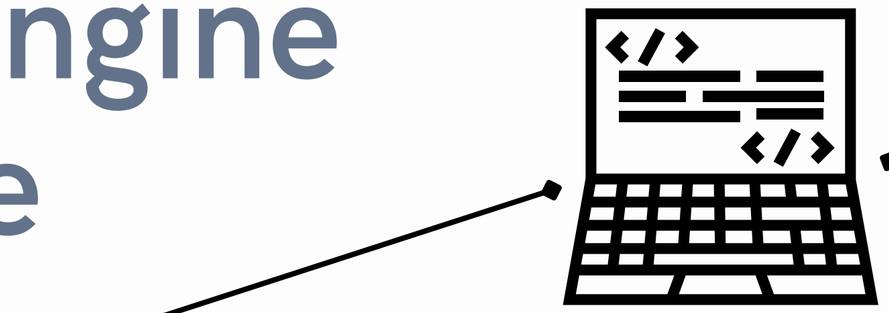
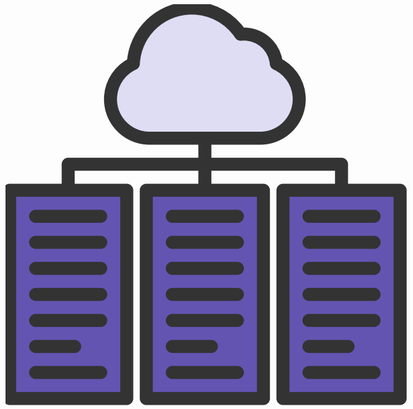
# Information Retrieval Project

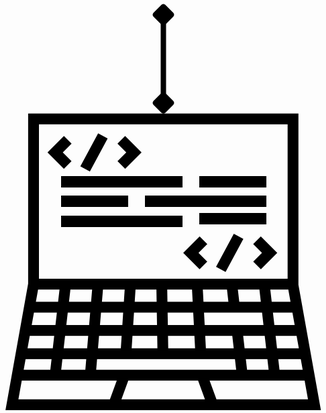
# Search Engine Structure



search_backend.py

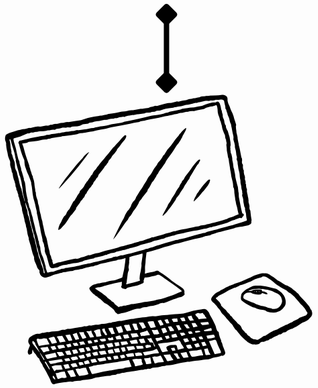


Google Storage Bucket



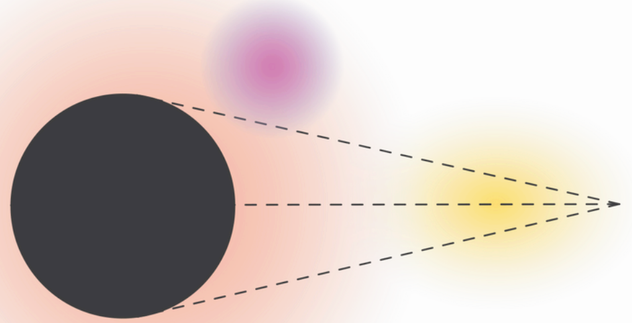
search_frontend.py

Virtual Machine



User

# What we did



## 1 Better Hardware

- e2-standard-2 instance (2 vCPUs, 8 GB RAM)

## 2 Multi-index approach

- Body + Title + Anchor Text + Doc length + PageRank + PageViews

## 3 Local SSD Caching

## 4 Scaling

- scaling *log10(1+x)* to anchor
- pruning (very) common words (high df)

## 5 Grid Search

- Over 108 hyperparameter combinations to maximize MAP@10

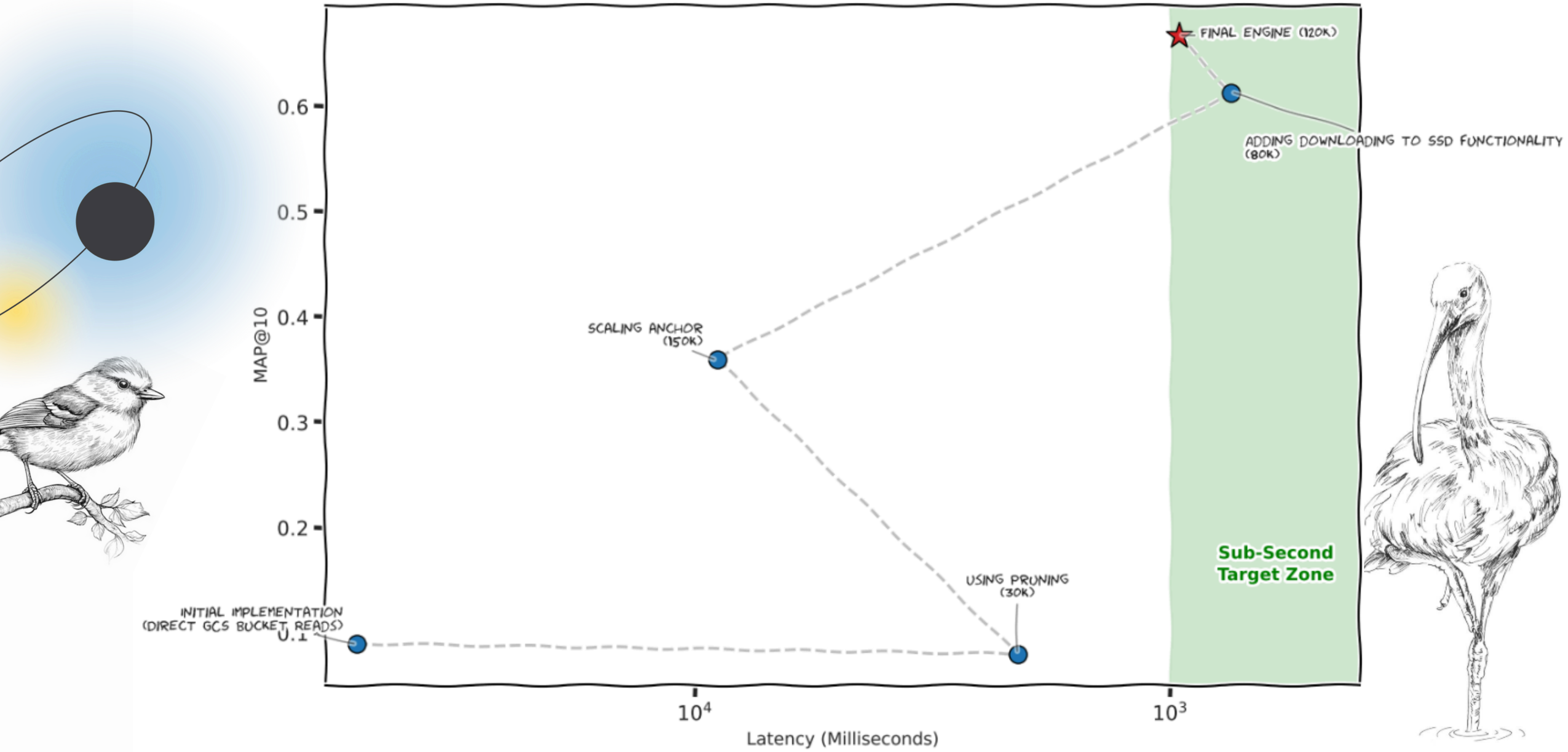
## 6 Testing

- Cosine Similarity / BM25
- Assignment 3 tokenizer / english stopwords tokenizer

## 7 Deployment

- tmux

# Experiments results



# Final Results

## Cosine Similarity

```
----------------------------------------
Total Queries:      30
Avg Latency:        766.46 ms
----------------------------------------
MAP@5 (Avg P@5):    0.5333
MAP@10 (Avg P@10):  0.3700
Avg F1@30:          0.1546
```

## BM-25

```
========================================
Total Queries:      30
Avg Latency:        959.14 ms
----------------------------------------
MAP@5 (Avg P@5):    0.7467
MAP@10 (Avg P@10):  0.6667
Avg F1@30:          0.3681
```