文本挖掘难点：

基于规则的特征词提取难点

文本本身问题：

1. 文本内容缺失。

从网络爬取的新闻中存在很多只有题目而没有新闻内容的文本，该类文本对突发事件的特征提取仅仅只依靠文章题目，容易造成提取缺失。

1. 文本内容嵌套。

同一条新闻中提及2-3条突发事件，形成新闻嵌套。

1. 文本内容过于冗杂。

文本描述过长，给时间和伤亡人数等类型的提取带来检索困难，增加错误率。

语料问题:

1. 使用'结巴分词'对文本进行分词时文本切分容错。

主要原因是，用于训练'结巴分词'软件的语料库为普适语料而非特定的新闻语料，导致对特定新闻类文本进行切分时，分词结果不够理想，使得后续检索任务加重。

1. 汉语的多元化及中文表达方式的多样化。

相比英文而言，汉语表达变化多端，识别难度增大。

特征词问题：

1. 突发时间特征词库缺乏。

在对事件类型进行识别、分类时，往往需要在待分类文本中检索应急类事件的特征词，从而确定类型。目前没有此类词库供使用，需要人工添加。

1. 事件特征词重叠。

不同主体出现同样事故类型，难以细化区分。比如：桥面崩塌和地面崩塌。根据所设规则，机器可以在切分后的文本中将'崩塌'识别为突发事件的关键词，但很难判断是哪里崩塌，也就是说，不知该输出桥梁相关事件还是道路突发事件。

传统机器学习方法的难点：

数据源产生的问题：

1. 语料库（训练数据）的数据量不够，涉及的事件类型单一，且数据质量差。
2. 人工分类耗时长，工作量大

机器学习最基础也最重要的一个方面是让机器'学习'，学习的内容由我们传输，我们需要告诉机器哪些东西是需要的，哪些是不需要的。当这些内容积累到一定程度后，我们希望机器可以得出他们为什么重要的原因，从而判断一个新的文本是不是重要，或者说是否为我们需要的内容。所以交给机器的内容就至关重要，这个内容就是已经分过类，质量好，且已经被分类处理过的事件文本。

模型问题：

1. 模型调参需要经验。
2. 基于不同的和数据情况如何选择分类器需要斟酌。
3. 使用传统的 TF 或 TF-IDF 生成的 VSM (向量空间模型)，很难应对语义问题、同义词无法进行准确分类
4. 传统机器学习方法，对于多类别分类存在很多局限性，待优化。