2023 年春季

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、 输入与输出格式

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、汉字范围

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，见附件 拼音汉字表/一二级汉字表 。训练语料中在该范围之外的汉字可一律不处理，测试语料中保证均为该范围内的汉字。

四、训练语料

1. 【必做】新浪新闻 2016 年的新闻语料库（见附件 语料库/sina_news_gbk）；
2. 【选做】微博情绪分类技术评测（SMP2020-EWECT）中通用训练集的微博语料库（见附件 语料库/SMP2020）；

3. 【选做】在 GitHub 项目 nlp_chinese_corpus 中选择一个语料库（见 https://github.com/brightmart/nlp_chinese_corpus ）；
4. 【选做】自己寻找其他中文语料资源。

其中，基于新浪新闻语料库的训练为必做，其他选做语料库可以与新浪新闻共同使用或单独使用，鼓励同学们针对语料库的选择和使用，进行定性或定量讨论。

五、输入法测试：

1. 提供用于自测拼音输入法效果的语料文件（见附件 测试语料），包含两个文件，input.txt 为输入的拼音，std_output.txt 为标准输出结果，共 500 个短语（句子）。也可自行构造其他测试例进行测试。

2. 需要汇报在测试语料上的字准确率和句准确率，此外，也可探究和讨论其他合理的评价指标。

注：1. 该样例集合包括以前选课同学众包的句子和助教提供的句子，可能存在错误，仅供参考，欢迎纠错。

2. 在本测试集上，无需在追求过高准确率上花费过多精力，本作业主要关注大家使用的方法与讨论，且助教处有其他测试集可验证模型效果。

六、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，要求：
   a) 需包含 README 文件和适当的注释；
   b) 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 或
      python pinyin.py ../data/input.txt ../data/output.txt
2. 完成实验报告。

六、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、语料库、模型参数进行实验分析。

七、实验报告要求

1. **PDF 格式**
2. **写明姓名、学号**
3. 包含的内容：
   a) 【必做】介绍基于字的二元模型算法的基本思路、公式推导和实现过程。
   b) 【必做】介绍实验环境。
   c) 【必做】介绍使用的语料库和数据预处理方法。
   d) 【必做】展示实验效果，在给定测试样例或自己构造的数据集上的准确率（包括字准确率和句准确率）。

e) 【必做】选取效果好和差的例子进行分析。

f) 【选做】介绍基于字的多元模型或基于词的模型的基本思路、公式推导和实现过程，以及测试准确率。

g) 【选做】对比参数选择，进行性能分析。

h) 【选做】对比不同的语料库的实验效果（定性或定量均可）。

i) 【选做】探究和讨论字准确率和句准确率以外的合理的评价指标。

j) 【选做】模型改进、模型对比、其他深入讨论等。

注：1. 此列表仅为展示实验报告需要包含的项目内容，不必按该顺序进行实验报告的撰写。

2. 追求作业高分的同学请注意，选做内容无需全部完成也可以获得很高的分数，鼓励大家针对某项或某几项内容进行深入思考和扩展讨论。

3. 本作业主要目的是提供一次人工智能相关实践的练手的机会，让大家亲自尝试实现一个可用的项目，请酌情分配时间和精力即可。

八、提交内容

1. **实验报告**（PDF 格式，**在网络学堂“第一次作业-实验报告”窗口中提交文件**）
2. **作业文件**（压缩文件，**在网络学堂“第一次作业-代码”窗口中提交文件**）
   - a 输入拼音文件（input.txt，置于 data 文件夹下）
   - b 转换结果文件（output.txt，置于 data 文件夹下）
   - c 源程序（全部放在 src 文件夹下）
   - d 与程序运行有关且小于 100M 的中间结果文件，如词频表等（置于 src 文件夹下，与源程序中调用位置一致）
   - e 可执行程序（例如 pinyin.exe，置于 bin 文件夹下）

     注：如果用 python 完成作业，无需提交可执行程序
   - f 说明文件（包含程序运行方式、文件结构，README.txt）

   注：新浪新闻语料库不必上传。
3. **其他补充材料**

如果有其他较大的补充材料需要上传，如自行构造的测试样例、额外使用的语料库等，请单独上传至清华云盘并在网络学堂“第一次作业-实验报告”窗口中**提交链接**（请勿与第 2 部分作业文件压缩到同一文件中）。如果补充材料与程序运行有关，请在云盘中提交 README 文件说明该补充材料应放在什么路径下。

九、其它

助教评分时，主要依据实验报告中的实验结果、讨论、个人思考和代码完成情况评分，同时综合考虑模型的准确度和效率。请先完成基本要求（即，必做部分），再进行扩展和创新。

实验的发挥空间大，鼓励创新和深入思考。创新性想法比提高准确率更受欢迎，鼓励将失败的尝试过程和结果在实验报告中记录下来。

增长知识，拒绝抄袭，如发现抄袭同届或往届作业，一律按零分处理。

2023/3/11