**Coursework 1：拼音（全拼）联想输入法**

自然语言处理课程作业一

2025年9月23日

要求：

以小组为单位，使用Python代码以及NLP相关开源工具包，以文件测试数据的形式输入多组拼音串（如，“jin tian tian qi bu cuo“），实现文字输入。每一个文本（测试）可以通过控制台输入调试。编写代码实现功能后，要写一份不少于1000字**作业报告**包括方法介绍、完整数学公式推导、实验分析等内容。可参照cnki较短的实验性论文编写格式。每个小组5分钟的”隐马尔科夫模型（HMM）”及2分钟的代码演示的**介绍视频**，以备后续各组互评。作业提交最后时间2025年10月20日23:00，作业（报告、视频、代码）上传到（不少于6个月可下载时长的）百度网盘汇总到学委，整理分组链接后发到QQ邮箱55354116@qq.com提交作业。

提示：

1, 语料数据，如《人民日报》语料。需要用到语料中，汉字的出现频次。也可能会用到词典等语料。

2, 拼音与汉字的匹配关系，需要拼音字典语料，构造输入拼音对应汉字的隐马尔科夫模型（HMM）。

3, 联想输入功能：可使用（动态规划等）算法求解“字串“形成的概率，提示出一组以概率排序的一个“字串”，取这个“字链”为输出的句子。

4, 结果分析：对自己实验中所使用的验证方法和结果进行分析，包含但不限于准确率。

5, 总结：写1000字以上的小组实验总结。如，有何收获，就这一方向未来还可以做些什么。

评分：

完成”隐马尔科夫模型（HMM）”的介绍、基本的功能实现和完整的报告文章最高可得90分（本次作业的成绩，满分为100分）。创新的功能增强和扩展，以及实现方法或评估方法的改进，都可能成为进一步加分的条件。同一小组每位成员将获得同等的分数，如果小组提出了每位成员作业贡献度的比例，将适当调整每位成员的分数。