Canonicalization of Media Entities

This exercise focuses on normalizing movie metadata and creating canonical versions of movies originating from different sources.
The file Movie.json contains data in JSON format for movie references. The reference_id uniquely identifies a reference. Design a scheme to create canonical movies from these references and implement it in Python or Java.
The output for a canonical should be in the following JSON format:
{
“canonical_id”: …,
— this is an unique ID assigned by you for this canonical
“references”: []
— these are the reference_ids of entities that are deemed similar enough to belong to this cluster. An array of one or more reference_ids.
“title”: …,
— the “best” title for this canonical derived from the titles of contributing references
You get to define what “best” means. Similarly
“description”:
“content_rating”:
“genre”:
“release_date”:
“cast_and_crew_all”:
}
Output result: a json file that contains movie canonicals with clustered references.
Please provide a zip or tarball of your python/java code and output file and command line to reproduce the result.
For example, if you did the exercise in Java:
%java -cp “.” myPackage.foo.Mainclass -inFile *inputFileName.json* -outFile *outputFile.json*

Also include a brief write-up of the approach taken, design choices and decisions made.

Good luck!