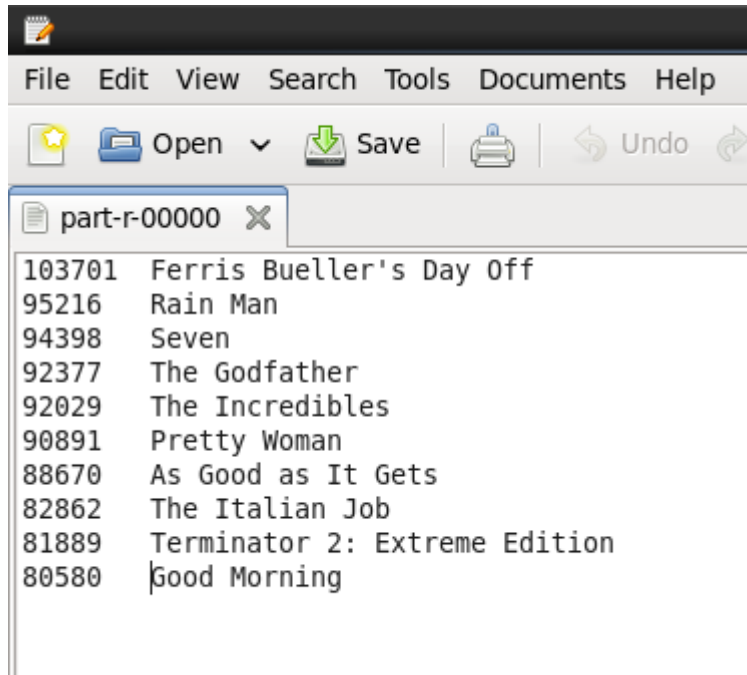


HOMEWORK 6

Problem 1: Top-10-List of Most Popular Movies (60%)

- (a) See java codes
- (b) The result of running implementation for $N = 10$ on the pseudo-cluster using the TrainingRatings.



The screenshot shows a Java IDE window titled 'part-r-00000'. The menu bar includes File, Edit, View, Search, Tools, Documents, and Help. The toolbar contains icons for Open, Save, Print, and Undo. The main text area displays a list of 10 movies with their corresponding IDs:

ID	Movie Title
103701	Ferris Bueller's Day Off
95216	Rain Man
94398	Seven
92377	The Godfather
92029	The Incredibles
90891	Pretty Woman
88670	As Good as It Gets
82862	The Italian Job
81889	Terminator 2: Extreme Edition
80580	Good Morning

Problem 2: Combiner (40%)

- (a) The respective line of code:

```
job.setCombinerClass(SumReducer.class);
```

Comparing with implementation without combiner, the result of job doesn't change.

- (b) Execution statistics of the job with Combiner:

FILE: Number of bytes read	28039	28033	56072
FILE: Number of bytes written	200045	171951	371996
HDFS: Number of bytes read	58524142	0	58524142
HDFS: Number of bytes written	0	21686	21686

Execution statistics of the job without Combiner:

FILE: Number of bytes read	50113669	50113663	100227332
FILE: Number of bytes written	100371136	50257412	150628548
HDFS: Number of bytes read	58524142	0	58524142
HDFS: Number of bytes written	0	21686	21686

The numbers of read and written in HDFS are same in both job. Because both jobs have the same Mapper input and Reducer output. So they have same operation on HDFS.

The numbers of read and written in FILE changed. Job without Combiner has much more number of bytes read and number of bytes written. Because the Combiner reduced the amount of intermediate data produced by Mappers to speed up data transfer to Reducers.

Combine input records	3255352	0	3255352
Combine output records	1822	0	1822

There are 3255352 Combine input records and 1822 Combine output records in total. So there are $3255352 - 1822 = 3253530$ key-value pairs that can be combined.

- (c) Because the Combiner acts as a mini Reducer so its function is similar as SumReducer. Input and output data types of Combiner and Reducer must be the same.

In this program, the values are ratings of movies so they are commutative and associative. We can sum up key-value pairs with same key just like what SumReducer does.

- (d) If the operation performed is not commutative or associative, we cannot use Reducer as Combiner. For example, the operation is to compute the norm for each

key. i.e. $[key, (value_1, value_2)] \rightarrow [key, \sqrt{value_1^2 + value_2^2}]$