2021-03-12

作者：郭礼华

时间：2021.03.12 19:30-21:00

参与人：程荣鑫、郭礼华、韩禧、庄子元

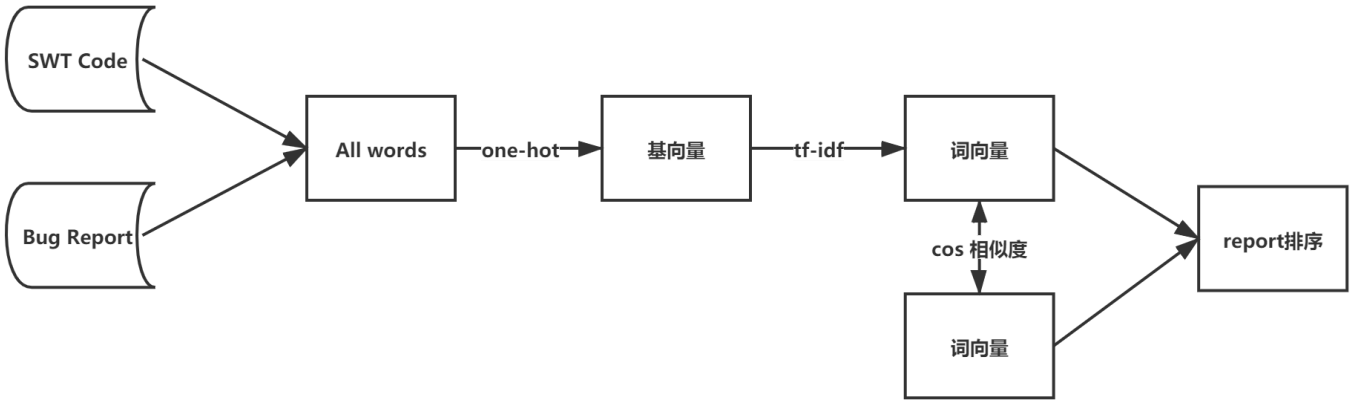
前情提要：前期已经确定了项目的需求以及约束

会议主题：项目迭代一实现方案及分工

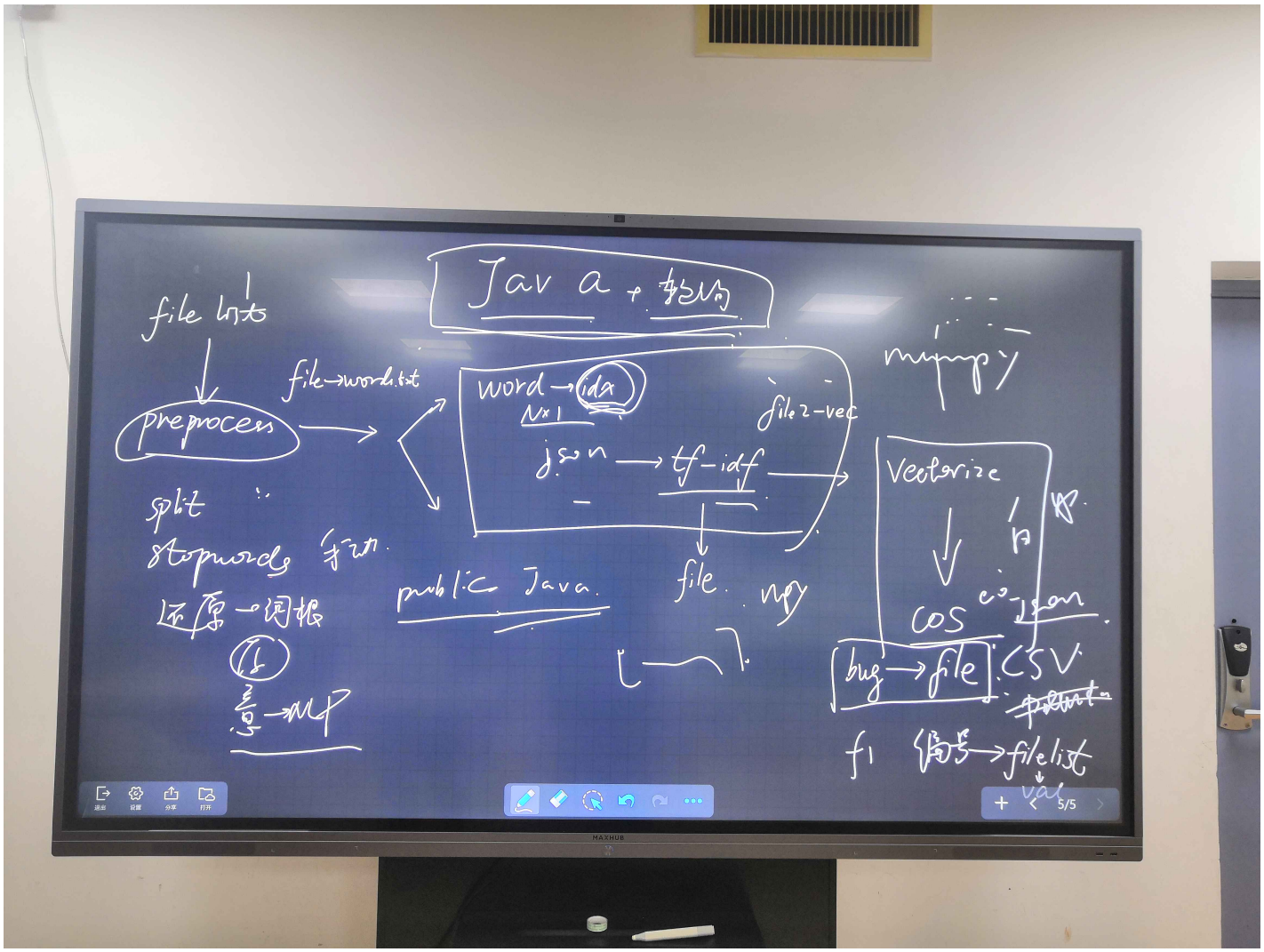
记录方式：拍照，录音以及根据录音事后整理的文档

要点：

1. 确定了项目的大致流程，理论模型如图所示



2. 项目具体的模块将分为以下几个部分



各模块描述及分工如下

- Java主模块 负责人：韩禧

  负责编写终端程序，能够根据用户的输入来执行相应地动作，具体的实现在各python文件中

- preprocess模块 负责人：庄子元

  负责swt源代码以及bug报告的预处理，具体为将各单词切分，并根据停用词进行词汇筛选以及词性还原

- word2idx并计算tf-idf 负责人：郭礼华

  负责计算每个文档的tf-idf值并转换成词向量存储

- vectorize并计算cos相似度 负责人：程荣鑫

  负责计算每个文档之间两两的cos相似度并进行bug报告与代码的相关性排序

3. 未来迭代中可能会用到的想法

   3.1 尝试用word2vector来压缩向量空间的大小

3.2 利用JavaParser来识别出Java源代码的变量名、函数名、类名、注释等关键信息，并给予他们较高的权重，对于如import进来的第三方包则给予较低的权重

3.3 在计算相似度时考虑引入矩阵M，相似度公式改为$sim(x1, x2) = x1 \cdot M \cdot x2 / (|x1| \cdot |x2|)$，可以对M进行学习并获得最优值