**IIIMMS NLP Open-Source Tools Assessment**

**HuYang 2014.06**

1. **Weka**

功能最为强大，几乎支持所有分类，聚类算法，java实现，为开发者提供api和详细的doc。但是有一个问题是，只处理自己定义的arff格式文件，目前已经找到通过txt文本文件库经过一系列预处理形成arff聚合文件的方法，不算太麻烦，过滤器强大，可以不借助lucene的过滤器，与lucene整合较好，资料较全。

推荐指数：4.5（5为满分）

1. **Mahout**

Hadoop旗下子项目，分类聚类性能良好，安装需要maven和hadoop，支持单机运行，可以建立在单个节点的hadoop伪集群上，更新频繁（所以需要maven，在项目上加上maven支持），不支持Knn算法，不支持DBScan算法（LDA方法实现得很好）。从相关论文的涉及方向来看，更多被用于推荐系统，可能推荐功能比较强大。对lucene的整合情况尚未研究，安装和使用应该比较方便，不用分散太多精力在hadoop上，资料较全，最大的价值是对于项目后期升级为分布式系统大有好处。

推荐指数：4.5

1. **Lingpipe**

不支持SVM算法，不支持DBScan算法（支持LDA方法），其他算法还算齐全，分类器的训练和预测模块实现超级简单（总共120行），让人对准确度产生怀疑，不过使用真的很方便，可以直接分析txt文本文件，与lucene整合较好，资料基本就是官方文档，英文，但是较详细，晚上看到一些人已经开始从lingpipe转入mahout，不知道本身是不是再走下坡路。可以当作辅助吧。

推荐指数：3

1. **carrot2**

开源软件，只有源码，没有api，资料缺乏，自己提供一个界面是基于eclipse的，但是不方便在eclipse中编译，而且很少被用于项目生产上，大多数是科研使用。

推荐指数：不推荐

1. **mallet**

java实现，没有过多了解，仅仅知道强项在于决策树算法，可以作为weka的补充。

推荐指数：2.5