



# CSCI 5410

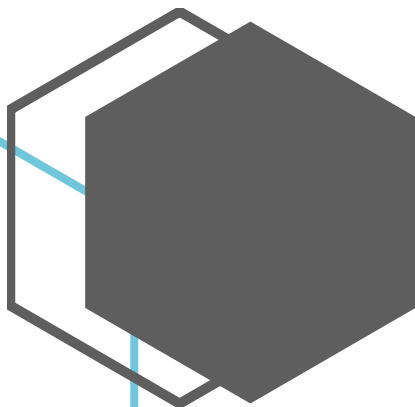
---

## Assignment 4 – Part B

Name: Benny Daniel Tharigopala

Banner ID: B00899629

GitLab URL: [https://git.cs.dal.ca/benny/csci5410\\_B00899629\\_Benny\\_Tharigopala](https://git.cs.dal.ca/benny/csci5410_B00899629_Benny_Tharigopala)



## Event-driven serverless application using GCP ML

## Creating a project on Google Cloud Platform (GCP)

The screenshot shows the GCP dashboard for project 'csci5410-assignment-4'. The top navigation bar includes 'Google Cloud', the project name, a search bar, and links to 'Products, resources, docs (/)'. Below the navigation bar are tabs for 'DASHBOARD', 'ACTIVITY', and 'RECOMMENDATIONS'. The main content area is divided into several sections:

- Project info:** Displays project details such as Project name (csci5410-assignment-4), Project number (429687142387), and Project ID (csci5410-assignment-4). It includes a link to 'ADD PEOPLE TO THIS PROJECT' and a button to 'Go to project settings'.
- Resources:** Lists various GCP services available to the project, including BigQuery, SQL, Compute Engine, Storage, Cloud Functions, and App Engine.
- RPI APIs:** A section for API requests, showing a graph of 'Requests (requests/sec)' over time. A message indicates 'No data is available for the selected time frame.' A button 'Go to APIs overview' is present.
- Google Cloud Platform status:** Shows the status of GCP services. A message indicates 'Data unavailable' and a link to 'Go to Cloud status dashboard'.
- Billing:** Displays estimated charges for the billing period Jul 1 - 9, 2022, as USD \$0.00. It includes a link to 'Take a tour of billing' and a button to 'View detailed charges'.
- Monitoring:** Provides options to 'Create my dashboard', 'Set up alerting policies', and 'Create uptime checks'. It also includes a link to 'View all dashboards' and a button to 'Go to Monitoring'.

## Creating a service account on GCP

The screenshot shows the 'Create service account' wizard in the Google Cloud IAM & Admin console. The left sidebar lists navigation options, with 'Service Accounts' selected. The main content area is divided into three steps:

- Service account details:** This step involves filling out the following information:
  - Service account name:** csci-5410-assignment-4
  - Service account ID:** csci-5410-assignment-4
  - Service account description:** Service Account for Assignment 4
 A 'CREATE AND CONTINUE' button is visible at the bottom of this step.
- Grant this service account access to project (optional):** This step is currently empty.
- Grant users access to this service account (optional):** This step is currently empty.

At the bottom of the wizard, there are 'DONE' and 'CANCEL' buttons.

## A new service account is created

The screenshot shows the 'Service accounts' page in the Google Cloud IAM & Admin console. The left sidebar lists navigation options, with 'Service Accounts' selected. The main content area displays a table of service accounts for project 'csci5410-assignment-4'.

**Service accounts for project "csci5410-assignment-4"**

A service account represents a Google Cloud service identity, such as code running on Compute Engine VMs, App Engine apps, or systems running outside Google. [Learn more about service accounts.](#)

Organization policies can be used to secure service accounts and block risky service account features, such as automatic IAM Grants, key creation/upload, or the creation of service accounts entirely. [Learn more about service account organization policies.](#)

Filter	Email	Status	Name	Description	Key ID	Key creation date	OAuth 2 Client ID	Actions
	csci-5410-assignment-4@csci5410-assignment-4.iam.gserviceaccount.com	✓	csci-5410-assignment-4	Service Account for Assignment 4	No keys	109685263419164228927		

## Generating a key to establish a connection to GCP with JAVA SDK

Google Cloud csci5410-assignment-4 Search Products, resources, docs (/)

IAM & Admin csci-5410-assignment-4

DETAILS PERMISSIONS KEYS METRICS LOGS

### Keys

Service account keys could pose a security risk if compromised. We recommend you avoid downloading service account keys and instead use the [Workload Identity Fe here](#).

Add a new key pair or upload a public key certificate from an existing key pair.

Block service account key creation using [organization policies](#).  
[Learn more about setting organization policies for service accounts](#)

ADD KEY ▾

Type	Status	Key	Key creation date	Key expiration date	
	Active	0356bf0945f38f4d56bfed730a2d1c3a3ad54ab4	Jul 9, 2022	Dec 31, 9999	

Left sidebar: IAM, Identity & Organization, Policy Troubleshooter, Policy Analyzer, Organization Policies, Service Accounts, Workload Identity Federat..., Labels, Tags, Settings, Privacy & Security, Identity-Aware Proxy, Roles

## Creating a Cloud Function with a Cloud Storage bucket as a Trigger

Google Cloud csci5410-assignment-4 Search cloud f

Cloud Functions Create function

### Basics

Environment: 1st gen

Function name \*: generateVector

Region: us-central1

### Trigger

**Cloud Storage**

Trigger type: Cloud Storage

Event type \*: On (finalizing/creating) file in the selected bucket

Bucket \*: sourcedatab00899629 [BROWSE](#)

☐ Retry on failure

[SAVE](#) [CANCEL](#)

Runtime, build, connections and security settings

**Try the new Cloud Functions 2nd gen!**

This next-generation of our Function-as-a-Service product comes with an advanced feature set giving you more powerful infrastructure, advanced control over performance and scalability, more control around the function's runtime, and triggers from over 90+ event sources via Cloud Audit Logs. [Learn more](#)

[MORE](#)

## New Cloud Function on GCP

(...) Cloud Functions Functions [CREATE FUNCTION](#) [REFRESH](#)

Filter Filter functions

<input type="checkbox"/>	Environment	Name ↑	Region	Trigger	Runtime	Memory allocated	Executed function	Last deployed	Authentication ?	Actions
<input type="checkbox"/>	1st gen	generateVector	us-central1	HTTP	Python 3.9	256 MB	hello_world	Jul 9, 2022, 7:36:26 PM		⋮

## Snapshot of Cloud Storage dashboard with no Buckets

Google Cloud csci5410-assignment-4

Search Products, resources, docs (/)

Cloud Storage

Browser [CREATE BUCKET](#) [DELETE](#) [REFRESH](#)

Filter Filter buckets

<input type="checkbox"/>	Name ↑	Created	Location type	Location	Default storage class ?	Last modified ?	Public access ?	Access control ?	Protection ?	Lifecycle rules ?	Labels ?	Requester P
No rows to display												

Store and retrieve your data

Get started by creating a bucket – a container where you can organize and control access to your data and files in Cloud Storage.

[CREATE BUCKET](#) [TAKE QUICKSTART](#)

## Driver Code for Creating Buckets

Assignment\_4 [CSCI 5410\Assignments\CSCI-5410-S2022-B00899629-Benny\_Daniel\Assignment - 4\Assignment\_4] - CreateBuckets.java

DRIVER

src > main > java > CreateBuckets > main

Project C:\Users\benny\OneDrive\Documents\Assignment\_4

```

9      String sourceBucket = "sourcedatab00899629";
10     String trainBucket = "traindatab00899629";
11     String testBucket = "testdatab00899629";
12     File parentDirectory = new File( pathname: "./src/data/Train");
13     File[] trainDataset = parentDirectory.listFiles();
14
15     try {
16         CreateBucketWithStorageClassAndLocation.createBucketWithStorageClassAndLocation(projectId, sourceBucket, trainDataset);
17         CreateBucketWithStorageClassAndLocation.createBucketWithStorageClassAndLocation(projectId, trainBucket);
18         CreateBucketWithStorageClassAndLocation.createBucketWithStorageClassAndLocation(projectId, testBucket);
19     } catch (IOException e) {
20         System.out.println(e);
21     }
22
23

```

Run: Driver

```

"C:\Program Files\jdk-17_windows-x64_bin\jdk-17.0.3.1\bin\java.exe" ...
Created bucket traindatab00899629
Created bucket testdatab00899629
Process finished with exit code 0

```

## Buckets for Source, Train and Test Files

Google Cloud

csci5410-assignment-4

Search

Products, resources, docs (/)

1

2

3

4

Cloud Storage

CREATE BUCKET

DELETE

REFRESH

HELP ASSISTANT

SHOW INFO PANEL

Browser

Monitoring

Settings

Filter

Filter buckets

<input type="checkbox"/>	Name <span>↑</span>	Created	Location type	Location	Default storage class <span>?</span>	Last modified <span>?</span>	Public access <span>?</span>	Access control <span>?</span>	Protection <span>?</span>	Lifecycle <span>?</span>
<input type="checkbox"/>	sourcedatab00899629	Jul 9, 2022, 6:55:05 PM	Multi-region	us	Standard	Jul 9, 2022, 6:55:05 PM	Subject to object ACLs	Fine-grained	None	None
<input type="checkbox"/>	testdatab00899629	Jul 9, 2022, 7:03:31 PM	Multi-region	us	Standard	Jul 9, 2022, 7:03:31 PM	Subject to object ACLs	Fine-grained	None	None
<input type="checkbox"/>	traindatab00899629	Jul 9, 2022, 7:03:30 PM	Multi-region	us	Standard	Jul 9, 2022, 7:03:30 PM	Subject to object ACLs	Fine-grained	None	None

## Empty Bucket for Train data

Cloud Storage	Bucket details	REFRESH	HELP ASSISTANT	LEARN
Browser	traindatab00899629			
Monitoring	Location	Storage class	Public access	Protection
Settings	us (multiple regions in United States)	Standard	Subject to object ACLs	None
	OBJECTS	CONFIGURATION	PERMISSIONS	PROTECTION
	LIFECYCLE			
	Buckets > traindatab00899629			
	UPLOAD FILES	UPLOAD FOLDER	CREATE FOLDER	MANAGE HOLDS
	DELETE	DOWNLOAD		
	Filter by name prefix only Filter objects and folders Show deleted data			
	Name	Size	Type	Created ?
				Storage class
				Last modified
				Public access ?
	No rows to display			

## Empty Bucket for Test data

Google Cloud csci5410-assignment-4	Search cloud	6	?	
Cloud Storage	Bucket details	REFRESH	HELP ASSISTANT	LEARN
Browser	testdatab00899629			
Monitoring	Location	Storage class	Public access	Protection
Settings	us (multiple regions in United States)	Standard	Subject to object ACLs	None
	OBJECTS	CONFIGURATION	PERMISSIONS	PROTECTION
	LIFECYCLE			
	Buckets > testdatab00899629			
	UPLOAD FILES	UPLOAD FOLDER	CREATE FOLDER	MANAGE HOLDS
	DELETE	DOWNLOAD		
	Filter by name prefix only Filter objects and folders Show deleted data			
	Name	Size	Type	Created ?
				Storage class
				Last modified
				Public access ?
	No rows to display			

## Upload “Train” Files to the Source Bucket

```

15 public class CreateBucketWithStorageClassAndLocation {
16     1 usage
17     @
18     public static void createBucketWithStorageClassAndLocation(String projectId, String bucketName, File[] dataset) throws IOException {
19         Credentials credentials = GoogleCredentials.fromStream(new FileInputStream("src/csci5410-assignment-4-0356bf0945f3.json"));
20         Storage storage = StorageOptions.newBuilder().setCredentials(credentials).setProjectId(projectId).build().getService();
21         Bucket bucket = storage.create(BucketInfo.newBuilder(bucketName).build());
22         System.out.println("Created bucket " + bucket.getName());
23
24         for(File file : dataset)
25         {
26             BlobId blobId = BlobId.of(bucketName, file.getName());
27             BlobInfo blobInfo = BlobInfo.newBuilder(blobId).build();
28             storage.create(blobInfo, Files.readAllBytes(file.toPath()));
29         }
30     }
31 }

```

Run: Driver x

```

File - 291.txt has been uploaded to bucket - sourcedatab00899629
File - 292.txt has been uploaded to bucket - sourcedatab00899629
File - 293.txt has been uploaded to bucket - sourcedatab00899629
File - 294.txt has been uploaded to bucket - sourcedatab00899629
File - 295.txt has been uploaded to bucket - sourcedatab00899629
File - 296.txt has been uploaded to bucket - sourcedatab00899629
File - 297.txt has been uploaded to bucket - sourcedatab00899629
File - 298.txt has been uploaded to bucket - sourcedatab00899629
File - 299.txt has been uploaded to bucket - sourcedatab00899629

```

Process finished with exit code 0

## Source Bucket with Files

Google Cloud

csci5410-assignment-4

Search

Products, resources, docs (/)

Cloud Storage

Browser

Monitoring

Settings

Bucket details

274.txt

2.2 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

275.txt

2.9 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

276.txt

3.2 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

277.txt

3.5 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

278.txt

2.7 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

279.txt

2.3 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

280.txt

3.3 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

281.txt

4.5 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

282.txt

3.3 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

283.txt

4.2 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

284.txt

975 B

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

285.txt

4.3 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

286.txt

4.8 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

287.txt

1.9 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

288.txt

4.6 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

289.txt

2.4 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

290.txt

1.9 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

291.txt

4.1 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

292.txt

2.4 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

293.txt

4 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

294.txt

4.5 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

295.txt

3.7 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

296.txt

3.8 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

297.txt

2.4 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

298.txt

2.5 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

299.txt

4.2 KB

application/octet-stream

Jul 9, 2022, 6...

Standard

Jul 9, 2022, 6...

Not public

—

Google-managed key

—

None

Rows per page: 50

251 ~ 299 of 299

## Automatically generated CSV file

Google Cloud csci5410-assignment-4

Search Products, resources, docs (/)

Cloud Storage

Bucket details

traindatab00899629

Location: us (multiple regions in United States) | Storage class: Standard | Public access: Subject to object ACLs | Protection: None

OBJECTS CONFIGURATION PERMISSIONS PROTECTION LIFECYCLE

Buckets > traindatab00899629

UPLOAD FILES UPLOAD FOLDER CREATE FOLDER MANAGE HOLDS DOWNLOAD DELETE

Filter by name prefix only Filter Filter objects and folders

Show deleted data

Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	Encryption	Retention expiration date	Holds
trainVector.csv	1.3 MB	text/csv	Jul 9, 2022, 9:15:42 PM	Standard	Jul 9, 2022, 9:15:42 PM	Not public	—	Google-managed key	—	None

## Enabling API for Using Managed Notebooks

Google Cloud csci5410-assignment-4

Search notebook

APIs & Services

Enabled APIs & services

Library

Credentials

OAuth consent screen

Domain verification

Page usage agreements

API/Service Details

DISABLE API

Notebooks API

Notebooks API is used to manage notebook resources in Google Cloud.

By Google Enterprise API

Service name: notebooks.googleapis.com | Type: Public API | Status: Enabled

LEARN MORE

METRICS QUOTAS CREDENTIALS COST

Select Graphs: 4 Graphs

Filters: Versions: v1 and v1beta1 | Credentials: App Engine default service... | Methods: 72 options selected

Traffic by response code

No data is available for the selected time frame.

Errors by API method

## Creating a new Notebook on Vertex AI

Vertex AI

Workbench

NEW NOTEBOOK

SHOW INFO PANEL

LEARN

MANAGED NOTEBOOKS

USER-MANAGED NOTEBOOKS

EXECUTIONS

SCHEDULES

Notebooks have JupyterLab 3 pre-installed and are configured with GPU-enabled machine learning frameworks. [Learn more](#)

Filter Enter property name or value

Notebook name	Zone	Auto-upgrade	Environment
assignment-4	us-west1-b	—	

Vertex AI

Dashboard

Datasets

Features

Labeling tasks

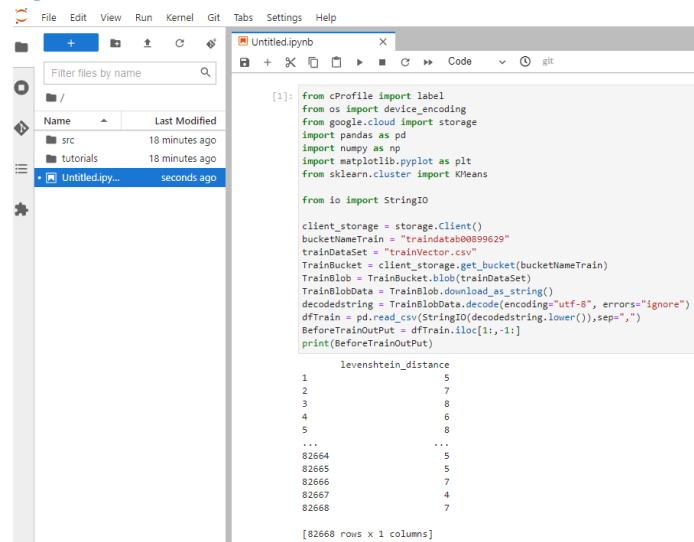
Workbench

Pipelines

Training

Experiments

## Peek at the Dataframe



The screenshot shows a Jupyter Notebook environment. On the left, a file explorer sidebar displays a directory structure with files like 'src', 'tutorials', and 'Untitled.ipynb'. The main area shows the code for loading data from Google Cloud Storage and calculating Levenshtein distances. The output of the code is a single column of Levenshtein distance values for 82668 rows.

```
[1]: from cProfile import label
from os import device encoding
from google.cloud import storage
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

from io import StringIO

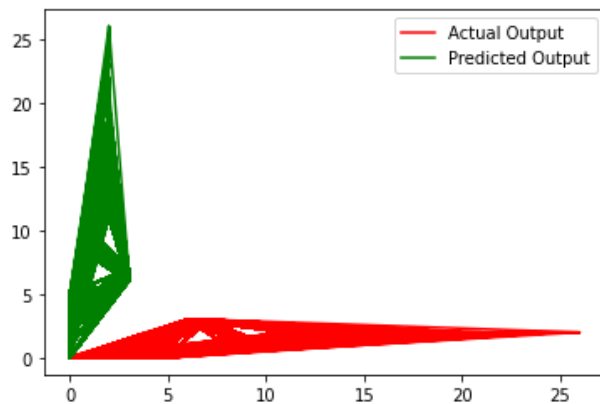
client_storage = storage.Client()
bucketNameTrain = "traindata00099629"
trainDataSet = "trainVector.csv"
TrainBucket = client_storage.get_bucket(bucketNameTrain)
TrainBlob = TrainBucket.blob(trainDataSet)
TrainBlobData = TrainBlob.download_as_string()
decodedstring = TrainBlobData.decode(encoding="utf-8", errors="ignore")
dfTrain = pd.read_csv(StringIO(decodedstring.lower()),sep=";")
BeforeTrainOutPut = dfTrain.iloc[1:,1:]
print(BeforeTrainOutPut)

levenshtein_distance
1      5
2      7
3      8
4      6
5      8
...
82664  5
82665  5
82666  7
82667  4
82668  7

[82668 rows x 1 columns]
```

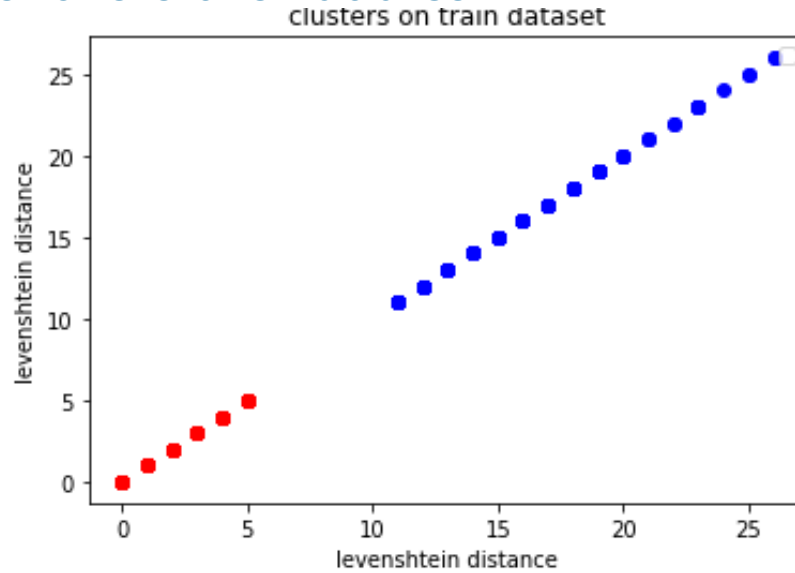
## Visualize the Plot for the Training Dataset

```
[9]: plt.plot(BeforeTrainOutPut, TrainOutput , label = "Actual Output",color="red")
plt.plot( TrainOutput,BeforeTrainOutPut , label = "Predicted Output",color="green")
plt.legend()
plt.show()
```





## Visualize the cluster for levenshtein distance



## Upload Files to the “Test” Bucket

```

Assignment_4 > src > main > java > CreateBuckets > main
CreateBucketWithStorageClassAndLocation.java x CreateBuckets.java x
10 String trainBucket = "trainatab00899629";
11 String testBucket = "testatab00899629";
12 File parentDirectory = new File( pathname: "./src/data/Train");
13 File[] trainDataset = parentDirectory.listFiles();
14
15 File testParentDirectory = new File( pathname: "./src/data/Test");
16 File[] testDataset = testParentDirectory.listFiles();
17
18 try {
19     CreateBucketWithStorageClassAndLocation.createBucketWithStorageClassAndLocation(projectId, sourceBucket, testDataset);
20 } catch (IOException e) {
21     System.out.println(e);
22 }
23
24
Run: Driver x
File - 393.txt has been uploaded to bucket - sourcedatab00899629
File - 394.txt has been uploaded to bucket - sourcedatab00899629
File - 395.txt has been uploaded to bucket - sourcedatab00899629
File - 396.txt has been uploaded to bucket - sourcedatab00899629
File - 397.txt has been uploaded to bucket - sourcedatab00899629
File - 398.txt has been uploaded to bucket - sourcedatab00899629
File - 399.txt has been uploaded to bucket - sourcedatab00899629
File - 400.txt has been uploaded to bucket - sourcedatab00899629
File - 401.txt has been uploaded to bucket - sourcedatab00899629
Process finished with exit code 0
  
```

## Empty Bucket for Test data

Google Cloud csci5410-assignment-4 Search cloud

Cloud Storage

Browser

Monitoring

Settings

Marketplace

Release Notes

Bucket details

REFRESH HELP ASSISTANT LEARN

Name	Size	Content type	Created	Actions
382.txt	3.5 KB	application/octet-stream	Jul 9, 202...	Download
383.txt	4 KB	application/octet-stream	Jul 9, 202...	Download
384.txt	3.7 KB	application/octet-stream	Jul 9, 202...	Download
385.txt	1.8 KB	application/octet-stream	Jul 9, 202...	Download
386.txt	4.4 KB	application/octet-stream	Jul 9, 202...	Download
387.txt	1.9 KB	application/octet-stream	Jul 9, 202...	Download
388.txt	3.5 KB	application/octet-stream	Jul 9, 202...	Download
389.txt	3.8 KB	application/octet-stream	Jul 9, 202...	Download
390.txt	2.1 KB	application/octet-stream	Jul 9, 202...	Download
391.txt	4.8 KB	application/octet-stream	Jul 9, 202...	Download
392.txt	2 KB	application/octet-stream	Jul 9, 202...	Download
393.txt	2.8 KB	application/octet-stream	Jul 9, 202...	Download
394.txt	4.8 KB	application/octet-stream	Jul 9, 202...	Download
395.txt	4.8 KB	application/octet-stream	Jul 9, 202...	Download
396.txt	5.9 KB	application/octet-stream	Jul 9, 202...	Download
397.txt	2.5 KB	application/octet-stream	Jul 9, 202...	Download
398.txt	2.2 KB	application/octet-stream	Jul 9, 202...	Download
399.txt	6.1 KB	application/octet-stream	Jul 9, 202...	Download
400.txt	2.3 KB	application/octet-stream	Jul 9, 202...	Download

Rows per page: 100 301 – 400 of 401

## Automatically Generated CSV file for Test Data

Google Cloud csci5410-assignment-4 Search Products, resources, docs (/)

Cloud Storage

Browser

Monitoring

Settings

Bucket details

testdatab00899629

Location: us (multiple regions in United States)

Storage class: Standard

Public access: Subject to object ACLs

Protection: None

OBJECTS CONFIGURATION PERMISSIONS PROTECTION LIFECYCLE

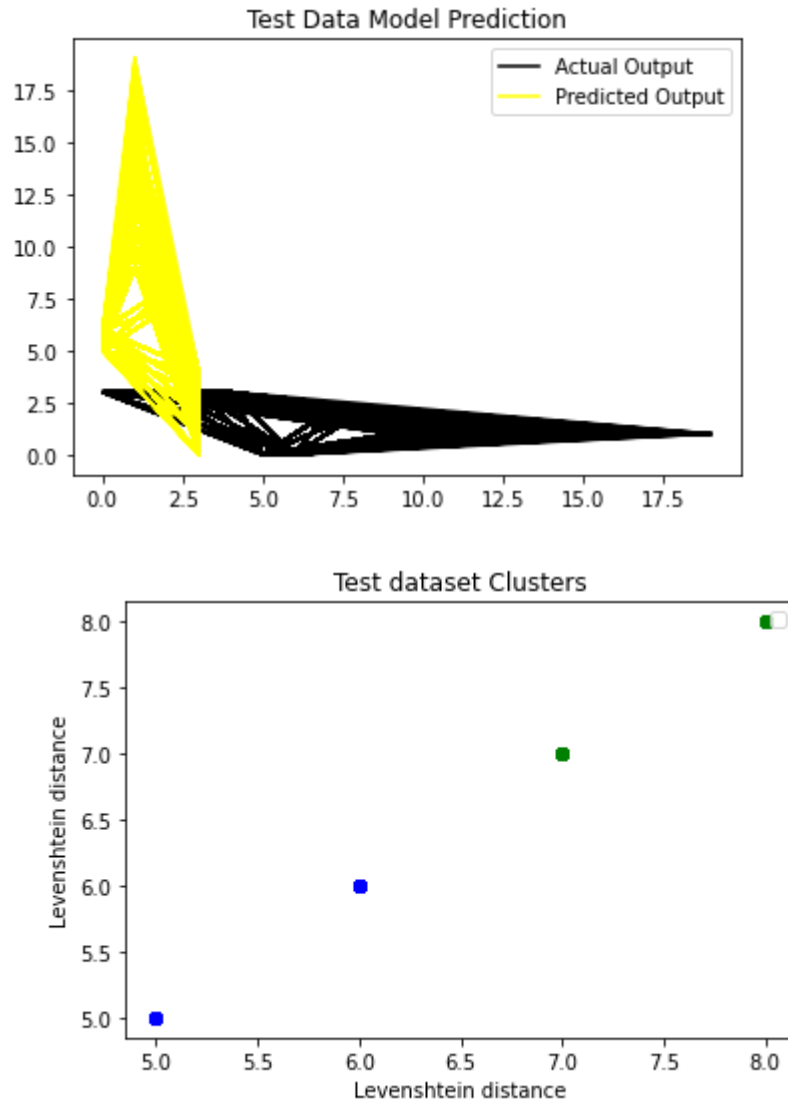
Buckets > testdatab00899629

UPLOAD FILES UPLOAD FOLDER CREATE FOLDER MANAGE HOLDS DOWNLOAD DELETE

Filter by name prefix only Filter objects and folders

Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	Encryption
testVector.csv	531.4 KB	text/csv	Jul 9, 2022, 9:17:25 PM	Standard	Jul 9, 2022, 9:17:25 PM	Not public	—	Google-ma

## Empty Bucket for Test data



## Code Blocks

### Create Buckets (Driver Class) [1-2]

```
import java.io.File;
import java.io.IOException;

public class CreateBuckets
{
    public static void main(String[] args)
    {
        String projectId = Credentials.PROJECT_ID;
        String sourceBucket = "sourcedatab00899629";
        String trainBucket = "trainatab00899629";
        String testBucket = "testatab00899629";
        File parentDirectory = new File("./src/data/Train");
        File[] trainDataset = parentDirectory.listFiles();

        File testParentDirectory = new File("./src/data/Test");
```

```

File[] testDataset = testParentDirectory.listFiles();

try {
CreateBucketWithStorageClassAndLocation.createBucketWithStorageClassAndLocation(projectId
,sourceBucket,trainDataset);
CreateBucketWithStorageClassAndLocation.createBucketWithStorageClassAndLocation(projectId
,trainBucket);
CreateBucketWithStorageClassAndLocation.createBucketWithStorageClassAndLocation(projectId
,sourceBucket);
    } catch (IOException e) {
        System.out.println(e);
    }
}
}

```

## Create Buckets and Upload Files (Driver Class) [\[1-2\]](#)

```

import com.google.auth.Credentials;
import com.google.auth.oauth2.GoogleCredentials;
import com.google.cloud.storage.BlobId;
import com.google.cloud.storage.BlobInfo;
import com.google.cloud.storage.Storage;
import com.google.cloud.storage.StorageOptions;
import java.io.File;
import java.io.FileInputStream;
import java.io.IOException;
import java.nio.file.Files;
import com.google.cloud.storage.BucketInfo;
import com.google.cloud.storage.Bucket;

public class CreateBucketWithStorageClassAndLocation {
    public static void createBucketWithStorageClassAndLocation(String projectId, String
bucketName, File[] dataset) throws IOException
    {
        Credentials credentials = GoogleCredentials.fromStream(new
FileInputStream("src/csci5410-assignment-4-0356bf0945f3.json"));
        Storage storage =
StorageOptions.newBuilder().setCredentials(credentials).setProjectId(projectId).build().g
etService();
        Bucket bucket = storage.create(BucketInfo.newBuilder(bucketName).build());
        System.out.println("Created bucket " + bucket.getName());
int count=0;
for(File file : dataset)
{
    if(count<10) {
        BlobId blobId = BlobId.of(bucketName, file.getName());
        BlobInfo blobInfo = BlobInfo.newBuilder(blobId).build();
        storage.create(blobInfo, Files.readAllBytes(file.toPath()));
        System.out.println("File - " + file.getName() + " has been uploaded to
bucket - " + bucketName);
        count++;
    }
    else {
        break;
    }
}
}

```

```

    }

    public static void createBucketWithStorageClassAndLocation(String projectId, String
bucketName) throws IOException
    {
        Credentials credentials = GoogleCredentials.fromStream(new
FileInputStream("src/csci5410-assignment-4-0356bf0945f3.json"));
        Storage storage =
StorageOptions.newBuilder().setCredentials(credentials).setProjectId(projectId).build().g
etService();
        Bucket bucket = storage.create(BucketInfo.newBuilder(bucketName).build());
        System.out.println("Created bucket " + bucket.getName());
    }
}

```

## Cloud Function to extract words and Levenshtein distance [3-5]

```

file = event
bucket = storage_client.get_bucket(file['bucket'])
uploadbucket = storage_client.get_bucket('traindata00899629');
print('traindata00899629')
blob = bucket.blob(file['name'])
contents = blob.download_as_string()

decodedstring = contents.decode(encoding="utf-8", errors="ignore")
decodedstring = decodedstring.replace("\n", " ")
words = decodedstring.split(" ")
length = len(words)
uploadCsv = 'trainVector.csv'
StoreInCsv = ""
blob = uploadbucket.blob(uploadCsv)

```

```

def distance(s1,s2):
    if len(s1) > len(s2):
        s1,s2 = s2,s1
    l1 = len(s1) + 1
    l2 = len(s2) + 1
    dp = {}
    for i in range(l1):
        dp[i,0] = i
    for j in range(l2):
        dp[0,j] = j
    for i in range(1,l1):
        for j in range(1,l2):
            cost = 0 if s1[i - 1] == s2[j - 1] else 1
            dp[i,j] = min(dp[i,j - 1] + 1, dp[i - 1,j] + 1, dp[i - 1,j - 1] + cost)
    return dp[l1 - 1,l2 - 1]

```

```

for i in range(length - 1):
    s1 = re.sub('[^A-Za-z]+', '', words[i])
    s2 = re.sub('[^A-Za-z]+', '', words[i + 1])
    if (s1.lower() not in stopWords) and (s2.lower() not in stopWords):
        distance = distance(s1.lower(), s2.lower())

```

```
StoreInCsv += s1 + "," + s2 + "," + str(distance) + '\n'
print(StoreInCsv)
```

## K-Means Cluster and Plots [6-7]

```
client_storage = storage.Client()
bucketNameTrain = "traindata00899629"
trainDataSet = "trainVector.csv"
TrainBucket = client_storage.get_bucket(bucketNameTrain)
TrainBlob = TrainBucket.blob(trainDataSet)
TrainBlobData = TrainBlob.download_as_string()
decodedstring = TrainBlobData.decode(encoding="utf-8", errors="ignore")
dfTrain = pd.read_csv(StringIO(decodedstring.lower()), sep=",")
BeforeTrainOutPut = dfTrain.iloc[1:,-1:]
print(BeforeTrainOutPut)
```

```
kmeans = KMeans(n_clusters=4, init='k-means++', max_iter=300)
TrainOutput = kmeans.fit_predict(BeforeTrainOutPut)
print(TrainOutput)

plt.plot(BeforeTrainOutPut, TrainOutput, label = "Actual Output")
plt.plot( TrainOutput, BeforeTrainOutPut, label = "Predicted Output")
plt.legend()
plt.show()
```

```
plt.scatter(BeforeTrainOutPut[TrainOutput == 0], BeforeTrainOutPut[TrainOutput == 0], color =
'red')
plt.scatter(BeforeTrainOutPut[TrainOutput == 2], BeforeTrainOutPut[TrainOutput == 2], color =
'blue')
plt.scatter(BeforeTrainOutPut[TrainOutput == 4], BeforeTrainOutPut[TrainOutput == 4], color =
'green')
plt.scatter(BeforeTrainOutPut[TrainOutput == 6], BeforeTrainOutPut[TrainOutput == 6], color =
'yellow')
plt.show();
```

```
bucketNameTest = "testdata00899629"
testDataSet = "testVector.csv"
TestBucket = client_storage.get_bucket(bucketNameTest)
TestBlob = TestBucket.blob(testDataSet)
TestBlobData = TestBlob.download_as_string()
decodedstring = TestBlobData.decode(encoding="utf-8", errors="ignore")
dfTest = pd.read_csv(StringIO(decodedstring.lower()), sep=",")
BeforeTestOutPut = dfTest.iloc[1:,-1:]

TestOutput = kmeans.fit_predict(BeforeTestOutPut)
```

```
plt.plot(BeforeTestOutPut, TestOutput , label = "Actual Output")
plt.plot( TestOutput,BeforeTestOutPut , label = "Predicted Output")
plt.show()
```

```
cluster_centres = kmeans.cluster_centers_

plt.scatter(BeforeTestOutPut[TestOutput == 0],BeforeTestOutPut[TestOutput == 0],color =
'red')
plt.scatter(BeforeTestOutPut[TestOutput == 2],BeforeTestOutPut[TestOutput == 2],color =
'blue')
plt.scatter(BeforeTestOutPut[TestOutput == 4],BeforeTestOutPut[TestOutput == 4],color =
'green')
plt.scatter(BeforeTestOutPut[TestOutput == 6],BeforeTestOutPut[TestOutput == 6],color =
'yellow')

plt.title("Test dataset Clusters")
plt.xlabel("Levenshtein distance")
plt.ylabel("Levenshtein distance")
plt.legend()
plt.show();
```

## Citations

- [1] “Cloud Storage client libraries,” *Google Cloud*, May 12, 2017. <https://cloud.google.com/storage/docs/reference/libraries> (accessed Jul. 06, 2022).
- [2] “Cloud Storage,” *Google Cloud*, May 26, 2019. <https://cloud.google.com/storage/> (accessed Aug. 13, 2016).
- [3] “Cloud Functions,” *Google Cloud*, Sep. 11, 2019. <https://cloud.google.com/functions> (accessed Jul. 06, 2022).
- [4] “Google Cloud Storage Triggers | Cloud Functions Documentation,” *Google Cloud*, Dec. 14, 2016. <https://cloud.google.com/functions/docs/calling/storage> (accessed Jul. 06, 2022).
- [5] Vatsal, “Text Similarity w/ Levenshtein Distance in Python,” *Medium*, May 15, 2015. <https://towardsdatascience.com/text-similarity-w-levenshtein-distance-in-python-2f7478986e75> (accessed Jul. 07, 2022).
- [6] R. Python, “K-Means Clustering in Python: A Practical Guide – Real Python,” *realpython.com*, Dec. 12, 2012. <https://realpython.com/k-means-clustering-python/> (accessed Nov. 12, 2016).
- [7] “Create a user-managed notebooks instance | Vertex AI Workbench,” *Google Cloud*, May 16, 2016. <https://cloud.google.com/vertex-ai/docs/workbench/user-managed/create-new> (accessed Jul. 04, 2022).