Search Technology for Media and Web (WiSe 2024/2025) - Programming Assignment 02

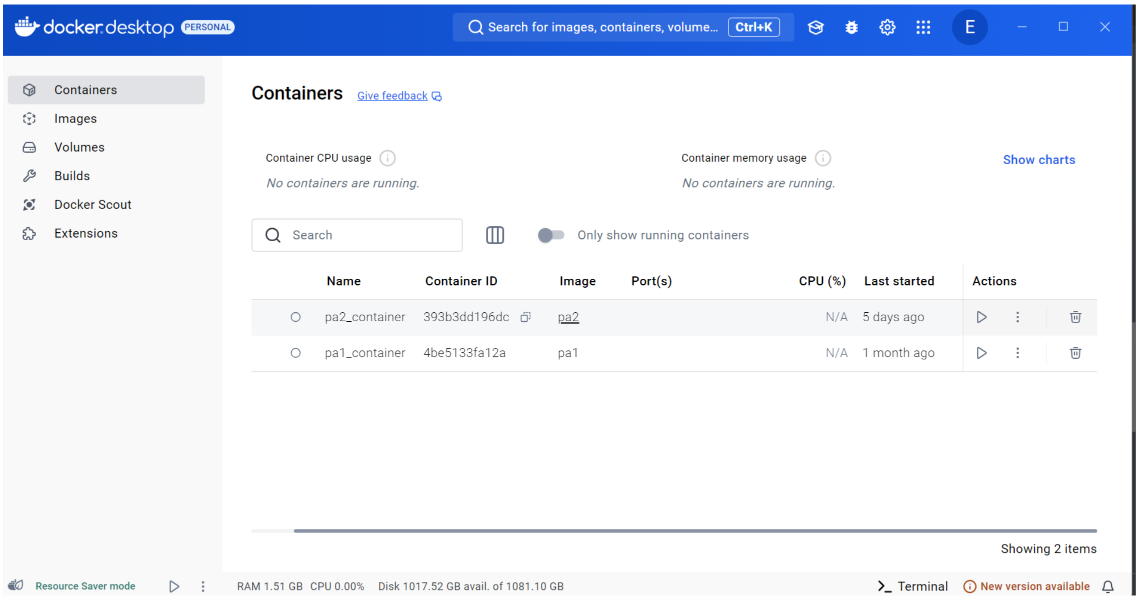
Name:
Enlik Enlik

Date:
19 February 2025

Presentation Guidelines:
Duration: 7 minutes for the presentation, followed by a 3-minute discussion.
Content: Focus on what you did and why, outlining how you approached solving the assignment.

Setting Up PA2 Docker Container in Windows 11

- Start `Docker Desktop` with administrator privilege



Containers

Container CPU usage
No containers are running.

Container memory usage
No containers are running.

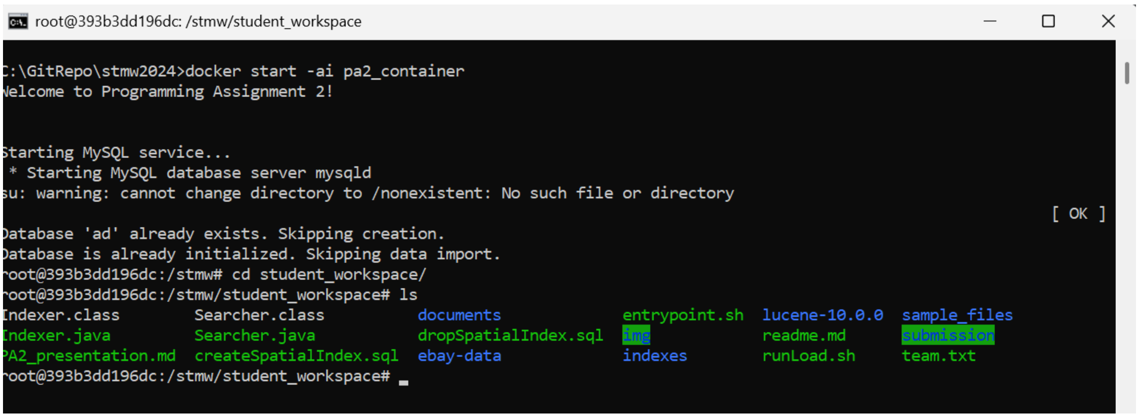
	Name	Container ID	Image	Port(s)	CPU (%)	Last started	Actions
○	pa2_container	393b3dd196dc	pa2		N/A	5 days ago	
○	pa1_container	4be5133fa12a	pa1		N/A	1 month ago	

Showing 2 items

RAM 1.51 GB CPU 0.00% Disk 1017.52 GB avail. of 1081.10 GB

- Open docker container with this command
  `docker start -ai pa2_container`
- Go to directory where the batch file stored

  `cd student_workspace`



root@393b3dd196dc: /stmw/student_workspace

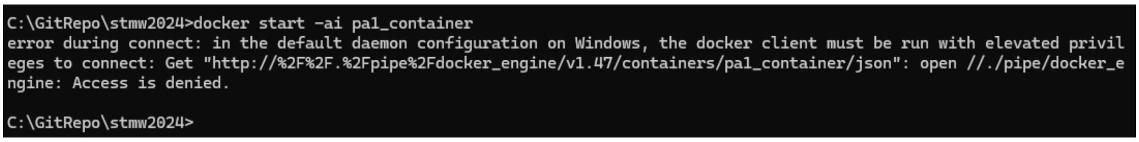
```
C:\GitRepo\stmw2024>docker start -ai pa2_container
Welcome to Programming Assignment 2!


Starting MySQL service...
 * Starting MySQL database server mysqld
su: warning: cannot change directory to /nonexistent: No such file or directory
                                                                                         [ OK ]
Database 'ad' already exists. Skipping creation.
Database is already initialized. Skipping data import.
root@393b3dd196dc:/stmw# cd student_workspace/
root@393b3dd196dc:/stmw/student_workspace# ls
Indexer.class        Searcher.class           documents             entrypoint.sh  lucene-10.0.0  sample_files
Indexer.java         Searcher.java            dropSpatialIndex.sql  img            readme.md      submission
PA2_presentation.md  createSpatialIndex.sql   ebay-data             indexes        runLoad.sh     team.txt
root@393b3dd196dc:/stmw/student_workspace# _
```

- Voila, setup is done for both docker pa2_container

Error found during setting up:

- Can't run docker without administrator privilege



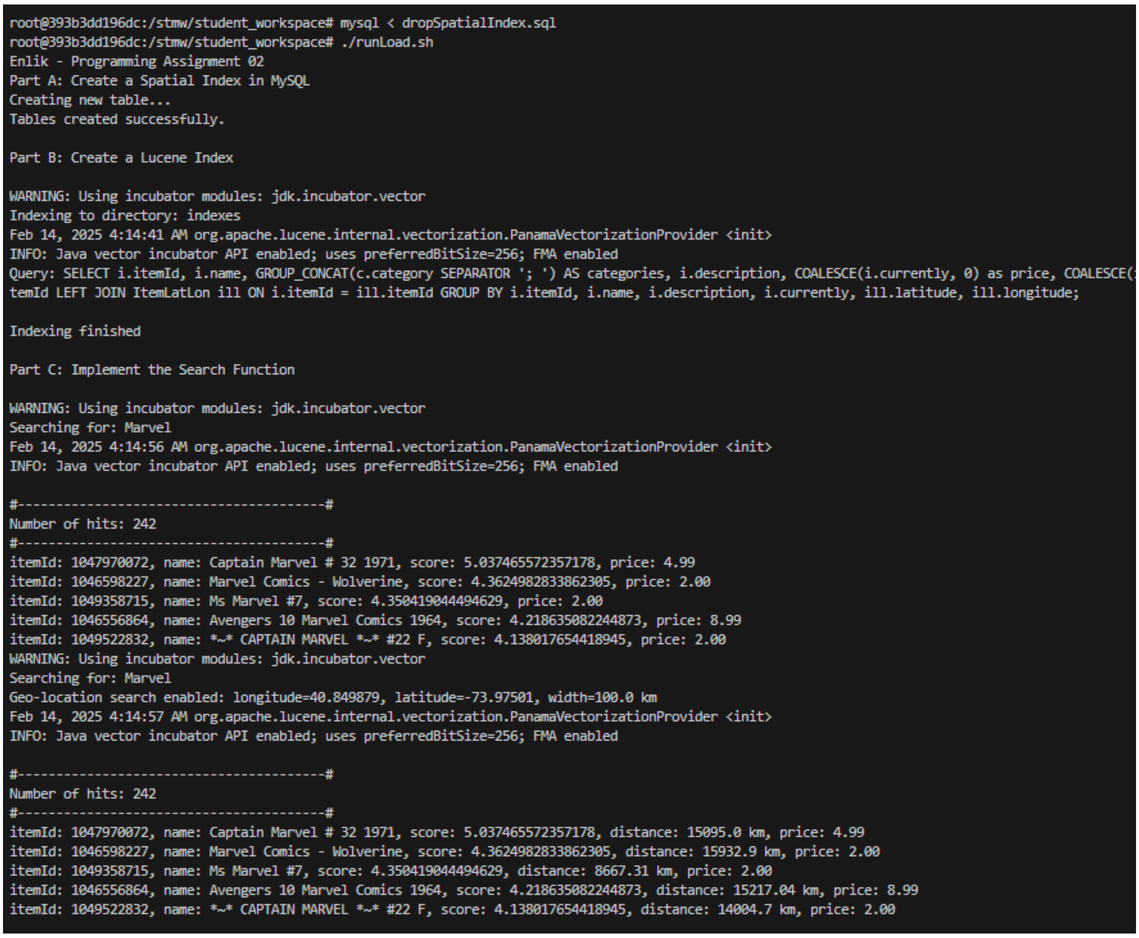
```
C:\GitRepo\stmw2024>docker start -ai pa1_container
error during connect: in the default daemon configuration on Windows, the docker client must be run with elevated privil
eges to connect: Get "http://%2F%2F.%2Fpipe%2Fdocker_engine/v1.47/containers/pa1_container/json": open //./pipe/docker_e
ngine: Access is denied.

C:\GitRepo\stmw2024>
```

Main Process using Shell Script

- Run the shell script using this command:

  `./runLoad.sh`



```
root@393b3dd196dc:/stmw/student_workspace# mysql < dropSpatialIndex.sql
root@393b3dd196dc:/stmw/student_workspace# ./runLoad.sh
Enlik - Programming Assignment 02
Part A: Create a Spatial Index in MySQL
Creating new table...
Tables created successfully.

Part B: Create a Lucene Index

WARNING: Using incubator modules: jdk.incubator.vector
Indexing to directory: indexes
Feb 14, 2025 4:14:41 AM org.apache.lucene.internal.vectorization.PanamaVectorizationProvider <init>
INFO: Java vector incubator API enabled; uses preferredBitSize=256; FMA enabled
Query: SELECT i.itemId, i.name, GROUP_CONCAT(c.category SEPARATOR '; ') AS categories, i.description, COALESCE(i.currently, 0) as price, COALESCE(
temId LEFT JOIN ItemLatLon ill ON i.itemId = ill.itemId GROUP BY i.itemId, i.name, i.description, i.currently, ill.latitude, ill.longitude;

Indexing finished

Part C: Implement the Search Function

WARNING: Using incubator modules: jdk.incubator.vector
Searching for: Marvel
Feb 14, 2025 4:14:56 AM org.apache.lucene.internal.vectorization.PanamaVectorizationProvider <init>
INFO: Java vector incubator API enabled; uses preferredBitSize=256; FMA enabled

#---------------------------------------#
Number of hits: 242
#---------------------------------------#
itemId: 1047970072, name: Captain Marvel # 32 1971, score: 5.037465572357178, price: 4.99
itemId: 1046598227, name: Marvel Comics - Wolverine, score: 4.3624982833862305, price: 2.00
itemId: 1049358715, name: Ms Marvel #7, score: 4.350419044494629, price: 2.00
itemId: 1046556864, name: Avengers 10 Marvel Comics 1964, score: 4.218635082244873, price: 8.99
itemId: 1049522832, name: *~* CAPTAIN MARVEL *~* #22 F, score: 4.138017654418945, price: 2.00
WARNING: Using incubator modules: jdk.incubator.vector
Searching for: Marvel
Geo-location search enabled: longitude=40.849879, latitude=-73.97501, width=100.0 km
Feb 14, 2025 4:14:57 AM org.apache.lucene.internal.vectorization.PanamaVectorizationProvider <init>
INFO: Java vector incubator API enabled; uses preferredBitSize=256; FMA enabled

#---------------------------------------#
Number of hits: 242
#---------------------------------------#
itemId: 1047970072, name: Captain Marvel # 32 1971, score: 5.037465572357178, distance: 15095.0 km, price: 4.99
itemId: 1046598227, name: Marvel Comics - Wolverine, score: 4.3624982833862305, distance: 15932.9 km, price: 2.00
itemId: 1049358715, name: Ms Marvel #7, score: 4.350419044494629, distance: 8667.31 km, price: 2.00
itemId: 1046556864, name: Avengers 10 Marvel Comics 1964, score: 4.218635082244873, distance: 15217.04 km, price: 8.99
itemId: 1049522832, name: *~* CAPTAIN MARVEL *~* #22 F, score: 4.138017654418945, distance: 14004.7 km, price: 2.00
```

This shell script contains the list of bash commands with following step-by-step process:

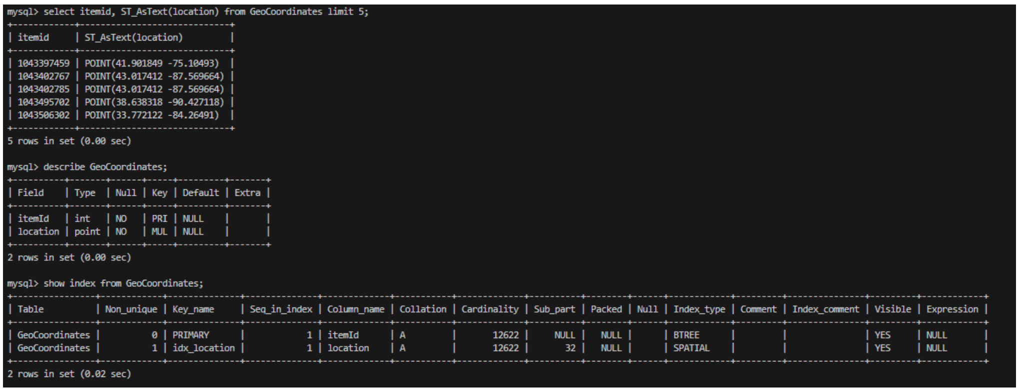
1. Create new table and a spatial index via `createSpatialIndex.sql`

```sql
CREATE TABLE IF NOT EXISTS GeoCoordinates (
    itemId INT NOT NULL,
    location POINT NOT NULL,
    PRIMARY KEY (itemId)
) ENGINE=InnoDB;

INSERT INTO GeoCoordinates (itemId, location)
    SELECT itemId, ST_GeomFromText(CONCAT('POINT(', latitude, ' ', longitude, ')'))
    FROM ItemLatLon
WHERE latitude IS NOT NULL AND longitude IS NOT NULL;

CREATE SPATIAL INDEX idx_location ON GeoCoordinates (location);
```

- MySQL Validation



```
mysql> select itemid, ST_AsText(location) from GeoCoordinates limit 5;
+------------+---------------------------+
| itemid     | ST_AsText(location)       |
+------------+---------------------------+
| 1043397459 | POINT(41.901849 -75.10493)  |
| 1043402767 | POINT(43.017412 -87.569664) |
| 1043402785 | POINT(43.017412 -87.569664) |
| 1043495702 | POINT(38.638318 -90.427118) |
| 1043506302 | POINT(33.772122 -84.26491)  |
+------------+---------------------------+
5 rows in set (0.00 sec)

mysql> describe GeoCoordinates;
+----------+-------+------+-----+---------+-------+
| Field    | Type  | Null | Key | Default | Extra |
+----------+-------+------+-----+---------+-------+
| itemId   | int   | NO   | PRI | NULL    |       |
| location | point | NO   | MUL | NULL    |       |
+----------+-------+------+-----+---------+-------+
2 rows in set (0.00 sec)

mysql> show index from GeoCoordinates;
+----------------+------------+--------------+--------------+-------------+-----------+-------------+----------+--------+------+------------+---------+---------------+---------+------------+
| Table          | Non_unique | Key_name     | Seq_in_index | Column_name | Collation | Cardinality | Sub_part | Packed | Null | Index_type | Comment | Index_comment | Visible | Expression |
+----------------+------------+--------------+--------------+-------------+-----------+-------------+----------+--------+------+------------+---------+---------------+---------+------------+
| GeoCoordinates |          0 | PRIMARY      |            1 | itemId      | A         |       12622 |     NULL |   NULL |      | BTREE      |         |               | YES     | NULL       |
| GeoCoordinates |          1 | idx_location |            1 | location    | A         |       12622 |       32 |   NULL |      | SPATIAL    |         |               | YES     | NULL       |
+----------------+------------+--------------+--------------+-------------+-----------+-------------+----------+--------+------+------------+---------+---------------+---------+------------+
2 rows in set (0.02 sec)
```

2. Create a Lucene index using `Indexer.java`, and put this SQL Query inside java file

```sql
SELECT i.itemId, i.name, GROUP_CONCAT(c.category SEPARATOR '; ') AS categories, i.description, COALESCE(i.currently, 0) as price, COALESCE(ill.latitude, 0.00) as latitude, COALESCE(ill.longitude, 0.00) as longitude FROM Items i INNER JOIN Categories c ON i.itemId = c.itemId LEFT JOIN ItemLatLon ill ON i.itemId = ill.itemId GROUP BY i.itemId, i.name, i.description, i.currently, ill.latitude, ill.longitude;
```

Explanation:

- Getting `itemId`, `name`, `description`, `price` values from table `Items`
- Getting `category` values from table `Categories`
- Getting `latitude`, `longitude` values from table `ItemLatLon`
- Using `COALESCE`, we're replacing NULL value(s) with 0, for normalization of column `price`, `latitude`, `longitude`

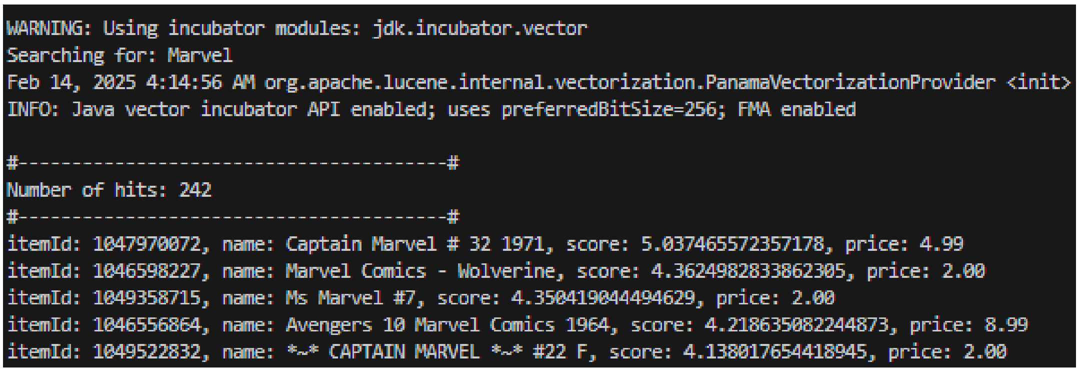
3. Implement the Search Function in `Searcher.java`

- Used {"name", "categories", "description"}`as the fields for text searching process using Lucene`SimpleAnalyzer`and`MultiFieldQueryParser`
- Added all required output fields
- Added extra arguments for `latitude`, `longitude`, and `width`
- Used `haversineDistance` mathematical formula to calculate geo-location distance

   - Implemented ranking system using Java `sort` function

4. Example search:
   - Without geo-location data

```
java --enable-native-access=ALL-UNNAMED --add-modules
jdk.incubator.vector Searcher "Marvel" 5
```

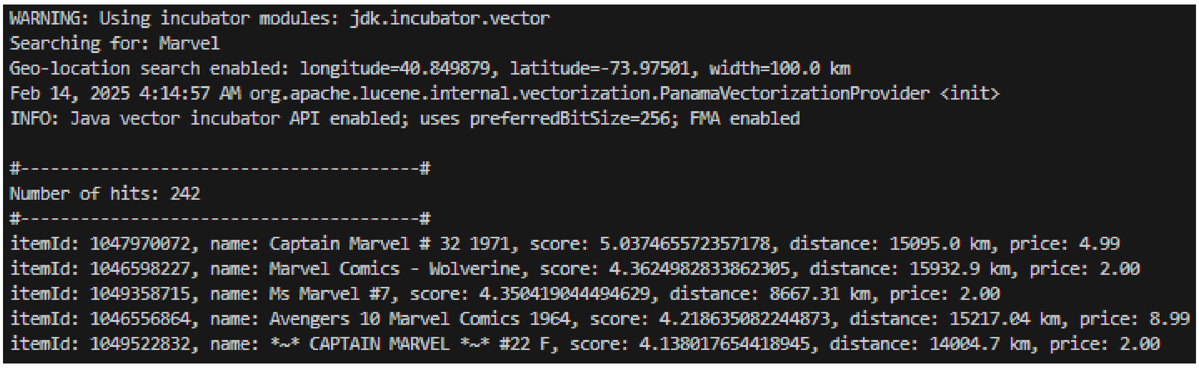


```
WARNING: Using incubator modules: jdk.incubator.vector
Searching for: Marvel
Feb 14, 2025 4:14:56 AM org.apache.lucene.internal.vectorization.PanamaVectorizationProvider <init>
INFO: Java vector incubator API enabled; uses preferredBitSize=256; FMA enabled

#----------------------------------------#
Number of hits: 242
#----------------------------------------#
itemId: 1047970072, name: Captain Marvel # 32 1971, score: 5.037465572357178, price: 4.99
itemId: 1046598227, name: Marvel Comics - Wolverine, score: 4.3624982833862305, price: 2.00
itemId: 1049358715, name: Ms Marvel #7, score: 4.350419044494629, price: 2.00
itemId: 1046556864, name: Avengers 10 Marvel Comics 1964, score: 4.218635082244873, price: 8.99
itemId: 1049522832, name: *~* CAPTAIN MARVEL *~* #22 F, score: 4.138017654418945, price: 2.00
```

   - With geo-location data

```
java --enable-native-access=ALL-UNNAMED --add-modules
jdk.incubator.vector Searcher "Marvel" 5 -x 40.84987900 -y
-73.97501000 -w 100
```



```
WARNING: Using incubator modules: jdk.incubator.vector
Searching for: Marvel
Geo-location search enabled: longitude=40.849879, latitude=-73.97501, width=100.0 km
Feb 14, 2025 4:14:57 AM org.apache.lucene.internal.vectorization.PanamaVectorizationProvider <init>
INFO: Java vector incubator API enabled; uses preferredBitSize=256; FMA enabled

#----------------------------------------#
Number of hits: 242
#----------------------------------------#
itemId: 1047970072, name: Captain Marvel # 32 1971, score: 5.037465572357178, distance: 15095.0 km, price: 4.99
itemId: 1046598227, name: Marvel Comics - Wolverine, score: 4.3624982833862305, distance: 15932.9 km, price: 2.00
itemId: 1049358715, name: Ms Marvel #7, score: 4.350419044494629, distance: 8667.31 km, price: 2.00
itemId: 1046556864, name: Avengers 10 Marvel Comics 1964, score: 4.218635082244873, distance: 15217.04 km, price: 8.99
itemId: 1049522832, name: *~* CAPTAIN MARVEL *~* #22 F, score: 4.138017654418945, distance: 14004.7 km, price: 2.00
```

Known Issues

- The geo-location distance calculation using `haversineDistance` somehow giving wrong calculation, need to be updated