1. 获取关键词，返回词典（key, value）格式数据，key为关键词，value为出现次数；

2. 获取所有新闻id和content，返回词典（key, value）格式数据，key为id，value为content；

3. 对关键词排序，统计出现频率大于1的关键词

4. 以关键词为行，新闻id为列构建二维数组

5. 统计关键词在content中出现的次数，完善二维数组，结果如图1所示

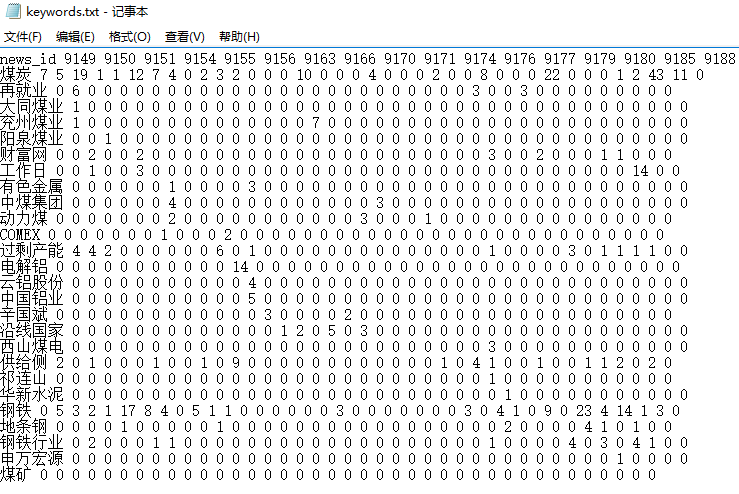


图1 词频统计数据

6. 将二维数组看作一个mxn的矩阵，将二维数组进行奇异值分解，得到u, sigma, vt三个数组，u为mxr的矩阵，sigma为rxr的矩阵，vt为rxn的矩阵，奇异值分解如图2所示，其中sigma除对角线外其他值都为0，sigma数组对角线的值如图3所示

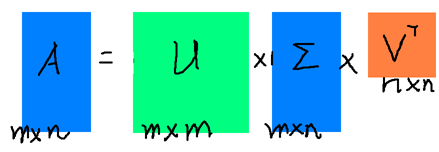
[](http://images.cnblogs.com/cnblogs_com/LeftNotEasy/201101/201101192226341537.png)

图2 奇异值分解

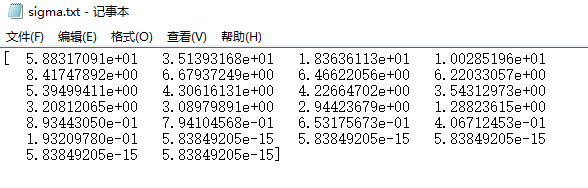


图3 sigma矩阵

7. 简化sigma矩阵，sigma矩阵变为3x3的矩阵sigma3，u矩阵变为mx3的矩阵u3，将vt矩阵变为3xn的矩阵vt3，其中u3矩阵如图4所示

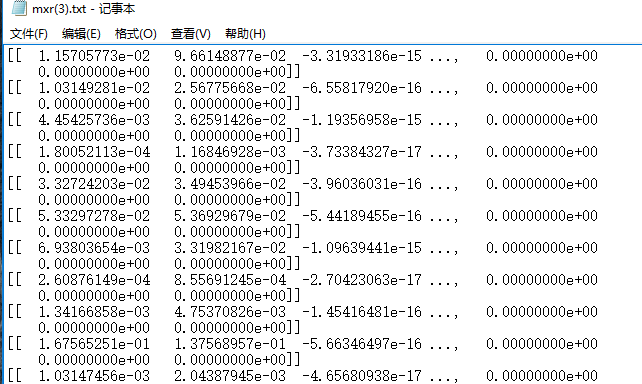


图4 u3矩阵

8. 使用excel，将u3矩阵转化为散点图，散点图如图5所示

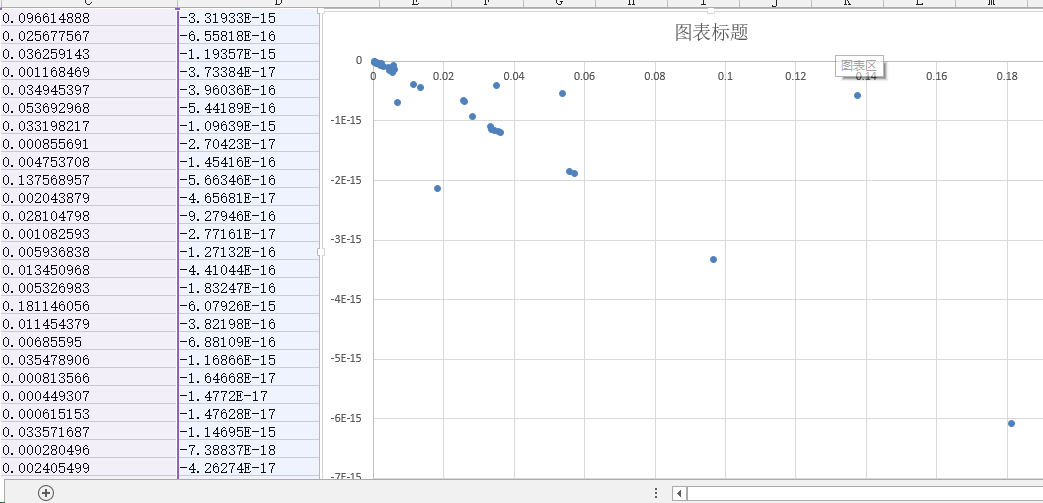


图5 散点图

9. 使用kmeans算法聚类，结果如图6所示

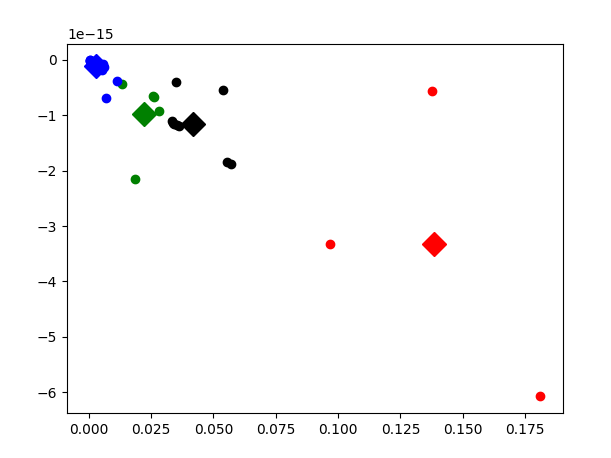


图6 k-means聚类结果