本次例会讨论了是否采用基于深度学习方式进行聚类？感觉可以尝试，但是可能我们拥有的数据量导致不一定能训练出比较好的性能

**预处理组：**

孙浩然通过增加地名集提高了命名实体识别的效果，建议后续采用基干词+前缀词的方法。关于如何确定前缀词的范围，建议采用分词基础上的“共现->固定词组”的方法

杨迈尝试：

同时采用多种聚类方法，提取其公共部分并增加其重复关键词权值。并减少部分其中的低频词汇的权值。然后再次尝试聚类

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：**

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

**以前讨论的总结：**

尽量减少人工介入而自动挖掘，例如自动产生关键词表，停用词表，同义词表。可以考虑使用聚类结果作为评价标准

对每段报警信息是给出单一的类型还是给出标签？初步考虑以聚类产生的层次性类型为主，对某些特别的报警信息可以再加上副标签

采用余弦距离比欧氏距离效果好