# 2024春人工智能导论

## 第一次作业-拼音输入法

**负责助教：王贝宁 wbn23@mails.tsinghua.edu.cn**

**附件下载：https://cloud.tsinghua.edu.cn/d/66b1261c781146d1b296/**

### 作业背景

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。

### 输入与输出格式

1. 输入
   - 多个拼音串，每行一个，储存在指定的文本文件中（如`input.txt`）
   - 同个拼音串的不同拼音之间用空格隔开，不含标点符号、阿拉伯数字和英文等内容
   - 每行为每句（或短语）的拼音串
   - 样例输入 `ren gong zhi neng`
2. 输出
   - 转换后的汉字串，储存在指定的文本文件中（如`output.txt`）
   - 汉字间没有空格，每行为对应的汉字串
   - 样例输出 `人工智能`

### 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.txt`和`一二级汉字表.txt`。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。

### 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】自己寻找其他中文语料资源，如在GitHub项目nlp_chinese_corpus中选择一个语料库（见https://github.com/brightmart/nlp_chinese_corpus ）

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。

## 作业内容及评分标准

本次作业分为两个实验，最终分数将按照实验分数和报告分数共同得出。

编写语言要求： `python` 或者 `C++` 。

其中，对于性能部分，主要使用两个指标：字准确率、句准确率来进行评判：
$precision_{字} = \frac{count_{正确的字}}{count_{所有字}}$，$precision_{句} = \frac{count_{正确的句}}{count_{所有句}}$

我们认为，每一行输入对应一个句子。

### 实验1-线上评测(40%)

请登录线上实验平台 oj.cs.tsinghua.edu.cn ，右上角登录选择清华ID登录，即可看见作业链接。

本实验要求在已给定的字一元和二元词汇表的基础上，完成拼音输入法的设计，如果正确完成拼音输入法算法设计，该部分分数应当为满分。对于本部分，要求$precision_{字} \geq 80\%$和$precision_{句} \geq 35\%$，达到该准确率即为满分，如果达不到将由助教视代码完成情况和性能酌情扣分。

具体要求详见线上实验平台。

> 旁听生同学如果需要OJ账号密码，请联系助教

### 实验2-性能测试(40%)

1. 【30%】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：
    - 需包含README文件和适当的注释
    - 支持命令行形式使用输入输出重定向输入文件名和输出文件名并运行程序，例如：`python pinyin.py <../data/input.txt >../data/output.txt` 或 `./a.out <../data/input.txt >../data/output.txt`
    - 测试文件为下发的包含两个文件， `std_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），需要汇报在测试语料上的字准确率和句准确率

        - 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
        - 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
    - 准确率达到OJ要求

2. 【10%】探索其他可以提高性能的方法，如：

    - 实现基于字的三元、四元模型
    - 实现基于词的语言模型
    - 对于未知效果方法的尝试

**实验报告(20%)**

1. PDF格式
2. 写明姓名、学号
3. 包含的内容：
    i. 【必做】介绍实验环境。
    ii. 【必做】介绍使用的语料库和数据预处理方法。
    iii. 【必做】介绍基于字的二元模型的拼音输入法的：
        a. 基本思路、公式推导和实现过程。
        b. 实验效果，包括在给定测试样例上的准确率（包括字准确率和句准确率），训练时间，生成一句话的平均时间，生成所有给定测试样例的总时间。
        c. 选取效果好和差的例子进行分析。
        d. 对比参数选择，进行性能分析。
        e. 时间和空间复杂度分析。估算所需进行的计算次数和预估时间，并与真实运行时间进行对比。
    iv. 【选做】介绍在性能测试中实现的其他模型或者算法，如基于字的多（三及以上）元模型或基于词的模型，要求同基于字的二元模型的拼音输入法。
        a. 如果实现了其他模型或者算法，要求最终使用清晰的形式（如图表）总结对比不同模型的实验效果。
    v. 【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。
    vi. 【选做1%】对实验的感受及建议，言之有理，助教认同即可得分。

注：

1. 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。
2. 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。

3. 完成所有【必做】部分 即可得到15%的分数，其余5%分数由助教酌情评定。

**提交方式**

- 实验1请在实验平台上提交
- 实验2请在网络学堂“第一次作业-代码”作业窗口中提交代码压缩包，代码压缩包必须严格按照以下格式提交，否则会影响该部分成绩

```
data/
    input.txt
    output.txt
src/
    <codes> //放置所有你的源代码
readme //包含程序运行方式、文件结构
```

  - 压缩包中**不得**(-10%)包括任何中间文件和语料库文件，请通过readme的方式说明语料库文件应当如何放置，且确保通过运行代码可以生成中间文件并得到结果
  - 压缩包命名为 学号-姓名，如 2023123456-张三
- 实验报告请在网络学堂“第一次作业-报告”作业窗口中提交报告pdf版，**不得**(-10%)提交其他文件格式的实验报告
- 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实验报告中提供下载链接（请勿与实验2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。

## 其他

- 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、代码出现雷同(我们将对每一次提交进行查重)等，一律按零分处理。
- 截止日期为2024.4.9（周二）晚23:59，每迟交一天扣除10%的分数，不足一天按照一天计算，迟交时间计算：OJ、“第一次作业-报告”、“第一次作业-代码”三者的最后一次提交时间