

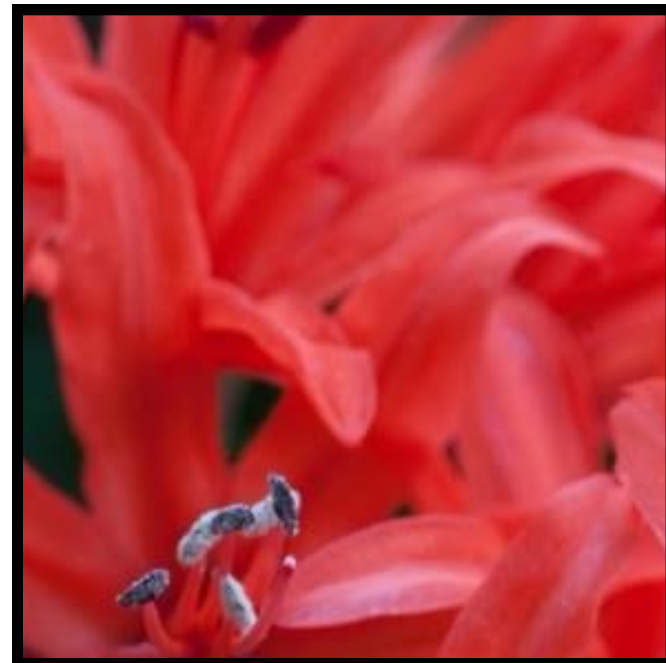
# Search for an image by its fragment

**Authors: Ilia Mistiurin, Nazgul Salikhova, Milyausha Shamsutdionova**

**Date: 17.07.2024**

# The Value Proposition

Matching image fragment to its original





# The Value Proposition

Finding highly similar images by image fragment





# The Value Proposition

Matching image fragment to its original or highly similar images

## Use-cases:

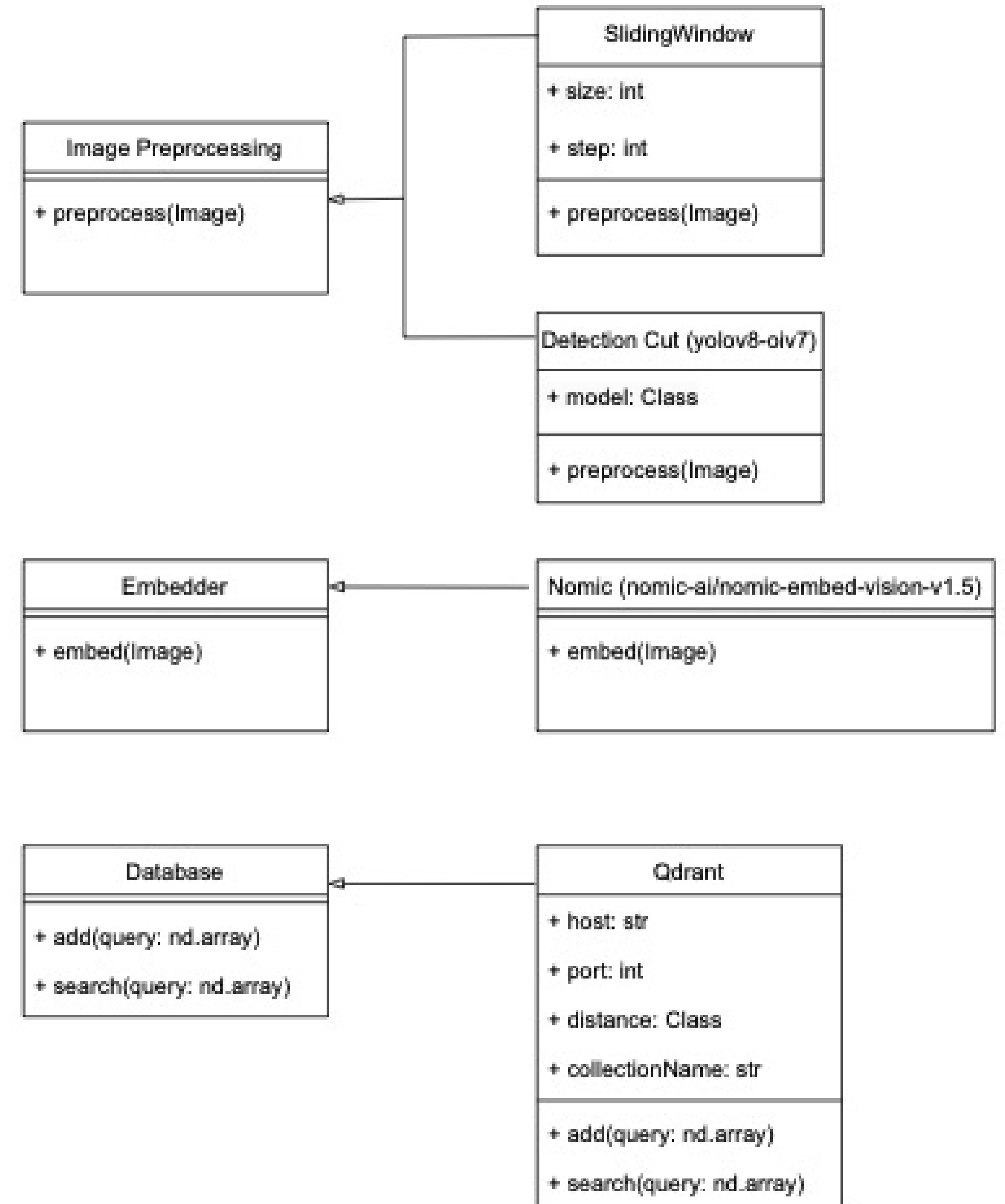
- Detection of plagiarism in images
- Product search in retail

# Architecture Design

1. Dataset preprocessing: Sliding window technique,  
Object detection model

2. Embedder

3. Database



## Stack and key technologies

- **Python 3.11**
- **Dataset: Open Images V7 (5000 pcs from train set → 59980 instances, class = “Flowers”)**
- **Object detection model: ultralytics/yolov8-oiv7**
- **Embedder: nomic-ai/nomic-embed-vision-v1.5**
- **Database: Qdrant (HNSW for indexing)**

## Evaluation dataset

**100 test images was taken from dataset Open Images V7 and fragments was created used the following methods:**

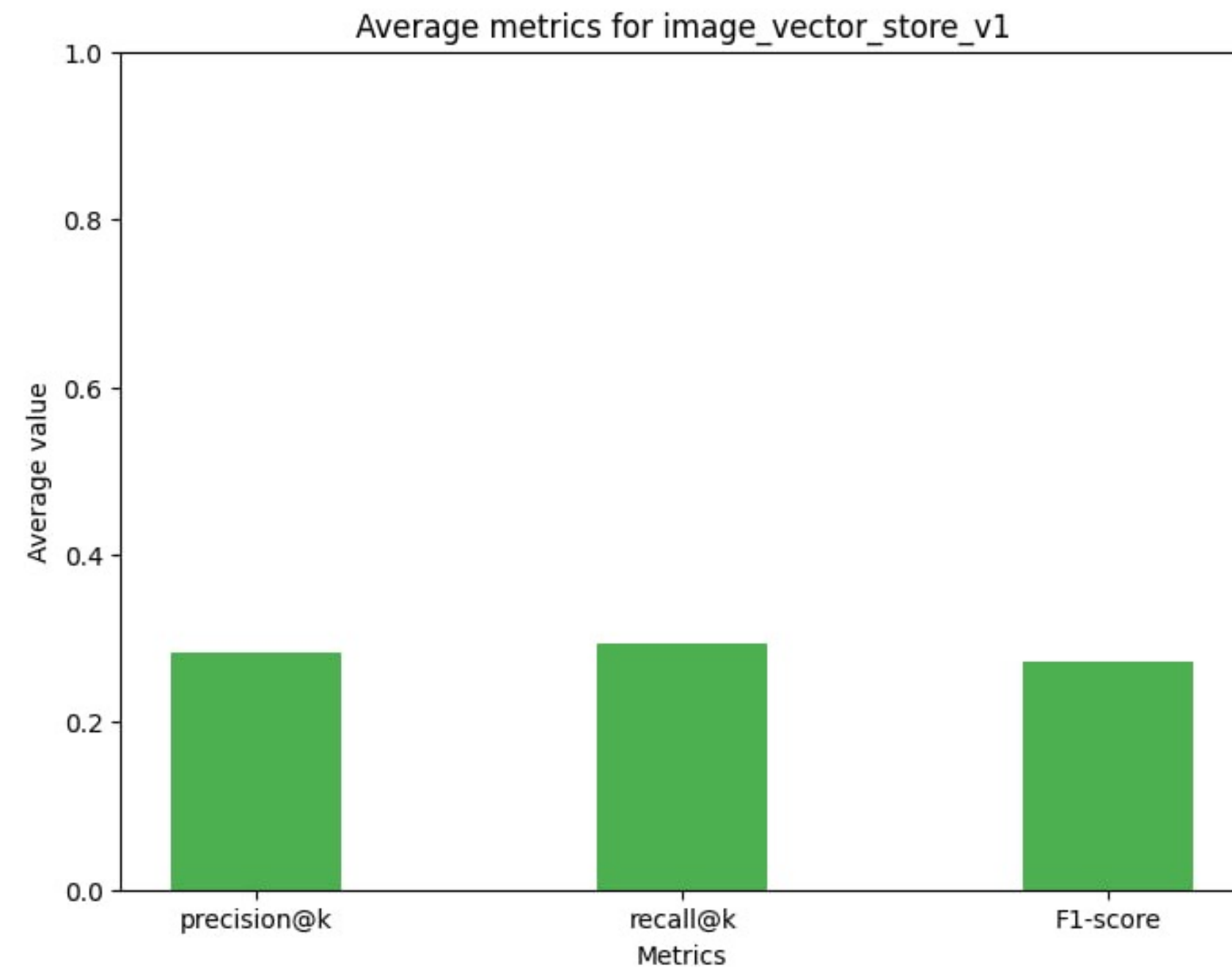
- **25 images from validation dataset cropped by flower detection model (Yolov8)**
- **25 images from train dataset cropped by detection boxes**
- **25 images from validation dataset cropped by sliding window method**
- **25 images from train dataset cropped by sliding window method**

# Experiments

Dimension Reduction Distance	method			
		AE (for size=526)	PCA	Gaussian
Euclidean				
Cosine				



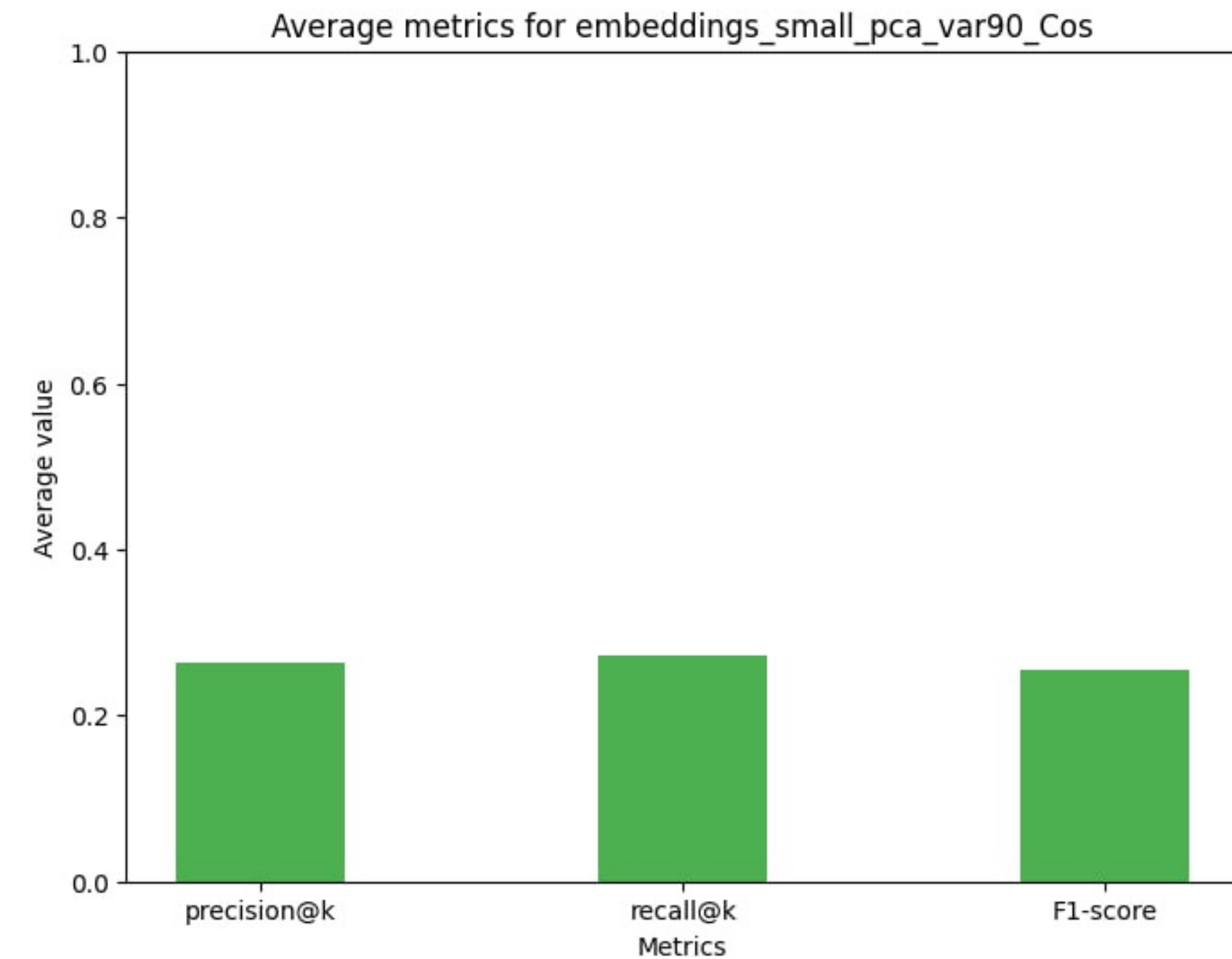
# Results



## Before dimension reduction

Size of embedding: 768

Memory consuming of all embeddings: 310 Mb



## After dimension reduction using PCA

Size of embedding: 223

Memory consuming of all embeddings: 140 Mb

Variance explained: 90%

# Minimum Hardware Requirements

**RAM: 1 Gb for database**

**Vram: 1.5 Gb for embedder (depend on batch size)**

**Disk space: 5 Gb for dataset**

# Thank you!

<https://github.com/MrIlyaneX/DLS-Project>