Homework 3 Report

Task 1.1

The 3 attributes I think that are most useful for record linkage are the movie's title, release year, and directors. It is rare for 2 movies to that are released in the same year to have the same title. In the rare occurrence that this scenario does happen, it is almost impossible for the 2 movies to have the same directors. However, when constructing the datasets, I used all available attributes so that it makes creating the ttl in task 2.2 easier.

Task 1.2

I blocked the data using the release year of the movies. I believe blocking using the year would yield the best possible result. Blocking by either the title or directors would result in too many blocks with a narrow scope. This could be solved by using a sub-string of the title or directors, but if I blocked using year, then doing entity linking, then that is one less attribute that will have to compared for each pair as all the entries are already grouped by year.

Task 1.3

There are 2 things that I compared to make sure if to entries are the same after blocking, the movie's title and directors. For title, I used the Jaro-Winkler similarity because for a movie title, the order of the words (letter) is of high importance. However, sometimes one website may add something to the title at the end to differentiate it from other movies of the same name or sequels, e.g. Intern, Intern(I). Jaro-Winkler similarity places more emphasis on matches that occur towards the beginning of the string than the end so I think this is a good choice. For directors, as there can be more than 1 director, I did a pairwise comparison of the directors using Jaro similarity. This is like the Jaro-Winkler only that the beginning of a string is not as heavily weighted. An example of why I chose comparison instead of previous one is John Snow and John Smith. Though their names are similar, they are different people. However, I did not do a exact string match because sometimes, directors have a middle name and some website may not include their middle name when listing them.

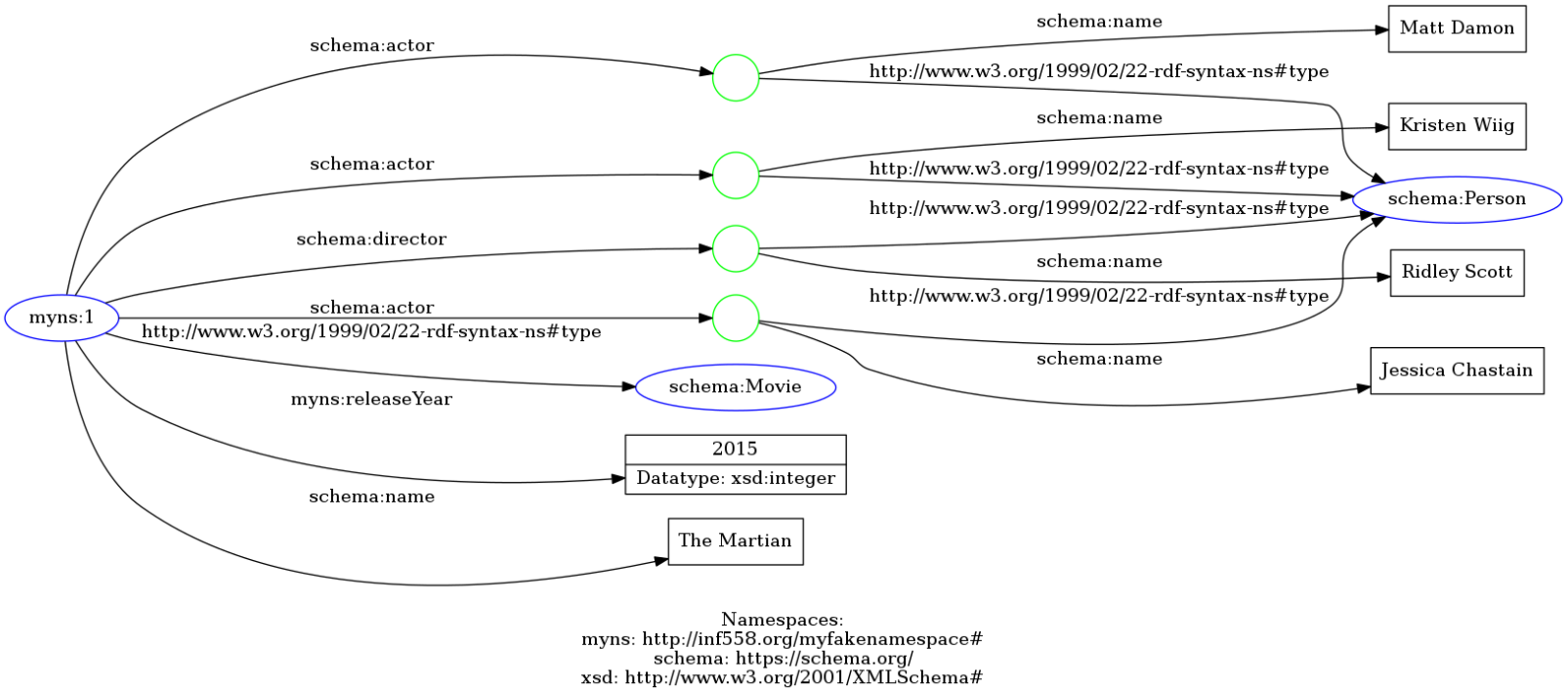
I weighted title/director match as 70/30 because having a matching title is a more indicative of whether two entries are the same. It is much easier to get false positives and negatives from director comparisons as there is more variation the comparisons.

The performance of my comparisons is not low.

Task 2.1
I created a property called releaseYear that denotes the year that the movie was made. The datatype that it expects is of xsd:integer.

Task 2.2

Task 2.3



Namespaces:
myns: http://inf558.org/myfakenamespace#
schema: https://schema.org/
xsd: http://www.w3.org/2001/XMLSchema#