搜索引擎第二次作业报告

基本任务

1. 跑通流程：学习了Tomcat，MyEclipse的使用，
2. 实现BM25评分算法
3. 实现VSM模型评分

拓展任务

1. 在ImageIndexer中同时对所附带的html文件进行解析并索引，并在ImageSearch中对html域也进行搜索。关键点在于读取html文件的编码并正确解码，同时提取内容并删掉标签和控制字段。

如下图，加入html解析后即使搜索没有在标签中出现的词汇也可以检索到相关结果。不过缺憾在于html文件无关内容较多，而正确识别与图片真正相关的字段比较困难，所以依赖于html文本检索到的图片精确度比较低

