

# Podrobná uživatelská dokumentace

Tato dokumentace obsahuje podrobný návod ke všem programům, které obsahuje příloha práce. Popis systémových požadavků, přístupu k programátorské dokumentaci a návod jak spustit ukázkou projektu obsahuje dokument `uživatelská dokumentace.pdf`.

Všechny cesty uvedené v tomto souboru jsou relativní k cestě složky, která obsahuje PDF soubor této dokumentace.

## Sestavení projektu

Spustíte skript `src/scripts/buildProject.py` pomocí interpretu Python verze 3.9. Tento skript vygeneruje virtuální prostředí, stáhne potřebné balíčky a zkompiluje knihovnu nové implementace indexu HNSW.

*OS Windows*

```
1 | py -3.9 .\src\scripts\buildProject.py
```

*OS Linux*

```
1 | python3.9 ./src/scripts/buildProject.py
```

## Virtuální prostředí

Pokud není uvedeno jinak, skripty uvnitř složky `src/scripts` vždy spouštějte pomocí vygenerovaného virtuálního prostředí. Pro prostředí aktivujete pomocí aktivačního skriptu ve složce `.venv/Scripts`. Výběr skriptu závisí na použitém OS a interpretu.

OS	Interpret	Cesta k aktivačnímu skriptu
Linux		<code>./venv/bin/activate</code>
Windows	Batch	<code>.\venv\Scripts\activate.bat</code>
Windows	Powershell	<code>.\venv\Scripts\Activate.ps1</code>

Pokud používáte OS Linux specifikujte zvolený interpret při volání aktivačního skriptu.

```
1 | source ./venv/bin/activate
```

## Seznam skriptů

Následuje seznam skriptů ve složce `src/scripts`.

Název skriptu	Stručný popis skriptu
buildProject	Vytvoří virtuální prostředí, nativní C++ řešení a jeho Python rozhraní.
clean	Odstraní vygenerované soubory a vrátí projekt do původního stavu.
datasetGenerator	Vygeneruje datové soubory pro debugování.
datasetToText	Převede datový soubor do textového formátu.
formatCMakeTemplates	Vygeneruje CMakeLists.txt.
generateTables	Vygeneruje LaTeX tabulky podobné těm, které jsou v bakalářské práci.
latexTable	Vygeneruje LaTeX tabulku na základě výsledků srovnání.
runBenchmarks	Spustí srovnání, vygeneruje a otevře webovou stránku s výsledky.
runRecallTable	Postaví nový index a zobrazí tabulku závislosti přesnosti na parametru vyhledávání <code>ef<sub>search</sub></code> .
SIMDCapability	Zobrazí SIMD rozšíření instrukční sady procesoru, která jsou k dispozici.

## Podrobný popis skriptů

U každého skriptu je uveden jeho účel, parametry a příklad spuštění. Pokud skript obsahuje alespoň jeden parametr, pak použitím parametru `--help` nebo `-h` zobrazíte nápovědu v anglickém jazyce.

### buildProject

Tento skript lze spustit bez virtuálního prostředí.

Vytvoří virtuální prostředí interpretu Python, stáhne potřebné softwarové balíčky, vygeneruje nativní C++ řešení pro knihovnu nového indexu, vytvoří rozhraní v jazyce Python pro nový index a otestuje funkčnost tohoto indexu spuštěním skriptu `runRecallTable`.

Parametr, zkratka	Význam
<code>--clean, -c</code>	Vrátí projekt do původního stavu před jeho opětovným sestavením.
<code>--cleanResults, -r</code>	Pokud je <code>--clean</code> nastaven, odstraní naměřené výsledky.
<code>--ignorePythonVersion, -i</code>	Umožňuje spustit skript s libovolnou verzí interpretu Python. Skript poté nemusí fungovat správně.

Příklad spuštění:

```
1 | py -3.9 buildProject.py --clean --cleanResults
```

## clean

Tento skript lze spustit bez virtuálního prostředí.

Odstraní datové soubory pro debugování, C++ nativní řešení a Python rozhraní. Pokud je spuštěn mimo virtuální prostředí, pak odstraní toto prostředí. Naměřené výsledky odstraněny nebudou, pokud o to uživatel nepožádá.

Parametr, zkratka	Význam
--results, -r	Odstraní naměřené výsledky srovnání, vygenerované grafy a tabulky.

Příklad spuštění:

```
1 | python clean.py --results
```

## datasetGenerator

Vygeneruje datové soubory pro debugování uvedené v konfiguračním souboru

`src/config/debugDatasets.json`. O konfiguraci tohoto skriptu se více dočtete v kapitole [Datové soubory pro debugování](#) níže v této dokumentaci.

Příklad spuštění:

```
1 | python datasetGenerator.py
```

## datasetToText

Převěde vybraný datový soubor ze složky `src/data` do textového formátu. Výstupní textový soubor zapíše pod jménem datového souboru do stejné složky.

Parametr, zkratka	Význam
--name, -n	Název datového souboru bez přípony. Pokud není uveden, výchozím souborem je <code>angular-small</code> .

Příklad spuštění:

```
1 | python datasetToText --name euclidean-medium
```

## formatCMakeTemplates

Vygeneruje soubor `src/index/CMakeLists.txt` a doplní do něj správnou definici maker tak, aby došlo pouze ke kompilaci těch funkcí, pro které je k dispozici vhodné SIMD rozšíření instrukční sady procesoru.

Příklad spuštění:

```
1 | python formatCMakeTemplates.py
```

## generateTables

Vygeneruje LaTeX tabulky podobné těm, které jsou v bakalářské práci, ale pouze v případě, že jsou pro ně dostupné naměřené výsledky. Tyto výsledky lze získat spuštěním následujících příkazů. Avšak tato měření mohou trvat více než 12 hodin.

```
1 python runBenchmarks.py -a ..\config\heuristic.yaml ..\config\naive.yaml -d lastfm-64-dot -r 5
2 python runBenchmarks.py -a ..\config\heuristic.yaml ..\config\prefetch.yaml -d glove-50-angular -r 5
3 python runBenchmarks.py -a ..\config\original.yaml ..\config\prefetch.yaml -d sift-128-euclidean -r 5
4 python generateTables.py
```

Vygenerované tabulky jsou dostupné ve složce `src/figures`.

## latexTable

Vygeneruje jednu LaTeX tabulku na základě výsledků srovnání implementací.

Parametr, zkratka	Význam
--algorithms, -a	Vyžadován. Seznam implementací oddělený mezerami.
--dataset, -d	Vyžadován. Název datového souboru.
--label, -la	Identifikátor tabulky.
--legend, -le	Názvy implementací v tabulce. Pokud není uveden, budou použity původní názvy.
--output, -o	Vyžadován. Cesta k výstupnímu souboru.
--percent, -p	Přidá do tabulky sloupec s procentuálním rozdílem časů stavby.
--recompute, -r	Znovu vypočítá výkonnostní metriky z naměřených výsledků. Tato operace může trvat více než 10 minut.

Příklad spuštění:

```
1 python latexTable.py -a new-prefetch original -d sift-128-euclidean -le "Nová impl." "Původní impl." -o ..\figures\table.tex -p
```

## runBenchmarks

*Před spuštěním se ujistěte, že je služba Docker zapnutá.*

Spustí srovnání implementací v jednom nebo více Docker kontejnerech, vypočítá výkonnostní metriku, vygeneruje webovou stránku s výsledky a otevře ji v nové kartě internetového prohlížeče. Kód vygenerované stránky lze poté najít ve složce `src/website` a můžete ji opětovně zobrazit otevřením souboru `index.html`.

Parametr, zkratka	Význam
--algoDefPaths, -a	Vyžadován. Seznam cest ke konfiguračním souborům oddělených mezerami. O konfiguraci se více dočtete v kapitole <a href="#">Konfigurace srovnání</a> .
--datasets, -d	Vyžadován*. Seznam datových souborů oddělených mezerami.
--datasetsPath, -p	Vyžadován*. Cesta k textovému souboru se seznamem datových souborů.
--force, -f	Spustí již provedená měření znovu.-
--runs, -r	Počet opakování měření. Výchozí hodnota je 1.
--workers, -w	Počet paralelně spuštěných Docker kontejnerů. Výchozí hodnota je 1.

Datové soubory využitě ke srovnání nejsou ty samé, které jsou využívány k debugování. Jejich seznam najdete v kapitole [Testované datové soubory](#).

\* Pouze jeden z parametrů označených hvězdičkou by měl být uveden.

Příklad spuštění:

```
1 python runBenchmarks.py -a ..\config\nobit.yaml -f -p ..\config\datasets.txt
  -r 5 -w 2
```

## runRecallTable

Postaví index nové implementace a vyhledá v něm nejbližší sousedy s různými hodnotami parametru vyhledávání  $ef_{search}$ . Poté vypíše tabulku závislosti přesnosti vyhledávání na tomto parametru. Konfigurace tohoto skriptu se nachází v souboru `src/config/recallTable.json` a více se o ní dočtete v kapitole [Konfigurace programů recallTable](#).

Příklad spuštění:

```
1 python runRecallTable.py
```

## SIMDCapability

Zobrazí SIMD rozšíření instrukční sady procesoru, která jsou k dispozici. Využívána ostatními skripty pro vygenerování správných maker v jazyce C++.

Příklad spuštění:

```
1 python SIMDCapability.py
```

## Nativní knihovna

C++ řešení vygenerujete pomocí skriptu `RUNME.py` nebo `src/scripts/buildProject.py`. Řešení bude vytvořeno ve složce `src/cmakeBuild`. V každém systému vypadají soubory řešení jinak. Např. při použití Windows s Visual Studiemi je řešením `.sln` soubor a projekty jsou `.vcxproj` soubory. Pro spuštění projektů je doporučena konfigurace `Release`. Řešení obsahuje dva

projekty.

- *datasetToText* - Vypíše textový popis datové kolekce do souboru. Slouží pro ověření konzistence mezi binárními a HDF5 soubory. Název datového souboru je prvním parametrem programu. Výchozí hodnotou je `angular-small`.
- *recallTable* - Postaví HNSW index a vypíše tabulku závislosti přesnosti na parametru vyhledávání `efsearch`. Konfigurace programu se nachází v souboru `src/config/recallTable.json` a více se o ní dočtete v kapitole [Konfigurace programů recallTable](#).

## Datové soubory pro debugování

Pro vygenerování jiných datových souborů pro debugování změňte konfiguraci v souboru `src/config/debugDatasets.json` a spusťte skript `src/scripts/datasetGenerator.py`. Konfigurace je JSON soubor s polem objektů, kde každý objekt popisuje jeden datový soubor.

```
1 {  
2     "name": "angular-small",  
3     "angular": true,  
4     "dim": 25,  
5     "k": 10,  
6     "testCount": 200,  
7     "trainCount": 20000,  
8     "seed": 104  
9 }
```

Klíč	Typ hodnoty	Význam
name	string	Unikátní název souboru sloužící k identifikaci.
angular	boolean	Pokud je nastaven na <code>true</code> , využívá soubor kosinusové podobnosti. Jinak využívá Eukleidovské vzdálenosti.
dim	int	Počet dimenzí prostoru.
k	int	Počet hledaných nejbližších sousedů dotazovaného prvku.
testCount	int	Počet dotazů.
trainCount	int	Počet prvků použitých k sestavení indexu.
seed	int	Nastavení generátoru náhodných čísel.

## Konfigurace programů *recallTable*

Pro změnu sestavovaných konfigurací programy `recallTable.cpp` a `recallTable.py` upravte soubor `src/config/recallTable.json`. Konfigurace je JSON soubor s polem objektů, kde každý objekt popisuje jednu konfiguraci.

```

1 {
2     "dataset": "angular-small",
3     "efConstruction": 200,
4     "efSearch": [10, 15, 20, 40, 80, 120, 200],
5     "mMax": 16,
6     "seed": 200,
7     "SIMD": "best",
8     "template": "prefetching"
9 }

```

Klíč	Typ hodnoty	Význam
dataset	string	Identifikace datového souboru. Odpovídá klíči <code>name</code> v souboru <code>src/config/debugDatasets.json</code> .
efConstruction	int	Počet uvažovaných sousedů při vytváření nových hran v indexu.
efSearch	array	Pole hodnot parametru vyhledávání <code>ef<sub>search</sub></code> .
mMax	int	Maximální povolený počet sousedů jednoho prvku v indexu na vrstvě vyšší než vrstva 0.
seed	int	Nastavení generátoru náhodných úrovní v indexu.
SIMD	string	Upřednostňovaný typ SIMD instrukcí. Možnosti jsou <code>avx</code> , <code>avx512</code> , <code>best</code> , <code>null</code> , a <code>sse</code> .*
template	string	Šablona indexu. Možnosti jsou <code>Heuristic</code> , <code>Naive</code> , <code>NoBitArray</code> a <code>Prefetching</code> .

\* Zvolením hodnoty `best` zvolíte nejmodernější dostupné SIMD rozšíření. Hodnotou `null` zakážete použití SIMD instrukcí.

## Šablony nové implementace

Šablona	Metoda výběru sousedů	Seznam navštívených vrcholů	Asynchronní přístup do paměti
Heuristic	Heuristika	Bitové pole	Ne
Naive	Naivní algoritmus	Bitové pole	Ne
NoBitArray	Heuristika	Obyčejné pole	Při výpočtu vzdáleností Při načítání dat seznamu navštívených vrcholů
Prefetching	Heuristika	Bitové pole	Při výpočtu vzdáleností

## Testované datové soubory

Pro srovnání implementací je možno využít následujících datových souborů.

Název	Dimenze	Počet prvků při stavbě	Dotazy	Metrika
fashion-mnist-784-euclidean	784	60 000	10 000	Eukleidovská vzdálenost
glove-25-angular	25	1 183 514	10 000	Kosinusová podobnost
glove-50-angular	50	1 183 514	10 000	Kosinusová podobnost
glove-100-angular	100	1 183 514	10 000	Kosinusová podobnost
glove-200-angular	200	1 183 514	10 000	Kosinusová podobnost
lastfm-64-dot	64	292 385	50 000	Kosinusová podobnost
mnist-784-euclidean	784	60 000	10 000	Eukleidovská vzdálenost
nytimes-16-angular	16	290 000	10 000	Kosinusová podobnost
nytimes-256-angular	256	290 000	10 000	Kosinusová podobnost
random-s-100-angular	100	100 000	10 000	Kosinusová podobnost
random-s-100-euclidean	100	100 000	10 000	Eukleidovská vzdálenost
random-xs-20-angular	20	10 000	10 000	Kosinusová podobnost
random-xs-20-euclidean	20	10 000	10 000	Eukleidovská vzdálenost
sift-128-euclidean	128	1 000 000	10 000	Eukleidovská vzdálenost

Pro spuštění srovnání nad více soubory lze využít textového formátu, kde každý řádek reprezentuje jeden datový soubor. Řádky, které začínají znakem `#` jsou ignorovány. Příklad:

```
1 # glove-25-angular
2 nytimes-256-angular
3 sift-128-euclidean
```

## Konfigurace srovnání



Výběr implementací ke srovnání a jejich parametrů zprostředkovávají konfigurační soubory ve formátu YAML. Příklad takového souboru je `src/config/algos.yaml`. Ve složce `src/config` se nacházejí předem vytvořené konfigurace.

Následuje příklad konfigurace. Komentáře označené znakem `#` popisují jednotlivé klíče.

```
1  algos: # Povinný klíč.
2      # Povinné pole názvů porovnávaných implementací.
3      - original
4      - new-prefetch
5  build: # Povinný klíč.
6      # Povinné pole objektů konfigurace stavby.
7      # Pro každou konfiguraci musí být uvedeny hodnoty parametrů efConstruction
8      a mMax.
9      - efConstruction: 50
10     mMax: 4
11     - efConstruction: 100
12     mMax: 8
13 efsearch: [10, 12, 14, 16, 18, 20, 25, 30, 40, 80] # Povinné pole hodnot
14 parametru vyhledávání efsearch.
```

Definované implementace popisuje následující tabulka.

Název implementace	Druh implementace	Šablona	SIMD rozšíření
original	Původní hnsplib		Nejmodernější dostupné
new-avx	Nová	Heuristic	AVX
new-avx-512	Nová	Heuristic	AVX-512
new-heuristic	Nová	Heuristic	Nejmodernější dostupné
new-naive	Nová	Naive	Nejmodernější dostupné
new-no-bit	Nová	NoBitArray	Nejmodernější dostupné
new-no-simd	Nová	Heuristic	Žádné
new-prefetch	Nová	Prefetching	Nejmodernější dostupné
new-sse	Nová	Heuristic	SSE

Popis šablon obsahuje kapitola `Šablony nové implementace`.