**IIIMMS\_NLP模块7月工作总结**

——胡杨 2014.7.27

1. **文本分类器训练以及执行分类核心API实现**
2. 分类器训练及模型生成，使用开源自然语言处理工具weka[1]，扫描训练文本文件夹，生成model后缀名的分类模型在磁盘上。改进arff文件的生成形式，形成字符串类型到标称型数据和数字型数据的过滤转换[2]。
3. 分类器执行分类，分类结果实现获得最佳分类结果和某个文档对于每个类别的分类得分。
4. 分类任务日志机制的设计和实现，文本分类器执行分类任务，分类任务分成训练型任务和正式任务两种（训练型任务需要进行分类器评测，以便调整分类器参数优化分类器，正式任务可以直接对分类结果进行处理，但是不对分类器进行准确性评测），执行分类任务会记录分类任务信息和执行状态，分类的结果即时存入分类结果缓存sql数据表。
5. 分类语料的管理和结果对应机制，每次分类之前，要求从语料库中拷贝需要进行分类的测试语料集到分类测试文件夹，并在系统中建立相应的缓存池以记录测试文件和文件元信息的映射关系，这样在分类结果获得的时候，就能准确对应特定文本和结果的关系。
6. 分类器执行分类，将细化的分类器实现类整合到一个总类当中，并整合入日志管理和结果处理以及评测等功能，形成热插拔式调用方法，方便前台组员方便调用。
7. 分类器结果准确度评测，对于分类任务，如果属于训练型任务，会对分类的测试文档进行提前的标记，通过最终得出的结果标记和原先的标记进行比较实现对分类任务的准确度的评测，编写分类器测评类对分类结果进行统计计算，得到每个类别的查准率，查全率，F1测评值以及总体任务的以上三值，采用微平均计算法计算[3]。
8. 分类器训练任务和正是任务执行全部前台功能刘卓然实现。
9. **语料检索索引管理机制设计和内核实现**
10. 设计sql数据库表，实现对于索引文件的基本管理，可以自动扫描原始情报库中是否有新抓取的文件，如果发现有新增文件，自动对索引文件进行增量加入新文档信息，可以根据采集任务在索引中删除对应的文档信息，并且支持将已经删除的文档信息恢复到索引当中。
11. 实现对索引中有效文档数，和回收站中文档数的统计[4]。
12. 实现对于索引文件在磁盘上的根除，并且彻底重新建立新的索引。
13. 设计索引管理控制面板界面，索引管理前台功能由戴丹实现。
14. **分词器API优化，整合入同义词流**
15. 从sql数据库中读取同义词信息，转化成支持lucene的analyzer分词器数据流。整合入分词器工厂，在其他部分调用分词器工厂生产出对应的分词器的时候，就已经载入了数据库中所有同义词的信息，在全文检索的时候可以实现“中国”和“天朝”的同义词检索结果同时显示，这部分代码封装性极高，外面调用者不会看到内部变化。
16. **不良信息产品全文索引建立，全文检索核心API实现**
17. 从mongodb中读取所有的分类产品的内容和基本信息，建立lucene索引，并实现简单的索引管理功能（全部删除，重新建立）。
18. 编写全文产品lucene全文检索核心方法，返回对应产品的元信息映射对。
19. 分类产品全文检索前台功能由戴丹编写。
20. **分类器模型管理机制设计和内核实现**
21. 对于分类模型，设定有两种状态，冻结和激活，设计sql数据库表进行模型数据的管理，模型被冻结之后，在磁盘上不存在对应的模型实体，在sql数据表中保存记录，激活状态为为0。
22. 自动扫描训练文件夹中的文件是否发生数量和内容上的变化，如有变化，自动触发重新训练并建立模型的操作。
23. 分类模型固定设置有三种算法的不同实现（Knn算法，SVM算法和naïve bayes算法），由于不同算法适应于不一样的数据集（标称型和数字型）[2]，所以提供两种不同的数据性质arff数据集，在训练模型之前就已经由训练文本提炼并且准备好，编写两种数据集之间的切换方法，切换之后，模型在下一次训练的时候将使用新数据集。
24. 分类器模型管理前台功能由王秋森实现。
25. **文本情感倾向性功能探究简单倾向性分析服务API实现**
26. 情感分析核心程序使用开源自然语言处理工具Lingpipe中的相应模块，正负极性分类功能采用的算法是svm分支中的最大间距线性算法（LMCClassifier）[5]，单独设置极性分类训练语料库（固定）以及分类测试集文件夹，编写极性分类模型训练实现代码。
27. 编写极性分类执行分类操作类，获得最佳分类结果和正负极分值参数，参数的计算使用正向情感得分减负向情感得分，将此结果存入语料信息sql数据库表。最终结果如果为负就代表是消极文档，如果为正就代表积极文档。
28. 文本倾向性分析服务整合如主题预料管理界面，前台功能又戴丹实现。
29. **后期规划和遇到的问题**
30. 后期工作：（1）探究并实现全文检索的拼音检查和拼音提示以及同义词检索结果提示功能；（2）探究并实现文本倾向性分析的主客观分级功能；（3）进一步完善分类器准确度；
31. 遇到的问题：（1）现阶段最大的问题是语料匮乏，导致文本分类和文本倾向性分析无法正常进行，现在准备采用搜狗传统语料库测试分类器，但是中文正负极语料库暂时没有找到；（2）其次，Lingpipe和weka等自然语言处理工具使用还是不熟练，众多高级功能不知道如何使用，资料匮乏，多为英文，需要进一步研究。

**参考文献**

[1] 曲伟. Weka开发［-1］——在你的代码中使用Weka[CP/OL]. [2014-7-20]. http://quweiprotoss.blog.163.com/blog/static/40882883200931113314795/.

[2] Harrington P, 李瑞[译], 李鹏[译], 等. 机器学习实战[M]. 北京:人民邮电出版社, 2013.6:15-89.

[3] 刘挺, 秦兵, 张宇, 等. 信息检索系统导论[M]. 北京:机械工业出版社, 2008.12:194.

[4] McCandless M, Hatcher E, Gospodnetic O, etal. Lucene实战[M]. 北京:人民邮电出版社, 2011.6:35-44.

[5] Lingpipe. Sentiment Tutorial[CP/OL]. [2014.7.23]. http://alias-i.com/lingpipe/demos/tutorial/sentiment/read-me.html.