Homework2 报告

霍瑜 201605130119

实验结果：

输出包含指定单词的文档集：

```
fuck [428, 4378, 5237, 5755, 6877, 9261, 10304, 10727, 10883, 10894, 10962, 10981, 11042, 11090, 11179, 11370, 11708, 11868, 12356,
13075, 14054, 14222, 14719, 14720, 14854, 15428, 15631, 15647, 15648, 15676, 15677, 15802, 15914, 15915, 16024, 16025, 16046, 16047,
16419, 17792, 18348, 18501, 19535, 20196, 20312, 21533, 22241, 23160, 23363, 26878, 29142, 29847]
[Finished in 13.8s]
```

```
shit [98, 384, 428, 599, 813, 2025, 2462, 2827, 3778, 5308, 6254, 6978, 7243, 8050, 8165, 9152, 9167, 9314, 9329, 10352, 10883, 11378,
11692, 11909, 11995, 12076, 12620, 12869, 13159, 13215, 13216, 13691, 14854, 15480, 15585, 16092, 16095, 17082, 17662, 18020, 18091,
18139, 18218, 18540, 18654, 18817, 18912, 18935, 19103, 19844, 21013, 21734, 23605, 25082, 25741, 26764, 30392]
[Finished in 13.2s]
```

输出指定单词的交集操作并输出包含这些单词的原文：

```
shit & fuck [428, 10883, 14854]
the kardashians look like they would suicide bomb the fuck outta your shit sahadasjoahwjsehurshhggs i hate the entire internet

shoulda named the storm the kraken or some shit fuck a nemo

what the fuck kinda meat is in these hush puppies this shit do nt seem right mydinner
```

数据处理：

分词阶段：nlpk 库分词，并手动去标点以及大小转小写等基础操作

处理阶段：根据分词结果，建立（单词-文档）二元组，并用 set 去重，list 排序，排序后存入字典

存储阶段：数据写入文件，以便后续快速使用

检索：

基本操作：利用扫描线简单实现两个集合的并，交以及减操作

查询方式：‘&’表示交，‘|’表示补，‘-’表示减，操作符以及查询的单词由空格隔开并按顺序执行。例：shit & fuck 表示查询包含两个单词的所有文档

问题：

没什么大问题