StarfishDB-SVI Compilation Guide

Sami Hadouaj* Ouael Ben Amara† Niccolò Meneghetti‡

1 Introduction

Introduction

This guide details how to compile the experiments from our paper, "Variational Inference for StarfishDB." Reproducing these experiments is complex and time-consuming (taking several days) due to StarfishDB's intricate dependencies and its reliance on just-in-time compilation, which demands significant toolchain and environment customization. To aid reproducibility and troubleshooting, we've structured the process into modular scripts, allowing easier issue identification and resolution at each stage. We thank the reviewers for their efforts in verifying our work.

This repository provides the implementation and benchmarking suite used for our paper. The presented implementation features StarfishDB with its two inference engines:

- **Stochastic Variational Inference (SVI)**
- **Collapsed Gibbs Sampler (CGS)**

As well as scripts for training LDA using **Mallet** Collapsed Gibbs Sampler on the given datasets.

Hardware Requirements

Experiments require significant resources:

- 24 + CPUs
- 512 + GB RAM
- 600 + GB storage
- Red Hat Enterprise Linux 8.1

*shadouaj@umich.edu
†benamara@umich.edu
‡niccolom@umich.edu

2 Datasets

We evaluate our framework on three large-scale text corpora:

Dataset	Documents	Tokens	Vocabulary
NYTimes	299,752	100M	102,660
PubMed	8.2M	730M	141,043
Wikipedia	4M	1B	192,000

Table 1: Evaluation datasets

Running the LDA Experiments

Follow the steps below to build the development environment, fetch dependencies & datasets, compile the project, and run benchmarks.

1 Create and Enter the Docker Development Container

```
# Build the image
sudo ./scripts/build_devenv_docker_img.sh

# Create the container
sudo ./scripts/create_devenv_docker_cont.sh

# Attach via tmux (recommended to avoid disconnects)
tmux
sudo docker exec -it starfishdb_dev_env_container_ec2-user863a3113
    bash
# Note: The exact container name may vary.
```

2 Install Dependencies

```
source scripts/get_deps.sh 2>&1 | tee "logs/log_get_deps.txt"
```

3 Download & Preprocess Datasets

```
source scripts/get_uci_datasets.sh 2>&1 | tee "logs/
    log_get_uci_datasets.txt"
source scripts/get_wiki_dataset.sh 2>&1 | tee "logs/
    log_get_wiki_dataset.txt"
```

4 Build StarfishDB

```
cd build
cmake3 -G Ninja -DCMAKE_BUILD_TYPE=RelWithDebInfo \
  -DCMAKE_C_COMPILER="$(readlink -f ../libs/llvm-project-cxxjit/bin)/
    clang" \
  -DCMAKE_CXX_COMPILER="$(readlink -f ../libs/llvm-project-cxxjit/bin
    )/clang++" \
  .. 2>&1 | tee "logs/log_cmake3.txt"

ninja 2>&1 | tee "logs/log_ninja.txt"
```

5 Run LDA Benchmarks

Stochastic Variational Inference (SVI) Experiments were conducted on the NY-Times, PubMed, and Wikipedia datasets using 8, 16, and 24 threads.

```
# 50 topics
source "scripts/run_vi_benchmark50.sh" 2>&1 | tee "logs/
    log_run_vi_benchmark50.txt"

# 100 topics
source "scripts/run_vi_benchmark100.sh" 2>&1 | tee "logs/
    log_run_vi_benchmark100.txt"

# 200 topics
source "scripts/run_vi_benchmark200.sh" 2>&1 | tee "logs/
    log_run_vi_benchmark200.txt"
```

Collapsed Gibbs Sampling (CGS) Using both StarfishDB-CGS and Mallet:

```
# 50 topics
source "scripts/run_lda_benchmarksP50.sh" 2>&1 | tee "logs/
    log_run_lda_benchmarksP50.txt"

# 100 topics
source "scripts/run_lda_benchmarksP100.sh" 2>&1 | tee "logs/
    log_run_lda_benchmarksP100.txt"

# 200 topics
source "scripts/run_lda_benchmarksP200.sh" 2>&1 | tee "logs/
    log_run_lda_benchmarksP200.txt"
```

The results of the experiments will be in CSV format under `benchmarks/csv`.