基于维基百科标题的搜索引擎

1. 使用流式读取的方式将xml中所有<page>标签内的内容保存成独立的页面文档，同时附上该<page>对应的<title>，总页面数为3029940。文件数太多采用分级目录存储。

代码: xmlReader.py

1. 由于页面标题基本没有重复，所以建立标题到页面的映射表。在这之前，将标题中关于'Category', 'Wikipedia', 'MediaWiki', 'Template', 'File', 'Topic'等和知识内容无关的标题去掉，将重复的标题对应的页面合并。

代码：titleSearch.py

1. 对于查询语句，使用jieba分词将其拆分，用每一个分词查询映射表，如果映射表中有对应的条目，说明维基百科中有专门的页面对应到该分词，返回该页面。

代码：titleSearch.py

1. 从返回的页面中逐句搜索，记录每一句话中包含查询语句的所有分词的数目，重复出现只按一次计算，作为该句子的得分，最后对得分排序，得到包含查询语句中不同分词数目最多的句子。可以筛选前几名作为下一步解析答案的语料。

代码：titleSearch.py

1. 注：原本采用倒排索引的方式检索页面，无奈页面太多，中文繁简体分词数太多，内存占用和计算量都很大，每次查询返回结果约要半分钟，且结果也不好。而直接采用搜索标题的方式可以在一秒内得到结果。