INF 558 Final Project:

Korean-Pop WIKI

Xiaoyue Chen
Qiuke Wang

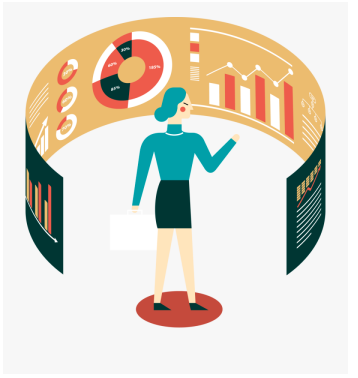
- Searching

- Recommending

- Popularity Prediction

Objective

- Presenting statistic and trends with data visualization.



Process & Tools

Data Crawling

Scrapy

BeautifulSoup

Spotify

Web API

import.io

Mapping to ontology

usc-isi-i2/rltk

Entity Linking

Regular expression

String similarity

Knowledge Graph

Jena

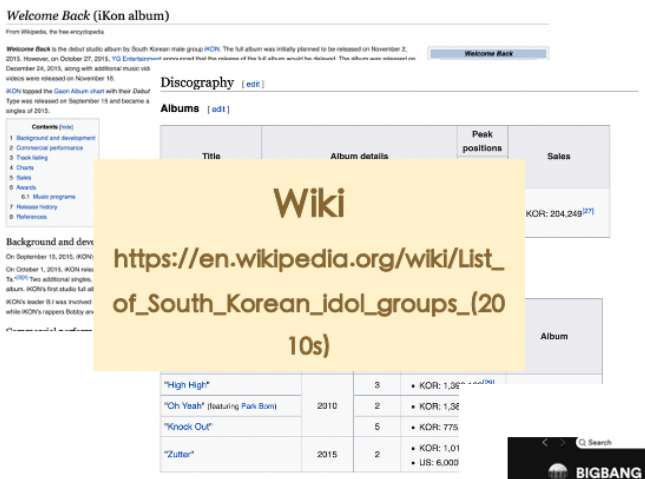
Query & Web

Apache Jena Fuseki

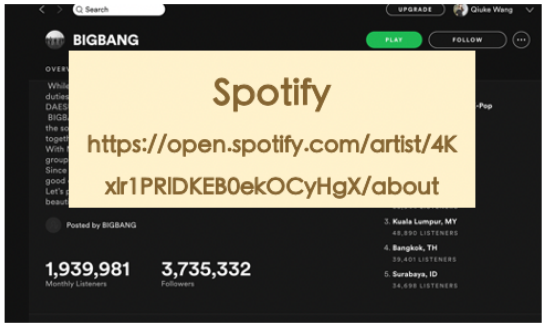
Flask
web development, one drop at a time

1.0.0

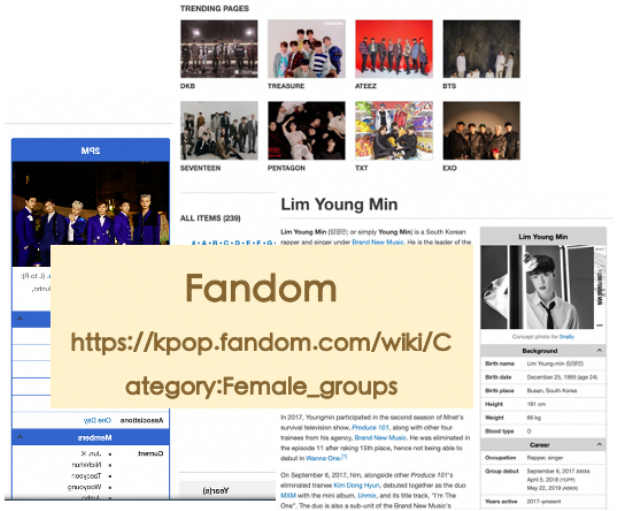
D3



Group label
Group genres



Monthly listener



Group name
Group gender
Group introduction
Group members
Group image
Member information
.....

Data Sources

Data Crawling

Challenge: inconsistent structure

```
<aside class="portable-infobox pi-background pi-theme-artist pi-layout-default">…</aside>
<b>BEBE6</b>
" (베베식스) is a six-member girl group under K12 Company. They debuted on August 10, 2017 with their first single "Shot Me".
"
```

```
</aside>
<p>
<b>Big Brain</b>
" (빅브레인) is a four-member R&B ballad group under "
<a href="/wiki/MAJOR9" title="MAJOR9">MAJOR9</a>
". They have been active since 2010 and officially debuted on October 22, 2015 with the single "Billionaire Sound" and their breakthrough song “Welcome”.
"
```

Members

- Sang Hoon (상훈)
- Jin Yong (진용)
- Byeong Eun (병은)
- Hong Hyun (홍현)

Members

Name	Position(s)	Year(s) active
Donghyun (동현)	Leader, Lead Vocalist	2011–2019
Hyunseong (현성)	Main Vocalist	2011–2019

Solution:

Optimizing spider and take more cases into consideration.

Data Crawling

Challenge: missing information

Group gender: Girl group or Boy group or Co_ed group? Number of group member?
How to measure popularity?

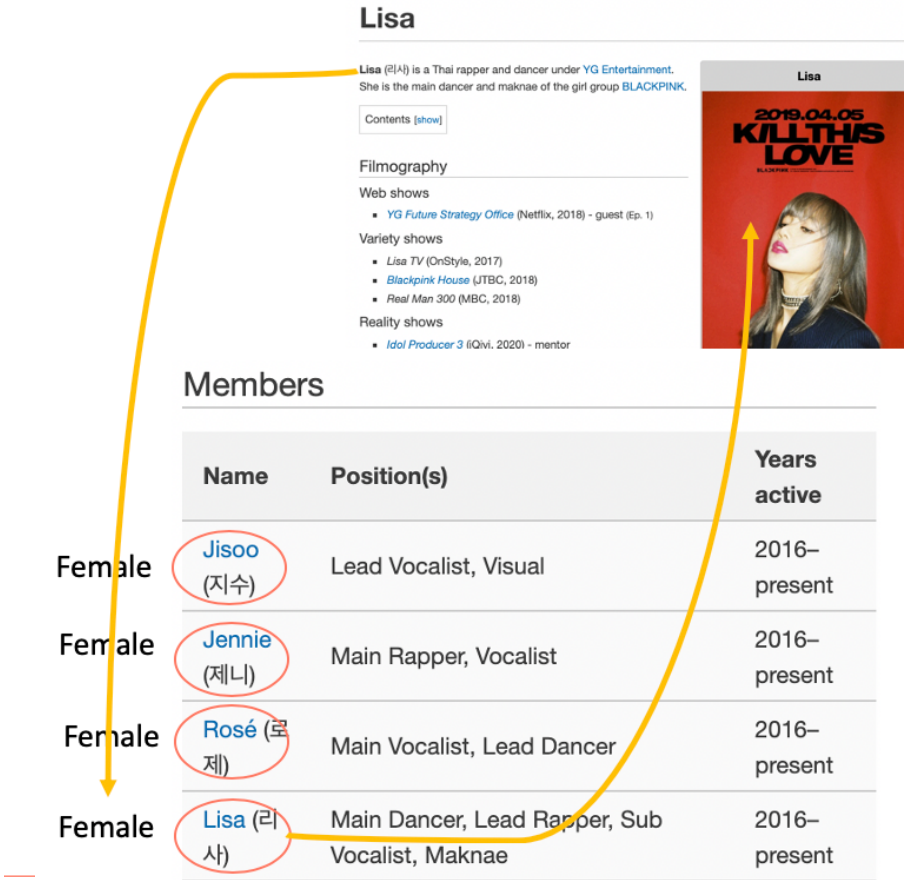
Solution:

Extracting desired information from unstructured data(introduction) or inferring from known knowledge.

BLACKPINK (블랙핑크; stylized as **BLΛƆKPIИK**) is a four-member girl group under YG Entertainment. They debuted on August 8, 2016 with their digital single album "Square One".

Count(members) = 4
Gender = 'Female'

```
Accuracy: 0.880000
Precision: 0.916000
Recall: 0.880000
F1 score: 0.894975
```



Lisa

Lisa (리사) is a Thai rapper and dancer under YG Entertainment. She is the main dancer and maknae of the girl group BLACKPINK.

Contents [show]

Filmography

Web shows
- *YG Future Strategy Office* (Netflix, 2018) - guest (Ep. 1)

Variety shows
- *Lisa TV* (OnStyle, 2017)
- *Blackpink House* (JTBC, 2018)
- *Real Man 300* (MBC, 2018)

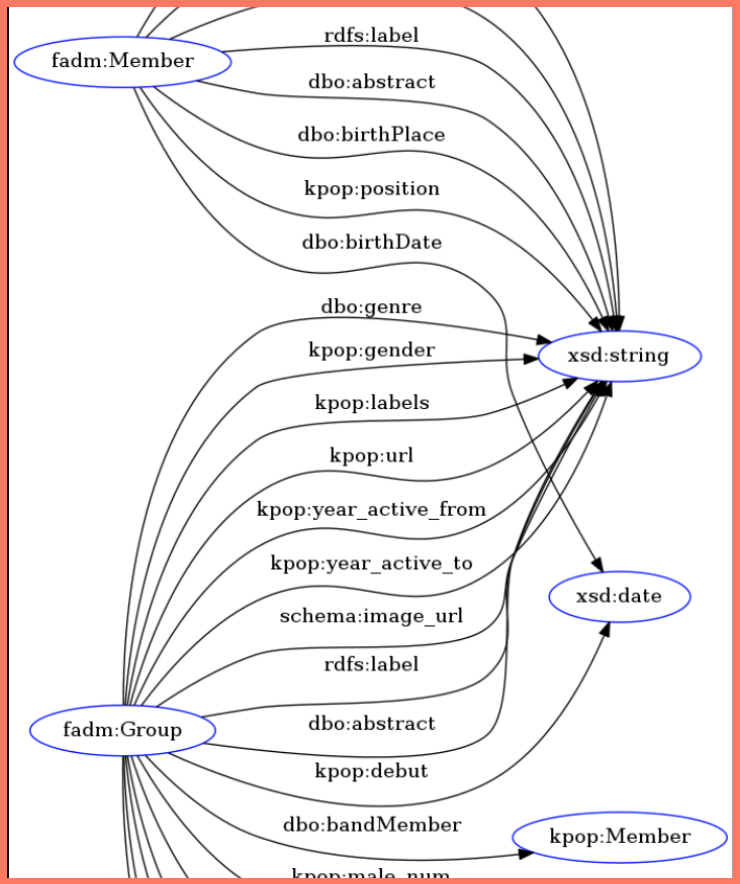
Reality shows
- *Idol Producer 3* (iQiyi, 2020) - mentor

Lisa

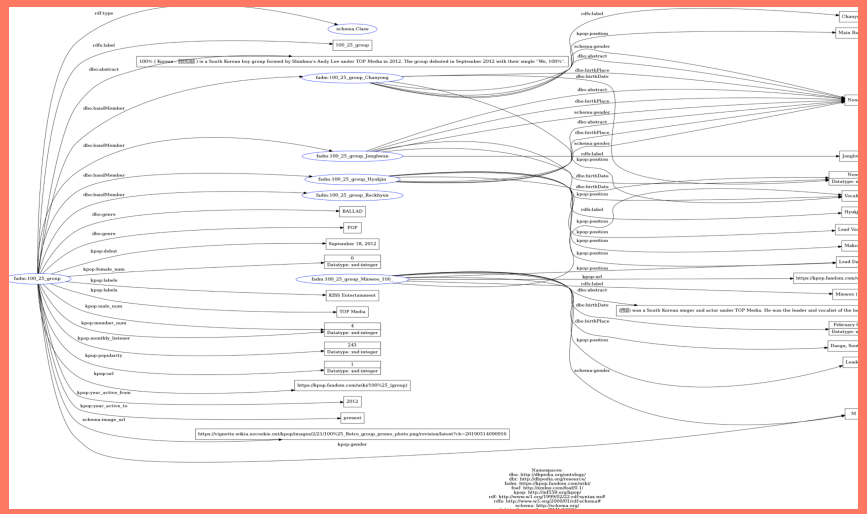
2019.04.05 KILL THIS LOVE

Members

	Name	Position(s)	Years active
Female	Jisoo (지수)	Lead Vocalist, Visual	2016–present
Female	Jennie (제니)	Main Rapper, Vocalist	2016–present
Female	Rosé (로제)	Main Vocalist, Lead Dancer	2016–present
Female	Lisa (리사)	Main Dancer, Lead Rapper, Sub Vocalist, Maknae	2016–present

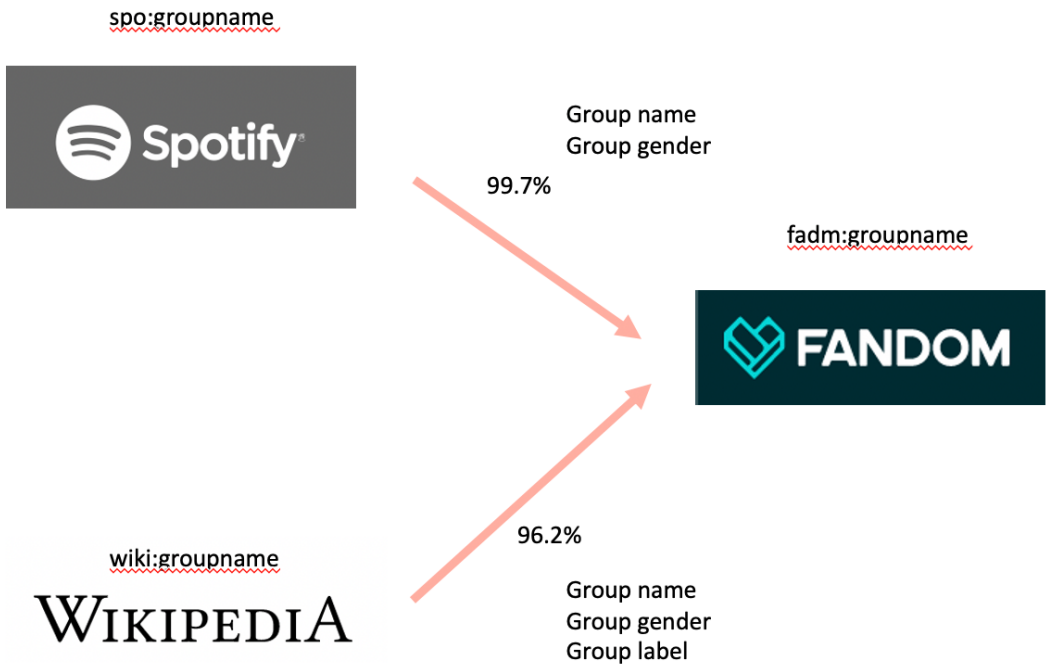


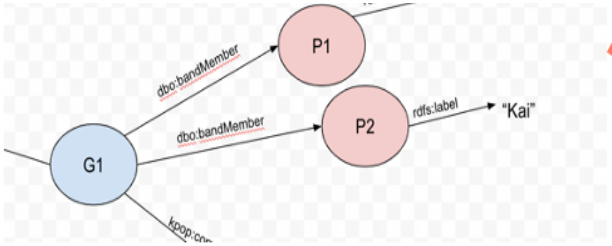

```
@prefix dbo: <http://dbpedia.org/ontology/> .
@prefix dbr: <http://dbpedia.org/resource/> .
@prefix fadm: <https://kpop.fandom.com/wiki/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix kpop: <http://inf558.org/kpop/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix schema: <http://schema.org/> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<https://kpop.fandom.com/wiki/100_25_group> a schema:Class ;
    rdfs:label "100_25_group" ;
    dbo:abstract "100% ( Korean : 백퍼센트 ) is a South Korean boy group formed by Shinhwa's And
    dbo:bandMember <https://kpop.fandom.com/wiki/100_25_group_Chanyong>,
        <https://kpop.fandom.com/wiki/100_25_group_Hyukjin>,
        <https://kpop.fandom.com/wiki/100_25_group_Jonghwan>,
        <https://kpop.fandom.com/wiki/100_25_group_Rockhyun> ;
    dbo:genre "BALLAD",
        "POP" ;
    kpop:debut "September 18, 2012" ;
    kpop:female_num 0 ;
    kpop:gender "M" ;
    kpop:labels "KISS Entertainment",
        "TOP Media" ;
    kpop:male_num 4 ;
    kpop:member_num 4 ;
    kpop:monthly_listener 243 ;
    kpop:popularity 1 ;
    kpop:url "https://kpop.fandom.com/wiki/100%25_(group)" ;
    kpop:year_active_from "2012" ;
    kpop:year_active_to "present" ;
```



Ontology

Entity Linking

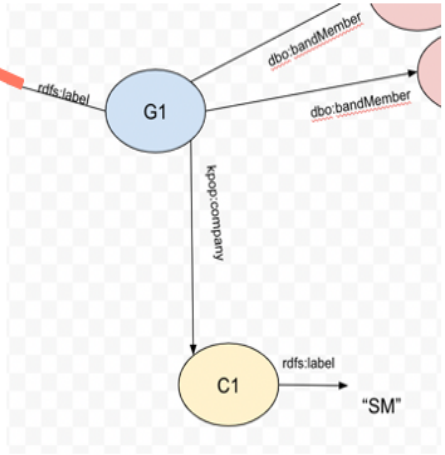




Kpop:company ------- rdfs:label

BNM
Brand New Music
브랜뉴 뮤직

BrandNew Music



Using dictionary to direct convert the different name to the same label.

rdfs:label ------- fandm:groupname

G-idle

(G)I-DLE
GI-DLE
GIDLE
GIDLE(group)

Lower case, delete all unnecessary punctuation, using gender label.

levenshtein similarity > 0.9

Entity Resolution

SPARQL

SPARQL query

Please input your sparql query here

```
SELECT distinct ?group ?groupname ?imageurl
WHERE{
 ?group a schema:Class.
 ?group schema:image_url ?imageurl.
 ?group rdfs:label ?groupname.
 ?group kpop:gender ?gender.
 filter regex(?gender,'M')
 ?group kpop:labels ?company.
 filter regex(?company,'')
 ?group kpop:member_num ?number.
 filter (?number)
}
LIMIT 5
```

QUERY

```
sparql = SPARQLWrapper("http://localhost:3030/kpop/query")
sparql.setReturnFormat(JSON)

prefix =  """
        PREFIX : <undefined>
        PREFIX kpop: <http://inf558.org/kpop/>
        PREFIX dbr: <http://dbpedia.org/resource/>
        PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
        PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
        PREFIX foaf: <http://xmlns.com/foaf/0.1/>
        PREFIX owl: <http://www.w3.org/2002/07/owl#>
        PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
        PREFIX dbo: <http://dbpedia.org/ontology/>
        PREFIX fadm: <https://kpop.fandom.com/wiki/>
        PREFIX schema: <http://schema.org/> """
```

groupname	imageurl
TVXQ!	https://vignette.wikia.nocookie.net/kpop/images/d/d4/TVXQ%21_XV_promo_photo_1.png/revision/latest?cb=20191104063447
SuperM	https://vignette.wikia.nocookie.net/kpop/images/2/2c/SuperM_Let%27s_Go_Everywhere_promo_photo_1.png/revision/latest?cb=20191119001756
SHINee	https://vignette.wikia.nocookie.net/kpop/images/d/d8/SHINee_Sunny_Side_promo_photo.png/revision/latest?cb=20180706154531
SUPER_JUNIOR	https://vignette.wikia.nocookie.net/kpop/images/f/fa/SUPER_JUNIOR_Timeless_group_promo_photo.png/revision/latest?cb=20200115010831
EXO	https://vignette.wikia.nocookie.net/kpop/images/1/16/EXO_Obsession_group_concept_teaser_photo_1.png/revision/latest?cb=20191127040511

```
    if num == '':
        queryline = "SELECT distinct ?group ?name WHERE{ ?group a schema:Class. ?group kpop:labels ?company.filter regex(?com
    else:
        queryline = "SELECT distinct ?group ?name WHERE{ ?group a schema:Class. ?group kpop:labels ?company.filter regex(?com
    sparql.setQuery(prefix + queryline)
    temp = sparql.query().convert()
    resultGroup = []
    if len(temp["results"]["bindings"]) > 0:
        keysGroup = temp["results"]["bindings"][0].keys()
        for i in range(len(temp["results"]["bindings"])):
            line = []
            for key in keysGroup:
                if temp["results"]["bindings"][i][key]["type"] == 'uri':
                        #need to replace link
                    line.append((temp["results"]["bindings"][i][key]["value"],True))
                else:
```

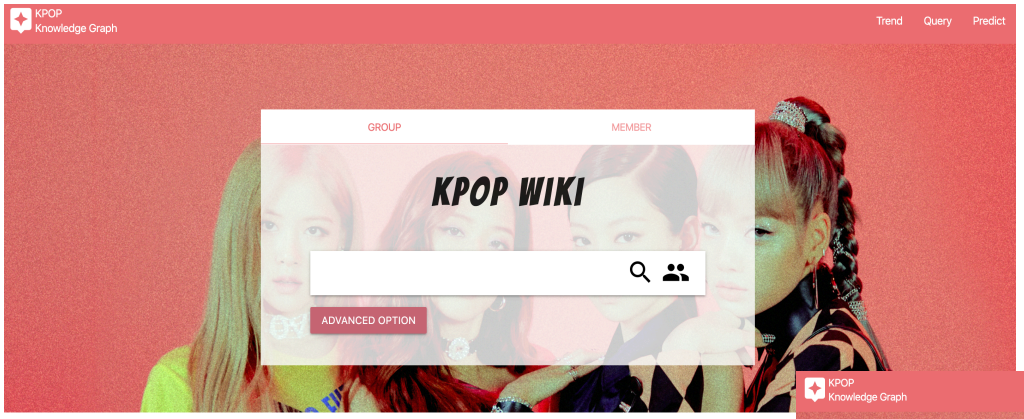
Choosing KG Plateforms

- Apach Jena provides Fuseki components.
- Fuseki is a SPARQL server that supports the SPARQL language for retrieval and runs efficiently on both stand-alone and server sides.
- For the web application we built, we do need more query functions than the visualization.

VS

Web

search & detail Page



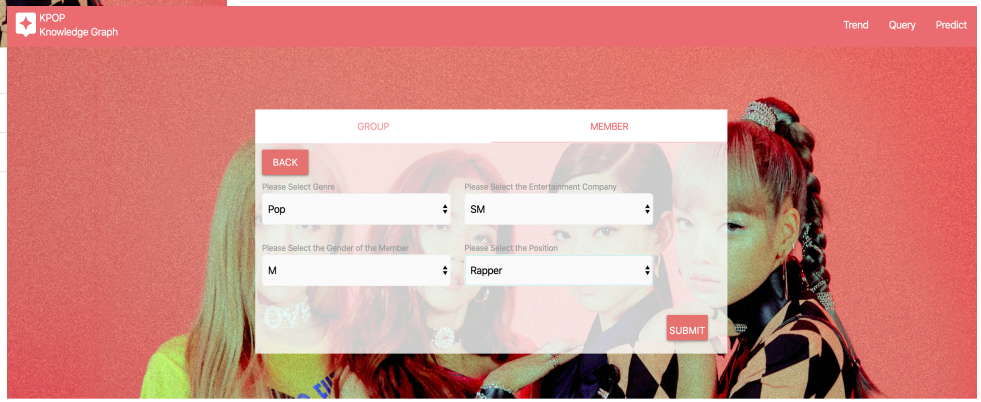
General Search(name)

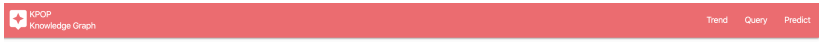
Advanced Search

group	name
https://kpop.fandom.com/wiki/1TEAM	1TEAM
https://kpop.fandom.com/wiki/1THE9	1THE9
https://kpop.fandom.com/wiki/2NE1	2NE1

Search Group

Search Member





WINNER

Winner (Korean : 위너), stylized as WINNER , is a South Korean boy group formed in 2013 by YGE through reality survival program WIN: Who Is Next . The group currently consists of four members, Jinwoo , Seunghoon , Mino and Seungyoon . Originally a five-piece group, Taehyun , departed from the group in November 2016 due to health and well-being. After several delays, the group officially debuted on August 17, 2014 with the release of studio album 2014 S/S ahead of official debut, claiming the number one spot on Gaon Album Chart and Billboard World Albums Chart upon release. The lead singles "Em...

BANDMEMBER:
https://kpop.fandom.com/wiki/Jinu_(WINNER) ,
https://kpop.fandom.com/wiki/Yoon_(WINNER) ,
https://kpop.fandom.com/wiki/Nam_Tae_Hyun ,
https://kpop.fandom.com/wiki/Mino ,
https://kpop.fandom.com/wiki/Hoony

DEBUT: August 12, 2014

GENDER: M

GENRE: POP

LABELS: YG

MEMBER_NUM: 5

MONTHLY_LISTENER: 823424

POPULARITY: 4

URL:
https://kpop.fandom.com/wiki/WINNER

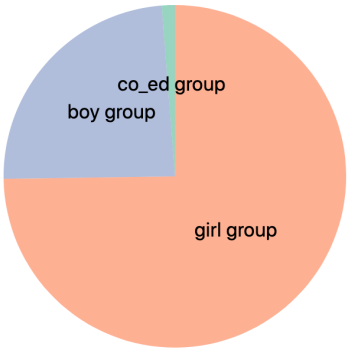
YEAR_ACTIVE_FROM: 2014

YEAR_ACTIVE_TO: present

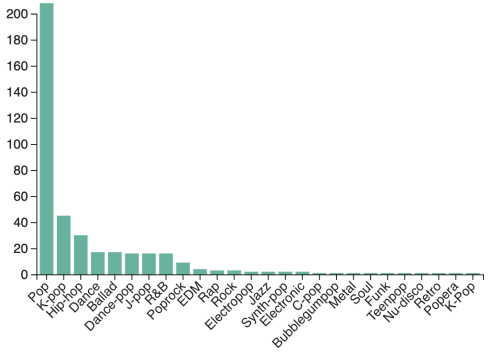
https://kpop.fandom.com/wiki/WINNER

Member	Detail
https://kpop.fandom.com/wiki/TVXQ_Park_Yu_Chun **NAME:**Park Yu Chun **Group:**TVXQ!	**BIRTHDATE:** June 4, 1986 . **GROUPNAME:** TVXQ! . **GROUP:** https://kpop.fandom.com/wiki/TVXQ! . **BIRTHPLACE:** Seoul, South Korea . **POSITION:** Main Rapper , Vocalist . **URL:** https://kpop.fandom.com/wiki/Park_Yu_Chun . **LABEL:** Park Yu Chun .

Trend Page



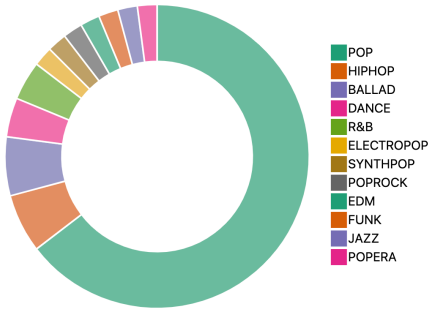
Group Gender Pie Chart



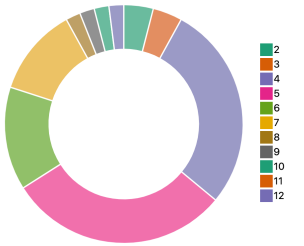
Group Genre Bar Chart

Trend Overview

Group Overview

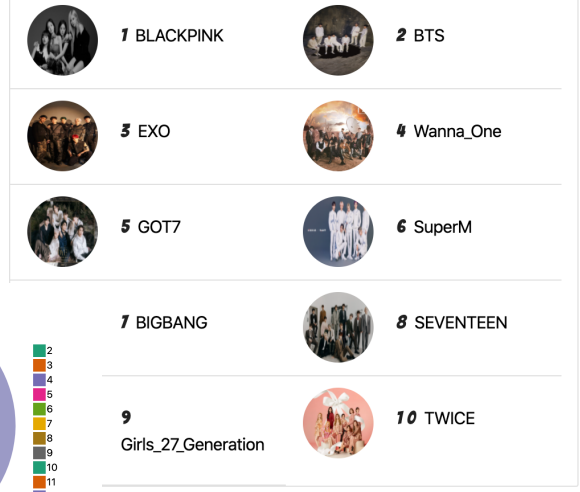


top 50 Most Popular K-Pop Grou

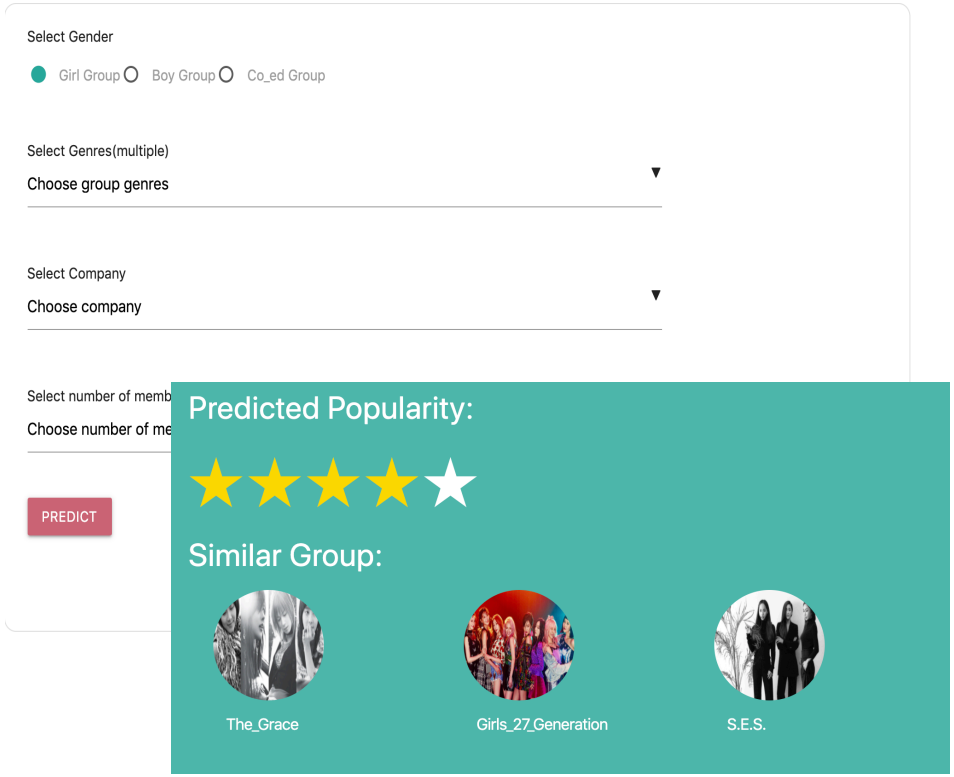


top 50 Most Popular K-Pop Groups Number of Members

TOP 10 MOST POPULAR K-POP GROUPS (2020)



Predict Popularity



How to predict popularity when we can't find similar group?

Solution:
Embedding the features(genre, gender, company and number of members). Train SVM + linear regression model. Query the avg popularity basing on different features, then add the predict score of our model. Using voting method to get the final score.

Machine Learning + Find Similarity + SPARQL

Evaluation

Data Crawling

Reviewing the data scraping from the web. Analyzing the wrong data and improve the spider.

Mapping to ontology

Using RDF validation tool.

Entity Linking

Labeling the test data manually and detect the accuracy.

Query & Web

Using SPARQL to test on the data and validate manually.

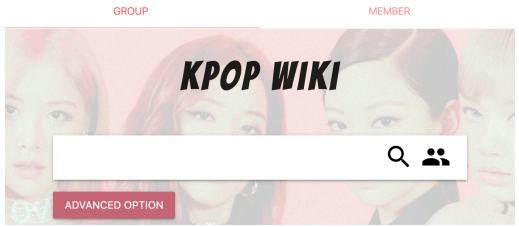
Improvement

We invited friends to help us test the whole application. We built a more user-friendly project according to their feedback.

GROUP: https://kpop.fandom.com/wiki/TXT .
BIRTHPLACE: Honolulu, Hawaii .
POSITION: Center , Maknae , Visual , Lead Vocalist .
URL: https://kpop.fandom.com/wiki/Hueningkai .

1. Add more hyper-links

In the member detail page, we add not only the attributes belonging to the member, but also the group information the member belongs to.



2. Change UI

After listening to users' advice, we combined the group search and member search together into one page using tab.

3. Retrieve more information

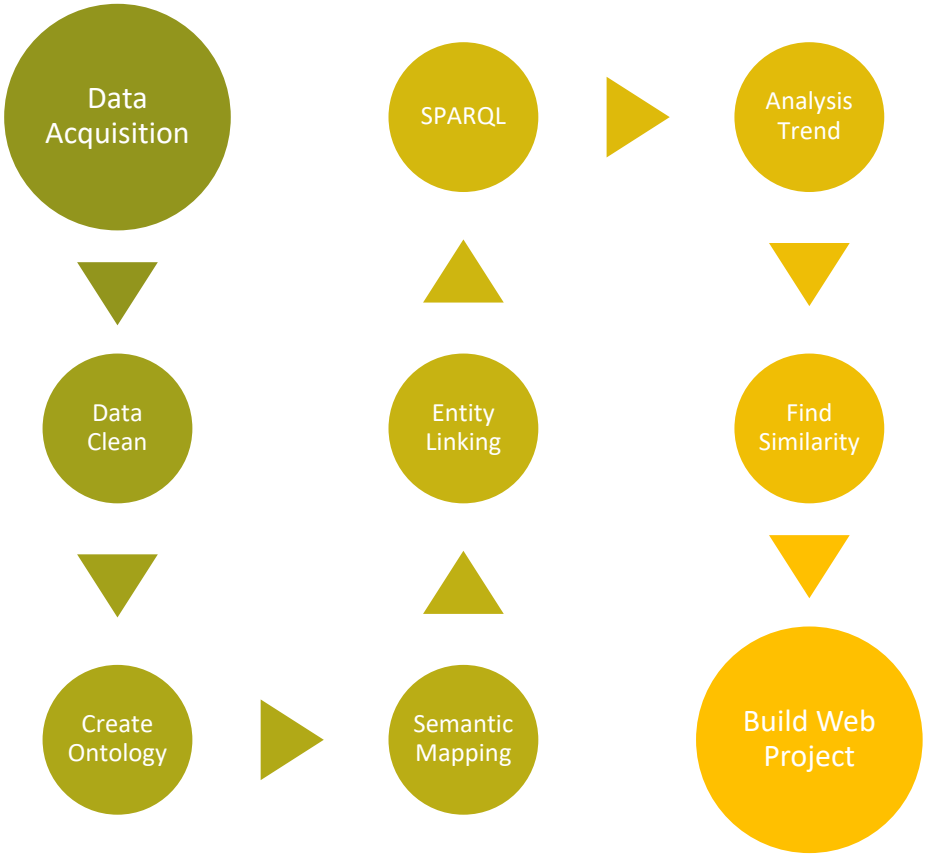
In order to present the images of each member and group., we crawled the data source again to make the description page looks better.

MAMAMOO

```
schema:image_url "https://vignette.wikia.nocookie.ne
```

Entertainment in 2014. The group officially debuted on June 18,

Conclusion



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

We learnt that knowledge graph is a better way when trying to analyze complicate relationships among different entities.

With the trend page and predict tool, we have a better understanding of data visualization and machine learning.

Thank You for Watching!