本次例会讨论确定尽量减少人工介入而自动挖掘

**预处理组：**

建议加上命名实体识别（主要是地名）

建议尝试是否可能直接用TF-IDF构建关键词表。个人感觉停用词对TF-IDF的干扰主要是因为IDF项是取对数的，如果直接限制IDF值可能就可以自动构建停用词表

建议尝试各种不同的构建同义词表的方法

建议尝试各种不同的构建固定词组表（自定义词典）的方法

**聚类和分类组：**

希望确定使用关键词表相对于直接用TF-IDF对聚类性能的影响到底怎么样

建议利用聚类尝试自动构建分类树（并对树的每个节点加关键词标签，对叶子节点加建议的处理方法）的各种方法