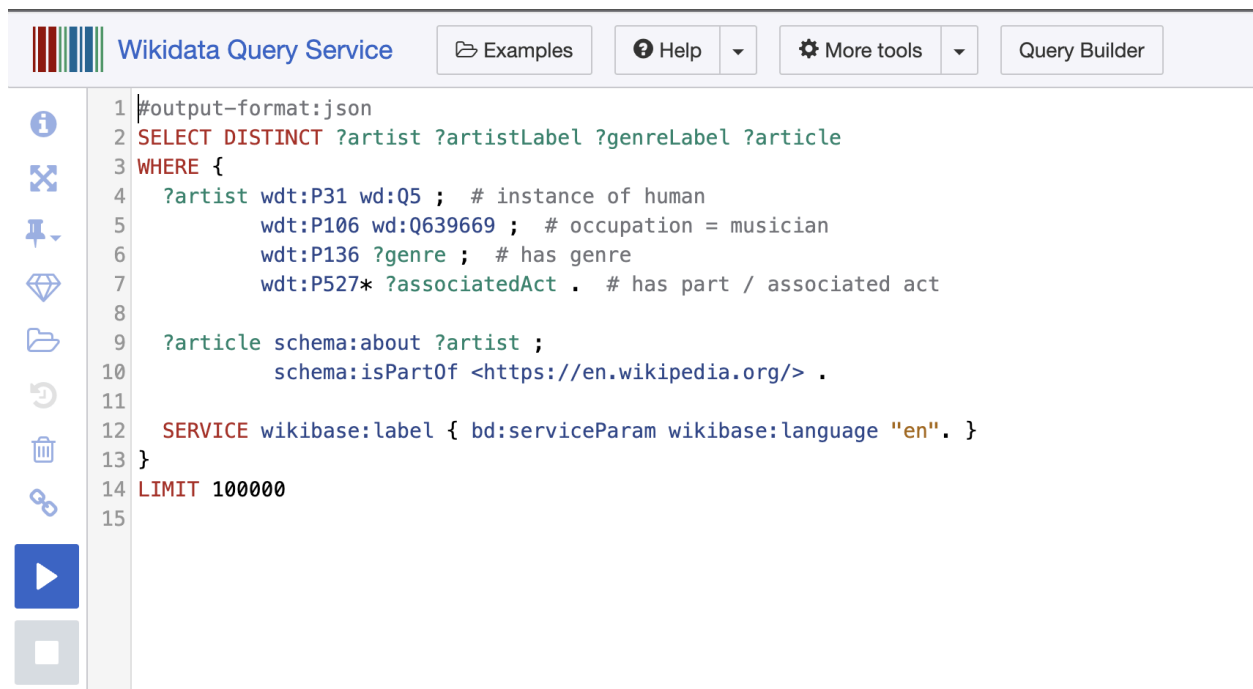


Artist Influence & Genre Graphing

Project Overview

Goal: Identify influential artists and highly connected genres from a music dataset using a graph-based model. Prioritize which artists to label by combining influence and community centrality.

Dataset



The screenshot shows the Wikidata Query Service interface. At the top, there's a header with the Wikidata logo, the text "Wikidata Query Service", and buttons for "Examples", "Help", "More tools", and "Query Builder". Below the header is a text area containing a SPARQL query. The query is as follows:

```
1 #output-format:json
2 SELECT DISTINCT ?artist ?artistLabel ?genreLabel ?article
3 WHERE {
4   ?artist wdt:P31 wd:Q5 ; # instance of human
5           wdt:P106 wd:Q639669 ; # occupation = musician
6           wdt:P136 ?genre ; # has genre
7           wdt:P527* ?associatedAct . # has part / associated act
8
9   ?article schema:about ?artist ;
10            schema:isPartOf <https://en.wikipedia.org/> .
11
12   SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
13 }
14 LIMIT 100000
15
```

On the left side of the query editor, there is a vertical toolbar with icons for information, expand/collapse, pin, diamond, folder, refresh, trash, and link. At the bottom left, there are buttons for running the query (a play button) and saving it (a square button).

- Source: [Wikidata query](#)

Data Processing

- Ran python script to pull length for each article from wikipedia api
- Trimmed strings and filtered entries with missing names or non-positive lengths
- Parsed genres from stringified lists (e.g. "[rock', 'pop']")
- Final size: 10549 x 3
- Contains: Artist name, genres (list), and length (Wikipedia page character count)

Code Structure

Modules

- **main.rs**: Executes overall workflow and prints key results
- **graph.rs**: Handles dataset parsing and builds the graph of artists and genres
- **community.rs**: Implements genre community detection, influence scoring, and labeling prioritization

Key Functions & Types

- **Graph struct**: Stores the graph (edges), artist lengths, and genre mappings
- **load_from_csv()**: Parses the CSV into the Graph structure
- **find_communities()**: Groups artists by genre into genre-based communities
- **top_influential_artists()**: Ranks artists by a log-scaled influence score
- **genre_connectivity_map()**: Calculates how interconnected genres are (via shared artists)
- **prioritized_labeling_targets()**: Combines genre connectivity with influence to rank artists for labeling

Main Workflow

1. Load the graph from CSV
2. Detect genre-based communities
3. Score and rank top artists by influence
4. Compute genre connectivity
5. Combine metrics to prioritize artists for labeling
6. Print top results for each category

Tests

1. **test_genre_connectivity_basic**
 - Verifies that genres sharing artists are properly linked in the connectivity graph
 - Confirms mutual connections (e.g., Rock connects to Pop, Pop connects to Jazz, etc.)
2. **test_prioritized_labeling_scoring**
 - Ensures that an artist with higher influence and more genre connections ranks higher
 - Verifies that the prioritized list returns expected artists in correct order.

```

running 2 tests
test test_prioritized_labeling_scoring ... ok
test test_genre_connectivity_basic ... ok

test result: ok. 2 passed; 0 failed; 0 ignored; 0 measured; 0 filtered out; finished in 0.00s

(base) ethanfreshman@crc-dot1x-nat-10-239-163-146 project1 %

```

Results

Sample Output

- Top 10 genres by connectivity
 - Genre connectivity is total number of unique genres connected to an artist
- Top 10 artists by influence score ($\log_{10}(\text{length}) * 10$ for smooth scaling of length)
- Top 10 artists prioritized for labeling (influence x genre connectivity)
 - Gives higher scores to artists that are both influential and genre-central

```

project1 -- -zsh -- 90x60
Compiling project1 v0.1.0 (/Users/ethanfreshman/Desktop/OS210/final/project1)
Finished 'release' profile [optimized] target(s) in 0.68s
Running 'target/release/project1'
Total artist-artist links created: 2577552
Top 10 communities by total influence score:
Genre: pop music with total influence score: 18935831.00
Genre: rock music with total influence score: 17615686.00
Genre: jazz with total influence score: 11899615.00
Genre: country music with total influence score: 18023147.00
Genre: alternative rock with total influence score: 9370401.00
Genre: hip-hop with total influence score: 7481859.00
Genre: blues with total influence score: 6551180.00
Genre: soul with total influence score: 6054955.00
Genre: folk music with total influence score: 5229457.00
Genre: classical music with total influence score: 5077530.00

Top 10 artists by influence score:
-----
Justin Bieber (length: 317943, score: 55.02)
Adele (length: 270680, score: 54.32)
Bruno Mars (length: 266992, score: 54.26)
Paul McCartney (length: 264094, score: 54.22)
David Bowie (length: 263581, score: 54.21)
Bob Dylan (length: 260730, score: 54.16)
Elton John (length: 253939, score: 54.05)
Christina Aguilera (length: 238055, score: 53.77)
Lauryn Hill (length: 237808, score: 53.76)
Amy Winehouse (length: 230169, score: 53.62)

```

```

Top 10 most connected genres:
-----
pop music connects to 243 other genres
rock music connects to 207 other genres
jazz connects to 201 other genres
alternative rock connects to 151 other genres
pop rock connects to 145 other genres
hip-hop connects to 125 other genres
electronic music connects to 118 other genres
classical music connects to 106 other genres
blues connects to 99 other genres
folk music connects to 99 other genres

Top 10 artists to prioritize for labeling:
-----
David Bowie (labeling priority score: 87158.38)
Frank Zappa (labeling priority score: 66289.62)
Christina Aguilera (labeling priority score: 55110.94)
Bob Dylan (labeling priority score: 49179.01)
Bruno Mars (labeling priority score: 47807.45)
Adele (labeling priority score: 47695.97)
Tina Turner (labeling priority score: 46285.74)
Ed Sheeran (labeling priority score: 42742.63)
Phil Collins (labeling priority score: 41464.30)
George Harrison (labeling priority score: 38128.35)

Total number of communities: 1017
Total number of artists: 10397
Total number of edges: 11414
Total number of genres: 18397
(base) ethanfreshman@crc-dot1x-nat-10-239-163-146 project1 %

```

Interpretation

- Genres like “rock” and “pop” tend to have high connectivity, serving as hubs
- High-priority artists often appear in multiple genres and have large length scores

Usage Instructions

cargo run --release

Expected Runtime: < 5 seconds