|  |  |  |  |
| --- | --- | --- | --- |
| 。。。 | 。。。 | 。。。 | 。。。 |
| 中 | 中 | 中 | 中 |
| 国 | 国 | 国 | 国 |
| 我 | 我 | 我 | 我 |
| 爱 | 爱 | 爱 | 爱 |
| 挨 | 挨 | 挨 | 挨 |
| 我 | 我 | 我 | 我 |
| 家 | 家 | 家 | 家 |
| 加 | 加 | 加 | 加 |
| 大 | 大 | 大 | 大 |
| 。。。 | 。。。 | 。。。 | 。。。 |

wo ai wo jia

训练：

将所有汉字视为隐状态，通过训练文本求转移概率，初始概率：训练语料可以使用cnews.test.txt或者toutiao\_cat\_data.txt（get\_data.py可以处理数据）中的任何一个即可，考虑到传输问题，上面两个语料规模均不大，同学们也可以自己下载并使用类似的其它语料。

发射概率（如多音字 P（le|乐） P（yue|乐）：可用语料“带拼音搜狗词库”或者“骆驼祥子”求此概率）

维特比搜索：

选择1：将所有可能汉字作为隐状态，该方法优势是模型结构稳定，对所有输入语句均适用。可用“常用2500汉字表.txt”。

选择2：只在输入拼音对应的 隐状态（字）集上进行寻优，可用语料pinyin2hanzi.txt求拼音可能对应的汉字。优势是速度快，但需要对每句话单独建模。