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 嘟哝 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。

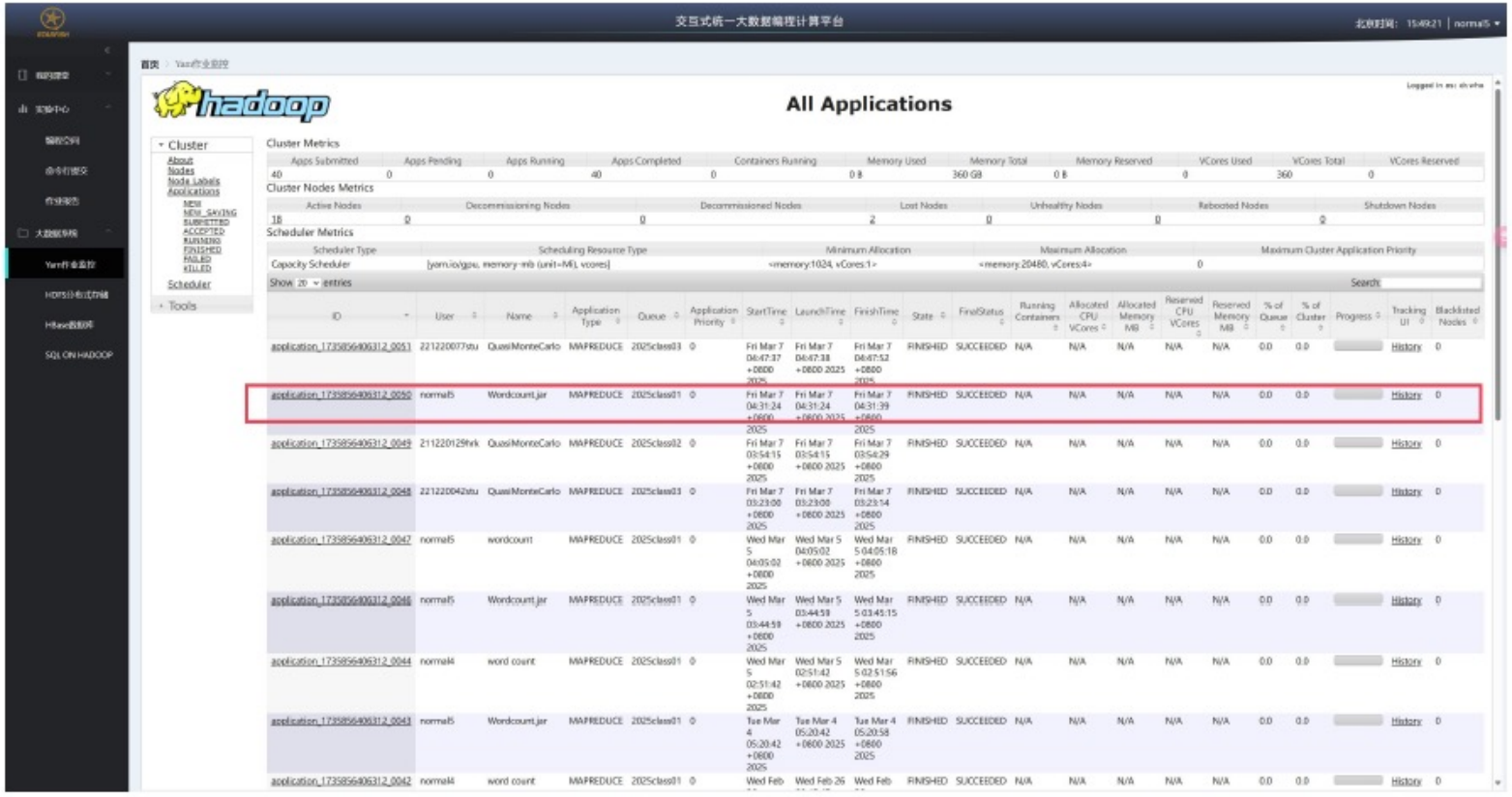
实验报告提交要求：

- 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：

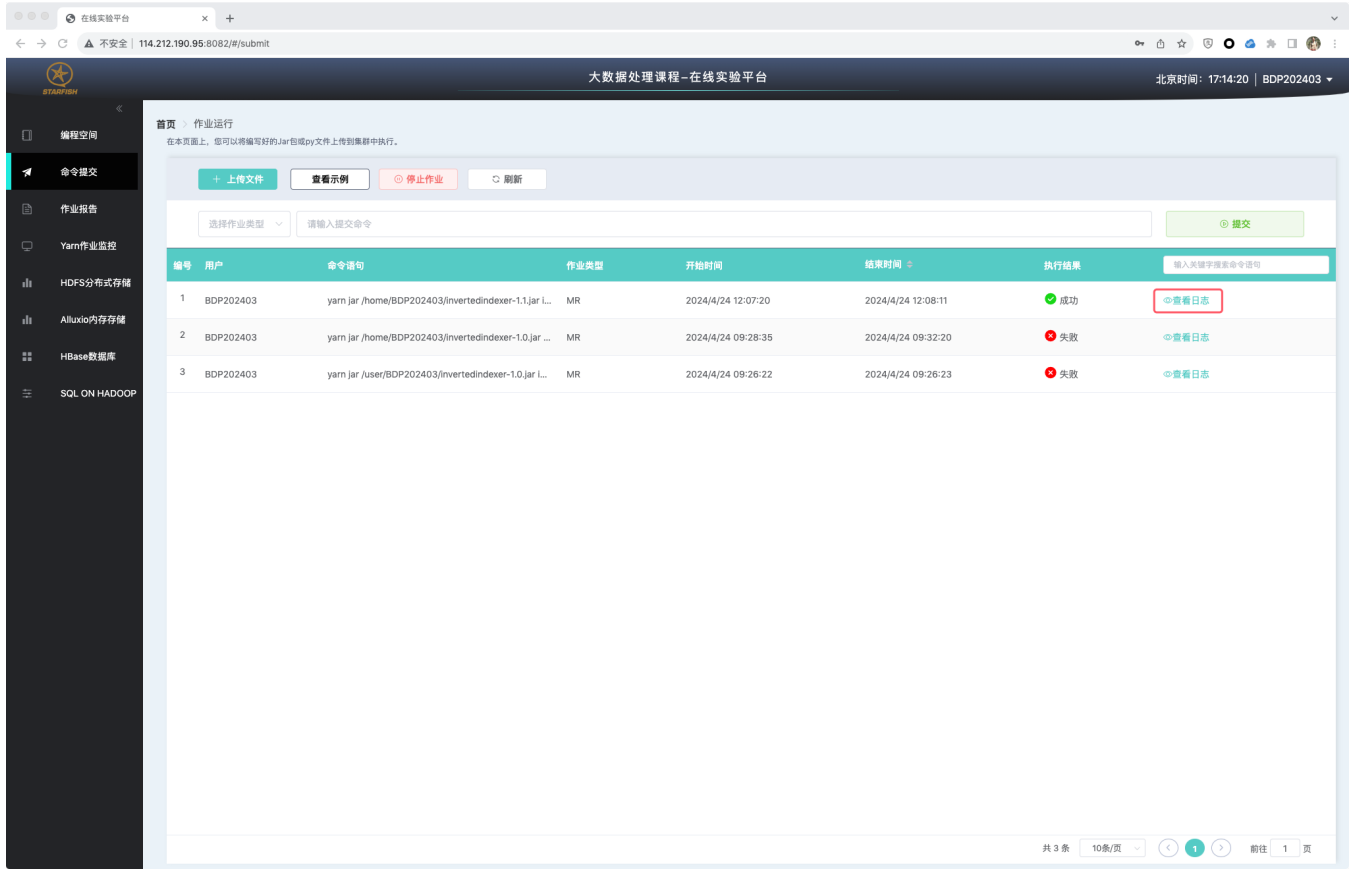
- 在以后的实验报告中，若需要在集群上执行 MapReduce Job，请在报告中附带上相关的 MapReduce Job 的执行报告，以作为评分依据。否则在评分时将认为该 MapReduce Job 没有在集群上执行，会影响实验得分。
- 校园网访问实验平台 http://114.212.130.202:8082/
- 参考资料：《大数据平台使用手册.pdf》
- 输入账户和密码，点击左侧栏“Yarn作业监控”，可以进入集群监控页面（见下图）。

WebUI执行结果：


All Applications
Cluster
Cluster Metrics
Cluster Nodes Metrics
Scheduler Metrics

WebUI执行结果：

- 或者在“命令行提交”页面“查看日志”，并截图



WebUI执行结果：

- 或者在“命令提交”页面“查看日志”，并截图