BU Spark Project: Heyrick Research and Fuzzy Matching
Deliverable 0

People

Client

Danielle Mendheim	danielle@heyrickresearch.org
Liz Dicus	liz@heyrickresearch.org

Project Managers

Rajat Tripathi	rajat08@bu.edu
Savannah Majarwitz	smajarwi@bu.edu

Team Members

Suzy Kirch	smkirch@bu.edu
Yang Hu	yhu112@bu.edu
Chengyang He	henryhcy@bu.edu
Haoran Kang	hrjkang@bu.edu

Background

Heyrick Research focuses only on the illicit massage industry, attacking it by understanding, disrupting, and defeating the business models of these illicit massage businesses, all by data-driven means. In this project, we are going to develop a data analysis tool to match illicit massage places with information on websites like Google Places and Yelp in order to identify illicit actions still ongoing.

Datasets

Already collected

1. From Google Places: List of all massage businesses (approx. 500k rows), which includes the name, address, and phone number of these businesses
2. From Rubmaps.ch: List of illicit massage businesses (approx. 30k rows), which includes the name, address, and phone number of these businesses

To be collected (using an API)

3. From Yelp: List of massage businesses, including name, address, and phone number of these businesses

Goals

1. Match between Rubmaps and Other Datasets
   → We are to use fuzzy matching to match between the Rubmaps data and Google Places and Yelp.
   → We can have a fuzzy match between addresses and names, but phone numbers should be an exact match.
   → Difficulties include:
      ↳ Phone numbers can be tied to more than one address, so while phone numbers need to be an exact match, it is considered a weak connection.
      ↳ There can be two locations with the same address, as they may have different suite numbers. There are also inconsistencies in the data for how the addresses are listed. That being said, addresses take precedence over phone and business name.
      ↳ Names of businesses are duplicated all over the country.
2. Matching Criterion
   → We are to identify an overall confidence score between potential matches and a confidence score for phone number match, address match, and business name match.
   → We are to determine what thresholds are for determining a match.
3. Create a Dynamic and Scalable Script
   → We do not have access to all of the datasets that will be used over time.
   → We are to create a model that can be used for any set of datasets.

Primary Question:

What is the confidence threshold for a datapoint match?

Sprint Schedule

1. Sprint 1: October 21
   → Gathering data from Yelp
   → Standardize phone numbers, apt/unit/suite
2. Sprint 2: Deliverable 1, October 28
   → Write up report
3. Sprint 3: November 9
   → Preliminary fuzzy matching model: address focused, primarily using Google Places
   → Preliminary match threshold

4. Sprint 4: Deliverable 2, November 23
    - → Full fuzzy matching model
    - → Tested with Yelp data
    - → Solidify match threshold
    - → If time: interactive GUI (map with color-coded location dots)
5. Sprint 5: Deliverable 3, November 30
    - → Draft of final report:
        - ↳ Match threshold breakdown with explanations
        - ↳ Have examples of matches and non-matches
        - ↳ Include code/model
    - → Have report reviewed by Danielle and Liz
6. Sprint 6: Presentation/Final draft, December 7
    - → Turn report into presentation
    - → Finalize report