行性能分析；
   d) 总结收获，提出改进方案。
3. 支持命令行形式提供输入文件名和输出文件名并运行程序，例如：
   pinyin ../data/input.txt ../data/output.txt

五、选做内容

1. 实现基于字的三元、四元模型，实现拼音到汉字的转换；
2. 实现基于词的二元、三元模型，实现拼音到汉字的转换；
3. 对不同的模型进行实验分析。

六、提交内容

1. 输入拼音文件（input.txt，置于 data 文件夹下）

2. 转换结果文件（output.txt，置于 data 文件夹下）
3. 源程序（请全部放在命名为 src 的文件夹下）
4. 可执行程序（例如 pinyin.exe，置于 bin 文件夹下）
5. 说明文件（readme）

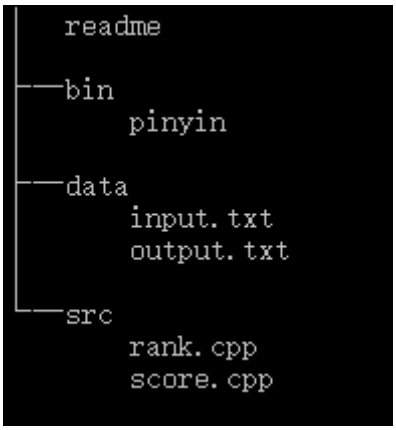

Figure 3 **提交内容格式示例**

七、其它

输入文件，国标汉字等文件会稍后发布，请同学们先思考实现方案，并根据现有的资源自行设计测试文件。

2017/3/18