1. **在Excel中选择出所有症状信息，对症状信息进行分词，具体操作如下(先将excel中隐藏的换行符去掉)：**

方法：CTRL+F调出查找替换对话框，选择替换  
查找中输入 ：按住ALT ，小键盘输入 10 ,然后松开ALT  
按替换

**2. 将末尾的标点符号去掉，将行中的标点符号用“ ”（空格）替换。将“腹泻”作为停用词去掉，将“全身症状与体征”、“性状”等无用词去掉。**

**3、发热程度划分**

以[口腔温度](http://www.a-hospital.com/index.php?title=%E5%8F%A3%E8%85%94%E6%B8%A9%E5%BA%A6&action=edit&redlink=1)为例，发热程度可划分为：

* [低热](http://www.a-hospital.com/w/%E4%BD%8E%E7%83%AD) 37.3~38℃（99.1~100.4F）
* [中热](http://www.a-hospital.com/w/%E4%B8%AD%E7%83%AD) 38.1~39℃ （100.6~102.2F）
* [高热](http://www.a-hospital.com/w/%E9%AB%98%E7%83%AD) 39.1~41℃ （102.4~105.8F）
* [超高热](http://www.a-hospital.com/w/%E8%B6%85%E9%AB%98%E7%83%AD) 41℃ （105.8F）及以上

[腋窝温度](http://www.a-hospital.com/index.php?title=%E8%85%8B%E7%AA%9D%E6%B8%A9%E5%BA%A6&action=edit&redlink=1) 分为低热型（＜38℃）中热型（38－39）高热型（39～40℃）、超高热型（＞40℃）。 人体最高的耐受温度为40.6~41.4℃（100.4~102.0F），[直肠温度](http://www.a-hospital.com/index.php?title=%E7%9B%B4%E8%82%A0%E6%B8%A9%E5%BA%A6&action=edit&redlink=1)持续升高超过41℃，可引起永久性的脑损伤；高热持续在42℃以上2~4H常导致[休克](http://www.a-hospital.com/w/%E4%BC%91%E5%85%8B)以严重并发症。体温高达43℃则很少存活。

本实验采用以下划分

* **发热程度划分**（注意，在正则表达式匹配时，.号要用反斜杠转义，\. 以下暂时没有转义，下次再处理时记得转义并验证）

以腋窝温度为例

低热（发热＜38℃）发热3[0-7]\.\*[0-9]\*°C （有的不带摄氏度符号，要去掉符号再重复一遍替换操作）

中热（发热38－39）发热38\.\*[0-9]\*°C

高热（发热39～40℃）发热39\.\*[0-9]\*°C

超高热（发热＞40℃）发热4[0-9]\.\*[0-9]\*°C

* **呕吐、腹泻程度划分**,记得考虑小数点,另外，记得先替换10次以上的，先替换频次中带小数点的，要不然像7.5次/天这样的频次，则会被[4-6]次/天匹配掉，这是错误的

超高频（10次及以上）[1-9][0-9][0-9]\*\.+[0-9]\*次/天 (先替换带小数点的低频次数，再替换不带小数点的) [1-9][0-9][0-9]\*次/天

低频（1-3次）[1-3]\.+[0-9]\*次/天 (先替换带小数点的低频次数，再替换不带小数点的，另外如果将.后面的+改成\*号，则会造成将10-39次之间的频次都换成低频) [1-3]次/天

中频（4-6次）[4-6]\.+[0-9]\*次/天 (先替换带小数点的低频次数，再替换不带小数点的) [4-6]次/天

高频（7-9次）[7-9]\.+[0-9]\*次/天 (先替换带小数点的低频次数，再替换不带小数点的) [7-9]次/天