简单中文输入法实验报告

MF1633019 经伟　　MG1633063 沈佳凯　　MF1633031 糜泽宇

一、设计思路

采用了一阶隐马尔可夫模型

> 输入的拼音序列---->观测状态O
> 汉字输出序列---->遗隐含状态S
> 短语中汉字的先后顺序---->隐含状态转移概率矩阵A
> 汉字的编码与拼音---->混淆矩阵B
> 短语的起始汉字---->初始状态概率矩阵π

然后在训练好的模型上利用viterbi算法计算最佳的输出序列

二、实现细节

实验设置

1. 由于每次使用之前都训练的话十分耗时，为了方便实际使用以及组员之间的数据沟通，实验利用数据库作为存储参数的手段。引入了sqlite3包来完成对数据库的创建，插入，查询等一系列操作。
2. 另一方面在频率值计算上面，如果计算普通的分数频率，在句子太长的时候，有可能因为频率值得乘积太小而导致计算机无法分辨。所以选择计算对数频率，在计算序列的频率是只要相加即可。
3. 要得到字典中汉字的拼音，本实验引入了pypinyin包，利用其中的pinyin方法可以得到汉字对应的拼音（有时不止一个），以在下一步计算混淆矩阵。

实验伪代码

初始状态概率表(starting)

> 遍历训练数据集
> 对于每一个词语：记录他的初始汉字和频数，置于临时字典charFreqMap中
> 将charFreqMap中key值相同的value叠加-->某个汉字初始的频数
> 计算频率以及log值放置到数据库的starting表中

隐含状态转移表(transition)

> 创建一个二维字典preBehMap,其key为汉字，value为该汉字下一个位置出现的汉字与概率
> 遍历训练数据集

> 对于每一个词语：记录他的后一个汉字与频数，叠加到对应的字典preBehMap[]中
> 在每个value的字典中，计算频率以及log值，放置到数据库的transition表中

混淆矩阵表(emission)

> 创建二维字典charPyMap,其中key为汉字，value为该汉字对应的拼音和概率
> 遍历训练数据集
> 对于每一个汉字：用pypinyin.pinyin方法得出对应的拼音
> 将拼音和频数叠加如charPyMap[]中去
> 在每个value的字典中计算频率的log值，放置到数据库的emission表中

viterbi算法

> pinYin是保存拼音的list，wordLists保存拼音对应的汉字list，一个拼音通常对应多个汉字
> 将probability矩阵的第一行设为wordLists第一行汉字对应的初始概率
> 计算前一个汉字到达当前汉字的最大概率
> 取概率矩阵最后一行概率最大的汉字作为输出序列的最后一个汉字
> 用保存在index中的数据回溯出前一个汉字，得到全部的序列

三、训练数据来源

来自github开源项目结巴分词的词库，每一行由三个元素构成“词语，频数，类型”，其中类型属性本实验中无用。其结构如下所示：

```
示例 83 n
示值 3 n
示功图 3 n
示好 3 v
示威 2256 n
```

对其写一个简单的生成函数即可以取得训练数据

```
def iter_dict():
    with open('dict.txt') as f:
        for line in f:
            word, freq, tag= line.split()
            yield word.decode('utf8'),int(freq)
```

此外，由于结巴词库的训练数据包含的大多是短语，因此训练得到的模型对于长句子的表现欠佳。

四、成果展示

```
type pinYin using space to separate = zhong hua ren min gong he guo
中华人民共和国 -14.3674560297
仲华人民共和国 -22.4963844633
忠华人民共和国 -23.7615413323
种花人民共和国 -27.7853165911
众化人民共和国 -28.2075276955
重化人民共和国 -29.242929633
钟花人民共和国 -30.4020574446
```

```
type pinYin using space to separate = qing hua da xue
清华大学 -16.4077045371
氢化大学 -17.399449706
轻化大学 -18.7982959103
庆华大学 -20.3534739068
青花大学 -20.9632287877
倾化大学 -21.2874406208
```

五、参考资料

https://github.com/LiuRoy/Pinyin_Demo

https://github.com/fxsjy/jieba