

作业 5

钟诚 16307110259

April 17, 2020

设计 MR 程序，实现具有 payload 的倒排文档索引

1. 数据预处理

通过观察，我们发现待处理的数据为 HTML 格式的数据，如图 1。而当我们把数据输入 MRjob 进行处理的时候，要求数据按行输入。为了整理原始数据的编码格式和输入格式，我们需要对数据进行预处理。我们通过 process 函数，寻找数据中的新闻标题和正文，并将其对齐到一行，以 utf-8 的编码格式输出。代码及输入输出样例如下：

process.py:

```
def process(line):
    for line in lines:
        if line[0:14:] == '<contenttitle>':
            contenttitle = line[14:-16]
        if line[0:9:] == '<content>':
            content = line[9:-11:] # 对齐标题与正文
            f = open('data.txt', 'a+', encoding='utf-8')
            writelines = contenttitle + '\t' + content + '\n' # 在同一行进行输出
            f.write(writelines)
            f.close()
```

原始数据:

```
1 <doc>
2 <url>http://news.sohu.com/20120612/n345428229.shtml</url>
3 <docno>c172394d49da2142-69713306c0bb3300</docno>
4 <contenttitle>公安机关销毁 1 0 余万非法枪支 跨国武器走私渐起</contenttitle>
5 <content>中广网唐山 6 月 1 2 日消息（记者汤一尧 庄胜春）据中国之声《新闻晚高峰》报道，今天（1 2 日）上午，公安机关 2 0 1 2 年缉枪制爆专项行动“统一销毁非法枪爆物品活动”在河
6 </doc>
7 <url>http://news.sohu.com/20120607/n344998325.shtml</url>
8 <docno>dbb4554e49da2142-69713306c0bb3300</docno>
9 <contenttitle>张绍刚发道歉信网友不认可：他的问题是俯视他人（图）</contenttitle>
10 <content>天津卫视求职节目《非你莫属》“晕倒门”事件余波未了，主持人张绍刚前日通过《非你莫属》节目组发出道歉信，称自己错在对留学生缺乏了解。但他的道歉，没有得到网友接受和
11 </doc>
12 <url>http://news.sohu.com/20120604/n344745879.shtml</url>
13 <docno>3fca104f49da2142-69713306c0bb3300</docno>
14 <contenttitle>#（关注夏收）（3）夫妻“麦客”忙麦收</contenttitle>
15 <content>临沂（山东），2 0 1 2 年 6 月 4 日 夫妻“麦客”忙麦收 6 月 4 日，在山东省临沂市郯城县郯城街道米顶村麦田间，范加江驾驶收割机在收获小麦。 三夏时节，山东小麦主产区
16 </doc>
17 <url>http://news.sohu.com/20120613/n345535702.shtml</url>
18 <docno>e4103f4f49da2142-69713306c0bb3300</docno>
19 <contenttitle>欧洲杯大战在即 荷兰葡萄牙面临淘汰将背水一战</contenttitle>
20 <content>中广网北京 6 月 1 3 日消息（记者王宇）据中国之声《新闻晚高峰》报道，明天凌晨两场欧洲杯的精彩比赛上演，死亡之组 B 组当中两支传统的强队荷兰队和葡萄牙队正面临着提前淘
21 </doc>
22 <url>http://news.sohu.com/20120601/n344598651.shtml</url>
23 <docno>ab18525249da2142-69713306c0bb3300</docno>
24 <contenttitle>扎克伯格携妻罗马当街吃 3 0 元麦当劳午餐（组图）</contenttitle>
25 <content>环球网记者李亮报道，正在意大利度蜜月的“脸谱”创始人扎克伯格与他华裔妻子的一举一动都处于媒体的追踪之下。5 月 3 1 日，在罗马这个拥有 1 3 家米其林星级餐厅的城市，身家
26 </doc>
27 <url>http://news.sohu.com/20120606/n344833606.shtml</url>
28 <docno>e3736c5949da2142-69713306c0bb3300</docno>
29 <contenttitle>“十分钟”“传声”“三公里”</contenttitle>
30 <content>本报记者 张忠德 本报通讯员 张艳 苏婧 城区“十分钟” 从 5 月中旬开始，在胶南市珠海街道办事处烟台东社区大舞台上，一台台京剧演出点燃了附近居民的热情，1 0 0 多
```

Figure 1: Rawdata

处理后数据，可见此时标题和正文已对齐到一行:

1 公安机关销毁10余万非法枪支 跨国武器走私渐起 中广网唐山6月12日消息(记者汤一亮 庄胜春)据中国之声《新闻晚高峰》报道,今天(12日)上午,公安机关2012年缴枪制爆
2 张绍刚发道歉信网友不认可:他的问题是俯视他人(图) 天津卫视求职节目《非你莫属》“晕倒门”事件余波未了,主持人张绍刚前日通过《非你莫属》项目组发出道歉信,称自己错在对留学生
3 #《关注夏收》(3)夫妻“麦客”忙麦收 临沂(山东),2012年6月4日 夫妻“麦客”忙麦收 6月4日,在山东省临沂市郯城县郯城街道米顶村麦田间,范加江驾驶收割机在收获小麦
4 欧洲杯大战在即 荷兰葡萄牙面临淘汰将背水一战 中广网北京6月13日消息(记者王宇)据中国之声《新闻晚高峰》报道,明天凌晨两场欧洲杯的精彩比赛上演,死亡之组B组当中两支传
5 扎克伯格携妻罗马当街吃30元麦当劳午餐(组图) 环球网记者李亮报道,正在意大利度蜜月的“脸谱”创始人扎克伯格与他华裔妻子的一举一动都处于媒体的追踪之下。5月31日,在罗马这
6 “十分钟”传“三公里” 本报记者 张忠德 本报通讯员 张艳 苏婧 城区“十分钟” 从5月中旬开始,在胶南市珠海街道办事处烟台东社区大舞台上,一台台京剧演出点燃了附近居民的
7 金正恩为朝少年团代表安排宴会 学生发誓坚决跟随(图) 中新网6月8日电 据朝中社报道,260多名朝鲜少年团代表7日参加了朝鲜劳动党第一书记、朝鲜国防委员会第一委员长金正
8 国资委回应央企审计报告 称97%问题完成修改 中新社北京6月2日电(记者 刘辰瑶)国资委2日在官网上回应了日前国家审计署发布的15家中央企业2010年财务收支等审计结果
9 证监会:重组中股票异常交易监管将加强 证监会近日召开新闻通气会,就《关于加强上市公司重大资产重组相关股票异常交易监管的暂行规定(征求意见稿)》向社会公开征求意见。 根
10 台媒曝岛内大学教授假发票案 涉案教授或达千人 中国台湾网6月15日消息 据台湾《中国时报》报道,岛内大学教授“假发票、真A钱”案,如滚雪球,愈演愈烈,目前共有台北、台中、
11 中石油创历史新低 大盘或开短线下跌空间 受到欧美股市大跌以及多部位联合吹风设立国际板的影响,今日沪深股市大幅跳空低开并持续震荡走低。截至收盘,上证指数报2308.55
12 中国驻英大使谈中英关系:尊重对方核心利益
13 “在野党”联手 台湾行政负责人施政报告受阻 中新网6月1日电 据台湾“中央社”报道,台当局行政机构负责人陈冲今天(6月1日)到台湾立法机构进行施政方针报告并备询,不过由于“
14 美国龙卷风最新消息 导读:4月27日晚龙卷风和强风暴横扫美国东南部,随后是九个小时,这场灾难造成的死亡人数可以说是美国历史上罕见的,已经超过1974年龙卷风爆发造成315人
15 美国大肠杆菌6州扩散 1名儿童死亡 中新社旧金山6月8日电(记者 刘丹)一种特殊的大肠杆菌在美国6个州引发14起病例,一名儿童死亡。美国疾病控制与预防中心官员表示正在着手
16 《文化》(2)上海老相机制造博物馆即将开门迎客 上海,2012年6月8日 上海老相机制造博物馆即将开门迎客 6月8日,技师在博物馆内的海鸥4A-109相机生产线上测试镜头
17 重大资产收购方案获通过 三公司今起复牌 恒泰艾普、*ST领先、轻纺城 三家上市公司重大资产交易方案获得中国证监会并购重组委有条件通过,三家公司股票将于今日复牌。恒泰艾
18 三部委决定在上海试行启运港退税政策 关于在上海试行启运港退税政策的通知 各省、自治区、直辖市、计划单列市财政厅(局)、国家税务局、海关总署广东分署、各直属海关,新疆生产建
19 银行员工倒苦水:没有存款资源就别进银行 6月5日,据中国证券报报道,多家银行的员工对目前高不可及的存款任务大倒苦水。“现在是没有存款资源就别进银行。”对于此前某外资银行要
20 梵高理财坠落轨迹 凶悍的人海战术
21 约定收益不断攀升 分级债基趁势吸金 富国天盈分级债基A份额日前开放申购,以4.9%的约定年化收益率吸引了超过40亿元资金的申购。而6月份还有万家添利、博时裕隆、长信利鑫等
22 煤企纷纷降价促销 煤炭“黄金十年”或已终结 目前国内外煤价倒挂达到了120多元,为给煤炭找销路,许多国内煤企已开始主动降价吸引买家,这是自2009年以来从未出现过的景象
23 财政部:政府性资金要对民间投资主体同等对待 中新网6月8日电 据财政部网站消息,财政部、国家发改委近日联合发布《关于安排政府性资金对民间投资主体同等对待的通知》。通知要求
24 全球股市暴涨A股熊样不改 市场人士一片哀鸿
25 李旭利:常理上不太可能选择工行建行做老巢 你好,我是李旭利。“2009年10月,上海,陆家嘴,花旗大厦8楼的重阳投资办公室,李旭利伸出了手与记者相握。这是记者最近一次与
26 拆迁款瘦身记:拨款1.71万支付1.4万到手3.1万 新一份的起诉书中,常恒光的“罪名”由涉嫌贪污更改为滥用职权。常恒光曾被河南省建设厅评为“拆迁管理先进个人”。在开庭审理
27 四大银行股价节节溃退 工行建行跌破汇金成本线 新快报讯 适合汇金公司增持四大行的时间窗口再次打开——四大行股价已经逼近今年一季度的最低价。也就是说,即使按照理论最低值计算
28 村民发现乌木值百万 当地政府夺走称属国有 2012年春节时,四川彭州市通济镇柳村农民吴高亮在自家承包地中,发现了一笔“横财”:长达3.4米,胸径约1.5米,重达6.0吨的乌
29 中国经济走过黄金十年 转型将造福世界经济 中国人民银行7日晚间决定,自8日起分别下调金融机构人民币存贷款基准利率,这是中国时隔三年半来首次降息,也是应对当前经济增速放缓
30 驾驶员不应转化为交强险中的“第三者”
31 10万7天理财收益仅百元 银行产品实际收益低 随着国民经济的增长,百姓手里的闲钱多了,如何让手里的钱保值甚至增值成为民众关注的焦点。面对纷繁复杂的投资方式,银行理财产品
32 中国单年捐赠过亿的十大富豪 余彭年:2010年4月,余彭年新捐3.2亿元投入“余彭年慈善基金会”。“余彭年慈善基金会”目前总价值已达8.2亿元,其中包括两部分,第一部分是银行存款
33 曝光朝鲜官员的隐秘特供商店 真令人失望
34 阳光人寿“龙凤宝贝计划之传家保单”上市 商报讯 (记者 楼志文 通讯员 陈佩云)正值儿童节,阳光人寿推出“龙凤宝贝计划之传家保单”。该计划由“传家保少儿年金保险(分红型)”
35

Figure 2: Processed data

2. 编写 MapReduce 程序并执行

随后我们对 MR 程序进行编写:

reducer.py:

```
from mrjob.job import MRJob
import jieba
class Reverse(MRJob):
    def mapper(self, _, line):
        contenttitle=line.split('\t')[0] # 提取题目
        SegmentDic={}
        content=line.split('\t')[1] # 提取正文
        segs=jieba.cut(content) # 对正文进行分词
        for seg in segs:
            SegmentDic[seg]=SegmentDic.get(seg, 0)+1 # 对所有的分词进行计数
        for key,value in SegmentDic.items():
            yield key, [contenttitle, value] #输出为 (词语, [题目, 出现次数])

    def reducer(self, key, values):
        p=[]
        for title, values in values:
            p.append((title,values)) # 对不同文章中出现的同一词语进行归并
        p.sort(key=lambda x: x[1], reverse=True) # 根据出现次数对文章标题进行排序
        yield key,p

if __name__ == '__main__':
    Reverse.run()
```

然后,我们打开 HDFS 并执行 MapReduce 程序

```
hadoop@guardianzc-VirtualBox:/usr/local/hadoop$ ./sbin/start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [guardianzc-VirtualBox]
```

Figure 3: 打开 HDFS

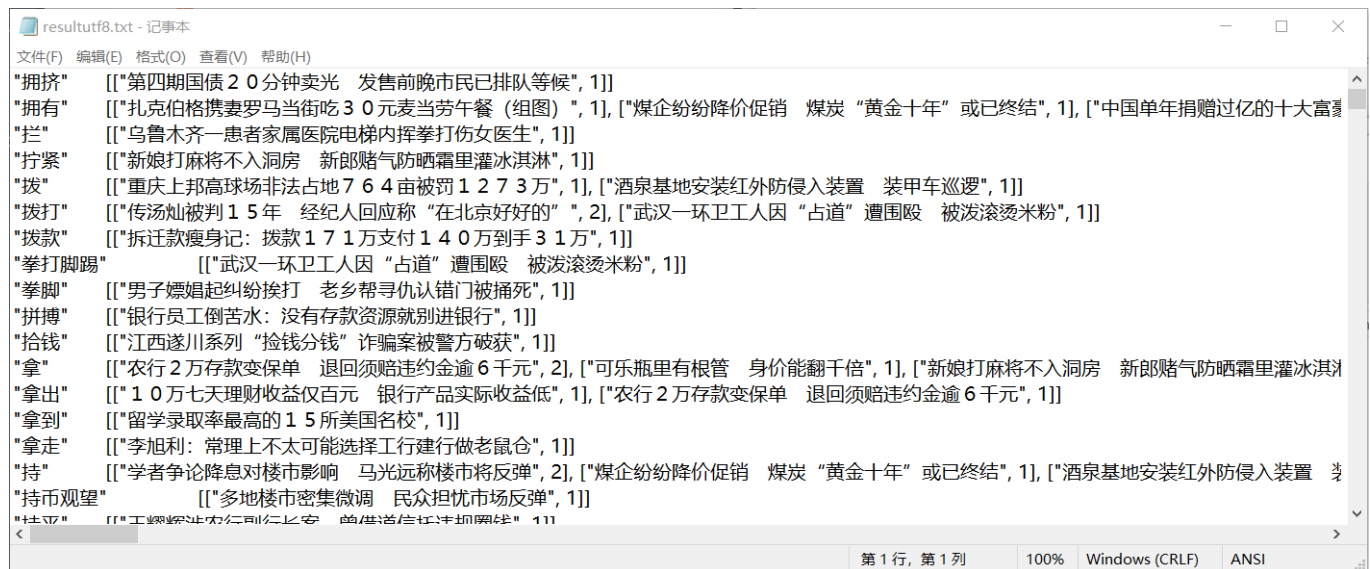


Figure 6: UTF-8 编码结果

这时，我们就可以对其中的词语进行倒排文档的查询

search.py

```
def search(word):
    file = open('resultutf8.txt', 'r')
    while 1:
        line = file.readline()
        if not line:
            break
        line_mark = line.split('\t')[0].rstrip('').lstrip('')
        if line_mark == word: # 如果能查询到，则输出倒排文档值
            print(line_mark, '-->', line.split('\t')[1])
            return 0
    print('Not Found.') # 如果一直没有查询到，则输出 "Not Found"
    return 0
```

以下为查询的示例结果：

```
(Normal) D:\Documents\课程介绍\大规模分布式系统\Project3\news_tensite_xml.smarty>D:/Application/Anaconda/envs/Normal/python.exe d:/Documents/课程介绍/大规模分布式系统/Project3/news_tensite_xml.smarty/transformer.py
控制 --> [["王耀辉涉农行副行长案 曾借道信托违规圈钱", 4], ["美国大肠杆菌 6 州扩散 1 名儿童死亡", 3], ["证监会：重组中股
票异常交易监管将加强", 1], ["重大资产收购方案获通过 三公司今起复牌", 1], ["河南漯河市省级文保旁建高楼 百年老房随时会坍塌",
1], ["缅甸若开邦骚乱局势总体得到控制", 1]]

(Normal) D:\Documents\课程介绍\大规模分布式系统\Project3\news_tensite_xml.smarty>D:/Application/Anaconda/envs/Normal/python.exe d:/Documents/课程介绍/大规模分布式系统/Project3/news_tensite_xml.smarty/transformer.py
接电话 --> [["可乐瓶里有根管 身价能翻千倍", 1]]
```

Figure 7: 查询结果

值得注意的是：

1. 这里对于单个词的查询，可以根据词频直接推出 TF-IDF 的值并排序，对于多个词的查询，则可以分别计算文档相似度并排序，与单个词的查询大致相同。

当然，如果需要对多个词的查询，我们则要对所有文档和所有分词结果构建 TF-IDF 矩阵，然后对查询词的 TF-IDF 进行排序，具体实现如下：

query.py

```
import numpy as np
from math import log
word_dic = {}
```



```

counts_art = {}
# 构建词语对于每篇文章的计数矩阵
with open("./resultutf8.txt", "r") as f:
    lines = f.readlines()
    idx = 0
    for line in lines:
        word, titles = line.split('\t')
        word = word.rstrip("\n").rstrip("\r")
        titles = titles.rstrip(']]\n').rstrip('[[').split(']', ['')
        counts_word = {}
        for title in titles:
            t, count = title.split(',')
            count = int(count)
            t = t.lstrip('[').rstrip(']')
            # 词语计数
            counts_word[t] = counts_word.get(t, 0) + count
            # 文章总词数
            counts_art[t] = counts_art.get(t, 0) + count
        word_dic[word] = counts_word

art_name = list(counts_art.keys())

# 构建TF-IDF矩阵
art_totalnumber = len(art_name)
TF_IDF = {}
for key, value in word_dic.items():
    word_TFIDF = {}
    word_totalInArt = len(value.keys())
    for title, counts in value.items():
        # 前半为TF值, 后半为IDF值
        word_TFIDF[title] = (counts / counts_art[title]) * log(art_totalnumber / (word_totalInArt + 1))

    TF_IDF[key] = word_TFIDF

query = input("请输入查询的词语(空格分割): ")
query = query.split()
# 计算并选出最高的五个结果
art_TFIDF = {}
for word in query:
    for art in art_name:
        art_TFIDF[art] = art_TFIDF.get(art, 0) + TF_IDF[word].get(art, 0)
Max10 = sorted(art_TFIDF.items(), key=lambda item: item[1], reverse=True)[0:5]
print("最接近的5条文章为:")
for high in Max10:
    print(high[0], '\t', high[1])

```

```

请输入查询的词语(空格分割): 改革
最接近的5条文章为:
日本首相野田组建新一届内阁 更换5名内阁大臣      0.03976166666174254
发改委连发两文赞阶梯电价: 实现富人补贴穷人      0.01645310344623829
留学录取率最高的15所美国名校      0.012965760867959523
中国经济走过黄金十年 转型将造福世界经济          0.006482880433979762
国资委回应央企审计报告 称97%问题完成修改        0.006117179486421929

```

Figure 8: 查询“改革”

```

请输入查询的词语(空格分割): 开放
最接近的5条文章为:
脱衣舞“疯狂”农村夏夜让人汗颜      0.04827542705784238
大连有望设立第五个国家新区 带动东北振兴      0.03186886382791551
第五个国家新区有望落户大连 10股最大赢家      0.030705218623362646
约定收益不断攀升 分级债基趋势吸金          0.03034776555440615
“在野党”联手 台湾行政负责人施政报告受阻      0.01088919407319753

```

Figure 9: 查询“开放”

请输入查询的词语(空格分割): 改革 开放
最接近的5条文章为:
脱衣舞“疯狂”农村夏夜让人汗颜 0.04827542705784238
日本首相野田组建新一届内阁 更换 5 名内阁大臣 0.03976166666174254
大连有望设立第五个国家级新区 带动东北振兴 0.03478536749503599
第五个国家级新区有望落户大连 1 0 股最大赢家 0.03351523040157767
约定收益不断攀升 分级债基趁势吸金 0.03034776555440615

Figure 10: 查询“改革开放”

显然，当查询的词语更多时，查询的结果更加准确。如果需要进一步提升查询的准确率，则要求我们对文章的分词进行更加合理的处理。至此，我们完成了整个倒排文档查询项目。