2025春人工智能导论

第一次作业-拼音输入法

实验说明

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

拼音输入法可以按照注音符号和汉语拼音这两种拼音方案分成两大类。汉语拼音输入法是依据汉语拼音（即汉字的读音）进行输入的一类中文输入法。早期的拼音输入法只有全拼输入方式，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进行简拼输入，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，我们要求同学们编程实现一个简单的汉语拼音输入法，实现从全拼拼音到汉字内容的转换。

输入与输出格式

1. 输入
    - 多个拼音串，每行一个
    - 同个拼音串的不同拼音之间用空格隔开
    - 保证输入拼音合法，且只存在小写英文字母和空格
    - 样例输入 `ren gong zhi neng`
    - 从标准输入中读入
2. 输出
    - 转换后的汉字串，每行一个
    - 汉字间没有空格，每行为对应的汉字串
    - 样例输出 `人工智能`
    - 输出到标准输出

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`，提示：该语料库部分格式没有对齐）
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

选做语料库可以单独使用、或与新浪新闻共同使用，鼓励针对语料库的选择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OJ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。

要求在已给定的字一元和二元词频表的基础上，完成拼音输入法的设计。
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**正确**完成拼音输入法算法设计，该部分分数应当为满分，否则助教将由代码完成情况和性能酌情给分，通常得分为$5 \times \lfloor precision_{句} \div 5 \rfloor$。

具体要求详见线上实验平台。

> 旁听同学如果需要OJ账号，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：
    - 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运行程序，例如：`python main.py <data/input.txt >data/output.txt` 或 `./a.out <data/input.txt`

`>data/output.txt`

- 测试语料为下发的包含两个文件，`input.txt`为输入的拼音，`answer.txt`为标准答案，共500句，需要汇报在测试语料上的字准确率和句准确率
  - 【bonus】该样例为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，欢迎纠错
  - 无需在追求过高准确率上花费过多精力
- 本地字、句准确率达到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
   - 实现基于词的模型
   - 针对已有模型进行优化（非调参！）
   - 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、院系
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、公式推导和算法原理
      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
   iv. 【必做】思考题
      a. 【语料编码】简述`gbk`和`utf-8`的区别。是否存在其他编码方式？若有，举例并说明应用场景。
      b. 【算法设计】假设输入句子由n个字组成，字库大小为V并且在读音上均匀分布，读音数量有限为常数。试分析`viterbi`算法的时间复杂度和空间复杂度。请使用大O记号表示。
      c. 【输入输出】假设你需要在OJ上提交题目，要求设计程序将接收到的输入，原封不动输出。以下是两段伪代码及对应终端输出。题目从标准输入中读入数据，输出到标准输出。仅从输入输出角度出发，不考虑数据规模，陈述两种代码的区别并分别判断正误。

- 代码A

```
while not EOF
    variable a
    a <- input
    print a
end while
```

输出

```
>>>1 //>>>代表输入，下同
1
>>>2
2
```

- 代码B

```
list l
while not EOF
    variable a
    a <- input
    l.append(a)
end while
for each in l
    print each
end for
```

输出

```
>>>1
>>>2
1
2
```

- 标准答案

```
1
2
```

v. 【选做】介绍在实验2中实现的其他模型或者算法

- 简述基本思路
- 选取例子进行分析，其相比二元模型好/差在哪
- 使用清晰的形式（如图表）总结对比不同模型（你实现的模型们和二元模型）的实验效

果、训练/预测时间
    - **不接受**使用大模型API、开源模型参数等需要运行中联网环境的方法

vi. 【选做】对你实现的方法进行调参，汇报结果，并且分析为何改动该参数会导致结果变化

vii. 【选做】探究和讨论字准确率和句准确率以外的评价指标，并且在你实现的方法上测试和分析

viii. 【选做1%】完成实验的时间/工作量，对实验的感受及建议，失败的尝试等等，言之有理，助教认同即可得分

提示：

1. 选做内容无需全部完成也可以获得高分
2. 报告分数根据要点、性能评判，与篇幅无关
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
4. 保持良好心态，正确看待作业的练习属性，合理分配时间和精力

提交方式

- 实验1请在OJ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

```
corpus/                 //助教会把sina_news_gbk/放在该目录下
data/
    answer.txt          //下发的标准答案
    input.txt           //下发的测试输入
    output.txt          //你程序的输出
    拼音汉字表.txt       //下发的拼音汉字表
    一二级汉字表.txt     //下发的一二级汉字表
src/                    //放置所有你的源代码
    <code>
<文件夹，若有>/          //其他你项目所需要的文件夹
    <文件夹，若有>/      //允许保留文件架构
                        //只允许空文件夹，不允许有中间文件
main.py                 //主程序入口
readme                  //包含程序运行方式等，格式建议为.txt/.md/.pdf
requirements.txt        //如果使用了第三方库
```

    - 助教会在评测时运行脚本：

```
testcase="助教的测例位置"
judgescript="助教的评测脚本"
# 解压缩
unzip 2023123456-张三.zip
```

```
# 将语料库放进文件夹
cp -r sina_news_gbk/ corpus/
# 将测例放进文件夹
cp $testcase data/input.txt
# 安装必要依赖
pip install -r requirements.txt
# 运行主程序
python main.py <data/input.txt >data/output.txt
# 进行评测
cp $judgescript ./judge.py
python judge.py
```

  - 要求通过运行上述脚本，可以重新生成**基于字的二元模型**、**使用新浪新闻作为语料库**的输出文件
    - 如果实现了其他模型和语料库，运行方式请在readme中说明，助教会按照其指示操作
    - 在按照readme修改前（即不参考readme的情况下），上述脚本必须可用有效，产生**和实验报告一致**的结果
  - 压缩包中**不得**(-10%)包括任何中间文件（如词频表）和语料库文件，若需要，请保留空文件夹维持项目架构
    - 如果有其他较大的补充材料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
  - 压缩包必须为 `.zip` 格式，**不得**为其他格式(-10%)

    > 助教使用 `unzip` 指令解压缩，感谢大家体谅助教：）

  - 压缩包命名为 学号-姓名
    - 如 `2023123456-张三.zip`

- 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
  - 实验报告命名为 学号-姓名
    - 如 `2023123456-张三.pdf`
  - 实验报告中通常**不应该**(-10%)出现代码，如极有必要，仅允许出现关键部分的若干行代码并配合详细注释进行阐释
  - 助教通过上传分数excel批改作业，不会看网络学堂提交作业栏的信息，请将想说的话写在实验报告中
  - 未提交实验报告按0分处理

其他

- 习题课，线上形式
  - 预计时间在发布作业后第二周周末，具体时间请关注网络学堂
  - 讲述作业完成思路并答疑
- 【bonus】为鼓励同学们互相讨论，我们建议积极帮助他人：）

- 【bonus】助教虚心接受大家对实验过程中任何问题（如typo）的发现和纠正：）
- 完成时间为三周，迟交惩罚为-10%/天，不足一天按照一天计算，具体截止日期同网络学堂作业截止时间
  - 补交方法：请将实验报告和代码压缩包附于标题为【拼音输入法作业补交】学号-姓名的邮件，发给助教
    - 如 `【拼音输入法作业补交】2024123456-张三`
  - 时间计算：邮件接收时间和OJ迟交窗口的最晚提交中，以较晚的时间为准

Honor Code

雷同

我们将对每一次提交进行查重，雷同一律按照0分处理

Plagiarism

> 请注意，抄袭他人及故意提供材料供他人抄袭都被视为作弊行为（因此不要将你的代码、结果、报告发送给他人），这将导致你的成绩被记为零分，或课程成绩不及格以及《清华大学学生纪律处分-管理规定实施细则》指导下的其他后果。这是一条“红线”，我们将严肃对待。

> Note that both copying from others and intentionally providing materials for others to copy are considered plagiarism (so do not send your codes, results, reports to others) and will result in 0 grades and/or failing the class and/or other consequences instructed by the school honor code 《清华大学学生纪律处分-管理规定实施细则》. It is a “red line” and we take it seriously.

作业不是为难大家，希望大家能通过本次作业学有所得：）