感谢杨迈和孙浩然的卓越工作，自动聚类和地名识别都已经有了较好的结果，可以此两项成果为基础，加上我们以前做的行为推荐，形成新一版原型系统。建议下次例会的时候集思广益，讨论一下这版原型系统做成什么样

**预处理组：**

杨迈：

在TF-IDF向量化不好前提下，采用多种聚类方法，通过研究多种聚类方法的公共部分尝试自动提取核心关键词，并以此为基础，修改TF-IDF矩阵权值。然后迭代一定次数或达到某些条件作为结束。对大部分类效果较好（只有others类效果不好）

后续建议：一方面尝试类内继续聚类以形成层次结构，另一方面尝试利用同义词/近义词减少错分到other类的样本。由于k-means等聚类方法的分类结果有偶发性，建议多次分类以后取稳定的结果

孙浩然：

继续提高了地名识别的效果（还是依赖于介词），建议后续对偏僻的地名尝试采用建议地名集的方法解决

聂志捷：

建议后续尝试做分类

建议尝试各种不同的构建同义词/近义词表的方法

建议尝试自动构建停用词表

建议尝试各种不同的构建固定词组表（自定义词典）的方法并和开源的分词方法结合

**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，例如自动产生分类标签词、关键词表、停用词表、同义词\近义词表、地名表。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