Assignment

Problem Description

People often have their favourite actor, director, or some other member of the film crew and they often love to watch the movies in which they appear. In order to enable these people to quickly find such movies, it is necessary to build an index that contains a list of movies in which each person has appeared.

All the movie data are collected in `csv` files that contain information about movies. Complex data structures are stored as JSON stings.

Since the number of movies is incredibly large, and the refresh of this data needs to be done in the shortest time possible, it is necessary to write a Python program, Jupyter notebook or a script that will use Python Spark to enable distributed data processing.

Algorithm

It is necessary to create a dictionary that contains an `id` as a value for each person.

Example:/

```
Tom Hanks, 31
Angelina Jolie, 11701
Kevin Spacey, 1979
```

Note: Information about each person is in the files, in the `cast` and `crew` columns.

Afterwards, it is necessary to make an index that contains a list of identifiers of movies for each person, in which a person appears.

Example:

```
(31, [862, 34996, 16295, …])
(11701, [9886, 10830, 11412, …])
(1979, [8391, 11448, 63020, …])
```

Note: Movie records are sometimes repeated. Keep this in mind when solving the task.

The solution should be able to process a huge amount of data. We need to be able to run an algorithm on a distributed system, so we are not limited by the amount of space, memory, or CPU of a single computer.