INF 558 Building Knowledge Graphs Project Summary: K-pop Wiki

Xiaoyue Chen and Qiuke Wang

University of Southern California, Los Angeles CA 90007, USA

1 Introduction

We all know that K-pop has become a truly global phenomenon thanks to its distinctive blend of addictive melodies, slick choreography, production values, and an endless parade of attractive South Korean performers who spend years in grueling studio systems learning to sing and dance in synchronized perfection. The K-pop Wiki is a knowledge graph project we implemented that captured the information about the Korean Pop music groups. Through our project we want to help you to find the K-pop group you might be interested and provide all the detail information about the K-pop groups. Besides, based on the information we retrieved and model we built, we are willing to help you predict whether one pre-debut group will be popular or not according to all the features they have and find similar debut groups.

2 Challenges

2.1 Data Crawl

We have met several challenges from data crawling to find similarity. Fortunately, according to the previous experience and the tutor of the class, we have successfully dealt with all the problems.

Our first challenge is how to crawling data using Scarpy. We found our data source(Fandom.com) has different formats to describe same feature. For instance, sometimes the member section in the website would be plain text with hyperlinks, however others would be a table using column index to distinguish the position, year active for the member. At first we didn't notice these differences until we found there are too many blank features in our data. So what we did was optimizing our spider, adding more if-else statements to help us distinguish the various formats. Every time we successfully crawled the data we would manually check the blank features and optimize the spider.

The second milestone was how to deal with missing data. The significant feature in our project would be the gender of the group. This would help us to distinguish different members with same name. However in the data we retrieved there seldom has direct label with gender. What we did was trying to extract desired information from unstructured data(introduction and abstract for the group). For example, the first sentence of the abstract would usually be "XXX

is a four-member girl group...", according to that sentence we could easily determine the gender of the group. Then we found the abstract was not always existing. The solution we dealt with that situation is we extracting the gender from the members and inferring that to our group gender.

2.2 Entity Resolution

For the difficulty we met on the entity resolution part was the data we crawled from different resources usually had different labels. For example, there is a company called "BrandNew Music", but in the data source, it has several names such as "BNM", "Brand New Music" or even in Korean. As for this part we created a dictionary to manually convert different labels to the same . As for the group name in different format, we converted all the name to lower case, deleted all unnecessary punctuation and using levenshtein similarity to compare.

2.3 Popularity Prediction

The highlight in our project is we want to predict the popularity of the predebut group based on their features(company, gender, genre, member numbers). At first we embedded the features and tried to find the similar groups in our knowledge graph. Then we found sometimes we could not find similar pairs of groups. The solution we came up with is after we embedding the features, we trained a SVM + linear regression model to do the machine learning prediction of the popularity score. Then we tried to find similarity based on SPARQL query on distinct features. If we couldn't query the similar groups, we would set the score of this feature equals to the average score. Then we using linear regression to get the weight of different scores, adding voting method to predict the final result.

3 Conclusions

This system helps us to digest all the knowledge we learnt from the class. From crawling data, entity resolution to building a full-stack web project.

In our project we have successfully finished the search, query, visualize and predict parts for our data using Flask. The project could not only search bu the member or groups' name to find the information, but also had a digest of the trending popular group features and analyse of the different features causing popularity. It would have a certain influence on the people who study K-pop.

From what we achieved in our project, We learnt that knowledge graph is a better way when trying to analyze complicate relationships among different entities. The traditional database usually could not have a good query function when deal with complicate predicates.

Fig. 1. Steps of the project