本次例会确定了以非结构化接警记录的结构化作为工作重点，智能推荐只是其一个典型应用

**关于预处理：**

目前我们没有做地名识别，这在数据结构化中还是比较重要的，建议后续加上

关键词表对性能影响很大

关键词表中的每个词需要扩充为同义词集

如果关键词表提取得比较好，是否就可以只用关键词而不用其他词（这样就不用建停用词表了）？

如何确定关键词表？可以考虑采用交互学习（人工先定一部分，再通过计算机推荐之后人筛选来不断添加）；也可以考虑根据聚类的效果来评价哪些关键词表效果好

如何确定关键词表中词的权重？建议采用TF-IDF的思路来算

建议加上一些固定词组（自定义词典）以应对一些意义比较固定的短语，需要考虑在分词前还是分词后做

如何生成自定义词典：可以采用类似关键词表的方法，也可以考虑利用共现信息

**关于分类：**

一些很强的词单独就可以决定分类，而不一定需要算向量的欧氏距离或者余弦距离

是否考虑采用决策树进行分类？

是否要层次式分类（先分大类再细分）？