**1. 拉数据（python爬虫）**

打开anaconda navigator，打开**JupiterLab**

选择已经写好的代码SteamComments.ipynb #需要预装selenium: conda install selenium

调整参数：



一个range7个，这里500个range就是3500个。

先实行这一步的代码，让网页把所有评论都刷出来；然后再用第二步的代码，把评论抓到表格里

**2. 洗数据，准备原始材料**

2.1 把每条数据用句号作为分隔，分成短句，数据-分列-按一定符号作为切分

2.2 把所有的数据排成一列：打开开发工具VBA，代码如下：

*Sub explicit()*

*Dim n As Long, i As Long*

*#将n、i定义为两个long（长整型）*

*Sheet2.Activate*

*Sheet2.Select*

*#将数据所在的Sheet2进行激活和选中*

*For n = 1 To 149*

*#n指表格中最长句被分割成的句数，也就是这张表最多的列数*

*For i = 1 To 1931*

*#i指表格的行数，也就是评论的数量*

*Sheet3.Cells(i + (n - 1) \* 1931, 1) = Sheet2.Cells(i, n)*

*#在sheet3中，最终会产生i+（n-1）\*1931个行，用两个循环，把sheet2中i\*n内所有的cell都放在sheet3中的一竖列里面*

*Next i*

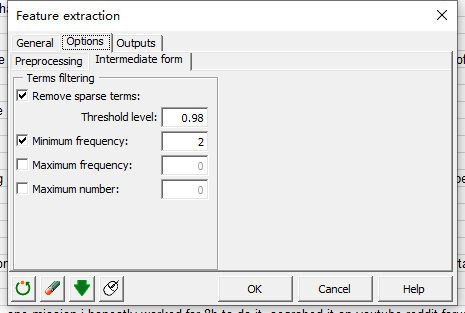
*Next n*

*End Sub*

**3. 整理数据**

3.1 词频分析

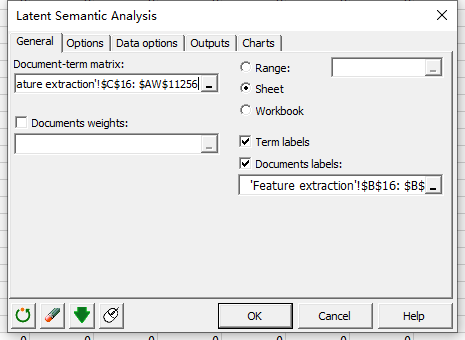
XLSTAT - Advanced features - Text mining - Feature extraction

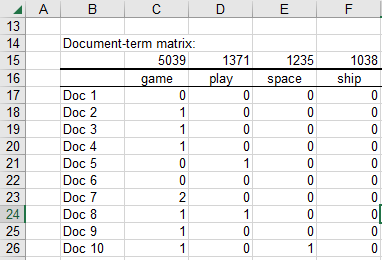


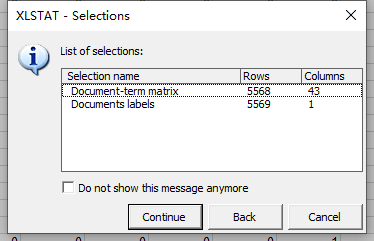
Remove sparse terms -> Threshold level 去除稀疏项

3.2 相关性分析

XLSTAT - Advanced features - Text mining - Latent semantic analysis







看到这个界面就是基本ok了！

Document-term matrix 要把整个范围拉下来，包括game、play这些关键词的第一行作为表头

Document labels 是每句话的label，这里拉的时候要多拉一行，从game、play的那一行开始拉，保证label的数量比matrix多一行，不然会报错

\*建议手动画框，打绝对位置容易打错

**4. 环境准备**

4.1 anaconda

下载anaconda后，下载 Chromedriver（[https://chromedriver.chromium.org/](https://chromedriver.chromium.org/%EF%BC%89%EF%BC%8C%E7%84%B6%E5%90%8E%E6%8A%8A%E5%AE%83%E6%94%BE%E5%9C%A8anaconda%E7%9A%84%E6%96%87%E4%BB%B6%E5%A4%B9%E9%87%8C%E9%9D%A2)），然后把它放在anaconda的文件夹里面

为了使用SteamComments.ipynb，需要在conda里面安装xlwt（方法：conda install xlwt）

4.2 XLSTAT

下载后就会出现在excel里面

Name: lavteam.org

Password: A53BD511-C14C1A0E-C050C6EF-B2CCF243