Task 1.1

Movie Title/Name: *Jaro Similarity Measure*
After inspecting both datasets, I noticed this attribute had the most discrepancies in the representations of the strings. There were more typos and OCR errors with this attribute and therefore, I chose this measure as I believed it worked best with cases of transpositions and typos.

Release Date/Year: *String Matchin*g
After inspecting both datasets, I noticed the commonality in this attribute between both datasets was the year. I decided to use regular expressions to extract the year from the AFI dataset's 'release_date' attribute to standardize it with IMDB's dataset. This allowed me to use a simple string matching, where I checked the two years were equal to one another.

Genre: *Jaccard Similarity Measure*
After inspecting both datasets, I noticed the categories for the genres were similar between both datasets. I converted the genre attribute for both datasets from strings to sets. I chose this measure, as I believed it was the best in finding similarities between two sets.

Task 1.2

The final scoring final that I created was:

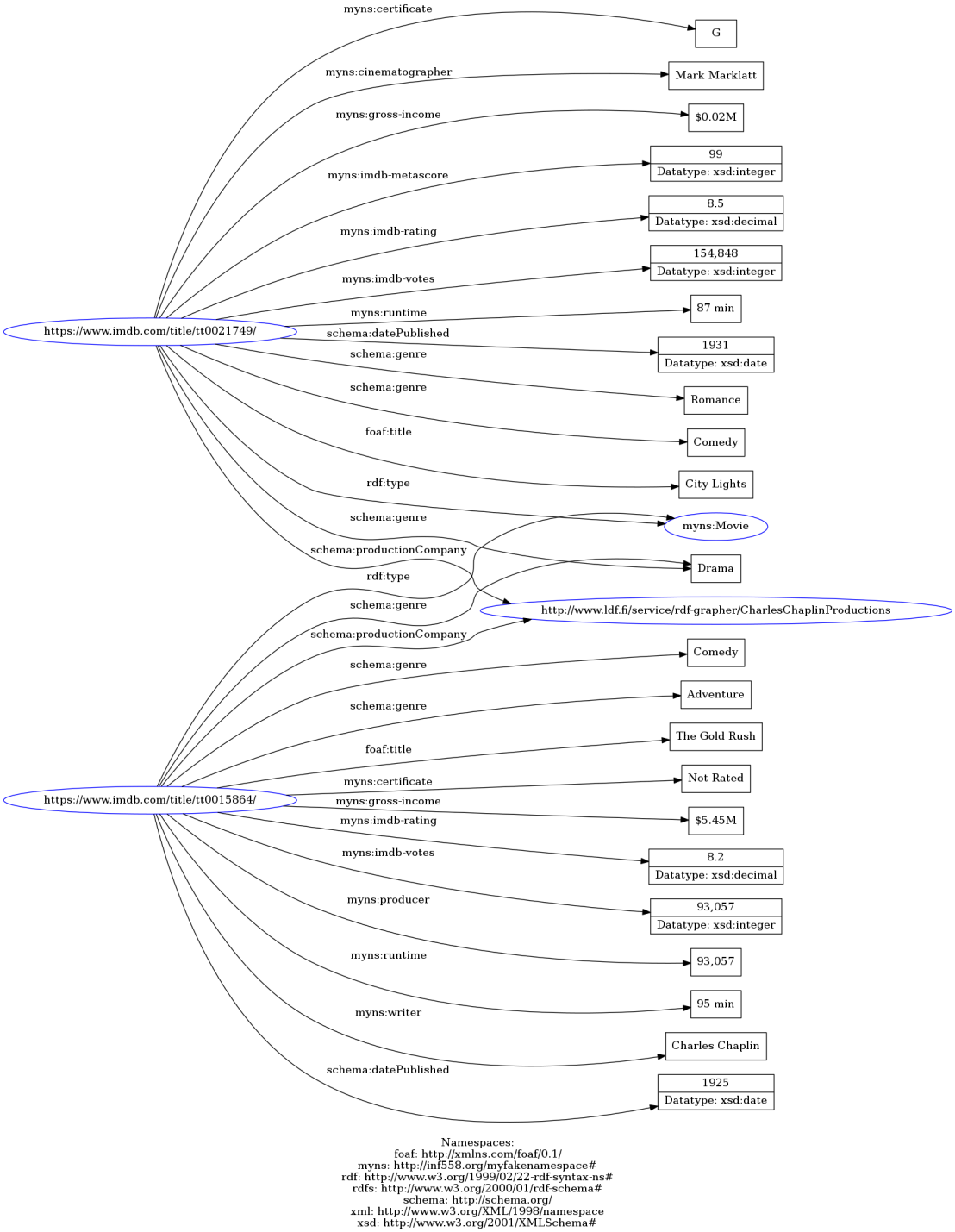
total = 0.7 * name_score + 0.1 * genre_score + 0.2 * year_score

I chose these weights based on the format of the data. The genre_score had the lowest weight, as it was the least specific of all the attributes would likely lead to false positives. Furthermore, I gave the year_score a weight that was slightly higher than that of the genre_score and marginally lower than that of name_score. This was because it had slightly more specificity than the genre but it had several missing values, especially in the AFI dataset. As a result, I chose to minimize its effect on the total. Finally, I gave the name_score the highest weight because it was the most specific attribute.

Task 2.1

<URI> <is a > <movie>.
<URI> <has a> <name>.
<URI> <was released on> <release-date>.
<URI> <has a certificate of> <certificate>.
<URI> <has a runtime of> <runtime>.
<URI> <has a genre of> <genre>.
<URI> <has a rating of> <imdb-rating>.
<URI> <has a score of> <imdb-metascore>.
<URI> <has> <imdb-votes>.
<URI> <grossed> <gross-income>.
<URI> <was produced by> <Producer>.
<URI> <was written by> <writer>.

<URI> <was shot by> <cinematographer>.
<URI> <was produced by> <production-company>.
<Production-company> <is an> <Organization>.



Namespaces:
foaf: http://xmlns.com/foaf/0.1/
myns: http://inf558.org/myfakenamespace#
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs: http://www.w3.org/2000/01/rdf-schema#
schema: http://schema.org/
xml: http://www.w3.org/XML/1998/namespace
xsd: http://www.w3.org/2001/XMLSchema#