**IIIMMS NLP And Correlation Function Initial Plan**

**HuYang 2014.06**

1. **前期采集器模块工作简述**
2. 前期的采集器任务已经可以对站点进行完整的爬取和解析，爬取和解析过程中需要用到的参数存放在mysql数据库当中，采集过程中采集到的主题内容存放在mysql数据库tbl\_theme表中，节点内容（不包括文章和回复，仅仅是以URL为主的基本节点信息）存放在mysql数据库tbl\_post表中，解析到的每个节点的具体文章和回复内容存放在mongodb中，存储结构尽可能还原网页原始的结构，即，在每个节点下面包含1或n个文章元素，每个文章元素下面包含0或n个回复元素。
3. 现已经完整对mongodb语料库的各种提取功能，能够实现较为完善的查看机制，可以单独取出对应节点的文章元素，取得文章元素后可以去的该文章下的回复元素。
4. **NLP引擎概述**

后期的任务思路首先从手头已经拥有的资料和元素入手，思考的重点是如何去处理和利用它。

1. 首先是对于不良信息的定义和分类，这一部分必须人工进行（question：没有任何机器代劳的简化手段么？），这一部分，必须完成：1.不良信息敏感词库的建立；2.不良信息分类训练语料库的建立。
2. 语料库的预处理：可能必须要将mongodb中的语料单独取出来，暂时存放为txt文件，作为训练语料库。txt文件的结构应该如下：
3. 对于不良信息敏感词库的建立，通过人工配置（question：有无更好的办法），在已经配置了一定量的敏感词之后，通过lucene在剩余语料中对敏感词进行搜索，检索之后发现出现频率高的文档重点检查，在其中挖掘出更多的敏感词。

Ps：敏感词库的要求：1.敏感词库也需要对应不良信息主题的类别来进行严格的分类（必要）2.敏感词库可以对每个敏感词定义对应类别的权值或者称为评分作为敏感程度的登记，作为后期为文档权值的设定做准备（非必要）。

1. 不良信息分类训练语料库的建立，通过人工配置，语料库的形式是2中描述的txt文档，存放在相应分类的文件夹中。不良信息主题的分类结构：最适合式采用两层模式的分类体系，第一层是对不良信息的大分类，在每个大分类下细化成若干小分类，再往下细分没有必要
2. 对于分类器的训练，训练过程可以借助开源工具，目前发现大多数开源工具训练过程自动化程度较高（question：自动化太高可靠度受怀疑，是否能够通过在预处理，特征选择，权值设定，相似度计算等方面加入更多的个性化，是否方便），训练完成后需要对分类器进行评价，如果不成功，需要为用户提供解决方案，最起码要提供操作规范和修改建议，分类器的训练分为手动训练和自动根据新语料分类。
3. 分类器尽量采用增量训练方式，（question：问题是哪些分类模型支持增量训练），分类器需要不断增强，用新的语料替换旧的语料（txt文本目录是随时更新不稳定的，而mongodb用于持久存储），（question：分类器训练语料库的更新策略和方法不明朗）。
4. 主题规化树根据不良信息分类结构建立，（question：用户自定义功能尚未很好的理解），我的设想，用户自定义主题规化树事先提供所有一级和二级的分类，用户可以自定义那些分类停用或使用，（question：用户自己定义新的分类暂时想不到后面对应的实现方案）。
5. 关于聚类（question：聚类服务功能意义不明朗）
6. **NLP主功能及其卫星功能的工作量规划**
7. 对敏感词的收集和定义，敏感词库的存储及管理，敏感词分类乃至分级（见：二3）等一系列元数据（数据的属性，下同）的数据库管理相关操作。
8. 文章相似度计算功能模块实现和文章特征提取功能模块实现。
9. 文本训练语料库的相关数据库管理相关操作；对训练语料库的升级刷新操作功能，在数据库中存储训练语料库的相关信息（比如：路径，等级，数量，所服务分类模型等等），训练语料库更新及其他操作的日志信息数据库管理。
10. 分类器信息的数据库信息管理（分类器可以在本地磁盘持久化保存），对分类器的选择。
11. 不良信息分类主题属性数据库信息管理，分类主题的激活和停用。
12. 对于停用词库和同义词库的建立和对应数据库信息的管理（涉及到语料，词库等记录数量巨大的情况，存储方式需要研究）。
13. 分类器训练以及对新加入文本的检验功能实现以及页面触发。
14. 对不良信息片段在语料中的全文检索，lucene配合分类器对中文进行分词（外接第三方分词器），lucene配合分类器训练。
15. 每次对新采集语料进行分类后分类结果的分析统计处理，对分类器的评估，得出评估报告，分类任务运行日志数据库的管理，分类任务成果日志数据库的管理，可以关联上tbl\_post表。
16. 以上功能的界面实现。
17. 其他功能及其界面实现。
18. **暂时想到的几个界面**
19. 敏感词库分类数据库管理界面
20. 训练语料库分类数据库管理界面
21. 训练语料库更新日志数据库管理界面
22. 分类器信息数据库管理界面
23. 不良信息主题数据库管理界面
24. 停用词库和同义词库数据库管理界面
25. 分类器训练控制界面（需要整合入分类器评估结果界面）
26. 使用训练好的分类器预测新语料类别的操作控制界面
27. 敏感词库，训练语料库元数据及更新日志数据库管理界面
28. 不良信息片段全文检索及结果展示界面
29. 分类结果分析统计界面（不了解，一句话带过）
30. 不良信息识别及分类结果展示页面
31. 敏感词库，训练语料库自动升级操作控制页面（不明朗，描述模糊）
32. 其他界面