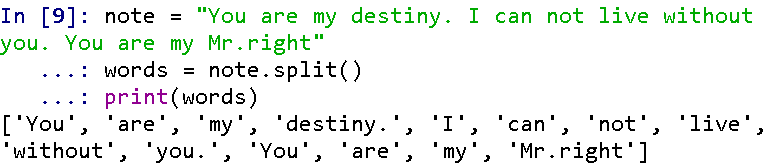
**使用朴素贝叶斯过滤垃圾邮件**

1. 准备数据

如何从文本文档中构建自己的词列表。

1. 对于一个文本字符串，可以使用string.split()方法将其切分，并保存为列表。

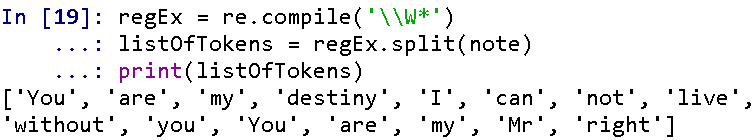


String.spilt()方法按照空格对字符串进行切分，可以看到标点符号被当成了词的一部分。

1. 可以使用正则表达式来切分，其中分隔符是除了单词、数字外的任意字符串。

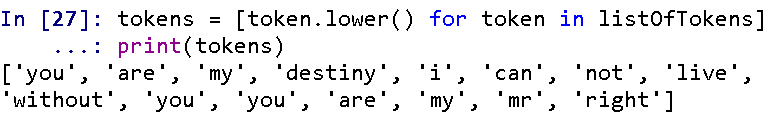
注意：使用正则表达式时，首先导入re包。

re模块中包含一个重要函数是compile(pattern [, flags]) ，该函数根据包含的正则表达式的字符串创建模式对象。可以实现更有效率的匹配。在直接使用字符串表示的正则表达式进行search,match和findall操作时，python会将字符串转换为正则表达式对象。而使用compile完成一次转换之后，在每次使用模式的时候就不用重复转换。当然，使用re.compile()函数进行转换后，re.search(pattern, string)的调用方式就转换为 pattern.search(string)的调用方式。

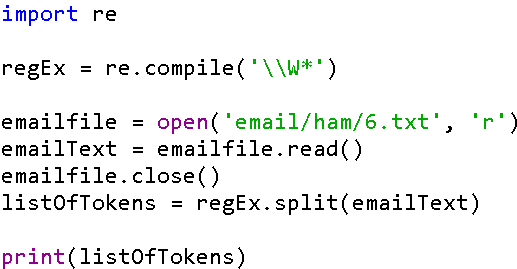


[\\W\*](file:///\\W*)匹配字母或数字或汉字

可以发现第一个单词是大写的，为了便于形式统一，使用string.lower()统一变成小写字母。

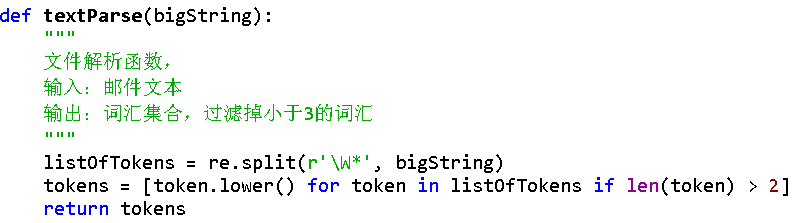


1. 将这个方法用于完整的电子邮件中查看实际处理的效果。

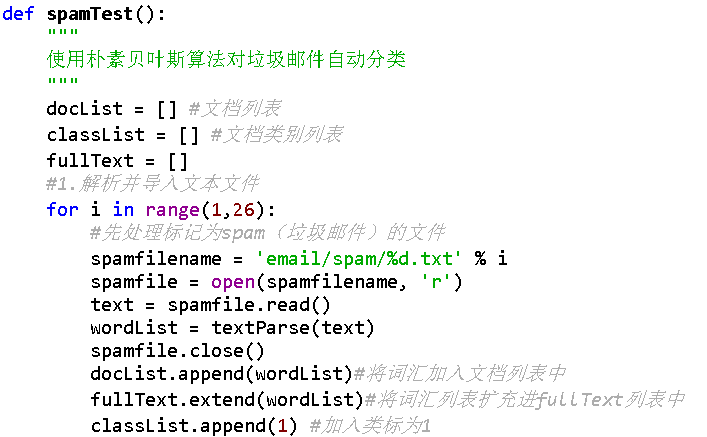


当前实现的结果，包含空字符串，而且形如：answer.py?hl=en&answer=174623的URL被解析为[“answer”, “py”, “hl”, “en”, “answer”, “174623”]，可以看到尝试对URL进行切分时，会得到很多词，可以过滤掉长度小于3的词汇。

1. 文件解析函数：



1. 垃圾邮件测试函数









运行结果为：

