



# 大数据处理综合实验课程 (2025)

课程实验2-文档倒排索引

南京大学      计算机学院



## 实验内容与要求（一）：

- 请实现课堂上介绍的“带词频属性的文档倒排算法”。
- 在统计词语的倒排索引时，除了要输出带词频属性的倒排索引，还请计算每个词语的“平均出现次数”并输出。“平均出现次数”在这里定义为：

$$\text{平均提及次数} = \frac{\text{词语在全部文档中出现的频数总和}}{\text{包含该词语的文档数}}$$

- 假如文档集中有四个文档：A、B、C、D。词语“同伴”在文档A中出现了100次，在文档B中出现了200次，在文档C中出现了300次，在文档D中没有出现。则词语“同伴”在该文档集中的“平均出现次数”为  $(100 + 200 + 300) / 3 = 200$ 。

## 实验内容与要求（一）：

- 对于每个词语，输出一个键值对，该键值对的格式为：  
[词语]\TAB 平均出现次数（保留2位小数），文档-1:词频; 文档-2:词频; ...; 文档-n:词频
- 下图展示了输出文件的一个片段（图中内容仅为格式示例）

```
[直拳]          4.00, 第五部-凤凰社:4  
[直指]          1.83, 第一部-魔法石:1; 第三部-阿兹卡班的囚徒:4; 第二部-密室:1; 第五部-凤凰社:2; 第六部-混血王  
子:2; 第四部-火焰杯:1  
[直挺挺]        1.25, 第七部-死亡神器:1; 第五部-凤凰社:1; 第六部-混血王子:1; 第四部-火焰杯:2  
[直接]          16.57, 第一部-魔法石:4; 第七部-死亡神器:24; 第三部-阿兹卡班的囚徒:14; 第二部-密室:13; 第五部-凤  
凰社:32; 第六部-混血王子:10; 第四部-火焰杯:19  
[直接参与]       1.00, 第七部-死亡神器:1  
[直朝]          1.50, 第七部-死亡神器:2; 第五部-凤凰社:1  
[直流]          1.00, 第七部-死亡神器:1; 第二部-密室:1; 第五部-凤凰社:1; 第六部-混血王子:1; 第四部-火焰杯:1
```

- 使用另外一个MapReduce Job对每个词语的平均出现次数进行全局排序，输出排序后的结果。



## 实验内容与要求（二）：

- 为每个作品计算每个词语的**TF-IDF**。TF定义为某个词语在该作品中的出现次数之和。IDF定义为：

$$\text{IDF}(\text{词语}) = \log \left( \frac{\text{语料库文档总数}}{\text{包含该词的文档数} + 1} \right)$$

$$\text{TF-IDF} = \text{TF} * \text{IDF}$$

- 输出格式：作品名称，词语，该词语的 TF-IDF。



## 实验内容与要求（三）：

- 去除停用词，重新计算每个词语的“平均出现次数”并输出，输出排序后的结果。
- 输出格式：参考实验内容与要求（一）



## 实验数据：

- 本次实验提供了《哈利·波特》系列小说共 7 部，每部小说对应一个文本文件。文本文件均使用 UTF-8 字符编码，并且已分词，两个汉语单词之间使用空格分隔。
- 本次实验提供了一个中文停词表（cn\_stopwords.txt）。
- 单机测试样例：提供《哈利·波特》全集作为单机测试样例，可在“本科教学支撑平台”中下载。该数据集主要供本地调试使用。
- 全部数据集：全部数据集位于集群的 HDFS 存储上，HDFS 存储位置为：  
[/user/root/Exp2](#)

# 实验数据：

- 输入数据情况如下如所示：



大难不死的男孩 家住 女贞路 四号 的 德思礼 夫妇 总是 得意 地说 他们 是 非常 规矩 的 人家 拜托 拜托 了 他们 从来 跟 神秘 古怪 的 事 不 沾 边 因为 他们 根本 不 相信 那些 邪门 歪道 弗农 德思礼 先生 在 一家 名叫 格朗宁 的 公司 做 主管 公司 生产 钻机 他 高大 魁梧 胖得 几乎 连 脖子 都 没了 却 蓄着 一脸 大胡子 德思礼 太太 是 个 瘦削 的 金发 女人 她 的 脖子 几乎 比 正常 人 长 一 倍 这样 每当 她 花 许多 时间 隔 着 篱墙 引颈 而望 窥探 左邻 右舍 时 她 的 长脖子 可 就 派 上 了 大用 场 德思礼 夫妇 有 一个 小儿子 名叫 达力 在 他们 看来 人世间 没有 比 达力 更好 的 孩子 了 德思礼 一家 什么 都 不 缺 但 他们 拥 有 一个 秘密 他们 最 害怕 的 就是 这 秘密 会 被 人 发现 他们 想 一旦 有人 发现 波特 一 家 的 事 他们 会 承 受 不 住 的 波特 太太 是 德思礼 太太 的 妹妹 不过 她们 已经 有 好 几 年 不 见 面 了 实际 上 德思礼 太太 佯装 自己 根本 没有 这么 个 妹妹 因为 她 妹妹 和 她 那 一 无 是 处 的 妹夫 与 德思礼 一 家 的 为 人 处 世 完 全 不 一 样 一 想 到 邻居 们 会 说 波特 夫妇 来 到 了 德思礼 夫妇 会 吓 得 胆战 心 惊 他们 知 道 波特 也 有 个 儿子 只是 他们 从 来 没 有 见 过 这 孩子 也 是 他们 不 与 波特 夫妇 来 往 的 一 个 很 好 的 借 口 他们 不 愿 让 达力 跟 这 种 孩子 厮 混 我 们 的 故 事 开 始 于 一 个 晦 暗 阴 沉 的 星 期 二 德思礼 夫妇 一 早 醒 来 窗 外 浓 云 低 垂 的 天 空 并 没 有 丝 毫 迹 象 预 示 这 地 方 即 将 发 生 神 秘 古 怪 的 事 情 德思礼 先生 哼 着 小 曲 挑 出 一 条 最 不 喜 欢 的 领 带 戴 着 上 班 德思礼 太太 高 高 兴 兴 一 直 絮 絮 叨 叨 把 唧 哇 乱 叫 的 达力 塞 到 了 儿 童 椅 里 他们 谁 也 没 留 意 一 只 黄 褐 色 的 猫 头 鹰 扑 扇 着 翅 膀 从 窗 前 飞 过 八 点 半 德思礼 先生 拿 起 公 文 包 在 德思礼 太太 面 颊 上 亲 了 一 下 正 要 亲 达力 跟 这 个 小 家 伙 道 别 可 是 没 有 亲 成 小 家 伙 正 在 发 脾 气 把 麦 片 往 墙 上 摔 臭 小 子 德思礼 先生 嘟 囔 了 一 句 咯 咯 笑 着 走 出 家 门 坐 进 汽 车 倒 出 四 号 车 道 在 街 角 上 他 看 到 了 第 一 个 异 常 的 信 号 一 只 猫 在 看 地 图 一 开 始 德思礼 先生 还 没 弄 明 白 他 看 到 了 什 么 于 是 又 回 过 头 去 只 见 一 只 花 斑 猫 站 在 女 贞 路 的 路 口 但 是 没 有 看 见 地 图 他 到 底 在 想 些 什 么 很 可 能 是 光 线 使 他 产 生 了 错 觉 吧 德思礼 先生 眨 了 眨 眼 盯 着 猫 看 猫 也 瞪 着 他 当 德思礼 先生 拐 过 街 角 继 续 上 路 的 时 候 他 从 后 视 镜 里 看 看 那 只 猫 猫 这 时 正 在 读 女 贞 路 的 标 牌 不 是 在 看 标 牌 猫 是 不 会 读 地 图 或 是 读 标 牌 的 德思礼 先生 定 了 定 神 把 猫 从 脑 海 里 赶 走 他 开 车 进 城 一 路 上 想 的 是 希 望 今 天 他 能 得 到 一 大 批 钻 机 的 定 单 但 快 进 城 时 另 一 件 事 又 把 钻 机 的 事 从 他 脑 海 里 赶 走 了 当 他 的 车 汇 入 清 晨 拥 堵 的 车 流 时 他 突 然 看 见 路 边 有 一 群 穿 着 奇 装 异 服 的 人 他 们 都 披 着 斗 篷 德思礼 先生 最 看 不 惯 别 人 穿 得 怪 模 怪 样 瞧 年 轻 人 的 那 身 打 扮 他 猜 想 这 大 概 又 是 一 种 无 聊 的 新 时 尚 吧 他 用 手 指 敲 击 着 方 向 盘 目 光 落 到 了 离 他 最 近 的 一 大 群 怪 物 身 上 他 们 正 兴 致 勃 勃 交 头 接 耳 德思礼 先生 很 生 气 因 为 他 发 现 他 们 中 间 有 一 对 根 本 不 年 轻 了 那 个 男 的 显 得 比 他 年 龄 还 大 竟 然 还 披 着 一 件 翡 翠 绿 的 斗 篷 真 不 知 羞 耻 接 着 德思礼 先生 突 然 想 到 这 些 人 大 概 是 为 什 么 事 募 捐 吧 不 错 就 是 这 么 回 事 车 流 移 动 了 几 分 钟 后 德思礼 先



## 实验报告提交要求：

- 实验报告要求提交一个压缩包，压缩包内除了包含源代码、JAR 包、JAR 包执行方式说明，还需要包含一个实验报告（pdf格式）。实验报告中包含：
  - Map 和 Reduce 的设计思路（含 Key、Value 类型）。
  - MapReduce 中 Map 和 Reduce 的伪代码（或者带注释的实际代码，如果使用实际代码，请做好排版）。
  - 输出结果文件的部分截图。输出结果文件在 HDFS 上的路径（某些情况下助教会检查 HDFS 上的输出文件）。
  - 在集群上执行作业后，Yarn Resource Manager 的 WebUI 执行报告内容（请完整包括执行报告内容，以表明该 Job 是在集群上实际执行过的，否则影响分数。每个 MapReduce Job 对应一个报告）。执行报告内容示例见下文。





## 实验报告提交要求：

- 实验报告压缩包文件名：学号\_姓名\_BG\_x，其中 x 表示第 x 次实验
- 实验报告压缩包提交至：本科教学支撑平台 <http://cslabcms.nju.edu.cn/>
- 实验提交截止日期：4月1日（包含当天，实验时长共2周）



## WebUI执行结果：

- 在以后的实验报告中，若需要在集群上执行 MapReduce Job，请在报告中附上相关的 MapReduce Job 的执行报告，以作为评分依据。否则在评分时将认为该 MapReduce Job 没有在集群上执行，会影响实验得分。
- 校园网访问实验平台 <http://114.212.130.202:8082/>
- 参考资料：《[大数据平台使用手册.pdf](#)》
- 输入账户和密码，点击左侧栏“Yarn作业监控”，可以进入集群监控页面（见下图）。



# WebUI执行结果：

交互式大数据编程平台

北京时间: 15:49:21 | norma5

Yarn作业监控

Cluster

Cluster Metrics

Apps Submitted	40
Apps Pending	0
Apps Running	0
Apps Completed	40
Containers Running	0
Memory Used	0 B
Memory Total	360 GB
Memory Reserved	0 B
Vcores Used	0
Vcores Total	360
Vcores Reserved	0

Cluster Nodes Metrics

Active Nodes	0
Decommissioning Nodes	0
Decommissioned Nodes	0
Lost Nodes	0
Unhealthy Nodes	0
Rebooted Nodes	0
Shutdown Nodes	0

Scheduler Metrics

Scheduler Type	Capacity Scheduler
Scheduling Resource Type	[yarn.io/gpu, memory-mb (unit-M), vcores]
Minimum Allocation	<memory:1024, vCores:1>
Maximum Allocation	<memory:20480, vCores:4>
Maximum Cluster Application Priority	0

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	Start Time	Launch Time	Finish Time	State	Final Status	Running Containers	Allocated CPU Vcores	Allocated Memory MB	Reserved CPU Vcores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
application_1735856406312_0051	2212200770u	QuasiMonteCarlo	MAPREDUCE	2025class03	0	Fri Mar 7 06:47:37 +0800 2025	Fri Mar 7 06:47:38 +0800 2025	Fri Mar 7 06:47:52 +0800 2025	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0		History	0
application_1735856406312_0050	norma5	Wordcount.jar	MAPREDUCE	2025class01	0	Fri Mar 7 04:31:24 +0800 2025	Fri Mar 7 04:31:24 +0800 2025	Fri Mar 7 04:31:39 +0800 2025	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0		History	0
application_1735856406312_0049	211220129fk	QuasiMonteCarlo	MAPREDUCE	2025class02	0	Fri Mar 7 03:54:15 +0800 2025	Fri Mar 7 03:54:15 +0800 2025	Fri Mar 7 03:54:29 +0800 2025	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0		History	0
application_1735856406312_0048	2212200420u	QuasiMonteCarlo	MAPREDUCE	2025class03	0	Fri Mar 7 03:23:00 +0800 2025	Fri Mar 7 03:23:00 +0800 2025	Fri Mar 7 03:23:14 +0800 2025	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0		History	0
application_1735856406312_0047	norma5	wordcount	MAPREDUCE	2025class01	0	Wed Mar 5 04:05:02 +0800 2025	Wed Mar 5 04:05:02 +0800 2025	Wed Mar 5 04:05:18 +0800 2025	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0		History	0
application_1735856406312_0046	norma5	Wordcount.jar	MAPREDUCE	2025class01	0	Wed Mar 5 03:44:59 +0800 2025	Wed Mar 5 03:44:59 +0800 2025	Wed Mar 5 03:45:15 +0800 2025	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0		History	0
application_1735856406312_0044	norma4	word count	MAPREDUCE	2025class01	0	Wed Mar 5 02:51:42 +0800 2025	Wed Mar 5 02:51:42 +0800 2025	Wed Mar 5 02:51:56 +0800 2025	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0		History	0
application_1735856406312_0043	norma5	Wordcount.jar	MAPREDUCE	2025class01	0	Tue Mar 4 05:20:42 +0800 2025	Tue Mar 4 05:20:42 +0800 2025	Tue Mar 4 05:20:58 +0800 2025	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0		History	0
application_1735856406312_0042	norma4	word count	MAPREDUCE	2025class01	0	Wed Feb 26 11:11:11 +0800 2025	Wed Feb 26 11:11:11 +0800 2025	Wed Feb 26 11:11:11 +0800 2025	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0		History	0



# WebUI执行结果：

- 或者在“命令行提交”页面“查看日志”，并截图

STATION

大数据处理课程-在线实验平台

北京时间: 17:14:20 | BDP202403

编程空间

命令提交

作业报告

Yarn作业监控

HDFS分布式存储

Alluxio内存存储

HBase数据库

SQL ON HADOOP

首页 作业运行

在本页面上，您可以将编写好的Jar包或py文件上传到集群中执行。

上传文件

查看示例

停止作业

刷新

选择作业类型

请输入提交命令

提交

编号	用户	命令语句	作业类型	开始时间	结束时间	执行结果	输入关键词搜索命令语句
1	BDP202403	yarn jar /home/BDP202403/invertedindexer-1.1.jar L...	MR	2024/4/24 12:07:20	2024/4/24 12:08:11	成功	查看日志
2	BDP202403	yarn jar /home/BDP202403/invertedindexer-1.0.jar ...	MR	2024/4/24 09:28:35	2024/4/24 09:32:20	失败	查看日志
3	BDP202403	yarn jar /user/BDP202403/invertedindexer-1.0.jar L...	MR	2024/4/24 09:26:22	2024/4/24 09:26:23	失败	查看日志

共 3 条 | 10条/页 | 1 | 前往 1 页



# WebUI执行结果：

- 或者在“命令提交”页面“查看日志”，并截图

