# Tab 1

# Quality of Embeddings & Triplet Loss


**Author: Atharv Katkar**
**Suggested by: Sandeep Giri, CloudxLab**

## OVERVIEW

In Natural Language Processing (NLP), embeddings transform human language into numerical vectors. These are usually arrays of multiple dimensions & have schematic meaning based on their previous training text corpus The quality of these embeddings directly affects the performance of search engines, recommendation systems, chatbots, and more.

But here's the problem:

Not all embeddings are created equal.

So how do we measure their quality?

**To Identify the quality of embeddings i conducted one experiment:**

I took 3 leading (Free) Text ➔ Embedding pretrained models which worked differently & provided a set of triplets and found the triplets loss to compare the contextual  importance of each one.

**1) Sentence-BERT (SBERT)**

Transformer-based Captures deep sentence-level semantics:

```
from sentence_transformers import SentenceTransformer

model = SentenceTransformer('all-MiniLM-L6-v2')
```

**2)Universal Sentence Encoder (USE)**

TensorFlow Encoder Good general-purpose semantic encoding:

```
import tensorflow_hub as hub

model2 = hub.load("https://tfhub.dev/google/universal-sentence-encoder/4")

embeddings = model2(["Cardiac arrest"])
```

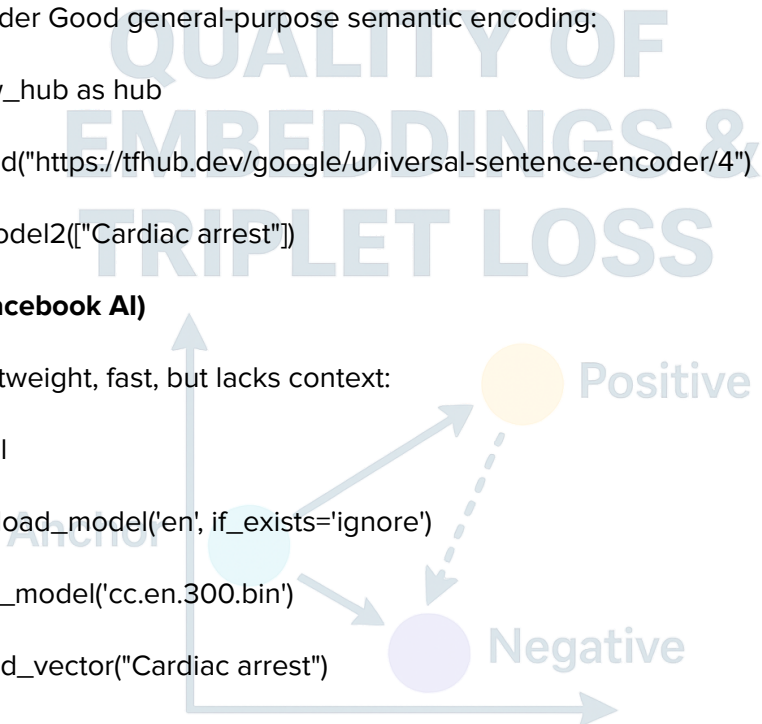
**3)FastText (by Facebook AI)**

Word-based Lightweight, fast, but lacks context:

```
import fasttext.util

fasttext.util.download_model('en', if_exists='ignore')

ft3 = fasttext.load_model('cc.en.300.bin')

vec = ft3.get_word_vector("Cardiac arrest")
```

when i compared the sizes of output produced by them are different for each one

**(384,), (1, 512), (300,)**




## GOALS

1. To compare them using a triplet-based evaluation approach using triplet loss.
2. Identify the Understanding of these around medical terminologies

## CONCEPTS

### What is Triplet Loss?

Triplet loss works with a 3-part input:

Anchor: The base sentence or phrase

Positive: A semantically similar phrase

Negative: A semantically absurd or unrelated phrase

| Anchor | Positive | Negative |
|---|---|---|
| tuberculosis | Lung infection | test tube accident |
| cardiac arrest | heart attack | cardi b attack |
| asthma | respiratory condition | Spiritual awakening |

The goal is to push the anchor close to the positive and far from the negative in embedding space.

**TripletLoss = max (d (a , p) − d (a , n) + margin , 0 )**

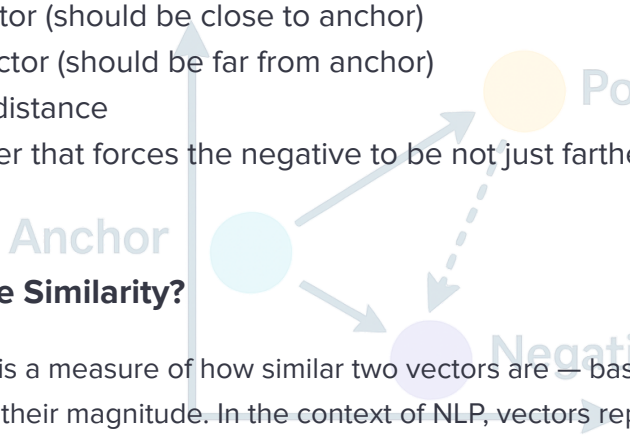
a = anchor vector

**p** = positive vector (should be close to anchor)

**n** = negative vector (should be far from anchor)

**d(x,y)** = cosine distance

**margin** = a buffer that forces the negative to be not just farther, but significantly farther




## What is Cosine Similarity?

Cosine similarity is a measure of how similar two vectors are — based on the angle between them rather than their magnitude. In the context of NLP, vectors represent words or sentences as embeddings.

$$\text{Cosine Similarity (A.B)} = \frac{A.B}{\|A\|.\|B\|}$$

**CosineDistance(A,B) = 1 − CosineSimilarity(A,B)**

## What is Margin?

The margin is a safety cushion.
If margin = 0.2, then even if the negative is slightly farther than the positive, the model still gets a penalty unless it's at least 0.2 farther.

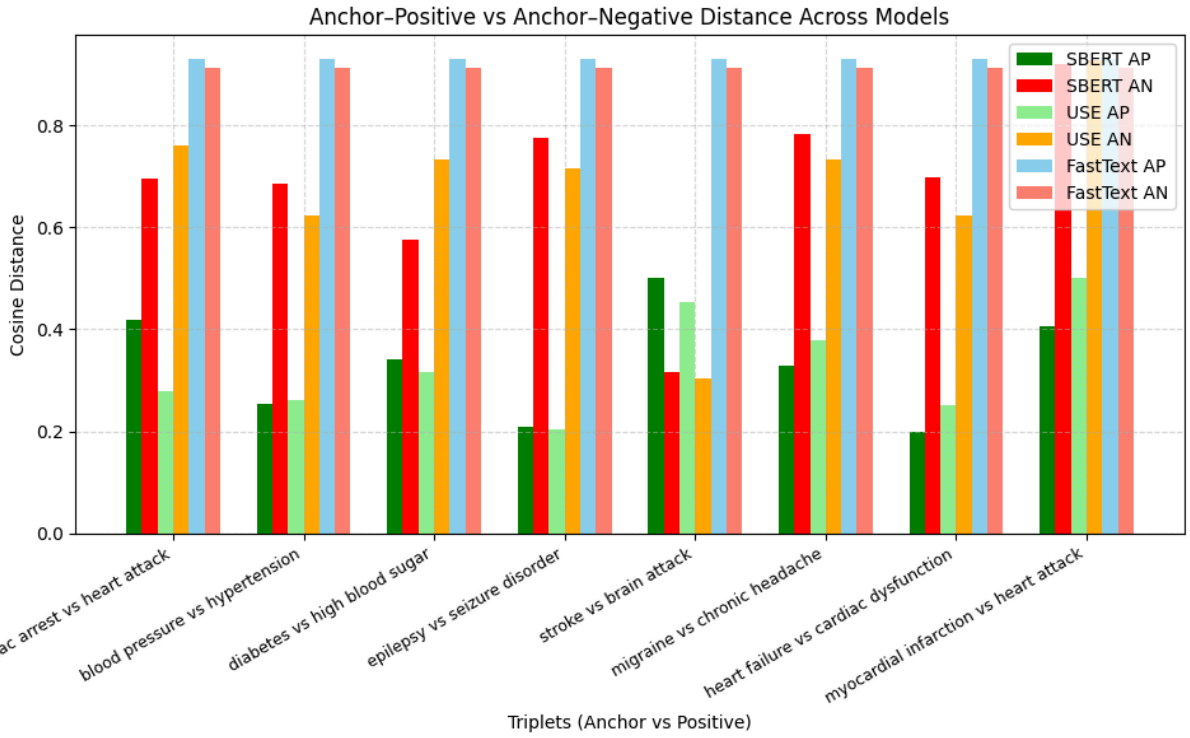
## Testing The accuracy

**TEST-SET**(click on test set see set)
We ran each model over a set of ~50 curated triplets
Calculated:
Anchor–Positive distance (AP)
Anchor–Negative distance (AN)
Triplet loss
Visualized both individual performance per triplet and overall averages

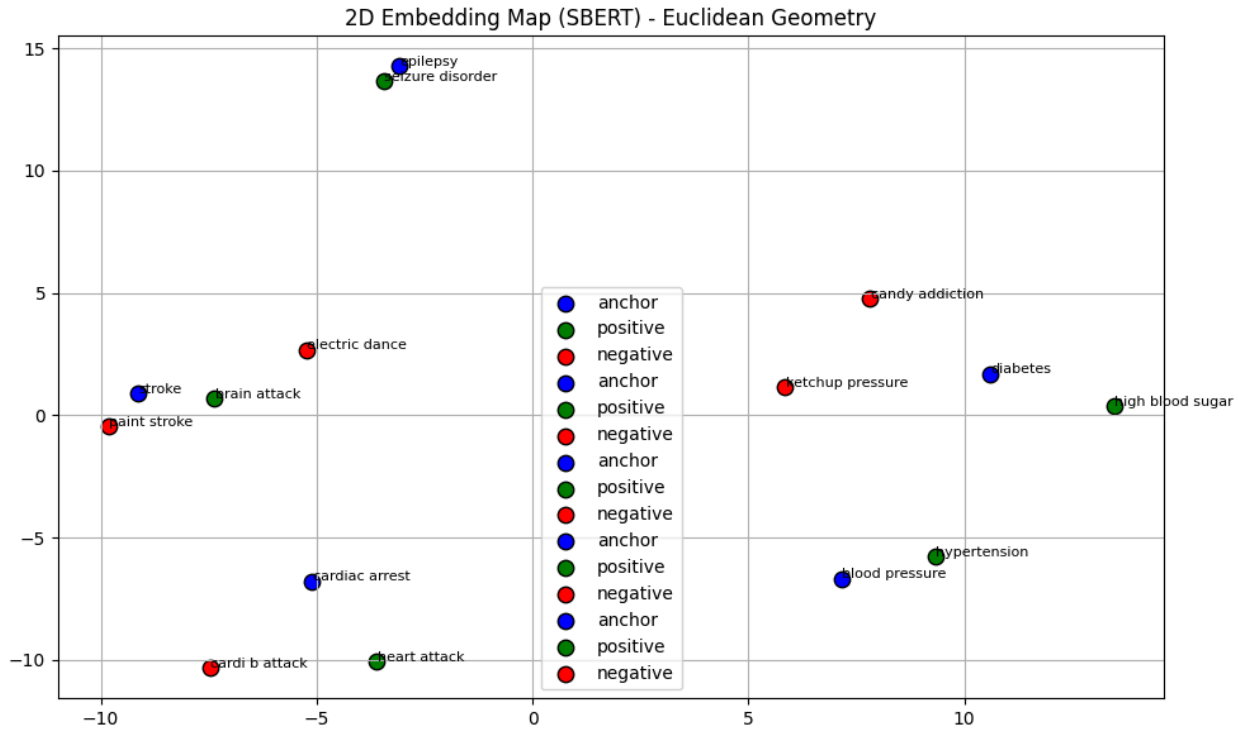
("asthma", "respiratory condition", "spiritual awakening"),
("pneumonia", "lung infection", "foggy window"),
("COPD", "chronic lung disease", "cop duty"),

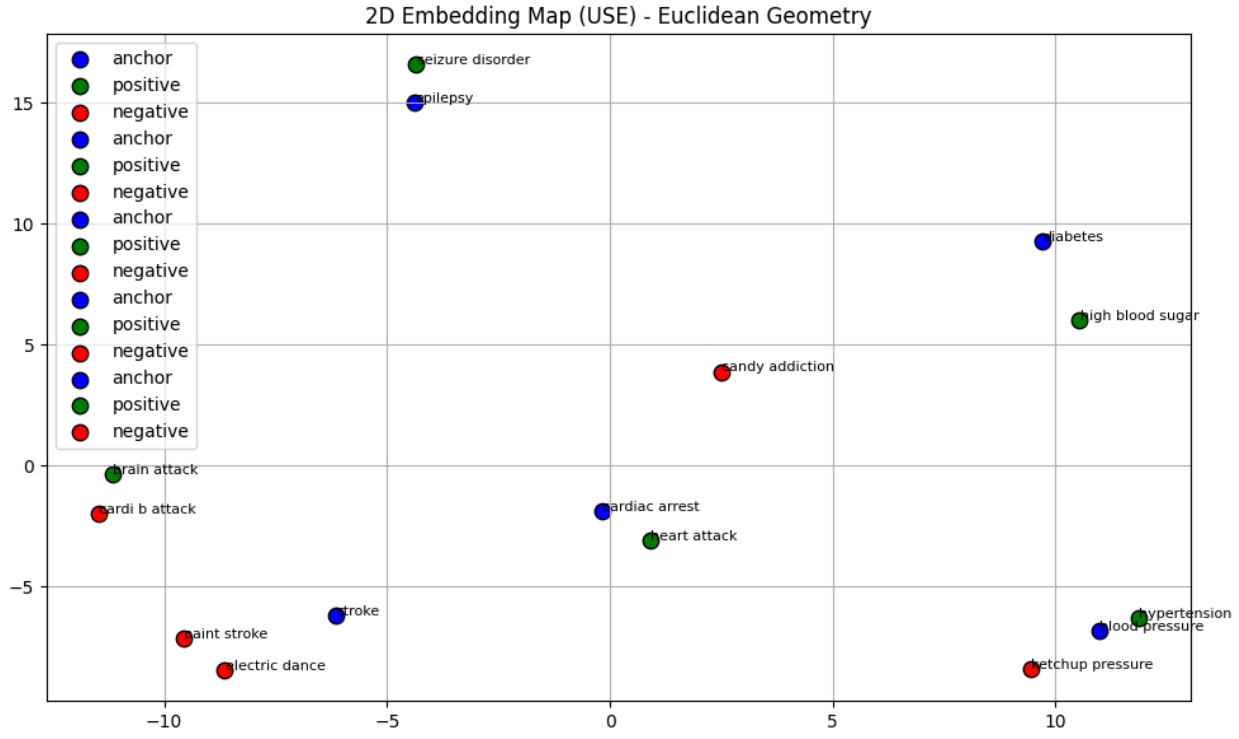
# General & Internal Medicine
("diabetes", "high blood sugar", "candy addiction"),
("anemia", "low red blood cells", "color fade"),
("arthritis", "joint inflammation", "rusty hinge"),
50+ such examples

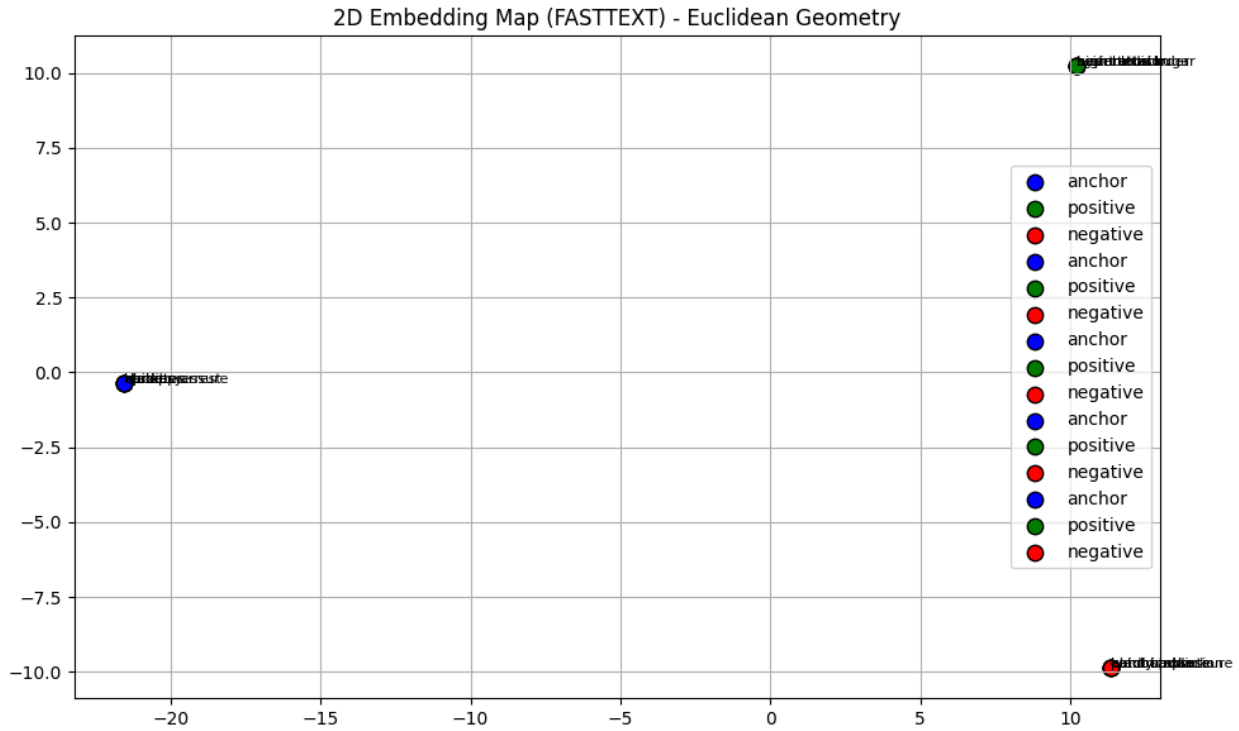




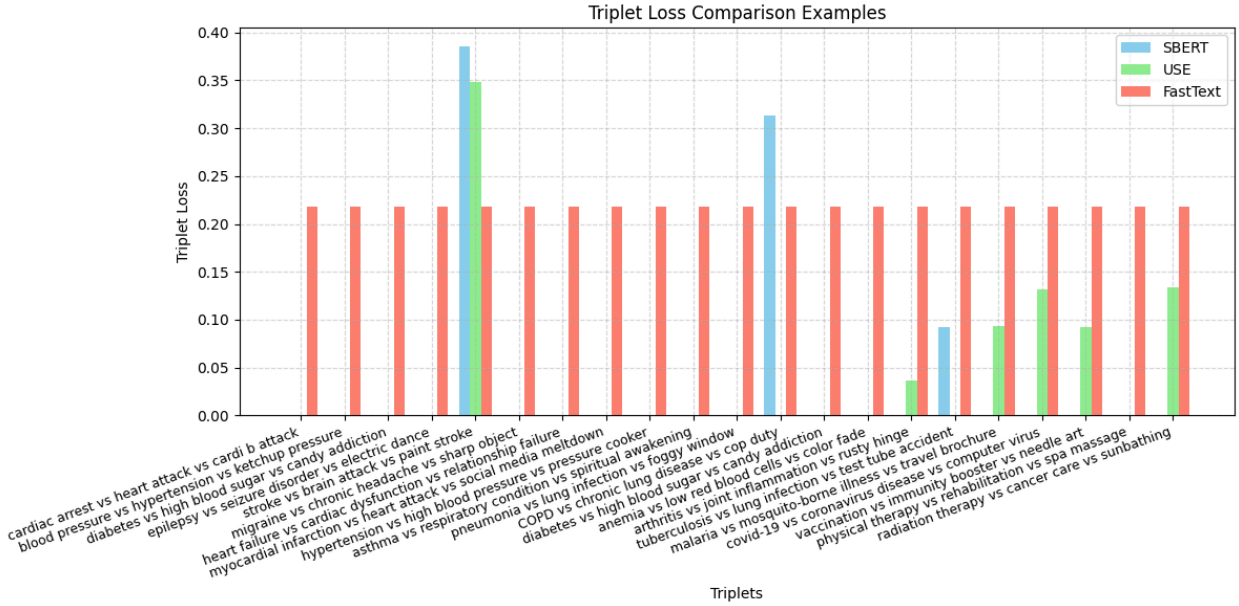

Using PCA, we visualized where each model placed the anchor, positive, and negative in space.
Insert 2D scatter plot
You can actually see the anchor and positive clustering together especially in SBERT's case while the negative floats far away.


2D Embedding Map (SBERT) - Euclidean Geometry
epilepsy
seizure disorder
candy addiction
electric dance
diabetes
ketchup pressure
stroke
brain attack
high blood sugar
paint stroke
hypertension
cardiac arrest
blood pressure
cardi b attack
heart attack
anchor
positive
negative


2D Embedding Map (USE) - Euclidean Geometry
seizure disorder
epilepsy
diabetes
high blood sugar
candy addiction
brain attack
cardi b attack
cardiac arrest
heart attack
stroke
hypertension
blood pressure
paint stroke
electric dance
ketchup pressure
anchor
positive
negative




Triplet Example wise Score




The SBERT & USE performed good as we can see using few interpretations and loss tracking of triplets

## CONCLUSION

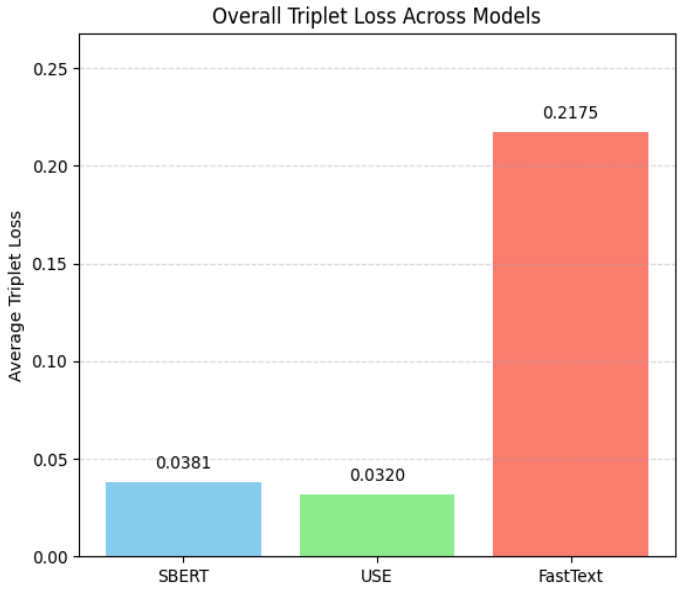
What We Learned

SBERT is highly reliable for understanding sentence-level meaning

USE performs reasonably well and is easy to use with TensorFlow

FastText, while fast, struggles with context and full sentences

**Visual Results**
**Triplet Loss (Lower = Better)**
**SBERT : 0.0381**
**USE : 0.0320**
**FastText : 0.2175**




If you're building:
Search engines
Recommendation systems
Chatbots ...or anything involving meaning, good embeddings are key.
Triplet loss is a simple yet powerful way to test how smart your model really is.

# Tab 2

```
    # Cardiovascular
    ("myocardial infarction", "heart attack", "stock crash"),
    ("angina pectoris", "chest pain", "paint bucket"),
    ("bradycardia", "slow heart rate", "slow motion"),
    ("tachycardia", "fast heartbeat", "speed limit"),
    ("arrhythmia", "irregular heartbeat", "funky music"),
    ("atherosclerosis", "artery blockage", "oil spill"),
    ("pericarditis", "inflammation of heart lining", "birthday card"),

    # Neurology
    ("epilepsy", "seizure disorder", "disco lights"),
    ("ischemic stroke", "brain clot", "paint stroke"),
    ("transient ischemic attack", "mini-stroke", "traffic jam"),
    ("meningitis", "brain membrane inflammation", "membrane cleaner"),
    ("encephalopathy", "brain dysfunction", "encyclopedia"),
    ("amyotrophic lateral sclerosis", "motor neuron disease", "muscle
builder"),
    ("hydrocephalus", "fluid in brain", "water balloon"),

    # Endocrinology
    ("hypothyroidism", "underactive thyroid", "sleepy panda"),
    ("hyperthyroidism", "overactive thyroid", "hyper student"),
    ("diabetes mellitus", "high blood sugar", "sugar craving"),
    ("acromegaly", "excess growth hormone", "tall tale"),
    ("cushing's syndrome", "high cortisol", "cushion pile"),
```

```
# Gastroenterology
("gastroesophageal reflux disease", "acid reflux", "cola fountain"),
("ulcerative colitis", "colon inflammation", "sad stomach"),
("crohn's disease", "digestive tract inflammation", "crayon box"),
("hepatitis C", "liver infection", "emoji virus"),
("pancreatitis", "pancreas inflammation", "kitchen pan"),

# Pulmonology
("chronic obstructive pulmonary disease", "COPD", "cop badge"),
("emphysema", "damaged lung sacs", "air freshener"),
("pleurisy", "chest lining inflammation", "pleura curtains"),
("asbestosis", "lung fibrosis", "building dust"),
("pulmonary embolism", "lung blood clot", "traffic block"),

# Nephrology
("nephrolithiasis", "kidney stones", "gem collection"),
("glomerulonephritis", "kidney inflammation", "glow ring"),
("uremia", "toxic blood condition", "urinal splash"),
("hydronephrosis", "kidney swelling", "hydration bottle"),

# Hematology
("thrombocytopenia", "low platelets", "thin pizza"),
("leukemia", "blood cancer", "lemonade brand"),
("hemophilia", "bleeding disorder", "movie drama"),
```

```
    ("polycythemia vera", "high red blood cells", "polymer cup"),

    # Infectious Disease
    ("tuberculosis", "lung infection", "test tube mix"),
    ("diphtheria", "throat infection", "dictionary word"),
    ("leprosy", "nerve and skin disease", "leopard spots"),
    ("schistosomiasis", "parasitic infection", "school science"),
    ("ebola", "viral hemorrhagic fever", "game controller"),

    # Rheumatology
    ("systemic lupus erythematosus", "autoimmune disease", "sunburn face"),
    ("rheumatoid arthritis", "joint inflammation", "joint account"),
    ("scleroderma", "hard skin disease", "glue bottle"),

    # Oncology
    ("lymphoma", "cancer of lymph", "light bulb"),
    ("carcinoma", "cancerous tumor", "car cinema"),
    ("melanoma", "skin cancer", "crayon mark"),

    # Dermatology
    ("psoriasis", "scaly skin disorder", "printer jam"),
    ("eczema", "itchy skin rash", "emoji explosion"),
    ("urticaria", "hives", "hotel towel"),
```

```
    # Pediatrics
    ("neonatal jaundice", "yellowing in newborn", "banana blanket"),
    ("croup", "child airway inflammation", "crow noise"),
    ("measles", "viral rash illness", "connect the dots"),
]
```