

## K-均值聚类

---

### 一、设计思路

#### (1) Mapper 类 (KMeansClusterMapper):

- Key 类型: 输入的键是 LongWritable, 表示输入文件中的行偏移量。
- Value 类型: 输入的值是 Text, 表示输入文件中的一行文本。

思路:

- 每个 Mapper 实例负责处理输入数据的一个片段。
- 在 setup 方法中, 从分布式缓存中读取初始簇中心。
- 在 map 方法中, 将每个输入行解析为向量, 并计算其到所有簇中心的距离, 找到最近的簇中心。
- 输出的键为簇中心的索引, 值为输入的向量。

#### (2) Reducer 类 (KMeansClusterReducer):

- Key 类型: 输入的键是 IntWritable, 表示簇中心的索引。
- Value 类型: 输入的值是 Text, 表示属于该簇的所有向量。

思路:

- 每个 Reducer 实例负责计算一个簇的新中心。
- 在 reduce 方法中, 遍历所有属于该簇的向量, 计算它们的平均值, 作为新的簇中心。
- 输出的键为簇中心的索引, 值为新的簇中心向量。

#### (3) 伪代码

```
class Mapper:
    setup(context):
        Load cluster centers from cache
        Initialize clusterCenters, clusterCount, vectorDim

    map(key, value, context):
        vector = parse value into array of doubles
        nearestCenter = find nearest cluster center to vector
```

```

        emit(nearestCenter, value)

    findNearestCenter(vector):
        minDistance = MAX_VALUE
        nearestCenterIndex = -1
        for each center in clusterCenters:
            distance = calculate Euclidean distance(vector, center)
            if distance < minDistance:
                minDistance = distance
                nearestCenterIndex = center index
        return nearestCenterIndex

    euclideanDistance(a, b):
        sum = 0
        for each dimension i:
            sum += (a[i] - b[i])^2
        return sqrt(sum)

```

```

class Reducer:
    reduce(key, values, context):
        Initialize newCenter as array of zeros
        count = 0
        for each value in values:
            vector = parse value into array of doubles
            Add vector to newCenter
            count++
        Divide newCenter by count to get average
        emit(key, newCenter as string)

    parseVector(string):
        Split string by commas into array of doubles
        return array

```

```

class KMeans:
    main(args):
        if args.length != 3:
            Print usage message
            Exit with error

        Create configuration with cluster count and vector dimension
        Create job with configuration and job name

```

```
Set job classes: Mapper, Reducer, output key/value types
Set input/output paths from args
Add initial centers file to job cache

Wait for job completion and exit
```

```
function parseVector(string):
    Split string by commas
    Convert each part to double
    return array of doubles
```

---


## 二、实验结果

输出文件路径： / Lab4-op/part-r-00000

### (1) 输出结果文件的部分截图

```
0
4.55032021957914,4.548788734635427,4.5553522415370535,4.532021957913998,4.550300330164286,4.5326981
98019014,4.550578781972234,4.537571104658101,4.543577707943832,4.53814789768885,4.528939098611719,4
.563884800509169,4.560324595250408,4.545327976450933,4.548331278093799,4.547615259159076
1
14.605699626123618,14.59191790629226,14.595974862779412,14.5991369023944,14.574815050513086,14.6174
528677114,14.607131493119084,14.598560178187892,14.619381910746958,14.611466868188689,14.603432503
380796,14.600946623180336,14.594662318033569,14.598659613395911,14.606912735661442,14.607032057911
065
2
25.31046274336771,25.314286764841064,25.27663486110345,25.287978232493153,25.40404096115309,25.3447
5024359752,25.36230764988142,25.323203353372676,25.279319030022247,25.279521261927087,25.278657180
151857,25.34552240177964,25.413490706524737,25.303421396135533,25.28200319894104,25.31645616163844
5
14
144.54048640915593,144.5618025751073,144.53588800326997,144.54522787655836,144.53016554261188,144.567
34109952995,144.5164316370325,144.53051297772328,144.57569997956264,144.5271816881259,144.56660535458
818,144.5124872266503,144.5624156958921,144.52094829348047,144.50815450643776,144.53359901900674
15
154.54537297361966,154.52995952224282,154.52742716994675,154.514964806285,154.5583737113916,154.54501
40575462,154.53982971426294,154.5308966919902,154.4969392434847,154.52888277402246,154.52746704951048
,154.52816494187553,154.53628043309206,154.51959083567627,154.53534326334469,154.53075711351718
16
164.74613099890573,164.71875488510238,164.71299046428013,164.76068860403313,164.73190558074097,164.71
033296857902,164.7002305768329,164.73659527903706,164.73706424886666,164.73536423323432,164.752911521
02548,164.73409410661247,164.7501172424574,164.69509144911677,164.7504689698296,164.70587775519775
17 174.86987518761356,174.89835295047,174.90579824630697,174.8579864128288,174.827948
```

## (2) Yarn Resource Manager 的 WebUI 执行报告内容



### Application application\_1720423563825\_5565

Logged in as: dr:who

Cluster

[About](#)  
[Nodes](#)  
[Node Labels](#)  
[Applications](#)  
[NEW](#)  
[NEW SAVING](#)  
[SUBMITTED](#)  
[ACCEPTED](#)  
[RUNNING](#)  
[FINISHED](#)  
[FAILED](#)  
[KILLED](#)  
[Scheduler](#)

Tools

#### Application Overview

User: 2024stu3\_12  
Name: BPE Char Pair Counter  
Application Type: MAPREDUCE  
Application Tags:  
Application Priority: 0 (Higher Integer value indicates higher priority)  
YarnApplicationState: FINISHED  
Queue: BD02202403  
FinalStatus Reported by AM: SUCCEEDED  
Started: Wed Jul 10 15:56:39 +0800 2024  
Launched: Wed Jul 10 15:56:39 +0800 2024  
Finished: Wed Jul 10 15:58:02 +0800 2024  
Elapsed: 1mins, 23sec  
Tracking URL: [History](#)  
Log Aggregation Status: SUCCEEDED  
Application Timeout (Remaining Time): Unlimited  
Diagnostics:  
Unmanaged Application: false  
Application Node Label expression: <Not set>  
AM container Node Label expression: <DEFAULT\_PARTITION>

#### Application Metrics

Total Resource Preempted: <memory:0, vCores:0>  
Total Number of Non-AM Containers Preempted: 0  
Total Number of AM Containers Preempted: 0  
Resource Preempted from Current Attempt: <memory:0, vCores:0>  
Number of Non-AM Containers Preempted from Current Attempt: 0  
Aggregate Resource Allocation: 213686 MB-seconds, 118 vcore-seconds, 0 yarn.io/gpu-seconds  
Aggregate Preempted Resource Allocation: 0 MB-seconds, 0 vcore-seconds

Show 20 entries

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1720423563825_5565_000001	Wed Jul 10 15:56:39 +0800 2024	<a href="http://slave019:8042">http://slave019:8042</a>	<a href="#">Logs</a>	0	0