文章查重系统设计

1.【文章文本数据爬取】

Python 多线程 + spark + MongDB存储

2.【文章文本处理】

Nltk 和 word2vec   
 （1）分词，词频统计

（2）词性标注，N-grame分块，关系抽取，语义识别

（3）文本相似度计算

（4）主题识别，PlA

3.【检索技术】

PageRank算法

4.【查重算法】

5.【前端展示】

Django + html + css

6.【盈利】

微信公众号开发