操作系统Project技术方案

组员：王楚然，陈佳辉，成昱霖，杨志昊，肖昂弘，许卓佳，黄冰，钟振远，陈家兴，钟声亮，曹丽洁

**基础方案**：基于概率模型的关键词检索

**实施：**

**文本预处理**（其流程如图1）：

1. 噪声去除
2. 词汇分析
3. 排除停用词
4. 词干提取

词序列

得到干净查询

排除停用

词汇分析

噪声去除

关键词

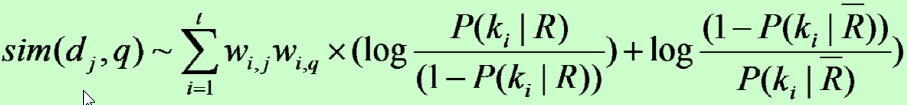
有用词序列

词干提取

**检索：**

1. 用布尔向量表示搜索语句和库中的描述文本；
2. 计算词汇的概率初值，根据相似度计算公式计算每个文本向量和查询向量的相似度，相似度公式为：
3. 按照文档向量和查询向量的相似度，排序输出初始排序结果集（即描述文本为键，case为值的键值对结果集）；
4. 在初始结果集，用户指定或按缺省约定选择相关文档，形成相关文档集合；
5. 根据初始概率的改进公式或其变形公式，计算初始概率分布或其变形公式，计算初始概率分布；
6. 重新计算各文档与查询的相似度，排序输出最终结果集（排序），根据键值导出case。

注，相似度公式为：



**检索思路2**：基于概率模型的关键词检索出候选API，通过case分析选出产生候选API的概率最大的case，成为候选case参与结果排序输出。

**进一步的方案:**

利用我们的基本方案积累大量的用户需求描述与case集合（用户需求为输入，case为输出的数据对）。通过这些输入输出样本，训练R3NN模型用于自动编程。

**具体做法如下**：

需求描述（自然语言）与输出通过IO Encoder（利用LSTM网络）来得到IO对，区分出哪些输出可以通过输入计算得到，哪些特征能帮助识别出有用的API。将IO对输入到R3NN模型进行训练，得到自动编程模型。（见流程图）

