A Decade of Song Covers Knowledge Graph

Duyen Nguyen (8386158893), Naicih Liou (4353720070)

1. Project Domain & Goals

We plan to build a knowledge graph about song covers. The knowledge graph will contain information about the interested song such as title, artists, language, gerne, release date, album, thumbnail, how many times it has been played,..etc. We will also show the artist's information such as their birthdates, photos, websites and their upcoming concerts in the U.S. A knowledge graph for this domain is necessary since it will help people discover more information about the songs they have been listening to, for example, they can find out more songs in the same album or more songs performed by the same artist that they might be interested in, as well as discover upcoming concert from their favorite artists. This knowledge graph will be used to visualize popular genres that have changed over the years and if time permits, we plan to use machine learning to build a song recommendation system based on similar artists, similar genres,... etc.

2. Datasets & Representation

Three data sources are applied into this project, which are data from Secondhandsongs (https://secondhandsongs.com/page/Introduction), Last.fm (https://www.last.fm/music/), and ticketmaster (https://www.ticketmaster.com/) separately.

Secondhandsongs will mainly be used for extracting information related to songs and their covers such as the original song title, writer, artist, language, genre, cover song title, cover song language, cover song artist,...etc. Due to large data scale, time constraints and computational capacity, we will only focus on songs within the past 10 years (2012-2022) and by doing that, we will be able to build a basic knowledge graph about covered songs. This source is the most important source for our final knowledge graph since it provides essential information about songs and artists, which later we will link with data from Last.fm and Ticketmaster to produce the final knowledge graph that provides users with song/cover information, artist information and artist's upcoming concerts.

Last.fm will be used for extracting artist information such as the artist's personal websites, music style tags, and their photos. This serves the purpose of providing users information about their interested artist, so that they can learn more about the artist.

Ticketmaster will be used for extracting event information such as artist, event location, event time, and ticket url. This serves the purpose of providing users with information about the future concerts of their favorite artists.

Approximately 10000 pages will be scraped from these 3 sources, with the smallest source ticketmaster of at least 100 pages, and the structured source Secondhandsongs of at least 100 records, as well as producing at least 8 semantic types (song title, artist, language, genre, artist website, artist photo, event time, event location,... etc). The data from these 3 sources should provide us with meaningful content for our knowledge graph and meet the project requirements.

To represent the knowledge, the ontology in Schema.org is a good reference with a full and well-built system for singers and songs. For the knowledge of events, we will also use Schema.org to represent. Moreover, we will use rdf and rdfs for further knowledge of the entities..

3. Technical Challenge

We will be likely to encounter dynamic websites while scraping for data. It will be more challenging compared to the homework assigned in class, which was about dealing with static websites. To tackle this problem, we will use Selenium, an automated web-browser, controlled by Python to perform data scraping. If the tools are used properly and effectively, we should be able to retrieve meaningful data for the project.
The other issue we had to struggle with is the data consistency. Because there are three data sources, each one may have their own data format. It is a challenge to match the same entity together across the data sources.We tried to solve this problem by taking the consideration of different cases like name abbreviations or the similarity of the song pairs.

For the knowledge graph evaluation, we will use the concepts of quality dimensions, specifically, timeliness, accessibility, and relevancy.

Since the song covers we chose are from 2012-2022 and 2022 is still going on, as well as that the data from ticketmaster is in real-time and many of the events are in the future, we need to make sure the data we use for our knowledge graph is sufficiently updated. The more current data are being used, the more up-to-date our knowledge graph is. Therefore, we plan to update our knowledge graph once in a while to maintain the timeliness of it.

To measure the accessibility, we will use the interlinking and check whether the data from different sources can be sufficiently connected. For example, we will check if the singer in Secondhandsongs can be found in the websites Last.fm and tickermaster. The union size of the singers from these sources will be the measure for accessibility.

Lastly, the metric of relevancy will also be used as one of the measurements in this project. If we are able to implement the knowledge graph and solve the problem in the end by using the data from these sources, then they satisfy the relevancy metric.