词库文档

词库，顾名思义，就是一个包含了大量词组，短句的文字库。它将拼音组合映射到中文词组，按照某种规则，通过输入法的其他模块，最终将符合的词组返回到用户界面，供用户使用。

在设计实现输入法的过程中，词库的设计是必须且重要的一个环节。词库的好坏直接影响着输入法的用户体验。一个好的词库，可以让用户在使用时更加得心应手，加快打字速度，并且拥有一个良好的工作心情。

在设计输入法的过程中，通过小组的内部讨论，我们结合我们所需要设计的输入法的需要，我们将词库的词条格式定义如下：

简单来说，词条的第一部分为拼音部分，每个音节间用符号“’”隔开，第二部分则为拼音部分所对应的中文词组（或者短句），第三部分则是这个词条的词频，即为一个权重值，值越大，说明该词在文章中出现的概率越高。在词条的不同部分间我们用制表符隔开。下面举个简单的例子：hai'shi'shan'meng 海誓山盟 1。

在各个不同的拼音间用符号“’”隔开，是为了防止出现歧义，例如xian与xi’an，如果没有符号进行标识，那么我们就无法快速识别这两者的区别。尽管有其它解决的方式，但在我们的设计中还是采用了这种解决方式。此外，词库中的词条是按照拼音的排序从a~z的顺序排列的，这样有利于后续工作的实现，减少查找的时间开销。

在我们的初始版本中，由于各种条件的限制，我们将各个单词的词频都初始化为1，因为我们无法获得一个较为准确的词频数据。在今后的版本中，我们会修改词频的初始值。我们的设想是通过对《人民日报》进行数据挖掘与分析，得到各个词出现的词频，并在我们的词库中加入新的单词。

在这次实验中，我们共设计了三个不同大小的词库，分别为8w，15w以及22w词汇规模的词库，以满足不同的需要。同时，在这些词库中，我们添加了现在流行的网络用语及新兴词汇，同时也添加了不少的唐诗、宋词等传统文学中的名言警句，来丰富我们的词库，期望得到良好的用户体验。在实现过程中，张绍文，张昊同学从网上寻找了现有的词库。张绍文同学主要负责对现有的词库进行格式的转换与修改，使得词库能符合我们自己设计的格式。张昊同学主要负责寻找现下流行的词汇短语，丰富我们的词库。