Robust Entity Resolution Through High Precision Proxy Labels

Chirag Nagpal, Kyle Miller, Benedikt Boecking and Artur Dubrawski

Auton Lab, School of Computer Science, Carnegie Mellon University

Objectives

Development of an entity resolution pipeline to identify sources of advertisements soliciting illegal activites on the internet.

- Utilising strong features like Phone Numbers as a Proxy Ground Truth.
- Training a Binary Classifier as a match function to detect related observations.
- Identifying a match function score to prevent breakdown of ER.
- Developing a heuristic based 'Blocking Scheme' ensuring high Precision rate among connected components.
- Applyling pipeline to escort ad data from 'Backpage.com', to counter human trafficking.

Match Function

To train our 'match function' to predict if two observations in our dataset are related, we train a Random Forest binary classifier on Features like:

- Time difference between posting of ads
- Length of longest common substring.
- Special Characters.
- Frequency, Standard Deviation of common Bi-grams, Uni-grams, Images.
- Features extracted from 'TJBatchExtractor', a RegEx based Information Extractor, like Hair Color, Price, Location etc.

The Match Function is generated using:

1. Resolving entites with Phone Numbers
2. For every phone number, sampling pairs of connected components.
3. Featurising pairs of components to generate positive and negative samples.
4. Training a Random Forest over the samples using Sci-Kit.[1]

References

[1] F. Pedregosa, G. Varoquaux, and A. et. al. Gramfort. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Performance

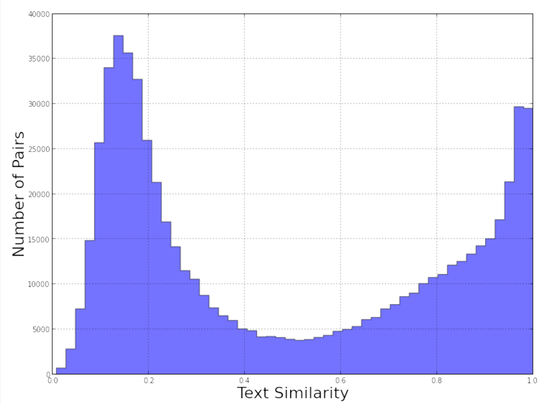

Figure 1: Text Similarty for Sampling Scheme

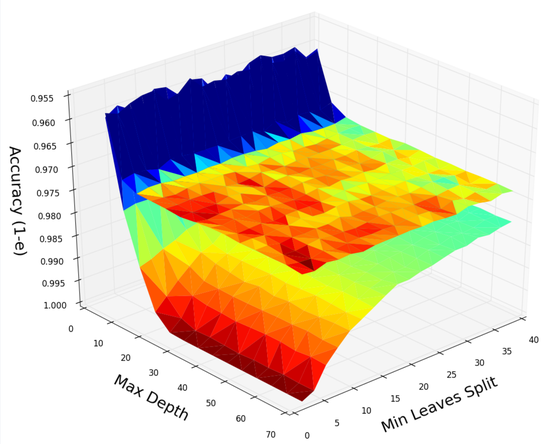

Figure 2: Parameter Tuning for Classifier

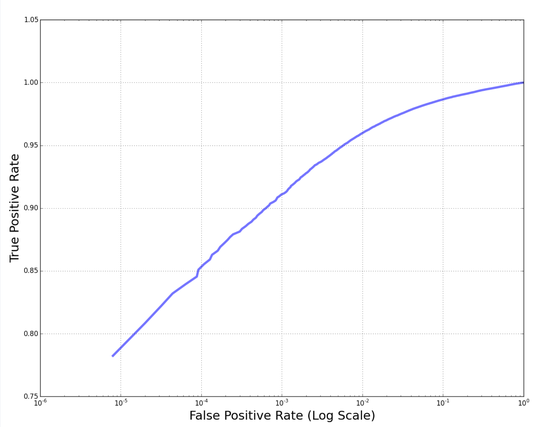

Figure 3: ROC plot for Match Function

Blocking Scheme

In order to prevent, naive n^2 comparisons across the entire corpus, we introduce a blocking scheme. The blocking scheme makes use of strong features, to reduce the size of comparisons to be carried out. Thus,

$$n_1{}^2 + n_2{}^2 + ... + n_m{}^2 << N^2 \qquad (1)$$

where,

n_m is size of block m and m is number of blocks. We use the following features to generate our blocks,

1. Images
2. Bi-Grams.
3. Uni-Grams.

ER over Blocks

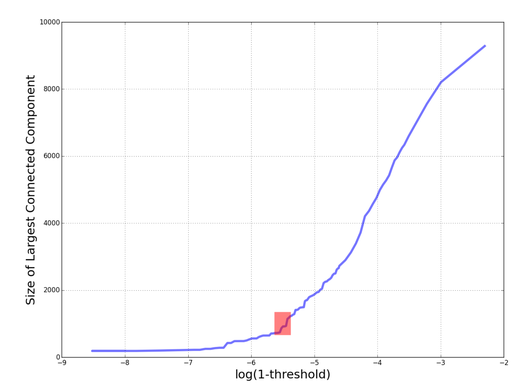

Figure 4: ER Breakdown (Size of Entity)

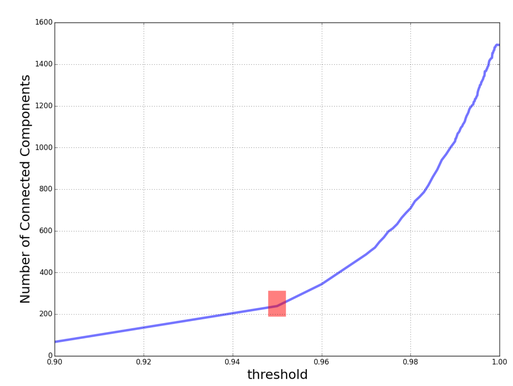

Figure 5: ER Breakdown (Number of Entities)

Representative Examples

[redacted] Couple Ready to Please!!! - 37
Hot New [redacted] Couple Ready for some fun!! One of a kind [redacted] couple ready to fulfill your wildest fantasies!! Lets have some fun tonight!$120hh (her) $200h (her) $200hh (both) $300h (both) Out Calls tonight!!!!! text only!!And please mention couple in your text.

Hot and New!! - 37
Hot new [redacted] couple! Ready to fulfill your wildest fantasies! $120hh (her) 200h (her) 200hh (both) 300h (both) Victoria and NickOut calls only!!!

[redacted] • ⓟⓗⓞⓣⓞ ⓕⓡⓔⓔ • || EXOTIC EROTIC [redacted] x || • [redacted] ⓑⓞⓓⓨ • ||Ready to Bl[redacted]w your mind x _THIS_IS_MY_SEPARATION_STRING_ [redacted] spectacular Little Treat ✘ Hi [redacted] my name is PARIS, I'm 20 years old & I'm looking for some company! I love to meet new people & have fun! & Baby, I promise to always aim for your complete satisfaction!!
5'6□130lbs□B34 ✓realfect party girl! ✓[redacted] ✓wingwoman service! ✓all safe service es! ✓no expertise!!Qv $70 $120HH $200FH NON-NEGOTIABLE□ ✘no explicit talk! ✘no black men! ✘ ♥ respectful & serious gentlemen only!! call now
[redacted] Wenatchee, COME & GET IT

||[redacted] • ⓟⓗⓞⓣⓞ ⓕⓡⓔⓔ • || EXOTIC EROTIC [redacted] || • [redacted] ⓑⓞⓓⓨ • ||
ready to Bl[redacted]w your mind x - 19 _THIS_IS_MY_SEPARATION_STRING_
[redacted] The Lady Of Your Dreams Is ready to relax your body & put your mind at ease as I turn ALL your dreams into A REAL LIFE FANTASY! ! !
*Beautiful play Mate ☆ №1 gentlemen choice available ηow ☆ [redacted] PARIS [509] 205×54O1 OUtCAlls 24/7!! [redacted] [[Genitlemen ☆ # ❶ Choice]]—♥—[[Vip playmate]]—♥—[[Special all Nite]] Tacoma, Ask about REVIEWS ♥
A Provider u ♥

Pensacola ,Classy Slender Beauty Touring Soon -
Pensacola, your upscale fantasy girlfriend is visiting soon! Interested in professional, mature, respectful acquaintances only. Simply visit my personal website to pre verify and book for my upcoming visit! I do welcome and appreciate your warm invitations! My website [redacted] comReserve Me Prefer longer engagements Preferred411 accepted Touring 03/10

10/30 Upscale Exotic Fantasy Companion Visiting Soon - 29
Gentlemen! Advance Bookings Welcome ! Your passion is my pleasure! I possess the wit ,charm and intuitive intellect to provide you with the utmost sensual,passionate ,tender companionship you deserve. Interested in meeting your sincere,polite acquaintances soon! Reserve Me www.sierra-dayna.com Twitter Follow Me! SierraDayna Vip/90 minutes/up engagements preferred Subscribe to my blog and newsletter xo xo xo Sierra Visiting 10/30!

????????????? 2 fiyah [redacted] exotic freaks? ???????????????? - 22 Hello Gentlemen???? ?????????????????????????? ?????????????? U?????li??v??????l?? 5?????????????? Ski?????s ???? Will drive you insane!???? ??? C??m?? 5???? 4 Y??urs?!!? ??? Bubbly, Sweet????, a little Fiery Treat???????? my body welcomes your desire ???? I want to help you relax, take the stress away from your long day. I'm made of ???Pure Delight???, a tasty Dessert.???? An EX??TiC Erotic Dream come true! Come on. ?????????????. K???????? ???????. ?????????? M?????????? black china && Starquality???? & No private/ explicit talking/N?? GAMES (ASK ABOUT THE 2 GIRLS SPECIAL)

PICS STOLEN I AM NOT in baton rouge anymore?????????????? - 22 Hey guys I'm Jasmine MY PICS HAVE BEEN STOLEN BY A GIRL NAMED BLACK CHINA & STAROUALITY THEY LOOK NOTHING LIKE ME SO IF U SEE A AD WITH ANY OF THESE PICS DONT BE FOOLED IM SURE BY NOW U ALREADY KNOW SHE IS CLEARLY NOT THE GIRL IN THE PICS LOL #DISQUALIFED THE AD UNDER 2 FIYAH [redacted] EXOTIC FREAKS HAVE STOLEN PICS OF ME HOW LAME

Future Work

- Isolating connected components and using rule learning to learn differences.
- Improving scalability of the entire system.
- More robust image analysis, to make better use of the images associated with ads.

Acknowledgements

The authors would like to thank Prof. S.R. Dhore, HoD AIT, University of Pune. We would also like to thank the team from CMU that made the Robotics Institute Summer Scholar's Program possible.

Auton Lab

THE ROBOTICS INSTITUTE

Carnegie Mellon University