Motivation for Dataset Creation

Nicolas Fan

EM 212 Applied Data Science

**Motivation for Dataset Creation**

Why was the dataset created?

Yelp Open Dataset is created for all-purpose learning. The dataset is a subset of Yelp businesses, reviews and user data for use in personal, educational, and academic purposes.

What tasks could the dataset be used for?

The dataset could be used for anything.

Has the dataset been used for any tasks already?

Yelp company encourages students to make use of the dataset to conduct research and analysis and share their discoveries with public. There are 11 round Yelp challenge and 11 winners so far. All the topic and information could be found at https://www.yelp.com/dataset/challenge/winners

Who funded the creation of the dataset?

Funding was provided by Yelp company, a local-search service powered by crowd-sourced review forum.

# Dataset Composition

What are the instances?

The instances are business’s information including location data, attributes, categories and so on.

Are relationships between instances made explicit in the data?

They are all businesses with a physical location which already registered on Yelp platform.

How many instances of each type are there?

There are 192,609 businesses in total at the dataset. Over 1.2 million business attributes like hours, parking, availability and ambience.

What data does each instance consist of?

Each instance consists of name, address, city, state, postal code, latitude, longitude, rating stars, review count, open or not, categories and hours.

Is everything included or does the data rely on external resources?

Everything is included in the dataset.

Are there recommended data splits or evaluation measures?

No. Since the basic data information is number and text, any proper attempts are doable.

What experiments were initially run on this dataset?

The dataset was originally released without any reported experiment but after that yelp challenge encourages researchers to run new experiments.

# Data Collection Process

How was the data collected?

The raw data for this dataset were obtained by the registration of users who are the business owner.

Who was involved in the data collection process?

Those business owners provide this information so that their business could be searched and rated by more people.

Over what time-frame was the data collected?

The data is cumulatively collected.

How was the data associated with each instance acquired?

It’s directly observable.

Does the dataset contain all possible instances?

The dataset does not contain all possible instances.

If the dataset is a sample, then what is the population?

The dataset should cover most of the business currently existing.

Is there information missing from the dataset and why?

Unknown.

Are there any known errors, sources of noise, or redundancies in the data?

For my project’s purpose, since I will only look at Boston area’s Chinese restaurant business, most of the data are irrelevant to my topic.

# Data Preprocessing

What preprocessing/cleaning was done?

Preprocessing done before is unknown. To processing my project, I need to first clean the irrelevant data. The following steps would be taken to process the data:

1. Eliminating businesses outside of Boston area.
2. Select restaurants from business at Boston area
3. Select Chinese style restaurants from restaurants at Boston area

Was the “raw” data saved in addition to the preprocessed/cleaned data?

Unknown.

Is the preprocessing software available?

Unknown.

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet?

Unknown.

# Dataset Distribution

How is the dataset distributed?

The dataset can be downloaded from https://www.yelp.com/dataset/download

When will the dataset be released/ first distributed?

The dataset was released in 2013.

What license is it distributed under?

Unknown.

Are there any fees or access/export restrictions?

No, there’s not.

# Dataset Maintenance

Who is supporting/hosting/maintaining the dataset?

The dataset is hosted at Yelp Company and you can find more information and example at its GitHub page: https://github.com/Yelp/dataset-examples

Will the dataset be updated?

Yes, the dataset usually got update after each round challenge ends.

If the dataset becomes obsolete how will this be communicated?

It will notice the public on its website.

Is there a repository to link to any/all papers/systems that use this dataset?

<https://scholar.google.com/scholar?q=citation%3A+Yelp+Dataset&btnG=&hl=en&as_sdt=0%2C5>

If others want to extend/ augment/build on this dataset, is there a mechanism for them to do so?

Unknown.