Unsupervised Wisdom: Mining Narrative For Insights

Mining the narrative text of medical records that describes the precipitating event of old age falls for useful insights is the central goal of this challenge. We leverage the embeddings of the primary narratives (OpenAI's text-embedding-ada-002) provided by the host. With it, we create a vector store using FAISS(Facebook Similarity Search), then perform a 'range search' for queries of interest. The narrative field of the query results is then cleaned removing attributes already tabulated in the given dataset for a concise description of the event before a fall. The narratives are structured according to a sequence of events as per coding manual guidelines. We exploit this fact along with the two highly relevant events; 'slipped' and 'tripped' that are explicitly and frequently used. As highlighted in our midpoint submission keyword search alone can lead to incorrect distribution statistics such as in the case of 'sleeping'.

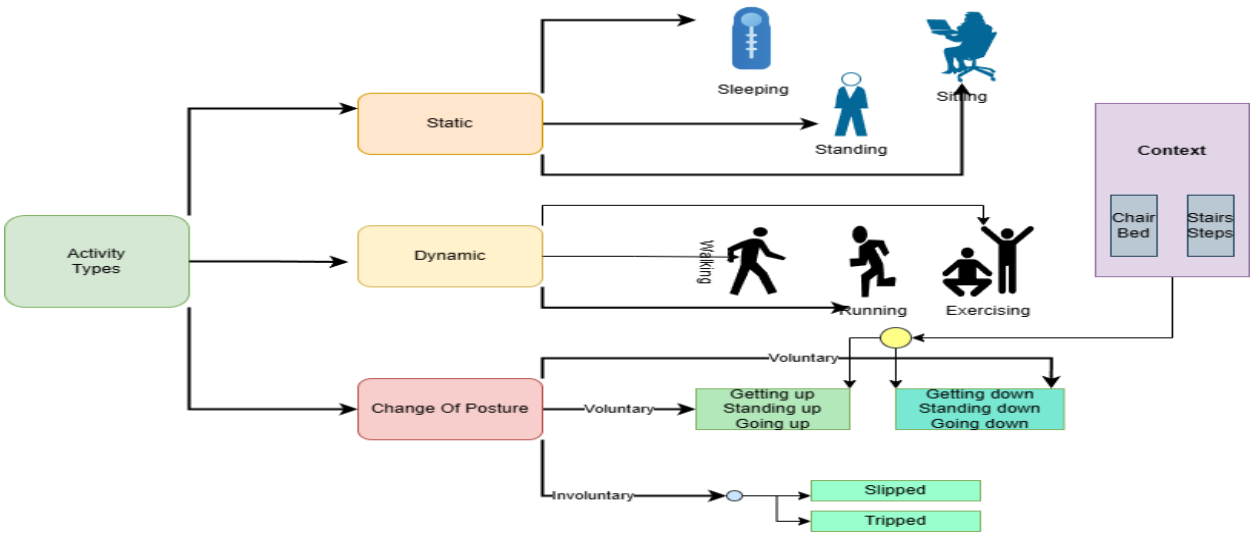

Figure 1, Categorizing activities broadly for query construction

A key element of our approach is ***query construction*** which should ideally be semantically rich and informative for semantic search but here we also need to understand how the limit-imposed narratives are written. Activities in general can be broadly categorized as shown in the figure above (Figure 1). An action alone describing a fall event such as 'getting up' is insufficient, a context in the form of the product involved is essential in our case. 'Getting up from a chair' is significantly different from 'Getting up a ladder' while still sharing 'Getting up' as a common word. We show through examples; query construction, searching and cleaning narratives, comparison with other queries (Figure 3), and appropriate cluster evaluation metrics such as 'silhouette score' and 'calinski_harabasz score', the viability and **validation** of our approach.

An ***alternate approach*** is to utilize OpenAI ChatGPT 3.5 through a prompt to extract the precipitating event for each narrative. This is time-consuming and rate-limited although the

returned description appears clean and concise. However, we do obtain ~5000 usable samples and compare the cosine similarity of the embeddings to our cleaned semantic search result embeddings, showing that our approach is just as effective. We also extract the 10 most used verbs describing a precipitating event, cluster them, and based on cross-validation and feature importance show this additional engineered feature to be relevant for predicting 'body_part' affected by a fall.

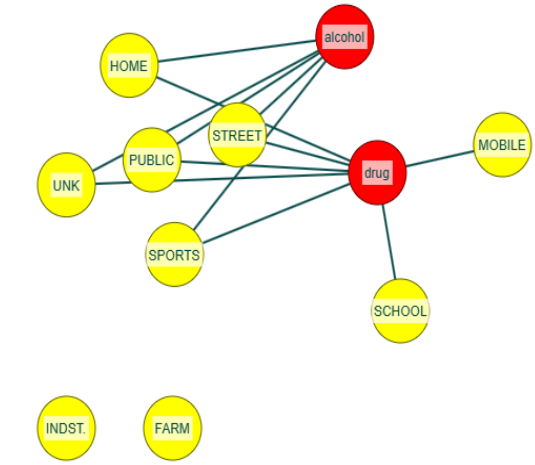

Figure 2 Drug and Alcohol in Locations

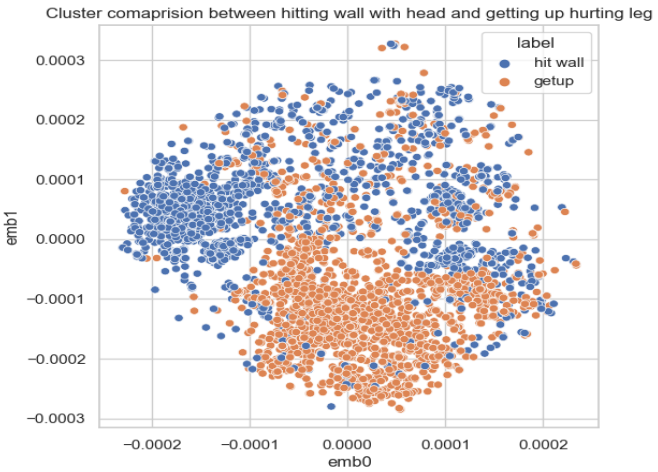

Figure 3 Demonstrating cluster cohesion

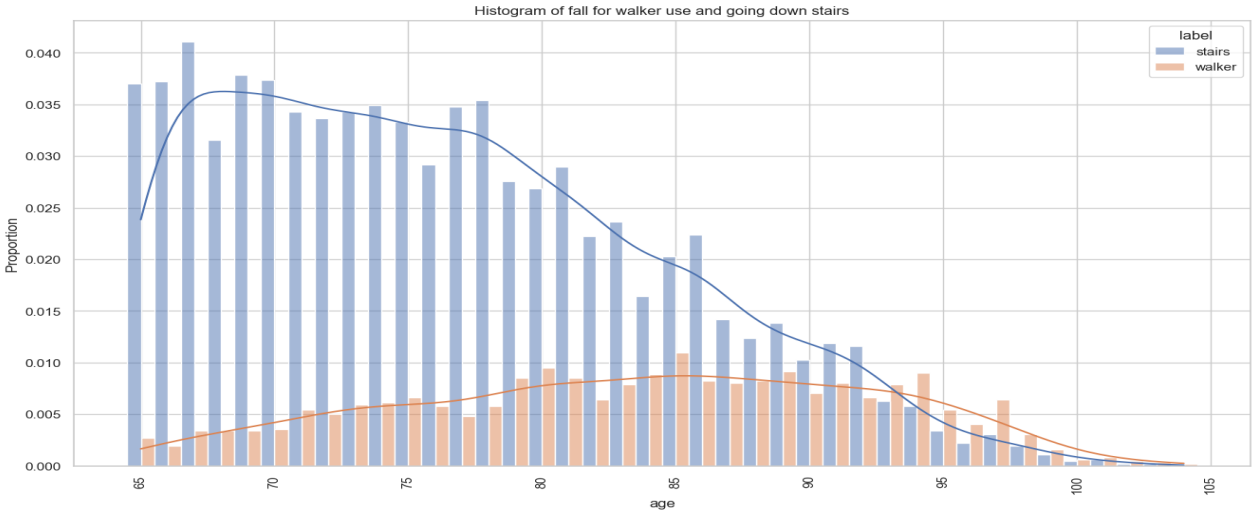

Figure 4, Falling from stairs compared to Fall when using walker

***Key* findings:**

1. Fall incidents while using a walker by age is gaussian which peaks at age 85, while falling down a stair drops as age progresses (long-tailed) as shown in Figure 4, and could be explained by the fact that as age progresses people will increase the use of walkers while they limit the use of stairs.

2. While both 'slipping' and 'tripping' injure the head most, 'tripping' causes injuries to the face twice as much, likely due to forward involuntary motion after tripping on an obstacle but without being explicitly mentioned in the narrative, we could not know which part (back or other) of the head, 'slip' causes injury. Also, comparatively more 'slip' events occur in winter months in the street while far fewer slips occur in rugs or carpets.

3. Through graph construction and inspection we found drugs (prescription medication perhaps) are involved in school, but not alcohol, and that neck injury occurs most in nerve damage (Figure 5).

4. Surprisingly falling from bed is the highest cause of fall incidents in a nursing home.

Limitation:

For a clean text of precipitating events, we rely on the structure of the narrative but this is not always followed as per coding guidelines along with spelling mistakes so we can have incorrect parsing. However clean text is for cluster evaluation and concise event description only and does not affect the conclusion of the underlying distribution.

When constructing a query, providing too much context in the form of product, body part, location, etc., is also not useful, as the small number of words in the narrative can cause many matches, compromising the specificity of the desired result.

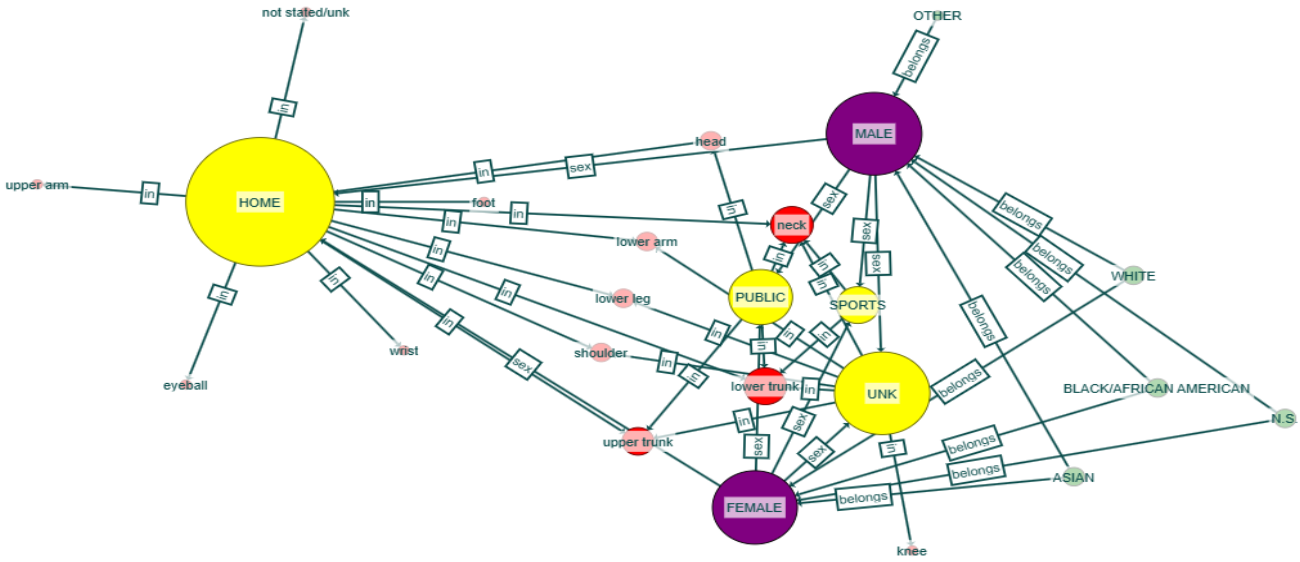

Figure 5, Knowledge Graph for a subset (nerve damage) only.