



MCIT-AWS- PDSA- Intake2

Date : 11 – 04 -2022

Data Deduplication – Fuzzy Matching- Use Case 1

Presented by : Abeer Nada

Problem Description:

Use Case Description Input: 1- Neat Dataset1 with some records (~2600). 2- Dataset2 with some records (~64,000). Messy, duplicated. 3- Matching records (labels) (~350).

Task: Using ML, perform de-duplication for the datasets

Keywords:

Data Deduplication

Semantic matching

Record linkage

Entity resolution

Approved Solving Methods:

- 1- Build your own model – using supervised ML Algorithm for clustering
- 2- Fuzzy Score ML Model
- 3- Python Libraries with built-in ML Algorithms
- 4- Using AWS services

Chosen methods for the solution:

- 1- AWS for cleaning and transforming datasets using AWS S3, AWS Glue databrew. (A problem happened with the AWS crawler that prevent from completing the tasks on AWS
See blow screenshots:

us-east-1.console.aws.amazon.com/iam/home#/users\$new?step=final&login&userNames=dldadmin&userNames=dldanlyst&passwordReset&passwordType=autogen&...

Success

You successfully created the users shown below. You can view and download user security credentials. You can also email users instructions for signing in to the AWS Management Console. This is the last time these credentials will be available to download. However, you can create new credentials at any time.

Users with AWS Management Console access can sign-in at: <https://952237412269.signin.aws.amazon.com/console>

Download .csv

	User	Password	Email login instructions
▶	dldadmin	lp5R(j#7&M34d56 Hide	Send email
▶	dldanlyst	mB=Xexl#T(9@%X+ Hide	Send email

Close

Feedback

© 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

new_user_credenti...csv

Show all

Type here to search

25°C 25°C صافى السماء

11:30 PM 4/6/2022

My Drive - Google Drive x DeduplicationUseCase1AbeerNa x S3 Management Console x

s3.console.aws.amazon.com/s3/buckets?region=us-east-1

Amazon S3 Buckets

Account snapshot

Storage lens provides visibility into storage usage and activity trends. [Learn more](#)

View Storage Lens dashboard

Buckets (5) Info

Buckets are containers for data stored in S3. [Learn more](#)

Find buckets by name

	Name	AWS Region	Access	Creation date
<input type="radio"/>	abeernadasecase1bucket	US East (N. Virginia) us-east-1	Objects can be public	April 11, 2022, 01:23:11 (UTC+02:00)
<input type="radio"/>	deduplicationusecase	US East (N. Virginia) us-east-1	Bucket and objects not public	April 9, 2022, 23:30:31 (UTC+02:00)
<input type="radio"/>	dlsourcedata	US East (N. Virginia) us-east-1	Objects can be public	April 6, 2022, 23:46:14 (UTC+02:00)
<input type="radio"/>	jsonbucket3	US East (N. Virginia) us-east-1	Bucket and objects not public	April 10, 2022, 21:54:15 (UTC+02:00)
<input type="radio"/>	jsonuniondatasets	US East (N. Virginia) us-east-1	Bucket and objects not public	April 10, 2022, 17:54:41 (UTC+02:00)

Feedback

© 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Type here to search

16°C مشمس

6:55 AM 4/11/2022

My Drive - Google Drive x DeduplicationUseCase1AbeerNa x dlsourcedata - S3 bucket x +

s3.console.aws.amazon.com/s3/buckets/dlsourcedata?region=us-east-1&tab=objects

AVIndustry IOT Fintech GovJobs 30DaysChallenge-K... VIP Movies DS&AI d-tools SW [SOLVED] HP Lapt... Official HP® Driver... CISCO Resume Builder · Re...

aws Services Search for services, features, blogs, docs, and more [Alt+S] Global AberNada

Amazon S3

Buckets

- Access Points
- Object Lambda Access Points
- Multi-Region Access Points
- Batch Operations
- Access analyzer for S3

Block Public Access settings for this account

▼ **Storage Lens**

- Dashboards
- AWS Organizations settings

Feature spotlight

► AWS Marketplace for S3

dlsourcedata

Objects Properties Permissions Metrics Management Access Points

Objects (5)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 Inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Upload

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	dataset1Mod.csv	csv	April 8, 2022, 00:39:14 (UTC+02:00)	331.7 KB	Standard
<input type="checkbox"/>	dataset2Mod.csv	csv	April 8, 2022, 00:39:23 (UTC+02:00)	8.2 MB	Standard
<input type="checkbox"/>	JoinedDataSets/	Folder	-	-	-
<input type="checkbox"/>	labels.csv	csv	April 8, 2022, 15:11:04 (UTC+02:00)	13.3 KB	Standard
<input type="checkbox"/>	output/	Folder	-	-	-

Feedback

© 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Type here to search

16°C مشمس 6:55 AM 4/11/2022

My Drive - Google Drive x DeduplicationUseCase1AbeerNa x deduplicationusecase - S3 bucket x +

s3.console.aws.amazon.com/s3/buckets/deduplicationusecase?region=us-east-1&tab=objects

AVIndustry IOT Fintech GovJobs 30DaysChallenge-K... VIP Movies DS&AI d-tools SW [SOLVED] HP Lapt... Official HP® Driver... CISCO Resume Builder · Re...

aws Services Search for services, features, blogs, docs, and more [Alt+S] Global AberNada

Amazon S3

Buckets

- Access Points
- Object Lambda Access Points
- Multi-Region Access Points
- Batch Operations
- Access analyzer for S3

Block Public Access settings for this account

▼ **Storage Lens**

- Dashboards
- AWS Organizations settings

Feature spotlight

► AWS Marketplace for S3

deduplicationusecase

Objects Properties Permissions Metrics Management Access Points

Objects (4)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 Inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	ETLDataSets/	Folder	-	-	-
<input type="checkbox"/>	JoinedDataSets/	Folder	-	-	-
<input type="checkbox"/>	JoinedDataSetsV2/	Folder	-	-	-
<input type="checkbox"/>	RawData/	Folder	-	-	-

Feedback

© 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Type here to search

16°C مشمس 6:56 AM 4/11/2022

My Drive - Google Drive x DeduplicationUseCase1AbeerNa x deduplicationusecase - S3 bucket x

s3.console.aws.amazon.com/s3/buckets/deduplicationusecase?region=us-east-1&prefix=RawData/&showversions=false

AVIndustry IOT Fintech GovJobs 30DaysChallenge-K... VIP Movies DS&AI d-tools SW [SOLVED] HP Lapto... Official HP® Driver... CISCO Resume Builder · Re...

Services Search for services, features, blogs, docs, and more [Alt+S]

Global AberrNada

Amazon S3

Buckets

- Access Points
- Object Lambda Access Points
- Multi-Region Access Points
- Batch Operations
- Access analyzer for S3

Block Public Access settings for this account

Storage Lens

- Dashboards
- AWS Organizations settings

Feature spotlight

AWS Marketplace for S3

RawData/

Copy S3 URI

Objects Properties

Objects (3)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Copy S3 URI Copy URL Download Open Delete Actions Create folder Upload

Find objects by prefix

	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	dataset1Mod09042022.csv	csv	April 9, 2022, 23:36:37 (UTC+02:00)	331.7 KB	Standard
<input type="checkbox"/>	dataset2Mod09042022.csv	csv	April 9, 2022, 23:36:52 (UTC+02:00)	8.2 MB	Standard
<input type="checkbox"/>	labelsMod09042022.csv	csv	April 9, 2022, 23:36:54 (UTC+02:00)	13.3 KB	Standard

Feedback

© 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Type here to search

16°C مشمس 6:56 AM 4/11/2022

My Drive - Google Drive x DeduplicationUseCase1AbeerNa x deduplicationusecase - S3 bucket x

s3.console.aws.amazon.com/s3/buckets/deduplicationusecase?region=us-east-1&prefix=JoinedDataSetsV2/&showversions=false

AVIndustry IOT Fintech GovJobs 30DaysChallenge-K... VIP Movies DS&AI d-tools SW [SOLVED] HP Lapto... Official HP® Driver... CISCO Resume Builder · Re...

Services Search for services, features, blogs, docs, and more [Alt+S]

Global AberrNada

Amazon S3

Buckets

- Access Points
- Object Lambda Access Points
- Multi-Region Access Points
- Batch Operations
- Access analyzer for S3

Block Public Access settings for this account

Storage Lens

- Dashboards
- AWS Organizations settings

Feature spotlight

AWS Marketplace for S3

JoinedDataSetsV2/

Copy S3 URI

Objects Properties

Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Copy S3 URI Copy URL Download Open Delete Actions Create folder Upload

Find objects by prefix

	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	TransformedDataSetsUnionFinal10042022_10Apr2022_1649555636923/	Folder	-	-	-
<input type="checkbox"/>	TransformedDataSetsUnionFinal10042022_10Apr2022_1649555650779/	Folder	-	-	-

Feedback

© 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Type here to search

16°C مشمس 6:56 AM 4/11/2022

My Drive - Google Drive x DeduplicationUseCase1AbeerNa x deduplicationusecase - S3 bucket x +

s3.console.aws.amazon.com/s3/buckets/deduplicationusecase?region=us-east-1&prefix=JoinedDataSetsV2/TransformedDataSetsUnionFinal10042022_10Apr2022_1649555636923/

AVIndustry IOT Fintech GovJobs 30DaysChallenge-K... VIP Movies DS&AI d-tools SW [SOLVED] HP Lapto... Official HP® Driver... CISCO Resume Builder - Re...

aws Services Search for services, features, blogs, docs, and more [Alt+S] Global AbeerNada

Amazon S3 Buckets

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

Access analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

AWS Organizations settings

Feature spotlight

AWS Marketplace for S3

Amazon S3 > Buckets > deduplicationusecase > JoinedDataSetsV2/ > TransformedDataSetsUnionFinal10042022_10Apr2022_1649555636923/

TransformedDataSetsUnionFinal10042022_10Apr2022_1649555636923/ Copy S3 URI

Objects Properties

Objects (1)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Copy S3 URI Copy URL Download Open Delete Actions Create folder Upload

Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	TransformedDataSetsUnionFinal10042022_10Apr2022_1649555636923_part00000.csv	csv	April 10, 2022, 03:54:11 (UTC+02:00)	9.3 MB	Standard

Feedback

Type here to search

© 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

16°C مشمس 6:57 AM 4/11/2022

My Drive - Google Drive x DeduplicationUseCase1AbeerNa x deduplicationusecase - S3 bucket x +

s3.console.aws.amazon.com/s3/buckets/deduplicationusecase?region=us-east-1&prefix=JoinedDataSetsV2/TransformedDataSetsUnionFinal10042022_10Apr2022_1649555650779/

AVIndustry IOT Fintech GovJobs 30DaysChallenge-K... VIP Movies DS&AI d-tools SW [SOLVED] HP Lapto... Official HP® Driver... CISCO Resume Builder - Re...

aws Services Search for services, features, blogs, docs, and more [Alt+S] Global AbeerNada

Amazon S3 Buckets

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

Access analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

AWS Organizations settings

Feature spotlight

AWS Marketplace for S3

Amazon S3 > Buckets > deduplicationusecase > JoinedDataSetsV2/ > TransformedDataSetsUnionFinal10042022_10Apr2022_1649555650779/

TransformedDataSetsUnionFinal10042022_10Apr2022_1649555650779/ Copy S3 URI

Objects Properties

Objects (1)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Copy S3 URI Copy URL Download Open Delete Actions Create folder Upload

Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	TransformedDataSetsUnionFinal10042022_10Apr2022_1649555650779_part00000.json	json	April 10, 2022, 03:54:13 (UTC+02:00)	12.8 MB	Standard

https://s3.console.aws.amazon.com/s3/object/deduplicationusecase?region=us-east-1&prefix=JoinedDataSetsV2/TransformedDataSetsUnionFinal10042022_10Apr2022_1649555650779/TransformedDataSetsUnionFinal10042022_10Apr2022_1649555650779_part00000.json

Type here to search

16°C مشمس 6:57 AM 4/11/2022

Glue Databrew:

My Drive - Google Drive x DeduplicationUseCase1AbeerNa x AWS Glue Databrew x +

us-east-1.console.aws.amazon.com/databrew/home?region=us-east-1#datasets

AVIndustry IOT Fintech GovJobs 30DaysChallenge-K... VIP Movies DS&AI d-tools SW [SOLVED] HP Lapto... Official HP® Driver... CISCO Resume Builder - Re...

aws Services Search for services, features, blogs, docs, and more [Alt+S]

DataBrew > Datasets

Datasets (13) Info View details Create project with this dataset Run data profile Actions Connect new dataset

Find datasets

	Dataset name	Data type	Data profile	Source	Location
<input type="checkbox"/>	FinalAllDatasets	csv	-	S3	s3://deduplicationusecase/JoinedDataSetsV2/TransformedDataSetsUnionFinal10042022_10Apr2022_1649555636923/
<input type="checkbox"/>	TransformedDataSet2	csv	-	S3	s3://deduplicationusecase/ETLDataSets/TransformDataset2_10Apr2022_1649549870267/TransformDataset2_10Apr2022_1649549870267_part00000.csv
<input type="checkbox"/>	TransformedDataSet1	csv	-	S3	s3://deduplicationusecase/ETLDataSets/ETLdataset_10Apr2022_1649549050999/ETLdataset_10Apr2022_1649549050999_part00000.csv
<input type="checkbox"/>	dataset2Mod09042022	csv	-	S3	s3://deduplicationusecase/RawData/dataset2Mod09042

Feedback © 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Type here to search 16°C مشمس 6:58 AM 4/11/2022

My Drive - Google Drive x DeduplicationUseCase1AbeerNa x AWS Glue Databrew x +

us-east-1.console.aws.amazon.com/databrew/home?region=us-east-1#projects

AVIndustry IOT Fintech GovJobs 30DaysChallenge-K... VIP Movies DS&AI d-tools SW [SOLVED] HP Lapto... Official HP® Driver... CISCO Resume Builder - Re...

aws Services Search for services, features, blogs, docs, and more [Alt+S]

DataBrew > Projects

Projects (14) Info Open project View job details View lineage Run job Actions Create project

Find projects

	Project name	Associated dataset	Attached recipe	Jobs	Create date	Created by	In use by	T
<input type="checkbox"/>	TransformedDataSetsUnionFinal10042022	TransformedDataSet1	TransformedDataSetsUnionFinal10042022-recipe	TransformedDataSetsUnionFinal10042022	a day ago April 10, 2022, 3:44:12 am	-	-	-
<input type="checkbox"/>	TransformedDataSetsJoinFinal	TransformedDataSet1	TransformedDataSetsJoinFinal-recipe	-	a day ago April 10, 2022, 3:19:03 am	-	-	-
<input type="checkbox"/>	TransformedDataSetsUnion3	TransformedDataSet2	TransformedDataSetsUnion3-recipe	TransformedDataSetsUnionFinal	a day ago April 10, 2022, 2:59:51 am	-	-	-
<input type="checkbox"/>	TransformedDataSetUnion	TransformedDataSet1	TransformedDataSetUnion-recipe	-	a day ago April 10, 2022, 2:49:04 am	-	-	-

https://us-east-1.console.aws.amazon.com/databrew/home?region=us-east-1#projects © 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Type here to search 16°C مشمس 6:58 AM 4/11/2022

My Drive - Google Drive x DeduplicationUseCase1AbeerNa x AWS Glue DataBrew x +

us-east-1.console.aws.amazon.com/databrew/home?region=us-east-1#project-workspace?project=TransformedDataSetsUnion3&view=grid

AVIndustry IOT Fintech GovJobs 30DaysChallenge-K... VIP Movies DS&AI d-tools SW [SOLVED] HP Lapto... Official HP® Driver... CISCO Resume Builder · Re...

aws Services Search for services, features, blogs, docs, and more [Alt+S]

N. Virginia AberNada

TransformedDataSetsUnion3

Dataset: TransformedDataSet2 Sample: First n sample (500 rows) Last job run a day ago, no job runs scheduled [Run job](#) [JOB DETAILS](#) [LINEAGE](#) [ACTIONS](#)

UNDO REDO FILTER SORT COLUMN FORMAT CLEAN EXTRACT MISSING INVALID DUPLICATES OUTLIERS SPLIT MERGE CREATE FUNCTIONS CONDITIONS NEST-UNNEST PIVOT GROUP JOIN UNION MORE RECIPE

Viewing 6 columns 500 rows [SAMPLE](#) GRID SCHEMA PROFILE

	Unique	Total	Distinct	Unique	Total	Distinct	Unique	Total
ABC authors	500	500	499	498	500	242	197	373
ABC venue								
key Road	1	0.2%	M Klein	2	0.4%	null	127	25.4%
n polystyrene	1	0.2%	QD Inc	1	0.2%	New Directions for Higher Education,	8	1.6%
g is a quantitative meth...	1	0.2%	AS Argon, JG Hannoosh	1	0.2%	SIGMOD Record,	8	1.6%
	497	99.4%	All other values	496	99.2%	All other values	357	71.4%
o Valley Road			QD Inc			San Diego,		
azes in polystyrene			AS Argon, JG Hannoosh			Phil. Mag,		
abelling is a quantitative method as ...			GH Hansen, LL Wetterberg, H Sjöström, O Norén			The Histochemical Journal,		
Infectious Disease Among Inmates a...			TM Hammett, P Harmon, W Rhodes			see		
culty Advising in Science and Engine...			JR Cogdell			NEW DIRECTIONS FOR TEACHING AND LEARNING,		
plicity of linear recurrence sequences			WM Schmidt			to		
ALIDITY OF KINDERGARTEN SCREEN...			RA Haats			null		
hatir Reaction Center			ID Noor, J Daisenhof			San Diego Academic		

Recipe (1)

TransformedDataSetsUnion3-recipe Working version [Publish](#) [More](#)

Applied steps (1) [Clear all](#)

1. Union TransformedDataSet1, TransformedDataSet2

Feedback © 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Type here to search 16°C مشمس 7:24 AM 4/11/2022

My Drive - Google Drive x DeduplicationUseCase1AbeerNa x AWS Glue DataBrew x +

us-east-1.console.aws.amazon.com/databrew/home?region=us-east-1#recipes

AVIndustry IOT Fintech GovJobs 30DaysChallenge-K... VIP Movies DS&AI d-tools SW [SOLVED] HP Lapto... Official HP® Driver... CISCO Resume Builder · Re...

aws Services Search for services, features, blogs, docs, and more [Alt+S]

N. Virginia AberNada

DataBrew > Recipes

What is a recipe

Recipes (3) Info [Download as YAML](#) [Download as JSON](#) [Create job with this recipe](#) [Actions](#) [Upload recipe](#)

[Find recipes](#) [Published](#) [1](#) [Settings](#)

	Recipe name	Version description	Associated projects	Published date	Published by	Tags
<input type="checkbox"/>	DataSetUnion09042022-recipe	-	DataSetUnion09042022	a day ago April 10, 2022, 2:06:29 am	-	-
<input type="checkbox"/>	DataSetsJoinFinalV2-recipe	-	DataSetsJoinFinalV2	2 days ago April 8, 2022, 8:54:03 pm	-	-
<input type="checkbox"/>	DeDeuplicationUseCase4-recipe	V1	DeDeuplicationUseCase4	3 days ago April 8, 2022, 6:52:32 pm	-	-

https://us-east-1.console.aws.amazon.com/databrew/home?region=us-east-1#recipes © 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Type here to search 16°C مشمس 7:25 AM 4/11/2022

us-east-1.console.aws.amazon.com/databrew/home?region=us-east-1#recipe-details?recipe=DataSetsJoinFinalV2-recipe&tab=steps

DataSetsJoinFinalV2-recipe

Project: DataSetsJoinFinalV2

Recipe steps | Data lineage

Recipe details

Recipe name DataSetsJoinFinalV2-recipe	Associated project DataSetsJoinFinalV2	Published date 2 days ago by arn:aws:iam::952237412269:root April 8, 2022, 8:54:03 pm
Version description -	Associated jobs -	

Recipe steps (1)

1. Union deduplicationv7dataset2, ApplyV5-08Apr2022-1649437402801-part00000

us-east-1.console.aws.amazon.com/databrew/home?region=us-east-1#jobs?tab=recipe

Recipe jobs (8) info

Find jobs

Show all

<input type="checkbox"/>	Job name	Status	Job input	Job output	Last run	Created on
<input type="checkbox"/>	TransformedDataSetsUnionFinal10042022	Succeeded	Transformed... (Transformed... + Transformed...) Project Dataset Recipe	2 outputs	a day ago April 10, 2022, 3:54:28 am	a day ago April 10, 2022, 3:50:00 am
<input type="checkbox"/>	TransformedDataSetsUnionFinal	Succeeded	Transformed... (Transformed... + Transformed...) Project Dataset Recipe	2 outputs	a day ago April 10, 2022, 3:06:19 am	a day ago April 10, 2022, 3:04:00 am

- 2- I generated one transformed csv and JSON file from the union of the two datasets (dataset1& dataset2) – please check attached files.

- 3- Using Dedupe Library in a python notebook to apply ML on the output dataset from the previous phase, please check attached notebook and file for the trained data.

```
[1] pip install pandas-dedupe
Downloading simplecosine-1.2-py2.py3-none-any.whl (3.2 kB)
Collecting Levenshtein-search==1.4.5
  Downloading Levenshtein_search-1.4.5-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux2014_x86_64.whl (71 kB)
7.9 MB/s
Collecting dedupe-variable-datetime
  Downloading dedupe_variable_datetime-0.1.5-py3-none-any.whl (4.8 kB)
Collecting categorical-distance>=1.9
  Downloading categorical_distance-1.9-py3-none-any.whl (3.3 kB)
Requirement already satisfied: numpy>=1.13 in /usr/local/lib/python3.7/dist-packages (1.19.5)
Collecting haversine>=0.4.1
  Downloading haversine-2.5.1-py2.py3-none-any.whl (6.1 kB)
Collecting Btrees>=4.1.4
  Downloading Btrees-4.10.0-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux2014_x86_64.whl (3.6 MB)
54.9 MB/s
Collecting zope.index
  Downloading zope.index-5.2.0-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux2014_x86_64.whl (105 kB)
65.9 MB/s
Requirement already satisfied: typing-extensions in /usr/local/lib/python3.7/dist-packages (3.7.4)
Collecting fastcluster
  Downloading fastcluster-1.2.6-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux2014_x86_64.whl (194 kB)
66.7 MB/s
Collecting doublemetaphone
  Downloading DoubleMetaphone-1.1-cp37-cp37m-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux2014_x86_64.whl (150 kB)
51.6 MB/s
Collecting affinegap>=1.3
0s completed at 6:45 AM
```

References:

Using the Dedupe Machine Learning Library for Cleaning and Matching Data

<https://www.azavea.com/blog/2019/08/30/using-the-dedupe-machine-learning-library-for-cleaning-and-matching-data/>

Solving Data- Duplication problem using Machine Learning Algorithm

<https://www.beyondkey.com/blog/solving-data-duplication-problem-using-machine-learning-algorithm/>

<https://stackoverflow.com/questions/16381133/using-machine-learning-to-de-duplicate-data>

Why Machine Learning Is the Smart Way to Dedupe Salesforce

<https://www.salesfix.com.au/blog/why-machine-learning-is-the-smart-way-to-dedupe-salesforce/>

Dedupe - How it works

<https://dedupe.io/documentation/how-it-works.html>

Using Machine Learning to Detect Duplicates: Some Real-Life Examples (Part II)

<https://dzone.com/articles/using-machine-learning-to-detect-dupes-some-real-l>