CSE304 Term Project

Box to Box: Analyzing CrossFit WOD Variability with Structured Embeddings

Sehyun Yun (20231233), nawhji@unist.ac.kr

Background

What is Crossfit?

- CrossFit is a high-intensity training system built on varied functional movements.
- Workouts (WODs) combine gymnastics, weightlifting, and cardio under time or repetition constraints.

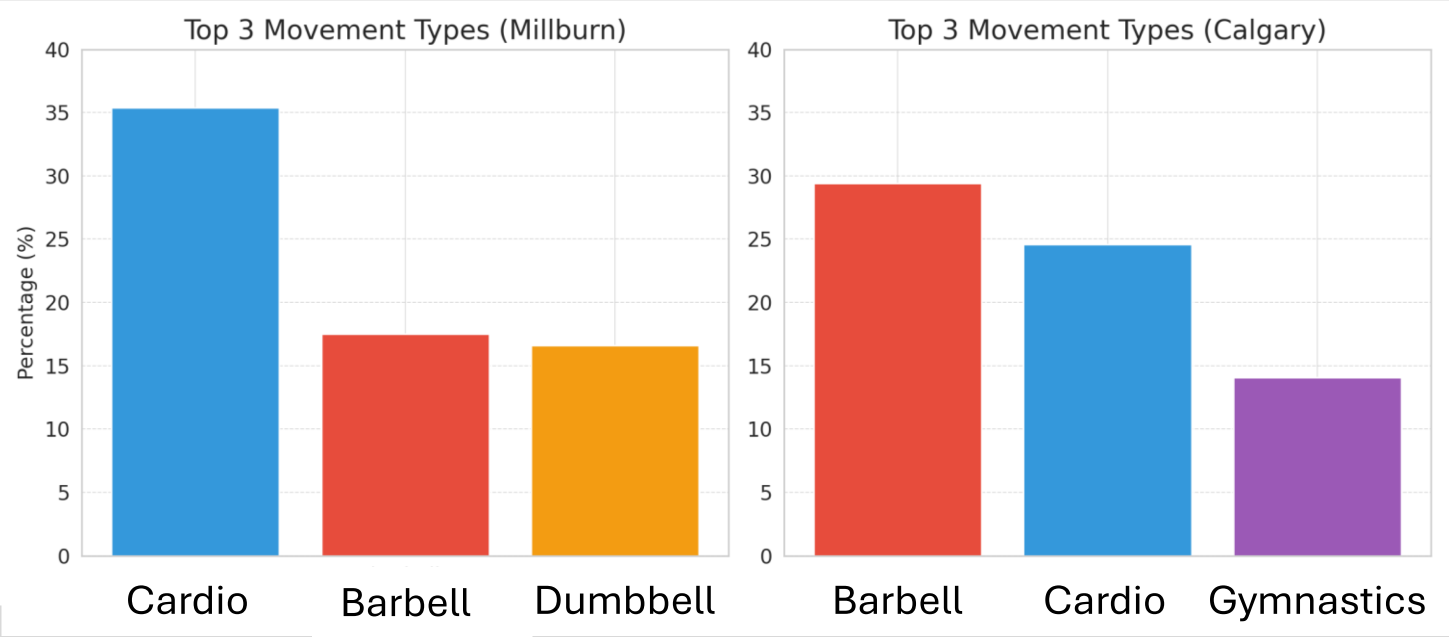
Challenges in WOD(Workout of the Day) Analysis

1. Crossfit gym(Box) level Variation

Each CrossFit box exhibits a unique WOD programming style.

- Millburn prefers cardio based workouts.

- Calgary emphasizes heavy barbell movements and strength training.



2. Textual Variability

Source A	Rephrased
For Time 2 Rounds 200 M DB/KB Farmers Carry 9 Power Cleans (115/85) 9 Pull Ups -into- 2 Rounds 200 M DB/KB Farmers Carry 7 Power Cleans (135/95) 7 Pull Ups -into- 2 Rounds 200 M DB/KB Farmers Carry 5 Power Cleans (155/105) 5 Pull Ups -Then- Buy Out: 400 Meter Run	**For time 6 Rounds:** 200m DB/KB Farmers Carry(50/35)(24k/16k) 9-9-7-7-5-5 Power Cleans, Pull ups (115/85 → 135/95 → 155/105) Buy Out: 400m Run

The same workout can be described in multiple ways, depending on the box or coach.

→ Difficult to parse and compare WODs systematically

Objective

This study aims to analyze the variability of CrossFit WODs by:

- Structuring raw WOD texts into a unified JSON format
- Generating semantically meaningful WOD vectors using fine-tuned Sentence-BERT
- Clustering WODs to reveal box-level programming tendencies and movement biases

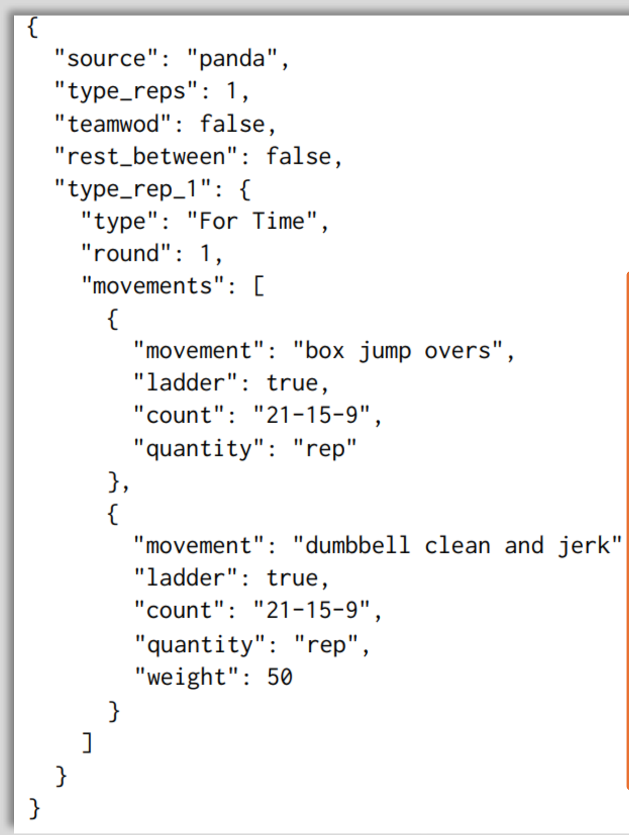
Methodology + Experiment

1. Data Collection

- **887 WODs** crawled from six CrossFit box websites
- Accessed posts via **date-based URL generation** using Selenium WebDriver
- Applied **site-specific parsing rules** with BeautifulSoup to extract only WOD content

2. Text Normalization

- Raw WOD texts were converted into a **structured JSON format** using **GPT-4 API**



```
{
  "source": "panda",
  "type_reps": 1,
  "teamwod": false,
  "rest_between": false,
  "type_rep_1": {
    "type": "For Time",
    "round": 1,
    "movements": [
      {
        "movement": "box jump overs",
        "ladder": true,
        "count": "21-15-9",
        "quantity": "rep"
      },
      {
        "movement": "dumbbell clean and jerk"
        "ladder": true,
        "count": "21-15-9",
        "quantity": "rep",
        "weight": 50
      }
    ]
  }
}
```

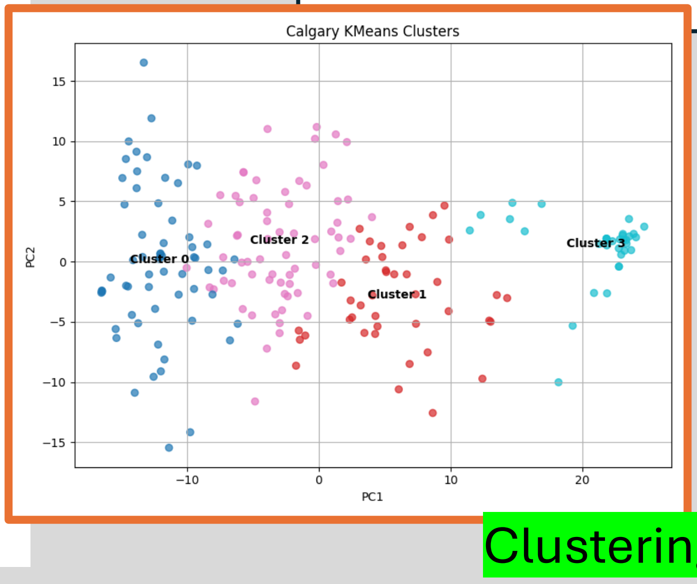
3. Vectorization

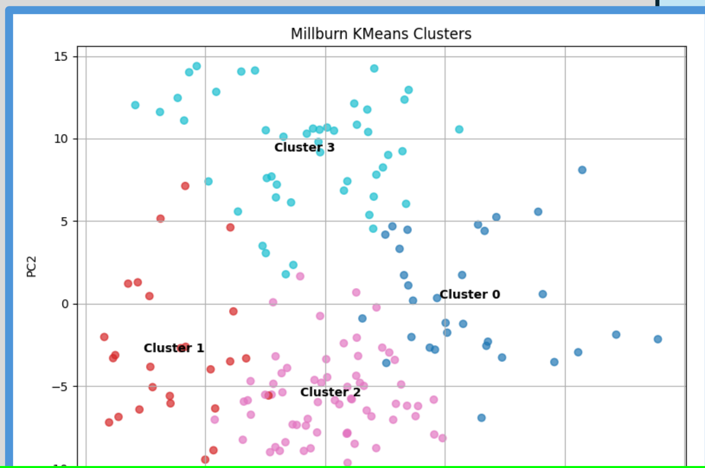
- Fine-tuned **Sentence-BERT** with triplet loss to embed movement names
- Built 397D WOD vectors using:
  - **Weighted average of embeddings** (log-scaled by reps)
  - **Structured features**: workout type, rounds, rest, etc.
  - **Weight scaling** for heavy equipment

➤ Each vector reflects both **movement meaning** and **program structure**

4. Clustering

- Applied **KMeans++** clustering to 397D WOD vectors
- Used **PCA** for dimensionality reduction + visualization
- Fixed **k = 4** across all boxes for interpretability

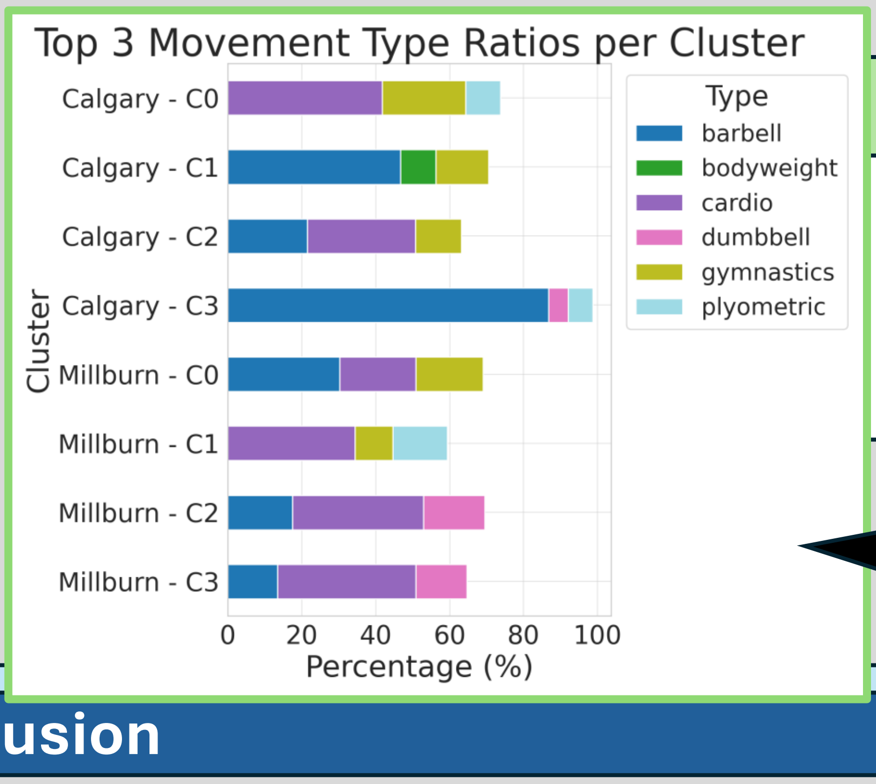




Clustering Results for Calgary and Millbu

5. Cluster Analysis

- For each cluster, analyzed:
  - **Dominant movement type**
  - **Top 15 movements**
  - **Average barbell weight**
  - **Movement type distribution**



Conclusion

- Proposed a pipeline that converts raw CrossFit WOD texts into structured vectors
- Fine-tuned embedding + structured features enabled meaningful WOD clustering
- Box-level analysis revealed distinct programming styles (e.g., strength vs. cardio bias)
- This method supports scalable comparison of unstandardized workouts

Future Work

- Add features for pacing, progression, and scaling options
- Develop WOD similarity search or recommendation systems