# 1 获取语料

## 爬取博客文章和论文

使用关键词“机器人故障”搜索相关博客文章和论文，爬取博客文章的标题、摘要、关键词、正文、时间、阅读量、点赞量（如果有）；爬取论文的标题、摘要、关键词、正文、参考文献名（如果有）。

爬取方法不限，推荐python+selenium；博客和论文网站不限，目前已有简书和维普期刊。

## 过滤无关语料

虽然是用关键词“机器人故障”搜索出来的文章，但爬取的内容还是有很多不相关的，需要过滤不相关内容。

### 1.2.1 规则过滤

在文章标题或摘要中，必需都出现过“机器人”和“故障”两个词语，否则视为不相关内容。

### 1.2.2人工过滤

经过规则过滤后，再进行人工过滤，去除不相关文章。

# 2 获取概念

## 2.1 分词

使用jieba对获取的语料进行分词。保存分词得到的词典以及分词结果。

## 2.2 统计词频

统计所有词语的信息，包括：（1）使用sklearn-CountVectorizer统计出语料中的关键词；（2）使用sklearn-TfidfTransformer统计出词语的TF-IDF；（3）统计所有词语各出现在多少篇文章中；（4）统计所有词语在语料中的出现次数。

## 2.3 获得候选概念

将所有论文的关键词以及上述4种统计方式各自得出的前100个词（共400个词）作为候选的概念。

## 2.4 过滤无关词语

候选概念主要是词频统计得到的，而词频高的词语中有很多是无关的常用词，如“可以”、“进行”等；此外，需要合并同义词。

### 2.4.1 过滤无关的常用词

使用大规模通用语料训练好的词性标注模型对词语进行标注，过滤非名词。

### 2.4.2 过滤同义词

过滤无关常用词后，使用word2vec在已获语料中训练词向量，计算词语相似度。相同语义的词只留下词频最高的。

### 2.4.3 人工过滤

最后进行人工过滤，去除非概念词语。

# 3 关系抽取

## 3.1 计算关系强度

统计每两个概念出现在同一句子中的次数，如果次数超过5次，则这两个概念之间有强的关系。

## 3.2 确定关系类型

对于有强关系的两个概念，选取同时出现这两个概念的句子，使用两种方法确定它们之间的关系。（1）基于规则：制定常见关系（如继承、组合等）的正则表达式，在选取的句子中进行正则匹配，看两个概念是否存在这些关系；（2）基于动词：使用大规模通用语料训练好的词性标注模型对选取的句子进行标注，将句子的动词作为这两个概念的关系。

## 3.3 去除关系同义词

使用word2vec在已获语料中训练词向量，计算词语相似度。相同语义的关系词统一改为词频最高的。

# 4 人工检查

经过概念获取和关系抽取后，得到初步的本体三元组。再进行人工检查，去除错误的三元组。