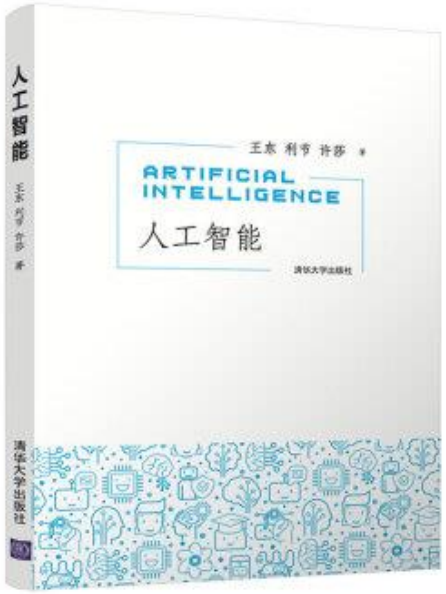
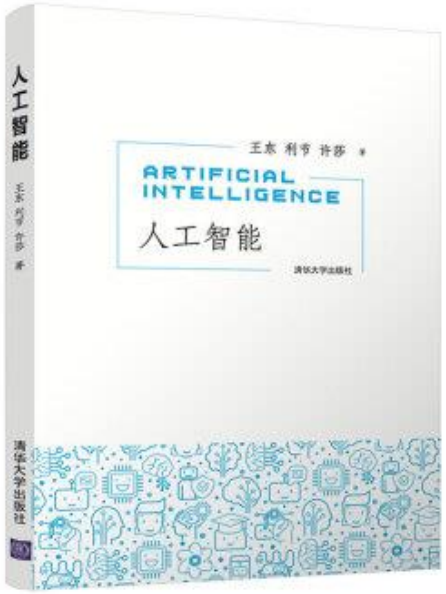


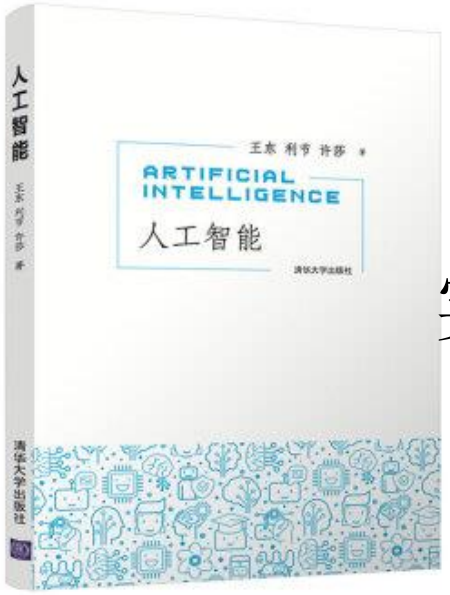
理解你的语言

利节



AIDemo示例

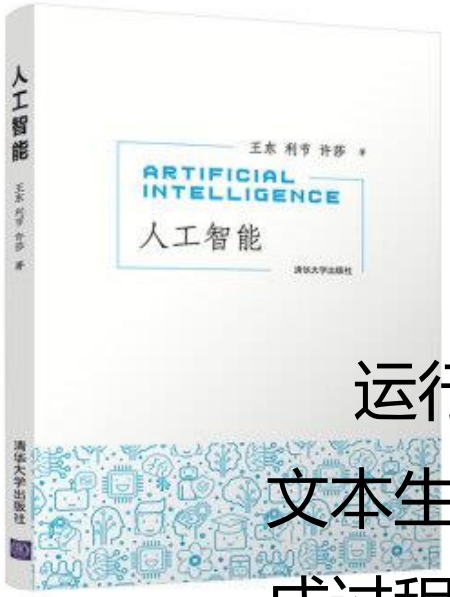
http://aibook.cslt.org



实验准备：

AIDemo 提供了一个**word2vec** 示例程序，帮助读者理解词向量技术。前面我们提到过，词向量是单词在一个连续语义空间上的映射，基于该映射，语义接近的单词所对应的向量间距离更小，从而实现词义的可计算化。

本实验利用**Google** 的**Word2Vet** 工具，计算中文词向量。



实验一：运行缺省配置

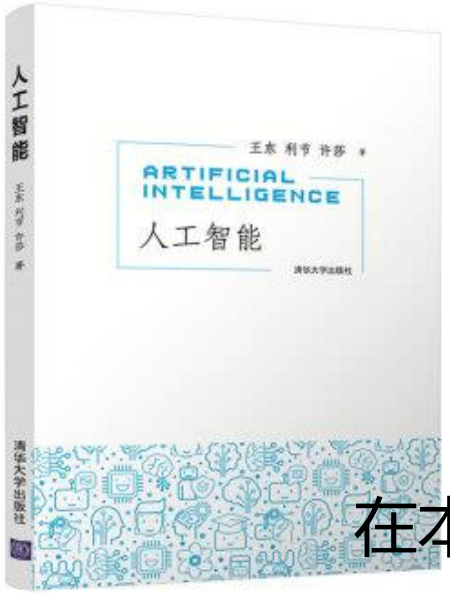
运行缺省配置的**run.sh**，将基于《射雕英雄传》中的文本生成词向量。观察**run.sh** 的内容，可以看到这一生成过程包括如下几个步骤：

（1）对《射雕英雄传》文本进行分词，生成**seg.txt** 文件。我们选择**Jieba** 分词工具。

（2）选择高频词（出现频率超过**10** 次的单词）组成词向量的词表。

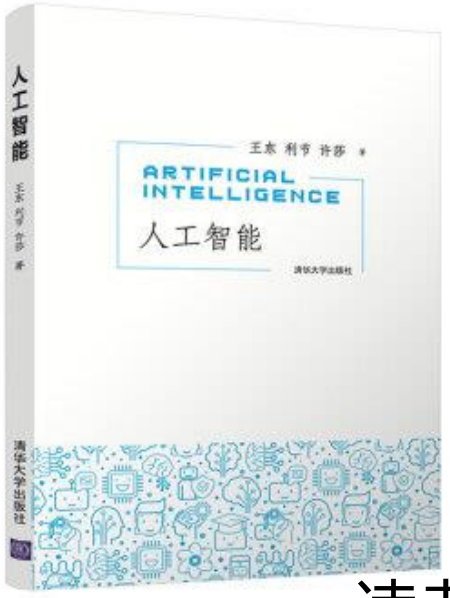
（3）利用**word2vec** 工具，生成词向量。

（4）利用**t-SNE** 工具，将词向量投影到二维空间，生成词向量表示图。



实验二：人名的词向量

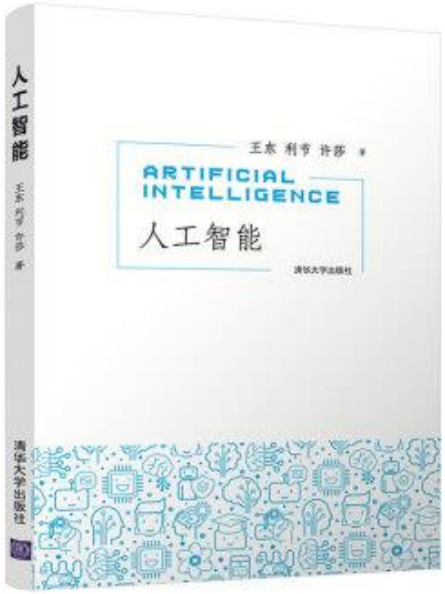
在本实验中，我们单独选出《射雕》中的人名，画出它们的词向量，以观察词向量对语义的表达能力。首先，在词表**dict.txt** 中进行筛选，保留人名，组成人名列表**name_list.txt**，然后从词向量文件**wordv** 中将这些人名对应的词向量筛选出来，组成词向量文件**wordv_name**，最后利用**t-SNE** 工具将这些人名的词向量投影到二维空间中。读者可参考**run-name.sh** 中的步骤来完成上述过程。



实验三：生成自己的词向量

读者可以选择自己感兴趣的任意文本，不论是四大名著，还是网络小说，生成词向量。将这些文本合并成一个文件，上传到**AIDemo**，基于实验一所述的步骤即可生成自己的词向量。读者还可以改变文本的领域，文本的

长度，观察基于不同数据源所生成的词向量的差别。



The end !