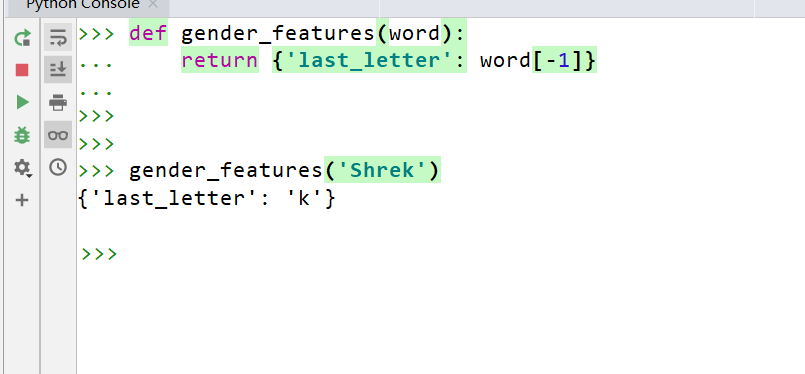
有监督分类

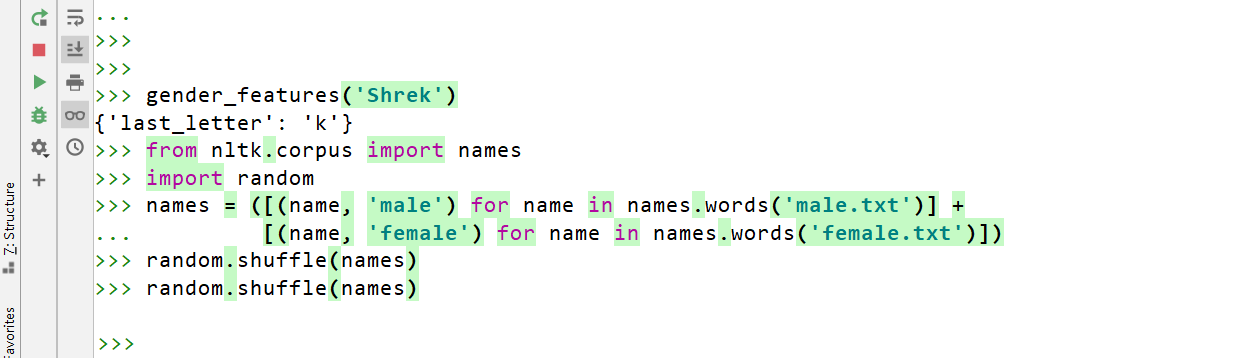
创建一个分类器的第一步是决定输入的什么样的特征是相关的，以及如何为那些特征编

码。在这个例子中，我们一开始只是寻找一个给定的名称的最后一个字母。以下特征提取器

函数建立一个字典，包含有关给定名称的相关信息：

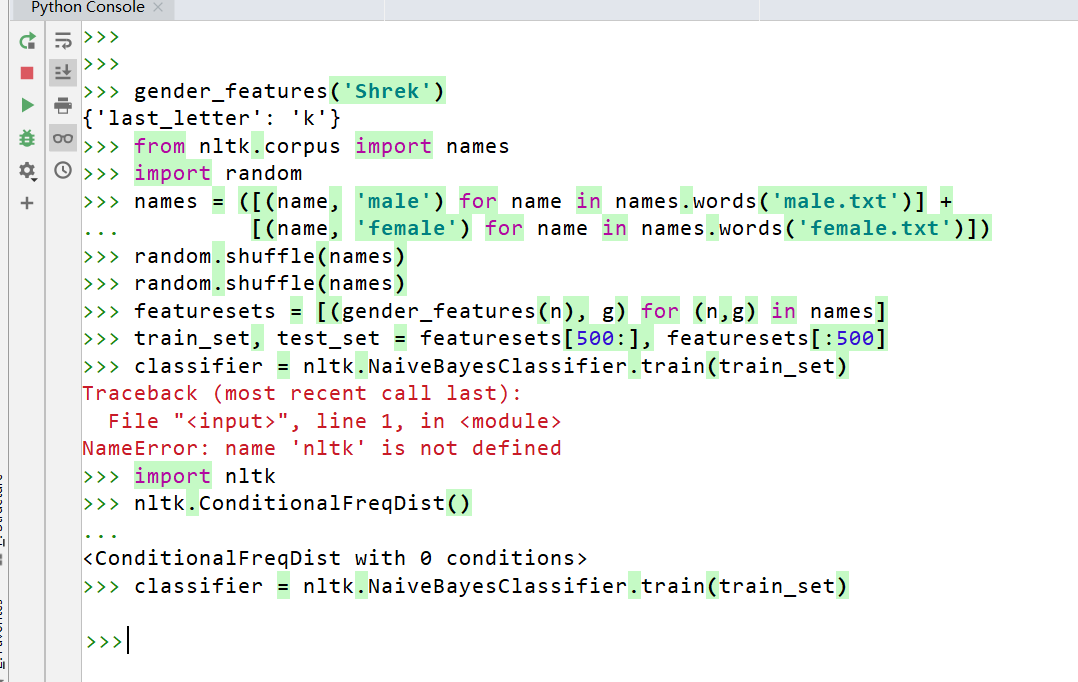


现在，我们已经定义了一个特征提取器，我们需要准备一个例子和对应类标签的链表



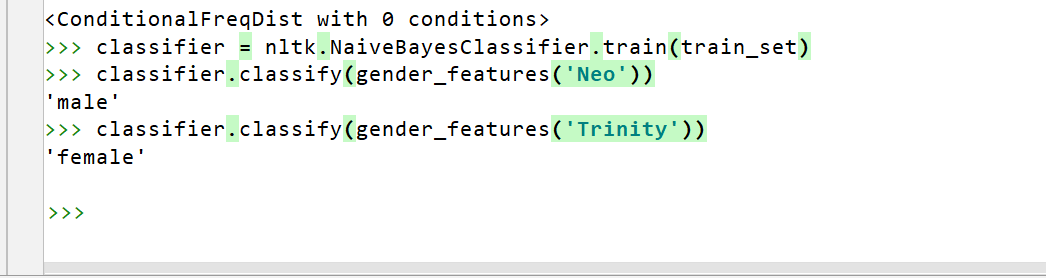
接下来，我们使用特征提取器处理名称数据，并划分特征集的结果链表为一个训练集和

一个测试集。训练集用于训练一个新的“朴素贝叶斯”分类器。



在本章的后面，我们将学习更多关于朴素贝叶斯分类器的内容。现在，让我们只是在上

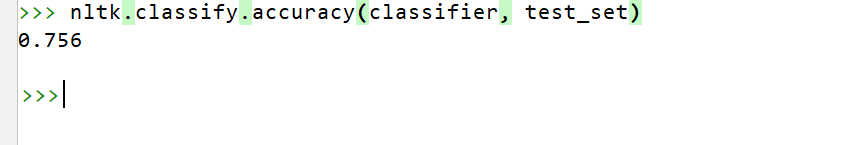
面测试一些没有出现在训练数据中的名字：



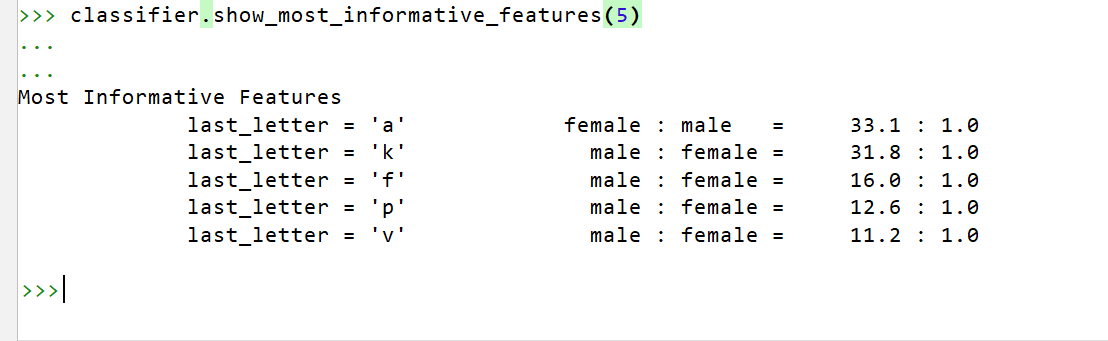
请看《黑客帝国》中这些角色的名字被正确分类。尽管这部科幻电影的背景是在 2199

年，但它仍然符合我们有关名字和性别的预期。我们可以在大数据量的未见过的数据上系统

地评估这个分类器

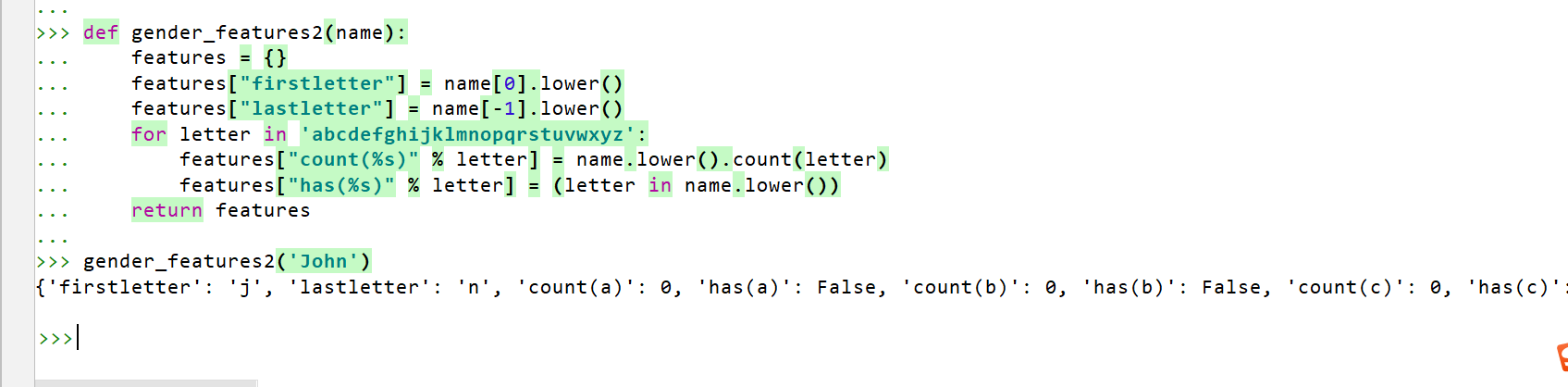


最后，我们可以检查分类器，确定哪些特征对于区分名字的性别是最有效的。



例 6-1. 一个特征提取器，过拟合性别特征。这个特征提取器返回的特征集包括大量指

定的特征，从而导致对于相对较小的名字语料库过拟合



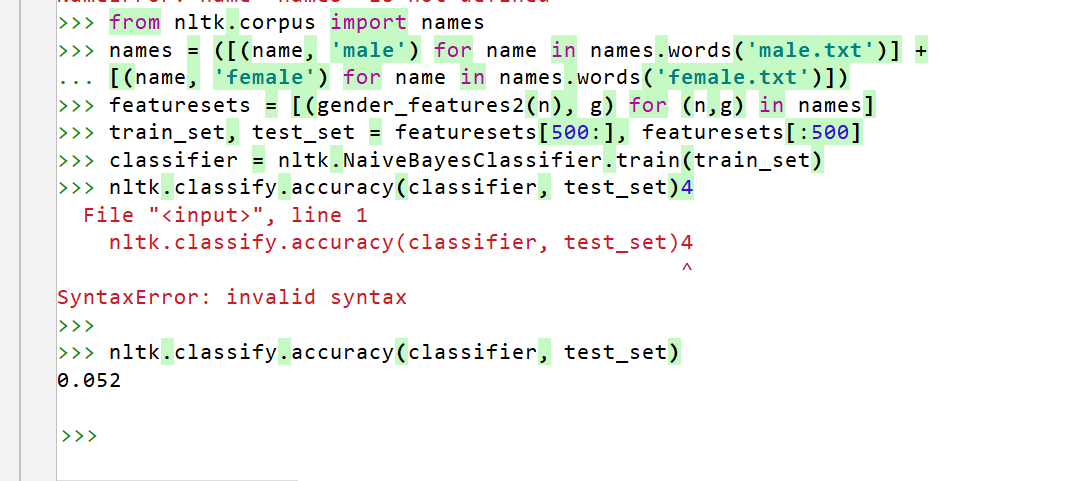
然而，你要用于一个给定的学习算法的特征的数目是有限的——如果你提供太多的特

征，那么该算法将高度依赖你的训练数据的特，性而一般化到新的例子的效果不会很好。这

个问题被称为过拟合，当运作在小训练集上时尤其会有问题。例如：如果我们使用例 6-1 中

所示的特征提取器训练朴素贝叶斯分类器，将会过拟合这个相对较小的训练集，造成这个系

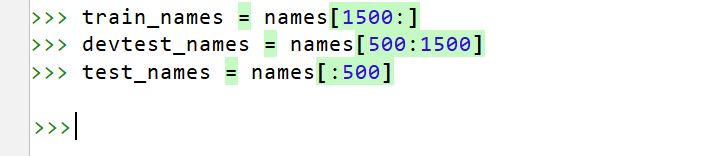
统的精度比只考虑每个名字最后一个字母的分类器的精度低约 1％。



一旦初始特征集被选定，完善特征集的一个非常有成效的方法是错误分析。首先，我们

选择一个开发集，包含用于创建模型的语料数据。然后将这种开发集分为训练集和开发测试

集



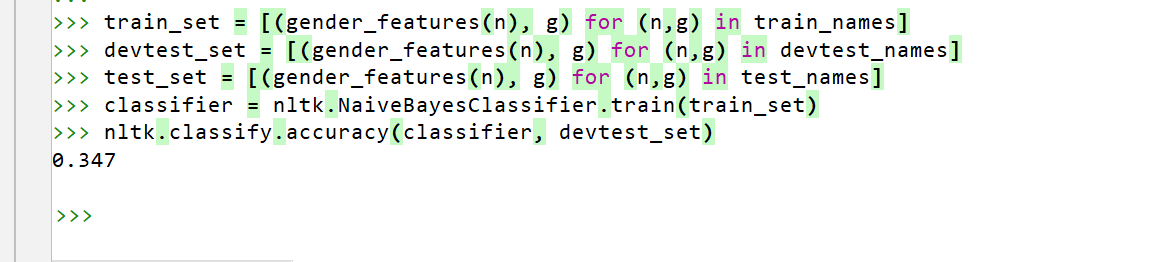
训练集用于训练模型，开发测试集用于进行错误分析，测试集用于系统的最终评估。由

于下面讨论的原因，我们将一个单独的开发测试集用于错误分析而不是使用测试集是很重要

的。在图 6-2 显示了将语料数据划分成不同的子集。

已经将语料分为适当的数据集，我们使用训练集训练一个模型�，然后在开发测试集上

运行



使用开发测试集，我们可以生成一个分类器预测名字性别时的错误列表。

