**分类器训练模块思考**

——胡杨 2014.7.12

分类器训练模块被我分成两个阶段：

1. 分类器测试文档准备和标记检查+分类器任务管理（新建任务+运行到对应阶段更新任务状态+任务状态切入各种情况处理）。
2. 分类器结果汇总和统计+分类器效果评测+分类器任务管理（任务总结）。

分类器任务难点：

1. Weka统一将测试文件夹中的所有文档读取，分类之后只能返回文档在文件夹中的编号和其对应的评分结果等，所以要将分类的结果和语料库中的语料准确地对应起来，因此建立测试文件夹文件对应信息缓存池，相关还要检查缓存池数据和测试文件夹文件以及两者关联关系的完整性，如果发生错误，分类任务无法进行。
2. 分类测试文档的元信息中，“labeledSetId”字段在训练型分类任务开始之前需要被检查，因为这个信息标记了这个文档在开始分类之前被标记的类别，只有这个信息存在，才能统计测评分类器的效果，所以开始训练型分类任务之前，需要检查所有测试文档关于这个信息的完整性，如果有标记信息不完整的文档，系统返回对应的元信息让用户进行标记，训练型分类任务才能继续。
3. 后期任务难点还没想到。

功能部分调整建议和问题：

1. 分类器训练模块在选择测试文本的时候（训练文本应当提前就准备好），可以设定使用两种方式：1从语料库中选择；2自己手动上传
2. 详细说明上面测试文档选择方式第二种：自己手动上传从用户方便的角度，不适合再转到其他页面进行上传，应该在本页面就地上传，但是还是存在“不良总类”文件夹中，在mysql当中的元信息也标注“blongSetId”属于“不良总类”，但是标记编号“labeledSetId”需要用户手动录入，用户也可以批量上传，批量录入标记信息。上传之后，自定义文档的名称仿照语料文档的格式“上传时间（精确到秒）”+“（diy）”，nodeId为“上传时间（精确到秒）-diy”，上传之后自动从“不良总类”文件夹中复制一份到测试文件夹，并自动加载测试文件信息缓存池。

Ps：在“不良总类”中的文档“blongSetId”必须是1，代表本身文件存在与“不良总类”中，但是“labeledSetId”可以是其他分类的值，代表已经被标记成某个类别的文档，方便于评估分类器。标记信息可以自由修改。

1. 训练型分类任务设定不能处理“未被标记”的文档（就是对应元信息中“labeledSetId”字段为1的文档），所以对检查到“未被标记”的文档需要弹出窗口让用户进行标记。

其他功能：

1. 对于分类产品**快速检索的拼音检查功能**：比如用户没有切换输入法，直接键入“zhongguo”，应当ajax提示用户“中国”，如果用户没有采用提示，而直接键入回车，在实际检索当中也应当有限反映出“中国”的检索结果。这个功能需要增加一个关联表存储“不良敏感词”的拼音信息。使用ajax实现动态提示，可以作为本项目的一个小亮点。
2. 基于同义词的“概念检索”：比如检索“岛国”，其实和“日本”是同义词，反映出“日本”的检索结果，这个功能已经在lucene索引流中实现，但是需要在用户检索的时候，反映给用户这个功能的体现，具体就是在检索结果页面上方提示“您搜索的是岛国，提供同义词日本的搜索结果”，这个功能可以使检索结果显得更加专业。

任务规划时间表：

1. 下周二前（2014.7.14）完成分类器训练第一阶段内核任务，完成任务管理机制和错误恢复机制。
2. 下周日前（2014.7.20）完成分类器训练模块所有内核任务。
3. 2014.7.22前完成分类系训练模块完整任务。
4. 2014.7.29前完成分类器正式任务完整任务。
5. 后期需要按照上面“其他功能”的想法升级分类产品的检索功能，完成分类任务服务的功能，加入初级的倾向性分析功能。