DSCI Homework 2 Report

1. For blocking, I tried a number of different attributes including year and differing versions of titles. What I settled on was a two attribute blocking strategy. The first attribute is the year of the movie's release. While this is a quality attribute to use, it does not shrink the dataset much. Thus, I used a second attribute that I called "important word" within the title of the movie. The definition of the important word in the longest word in the movie title, unless the movie title has a number within it. I figured this would be a good blocking technique because it would vastly shrink my data set while also contributing to the eventual linking of the movies.
2. I tried three different similarity measures – Levenshtein distance, Jaro-Winkler similarity, and Jaccard similarity. The Jaccard similarity performed poorly, and was quickly discarded. Both Levenshtein distance and JW similarity were quality measures for similarity of the blocked movie titles. However, I ended up using the JW similarity because of the normalized aspect of the score. From the measure between 0 and 1, I was able to provide a confidence number that felt more accurate than a rules-based confidence measure based on arbitrary Levenshtein distances.
3. My Knowledge graph has three main nodes – movie, director, and production company. Each of these nodes have unique URIs. The director and production company URIs are random identifiers, while the movie URI is the IMDB tconst. The director and production company nodes have the property name. The movie nodes have properties startYear, genres, runtimeMinutes, original_release_date, audience_rating, tomatometer_rating, and rotten_tomatoes_link.