Dataset1: Contains TripAdvisor Review Details for Attraction in California

Field Name	Description	An Example
attraction_in	Place where attraction is found	Los Angeles
no_of_attractions	Total no of attractions at the place(attraction_in)	519
no_of_reviews	No of Review comments for the attraction	9709
url	url link for TripAdvisor page for the attraction review comment	https://www.tripadvisor.com/Attraction_Review-g32655-d147966-Reviews-The_Getty_Center-Los_Angeles_California.html
reviewComment	One of Review comments for the attraction. There shall be multiple review comments	A Fantastic Way To Spend An Afternoon While visiting LA last month, I had the immense pleasure of visiting The Getty for the first time. The experience was outstanding from beginning to end: from the parking, to... read more Reviewed 2 days ago sox1fan , Weare, New Hampshire
activity_type	Type of Activity to do at the attraction : Boat Tours & Water Sports, Food & Drink, things to do, and so on…	things to do
Rank	Ranking for the attraction	1
max_rank	Overall ranking	505
attraction	Name of attraction	The Getty Center
contact_no	Phone no	13104407300
knownFor	Attraction knownFor tags	Specialty Museums , Museums
Address	Address of the attraction	1200 Getty Center Dr , Los Angeles , CA 90049-1657
pinCode	Pincode of attraction	90049-1657
imp_pinCode	5-digit pincode of attraction	90049

Preparing Dataset2 was quite involved task. As part of final project-Enroute Genie our goal is extract location mentions from blogs, and enrich it with TripAdvisor metadata along with multimedia image and youtube video. Details and scripts used to generate Dataset2 is mentioned README.txt

NOTE: **Dataset2, is not yet complete** with multimedia data. Also **of 630 combination** of location Files, **only 46 combination of location** were processed to collect geo co-ordinates and respective addresses. We can get only limited geo metadata using geopy, each day before our requests are blocked.

Dataset2: location metadata from 46 combinations of location mentions

Field Name	Description	An Example
Name	**Location mention** specified in a blog	getty center
locType	Type of location as extracted from geopy module for the given location mention	museum
address	Address as extracted from geopy module for the given location mention	"J. Paul Getty Museum, 1200, Getty Center Drive, Westgate Heights, Brentwood, LA, Los Angeles County, California, 90049, United States of America"
pinCode	pinCode extracted from address field using TextWrangler	90049

Found a python module **py_stringsimjoin** :
https://sites.google.com/site/anhaidgroup/projects/magellan/py_stringsimjoin

which is part of excellent research efforts by AnHai Group to build an Entity Management System :
https://sites.google.com/site/anhaidgroup/projects/magellan

This project seeks to build a Python software package that provides scalable implementation of string similarity joins over two tables, for commonly used similarity measures such as Jaccard, Dice, cosine, overlap, overlap coefficient and edit distance.

Guide: http://anhaidgroup.github.io/py_stringsimjoin/v0.1.x/singlepage.html

editDistSim.py: script leveraging py_stringmatching python module to compute edit_distance similarity between **Dataset1. Attraction and Dataset2. Name**

Results in: **results_editDistSim.csv**

Package offers multiple advantages : Profilers, tokenizer, Joins, Filters & Matchers, leveraging multi-core processor ability to efficiently perform Entity Matching.

Unfortunately, the package doesn't have a means to consider combinations of multiple fields from two datasets. They are yet to opensource it:
https://sites.google.com/site/anhaidgroup/projects/magellan/py_entitymatching

It was good learning experience.

Resorted back to FRIL.

Dataset1. Attraction and Dataset2. Name : A combination of sequence and set-based measures were used in order to match location mention (Dataset2.Name) with attraction name(Dataset1.Attraction), based on heuristic that people writing blogs may refer to attractions without complete name, say liberty statue, instead of Statue of Liberty.

sequence-based measure: Edit Distance Similarity Measure was used to join the Datasets with

Approve Level : 0.1, Disapprove level : 0.3 (As recommended by FRIL documentation)

Condition weight : **30**

set-based measure : Q-Gram(3-Gram)

Condition weight : **30**

Approve Level : 0.2, Disapprove level : 0.4(Default ones, worked well)

Dataset1. Imp_pinCode and Dataset2. pinCode : Numeric distance similarity measure was used to leverage heuristics from pinCode part of address, to enforce strict matches bw location mentions.

Condition weight : **40**

Sorted neighbourhood method was used

Manual Decision, User intervention was used to enable classification of records whose confidence values lies within 70-80%

84 records were matched with enriched metadata as expected with very intuitive results. i.e fields in datasets, which weren't considered in join conditions i.e knownFor, locType, also **show interesting similarity results**.

The following are interesting observations with apt review comments: **final_results.csv**

attraction@sourceA	knownFor@sourceA	name@sourceB	locType@sourceB	reviewComment@sourceA **(just one reviews shown here for reference**)
Anaheim Convention Center	Conference & **Convention Centers** , Traveler Resources	anaheim convention center	**convention_centre**	One of the best Great, big and perfect for big gatherings. I like the place a lot. Reviewed January 3, 2012 Lamboswede , California
Anaheim **Plaza**	N/A	anaheim plaza	**retail**	Good Shopping Available Here Short journey from the Disneyland area of Anaheim. Always worth a look around Walmart for a bargain or two. Forever 21 seems to be my daughter's favourite shop at the moment... read more Reviewed April 8, 2015

				Simz61 , Ashby de la Zouch, United Kingdom via mobile
California Science Center	Science **Museums** , Museums	california science center	**museum**	Amazing This place is amazing. So much to do for our kids who are aged 4&8. Arrive early as it can get busy (at least on the rainy weekend we went) Reviewed 3 days ago Snowflake000 , Perth via mobile

Evaluation: Our project domain problem doesn't map 1-1 record, our intention is to collect all review comments from dataset1, along with other metadata for location mention names in dataset2, so its m:1 mapping

	Record in dataset1 (if only 1 review record the corresponding row is indicated) else no of review records is shown	Record in dataset2	Record in final_results.csv	Matched?
Anaheim Convention Center	61 review records	dataset2:522	8 review records fetched	8/61
California Science Center	dataset1:8	dataset2:973	final_results:21	Yes
Hollywood Walk of Fame	dataset1:33	dataset2:732	24 review records fetched	Yes
Angel Stadium of Anaheim	107 review records	dataset2:114	8 review records fetched	8/107
Honda Center	39 review records	dataset2:563	8 review records fetched	8/39
Santa Barbara Museum of Art	30 review records	dataset2:986	8 review records fetched	8/30
Santa Barbara Botanic Garden	43 review records	dataset2:857	8 review records fetched	8/43
Diamond Valley Lake	dataset1:23681	Dataset2:859	final_results:51	Yes