Box to Box: Analyzing CrossFit WOD Variability with Structured Embeddings

Sehyun Yun (20231233)
UNIST
South Korea
nawhji@unist.ac.kr

1 Introduction

1.1 Crossfit and WODs

1.1.1 About Crossfit. CrossFit is a high-intensity fitness program that combines various types of exercise and was developed by American coach Greg Glassman. Since its establishment in 2000, CrossFit has grown significantly and now includes more than 10,000 officially registered CrossFit gyms (referred to as "boxes") around the world. In Korea, more than 300 affiliated gyms are currently registered and the number continues to increase annually.

CrossFit training is centered on a daily workout known as the 'Workout of the Day (WOD),' which typically includes a combination of strength, gymnastics, and cardiovascular exercises. The movements used in WOD are often based on familiar functional actions, such as deadlifts, jumps, push-ups, rowing, and pull-ups.

These movements are performed either under time constraints (e.g., AMRAP, EMOM) or for speed (e.g., For Time), with minimal or no rest between efforts. This format effectively improves overall physical fitness, including strength, endurance, flexibility, power, agility, and balance.

1.1.2 Drivers of CrossFit's Widespread Adoption. One reason for the rapid growth of CrossFit is its effectiveness in producing noticeable physical and mental benefits. Even a single high-intensity WOD can acutely stimulate the secretion of hormones such as adrenaline, noradrenaline, and growth hormone (GH), leading to increased mental focus and alertness [6]. In the long term, consistent participation has been shown to improve body composition (i.e., reduced body fat and increased lean mass) [9], as well as muscular endurance and power [2]. Repeated training also promotes faster muscle recovery and enhances immune response. Moreover, the high levels of enjoyment and engagement reported by participants contribute to sustained adherence over time [5].

1.2 Challenges in WOD analysis

1.2.1 Box-Level Differences in WOD Design. However, the training effects of CrossFit can vary depending on how each box or platform structures its WODs. Since CrossFit programming is typically left to the discretion of individual coaches, there is often a stylistic bias in the types of workouts favored by each box [4]. Previous studies have also pointed out that there is little standardization in WOD design, with each site or study using different training structures [7].

1.2.2 Expression Variability in WODs. In this context, WODs exhibit substantial variability not only in structure but also in textual representation. The same WOD can be described in multiple ways depending on the gym or platform (see Table **??**). For example, one version of an interval workout might use the format "Every 2 Min for 8 Rounds," while another describes it as "8 Rounds – 2:00 on / 1:00 off." These variable and inconsistent representations, coupled with stylistic programming biases, present significant challenges for systematic analysis of WOD data.

Source A, B (Original)	**(Rephrased)**
Every 2 Min for 8 Rounds: 1 Rope Climb 3 Bar Facing Burpees 5 Power Snatch 3 Bar Facing Burpees Max Bike Calories – Rest 1 min –	*8 Rounds - 2:00 on / 1:00 off* 1 Rope Climb 3 Bar Facing Burpees 5 Power Snatch 3 Bar Facing Burpees Max Bike Calories
For Time 2 Rounds 200 M DB/KB Farmers Carry (50/35)(24K/16K) 9 Power Cleans (115/85) 9 Pull Ups -into- 2 Rounds 200 M DB/KB Farmers Carry (50/35)(24K/16K) 7 Power Cleans (135/95) 7 Pull Ups -into- 2 Rounds 200 M DB/KB Farmers Carry (50/35)(24K/16K) 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

1.2.3 Prior Work on Short Text Clustering. While there is limited prior research on CrossFit-specific WOD analysis, related studies have explored the clustering of short, variable texts in other health-related domains. For example, Afrimi (2023) applied Sentence-BERT and other embedding methods to the 20 Newsgroups dataset. Similar techniques, such as TF-IDF, Word2Vec, and KMeans, have also been used to cluster brief clinical and wellness-related texts [8]. These approaches demonstrate the feasibility of semantically organizing diverse short expressions, motivating their application to WOD analysis.

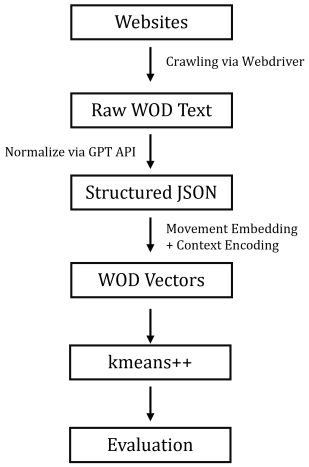


Figure 1: WOD Analysis Pipeline

1.3 Approach Overview

Building on prior work in short-text clustering, this study applies similar techniques to the domain of CrossFit workouts. The primary objective is to normalize unstructured WOD texts into a structured JSON format and derive semantic vector representations that reflect their content. Movement descriptions are embedded using a fine-tuned BERT model, while additional contextual information is encoded through one-hot or domain-specific encodings.

The resulting WOD vectors are clustered using the KMeans++ algorithm to identify characteristic patterns in programming across different boxes. The overall architecture is illustrated in Figure 1.

Clustering quality is assessed using custom metrics, including movement type distributions and average barbell weights. This framework supports both qualitative and quantitative comparisons of workout characteristics. The outcome provides a foundation for downstream applications such as WOD recommendation systems, automated classification, and comparative analysis of gym-level programming styles. To facilitate reproducibility and further exploration, the full codebase is available at public repository[1].

2 Related Work

Data-driven research related to CrossFit has primarily focused on performance prediction and statistical analysis using CrossFit Open scores. For example, *Normative Scores for CrossFit Open Workouts (2011–2022)* analyzes a decade of Crossfit Open data to establish benchmark scores categorized by sex, age, and division. Similarly, *Exploring CrossFit Performance Prediction and Analysis via Extensive Data and Machine Learning* proposes machine learning models that predict athlete performance based on demographic information and past scores. These studies rely on pre-structured numerical data such as competition scores and do not address the composition or semantics of the workout programs themselves.

In the general field of text clustering, BERT-based representations have been shown to offer improved semantic expressiveness compared to traditional approaches. For instance, *The performance of BERT as data representation of text clustering* evaluates BERT embeddings against TF-IDF and Word2Vec using clustering algorithms such as KMeans, DEC, and IDEC. The results suggest that BERT captures semantic similarity between texts more effectively, often yielding more coherent clusters. This finding indicates that BERT-based methods may also be suitable for representing workout descriptions, where contextual and compositional meaning plays a significant role.

There have also been a few efforts to analyze exercise-related text directly. *BERT-Based Semantic Similarity for the Clustering of Exercise Descriptions* explores the focuses on clustering short exercise descriptions, primarily collected from rehabilitation and physical therapy domains, using BERT embeddings to compute semantic similarity. This study demonstrates the potential of applying BERT to movement-level text, showing that demonstrating that semantically similar exercises can be effectively grouped based on their textual representations. In this study, BERT is also applied at the individual movement level, but the embeddings are subsequently combined to represent an entire workout program, which differs structurally from single-sentence clustering.

Collectively, prior studies have addressed performance analysis based on structured CrossFit scores, semantic clustering of single movement descriptions, or general text clustering techniques.

3 Problem Statement

WOD texts do not follow conventional sentence structures. Instead, they consist of fragmented commands and recurring format patterns. For example, expressions like “3 Rounds for Time” or “21-15-9 of Thrusters and Pull-ups” are not grammatically complete sentences, and their semantics depend more on structural conventions and domain-specific knowledge than on linguistic context.

Due to these characteristics, feeding entire WOD texts directly into sentence-based embedding models such as BERT can result in semantic distortion. BERT is optimized for interpreting meaning based on smooth sentence flow and syntactic cues, which are largely absent in WODs.

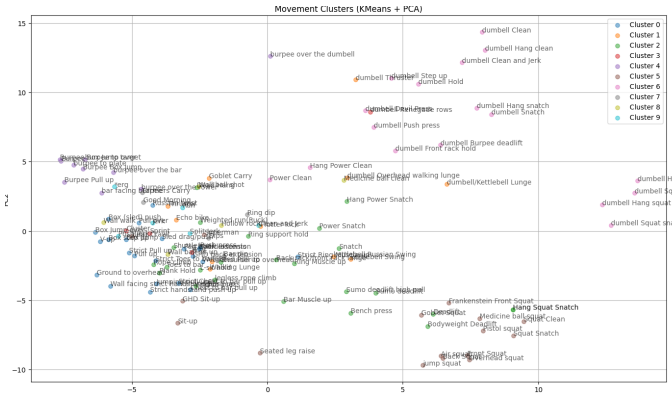


Figure 2: Movement Clusters with untuned model

Moreover, even when only movement names are used, domain-insensitive behavior is observed. As illustrated in Figure 2, movements such as *push-up* and *push press*, or *pull-up* and *V-up*, are often embedded into similar vector spaces simply due to the presence

[1] https://github.com/nawhji/crossfit_wod_analyzing

of shared tokens like "push" or "up"—despite being fundamentally different exercises in function and muscle engagement.

This problem stems from the fact that general-purpose language models like BERT lack explicit awareness of domain-specific semantics in fitness contexts. Terms that appear lexically similar may encode entirely different types of physical movement or intensity. Thus, embeddings generated without contextualization or disambiguation can compromise downstream applications such as clustering or classification.

Therefore, the core problem addressed in this study is as follows:

- Identify which components of WOD texts are essential for semantic preservation and which can be omitted without significant information loss.
- Address the limitations of general-purpose language models like BERT when applied to domain-specific terminology and workout descriptions.
- Abstract structural features such as repetition, load, round count, and timing into a separate, well-defined feature vector instead of embedding them as raw text.
- Ultimately, design a vector representation that balances semantic fidelity with computational efficiency.

4 Algorithm

4.1 Data Collection

WOD texts were collected from five affiliated CrossFit box websites and the official CrossFit homepage using automated web crawling. Posts were accessed using date-based URLs, and main workout contents were extracted with site-specific rules.

4.2 Data Preprocessing

The unstructured WOD texts were converted into structured JSON format using the GPT-4. A carefully crafted system prompt enforced a strict schema reflecting workout types, repetition patterns, weight values, time caps, and rest logic. Outputs were stored as individual `.json` files.

Example:

```
{
  "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
      }
    ]
  }
}
```

4.3 Movement Embedding and WOD Vectorization

To reflect domain-specific semantics, a Sentence-BERT model (`all-MiniLM-L6-v2`) was fine-tuned using triplet loss. Functional groups (e.g., *Core*, *Lifting*, *Gymnastics*) guided anchor–positive sampling, and manually constructed hard triplets helped distinguish semantically similar but functionally different movements.

Each WOD vector was constructed by combining:

- **Movement embedding:** Weighted average of movement vectors using log-scaled repetition counts.
- **Structural features:** Workout types, number of blocks, time caps, movement count, and progressive loading.
- **Rest encoding:** Categorical variable indicating presence of explicit or embedded rest.
- **Weight scaling:** Scalar boosting based on heavy barbell (135+ lbs), dumbbell (50+ lbs), or sandbag usage.

The final vectors were saved in both `.npy` and `.csv` formats for downstream clustering and analysis.

4.4 Clustering and Visualization

WOD vectors were standardized and clustered using the KMeans++ algorithm with $k = 4$, which was chosen based on interpretability and visual separability. PCA was applied for dimensionality reduction and 2D visualization.

4.5 Cluster Analysis

Each cluster was analyzed based on its movement composition and type distribution. Movements were classified into predefined types based on keywords, including equipment-based (*barbell*, *dumbbell*, *carry*, *odd object*) and functional categories (*cardio*, *plyometric*, *gymnastics*, *bodyweight*, *core*, *full_body*).

For each cluster, the top 15 most frequent movements, proportions of movement types, and the average barbell weight were computed. Summary statistics across all clusters were also recorded.

5 Experiments

5.1 Experimental Setup

A total of 887 WOD texts were collected and normalized from six CrossFit-related websites[2]. Data were collected using an automated crawler implemented with Selenium, and key workout content was extracted through BeautifulSoup and parsing.

Each WOD was converted into a structured JSON format using GPT-4 prompt. Movement names were embedded using the fine-tuned ‘sentence-transformers/all-MiniLM-L6-v2‘ model, yielding 384-dimensional vectors. Additional contextual information (e.g., workout type, number of rounds, presence of rest) was encoded

[2]https://www.crossfit.com/, https://crossfitcalgary.ca/, https://crossfitdfw.com/, https://crossfitmillburn.com, https://wods.crossfitpanda.com/, https://tamcrossfit.com/

using one-hot vectors and handcrafted domain features, resulting in a final input vector of 397 dimensions.

All features were standardized using StandardScaler. Principal Component Analysis (PCA) was applied to reduce dimensionality. Clustering was performed using the KMeans++ algorithm.

5.2 Number of Clusters

The number of clusters k was fixed to 4 across all experiments. Each box contained only 100–200 WODs, and setting a higher k resulted in overly fragmented clusters. Although silhouette scores were computed for $k = 2$ to 11, the differences were minor and did not suggest a clearly optimal value. Based on this observation, $k = 4$ was selected for consistency and to facilitate interpretability in the subsequent analysis.

5.3 Cluster Analysis

Clustering was conducted independently for each box with $k = 4$, and the characteristics of each cluster were evaluated using the metrics described in Section 4.5. Table 1 summarizes the cluster-wise results for *Calgary* and *Millburn*, including the number of WODs, dominant movement type, top 3 movement type ratios, and average barbell weight. For detailed cluster compositions and raw JSON representations, please refer to the public repository[3].

Table 1: Summary of Cluster Characteristics

Box	Cluster	# of WODs	Dominant Type	Top 3 Type Ratio	Barbell Avg
calgary	0	62	cardio (70 times)	cardio (41.7%); gymnastics (22.6%); plyometric (9.5%)	173.00
calgary	1	42	barbell (49 times)	barbell (46.7%); gymnastics (14.3%); bodyweight (9.5%)	137.12
calgary	2	66	cardio (68 times)	cardio (29.2%); barbell (21.5%); gymnastics (12.4%)	110.42
calgary	3	33	barbell (65 times)	barbell (86.7%); plyometric (6.7%); dumbbell (5.3%)	136.43
millburn	0	31	barbell (35 times)	barbell (30.2%); cardio (20.7%); gymnastics (18.1%)	111.56
millburn	1	26	cardio (44 times)	cardio (34.4%); plyometric (14.8%); gymnastics (10.2%)	111.25
millburn	2	65	cardio (111 times)	cardio (35.4%); barbell (17.5%); dumbbell (16.6%)	119.91
millburn	3	47	cardio (95 times)	cardio (37.5%); dumbbell (13.8%); barbell (13.4%)	120.59
crossfit.com	0	51	barbell (49 times)	barbell (34.0%); gymnastics (18.1%); dumbbell (13.2%)	143.75
crossfit.com	1	46	cardio (42 times)	cardio (31.8%); gymnastics (18.9%); barbell (14.4%)	103.21
crossfit.com	2	21	barbell (36 times)	barbell (90.0%); dumbbell (5.0%); gymnastics (2.5%)	170.94
crossfit.com	3	26	cardio (47 times)	cardio (61.8%); gymnastics (13.2%); core (7.9%)	137.50
dfs	0	26	gymnastics (30 times)	gymnastics (25.9%); barbell (19.8%); cardio (17.2%)	96.00
dfs	1	7	barbell (15 times)	barbell (62.5%); cardio (16.7%); gymnastics (12.5%)	135.00
dfs	2	16	cardio (29 times)	cardio (48.3%); barbell (11.7%); full_body (10.0%)	N/A
dfs	3	20	barbell (41 times)	barbell (50.0%); gymnastics (13.4%); cardio (12.2%)	185.00
panda	0	37	barbell (54 times)	barbell (43.2%); dumbbell (24.0%); gymnastics (10.4%)	127.07
panda	1	38	cardio (42 times)	cardio (28.6%); dumbbell (15.0%); gymnastics (13.6%)	87.14
panda	2	49	barbell (42 times)	barbell (27.3%); gymnastics (16.2%); cardio (14.9%)	108.38
panda	3	38	dumbbell (51 times)	dumbbell (40.5%); cardio (17.5%); barbell (15.1%)	77.22
tam	0	40	barbell (53 times)	barbell (40.8%); cardio (29.2%); dumbbell (16.2%)	115.11
tam	1	43	barbell (35 times)	barbell (22.6%); gymnastics (21.3%); cardio (18.1%)	110.30
tam	2	35	cardio (53 times)	cardio (50.0%); gymnastics (16.0%); full_body (11.3%)	81.25
tam	3	22	cardio (34 times)	cardio (51.5%); gymnastics (9.1%); barbell (7.6%)	61.67

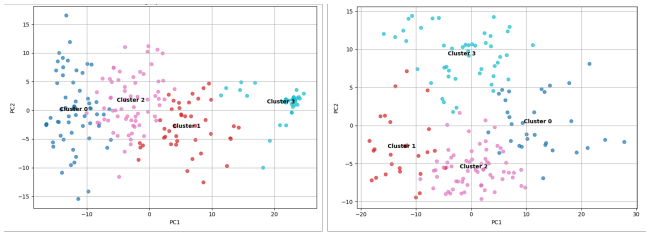
5.4 Cluster Interpretation: `calgary` and `millburn`

`calgary`. Cluster 0 is dominated by cardio, followed by gymnastics and plyometric movements. This cluster mostly consists of aerobic and indoor WODs, such as workouts combining machines, double unders, and weighted lunges, or sprint-based conditioning. Cluster 1 is barbell-dominant with a relatively heavy average weight. It also includes some gymnastics and bodyweight elements. Representative WODs include clean + toes-to-bar and overhead squat + ring dip combinations. Cluster 2 contains a balanced mix of cardio, barbell, and gymnastics movements. This reflects a typical metcon style with lower barbell loading. Examples include push press + box jump + sumo deadlift high pull + wall ball, or clean and jerk + rowing WODs. Cluster 3 features both a high average barbell weight and a barbell movement ratio over 80%, clearly indicating a strength-focused group. WODs such as Heavy Grace (30 clean and jerks) and Triple Isabel (90 snatches) were frequently observed.

`millburn`. Cluster 0 is centered around barbell and cardio movements, with moderate weights. It follows a metcon style, exemplified by rowing + toes-to-bar + thrusters or shoulder-to-overhead + double unders. Cluster 1 is strongly aerobic, dominated by cardio, plyometric, and gymnastics movements. WODs like box jump overs + sit-ups + machine work, or wall-ball based workouts were common. Cluster 2 shows a mix of cardio, barbell, and dumbbell components—again reflecting a metcon preference. Examples include rowing + weighted squats + toes-to-bar, and machine + hang power clean + rope climb workouts. Cluster 3 is similarly composed of cardio, barbell, and dumbbell elements. WODs like box jump over + cluster + running, or burpee-to-target + deadlift were observed.

Overall, `calgary` tends to favor strength-oriented programming, as shown by its multiple barbell-dominant clusters and the frequent presence of heavy lifting WODs. In contrast, `millburn` primarily emphasizes cardio-based or metcon-style workouts. While barbell and dumbbell elements appear across all clusters, they are generally used in supportive roles, and high-load strength-focused WODs were rarely found. Figure 3 visualizes the cluster distributions of the `calgary` and `millburn` boxes in 2D PCA space, which correspond to the detailed interpretations discussed above.

Although this section focuses on the detailed interpretation of `calgary` and `millburn`, cluster-level summaries for other boxes are provided in Table 1 for completeness.



(a)Calgary Clusters (b)Millburn Clusters

Figure 3: `calgary` and `millburn` clusters

6 Conclusion

This study introduced a pipeline to convert unstructured CrossFit WOD texts into structured JSON representations and semantically interpretable vectors. By clustering workouts for each affiliate, distinct programming characteristics were revealed, such as tendencies toward cardio-intensive or strength-focused training styles.

The analysis primarily focused on movement-based features. Future extensions may include finer-grained attributes—such as pacing, progression, and scaling—in order to enhance the expressiveness and applicability of WOD representations.

[3]https://github.com/nawhji/crossfit_wod_analyzing

References

[1] Daniel Afrimi. Text clustering using nlp techniques. https://medium.com/@danielafrimi/text-clustering-using-nlp-techniques-c2e6b08b6e95, 2023. Accessed: 2025-05-27.

[2] David Bellar, Andrew Hatchett, Lawrence W Judge, ME Breaux, and Lena Marcus. The relationship of aerobic capacity, anaerobic peak power and experience to performance in in crossfit exercise. *Biology of sport*, 32(4):315–320, 2015.

[3] CrossFit, LLC. Affiliate map. https://www.crossfit.com/map, 2025. Accessed: 2025-05-27.

[4] Greg Glassman. Understanding crossfit. https://library.crossfit.com/free/pdf/56-07_Understanding_CF.pdf, 2007. Accessed: 2025-05-27.

[5] Katie M Heinrich, Pratik M Patel, Joshua L O'Neal, and Bryan S Heinrich. High-intensity compared to moderate-intensity training for exercise initiation, enjoyment, adherence, and intentions: an intervention study. *BMC public health*, 14:1–6, 2014.

[6] Nacipe Jacob, Jefferson S Novaes, David G Behm, João G Vieira, Marcelo R Dias, and Jeferson M Vianna. Characterization of hormonal, metabolic, and inflammatory responses in crossfit® training: A systematic review. *Frontiers in physiology*, 11:1001, 2020.

[7] Jena Meyer, Janet Morrison, and Julie Zuniga. The benefits and risks of crossfit: a systematic review. *Workplace health & safety*, 65(12):612–618, 2017.

[8] Enas Saad, Tamer Elsayed, and Walid Magdy. Evaluation of clustering and topic modeling methods over health-related tweets and emails. *Artificial Intelligence in Medicine*, 127:102179, 2022.

[9] Michael M Smith, Allan J Sommer, Brooke E Starkoff, and Steven T Devor. Crossfit-based high-intensity power training improves maximal aerobic fitness and body composition [retracted]. *The Journal of Strength & Conditioning Research*, 27(11):3159–3172, 2013.