Yelp Dataset Challenge

Round 9 Of The Yelp Dataset Challenge: Our Largest Yet!
We've had 8 rounds, over $50,000 in cash prizes awarded, hundreds of academic papers written, and we are excited to see round 9.

Our dataset has been updated for this iteration of the challenge - we're sure there are plenty of interesting insights waiting there for you. This set includes information about local businesses in 11 cities across 4 countries.

This round also includes photos! These photos nicely complement reviews, business attributes, check-ins, and tips, and open the door to even more exciting research. An auxiliary file has been provided for download (see the "Get the Data" link on this page), containing 200,000 pictures from 85,901 businesses described in the main dataset. The photo archive includes a json file linking each photo to its corresponding business in the dataset, and listing its caption (if any), and type of content as determined by our image classifier (we currently only list labels for some restaurants).

This treasure trove of local business data is waiting to be mined and we can't wait to see you push the frontiers of data science research with our data.

The Challenge Dataset:

- **4.1M** reviews and **947K** tips by **1M** users for **144K** businesses
- **1.1M** business attributes, e.g., hours, parking availability, ambience.
- Aggregated check-ins over time for each of the **125K** businesses
- **200,000** pictures from the included businesses

Cities:

- U.K.: Edinburgh
- Germany: Karlsruhe
- Canada: Montreal and Waterloo
- U.S.: Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison, Cleveland

The Challenge

Not only would we like to give you our data, we'd also like to announce the ninth round of the **Yelp Dataset Challenge**. We challenge students to use this data in an innovative way and break ground in research. Here are some examples of topics we find interesting, but remember these are only to get you thinking and we welcome novel approaches!

Cultural Trends: By adding a diverse set of cities, we want participants to

The Awards

If you are a student and come up with an appealing project, you'll have the opportunity to win one of ten Yelp Dataset Challenge awards for $5,000. Yes, that's $5,000 for showing us how you use our data. We'll judge submissions on their technical depth and rigor, the relevance of the results to Yelp, our users, or the field, and finally their novelty, uniqueness, and yes, their Yelpy-ness. Please note: if contest participants work with professors or other non-student advisors to craft a submission, only the

compare and contrast what makes a particular city different. For example, are people in international cities less concerned about driving to a business, indicated by their lack of mention about parking? What cuisines do Yelpers rave about in these different countries? Do Americans tend to eat out late compared to those in Germany or the U.K.? In which countries are Yelpers sticklers for service quality? In international cities such as Montreal, are French speakers reviewing places differently than English speakers?

Location Mining and Urban Planning: How much of a business' success is really just location, location, location? Do you see reviewers' behavior change when they travel?

Seasonal Trends: What about seasonal effects: Are HVAC contractors being reviewed mainly during winter, and manicure salons over the summer? Are there more reviews for sports bars on major game days and if so, could you predict that?

Infer Categories: Do you see any non-intuitive correlations between business categories e.g., how many karaoke bars also offer Korean food? What businesses deserve their own subcategory (i.e., Szechuan or Hunan versus just "Chinese restaurants"), and can you learn this from the review text?

Natural Language Processing (NLP): How well can you guess a review's rating from its text alone? What are the most common positive and negative words used in our reviews? Do Yelpers typically use sarcasm And what kinds of correlations do you see between tips and reviews: could you extract tips from reviews?

Changepoints and Events: Can you detect when things change suddenly (e.g., a business coming under new management)? Can you see when a city starts going nuts over cronuts?

Social Graph Mining: Can you figure out who the trend setters are and who found the best waffle joint before waffles were cool? How much influence does my social circle have on my consumer choices and my ratings?

students are eligible for prize consideration--any non-student advisors must be credited only in an advisory capacity and not as co-authors.

Additionally, if you publish a research paper about your winning research in a peer-reviewed academic journal, then you'll be awarded an additional $1,000 as recognition of your publication. If you are published, Yelp will also contribute up to $500 to travel expenses to help you present your research using our data at an academic or industry conference.

The deadline for the ninth round of the Yelp Dataset Challenge is **June 30, 2017**. Submit your project to Yelp by visiting yelp.com/challenge/submit. You can submit a research paper, video presentation, slide deck, website, blog, or any other medium that conveys your use of the Yelp Dataset Challenge data. If you have any questions regarding the challenge, feel free to contact dataset@yelp.com.

Round Seven Challenge Winners

From the completed entries we received, a team of our data scientists and data mining engineers selected the following entry as the grand prize winner:

- "Semantic Scan: Detecting Subtle, Spatially Localized Events in Text Streams" Abhinav Maurya, Kenton Murray, Yandong Liu, Chris Dyer, William W. Cohen, and Daniel B. Neill from the Event and Pattern Detection Laboratory, H.J. Heinz III College and School of Computer Science at Carnegie Mellon University, and University of Notre Dame in Indiana.

Round Six Challenge Winners

From the completed entries we received, a team of our data scientists and data mining engineers selected the following entry as the grand prize winner:

- "Topic Regularized Matrix Factorization for Review Based Rating Prediction" Jiachen Li, Yan Wang, Xiangyu Sun, Chengliang Lian, and Ming Yao, from the Language Technologies Institute, School of Computer Science, at Carnegie Mellon University.

Round Five Challenge Winners

From the completed entries we received, a team of our data scientists and data mining engineers selected the following entry as the grand prize winner:

- "From Group to Individual Labels Using Deep Features" Dimitrios Kotzias (University of California, Irvine), Misha Denil (University of Oxford, UK), Nando

De Freitas (University of Oxford, UK, and Canadian Institute for Advanced Research), and Padhraic Smyth (University of California, Irvine).

Round Four Challenge Winners

From the completed entries we received, a team of our data mining engineers selected three entries as grand prize winners (in alphabetical order by entry name):

- "Collective Factorization for Relational Data: An Evaluation on the Yelp Datasets" Nitish Gupta, Indian Institute of Technology, Kanpur and Sameer Singh, University of Washington.
- "Mining Quality Phrases from Massive Text Corpora" Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, Jiawei Han, University of Illinois, Urbana Champaign.
- "Oversampling with Bigram Multinomial Naive Bayes to Predict Yelp Review Star Classes" Kevin Hung and Henry Qiu, University of California, San Diego.

Round Three Challenge Winners

From the completed entries we received, a team of our data mining engineers selected two entries as grand prize winners (in alphabetical order by entry name):

- "On the Efficiency of Social Recommender Networks." Felix W. Princeton University.
- "Personalizing Yelp Star Ratings: a Semantic Topic Modeling Approach." Jack Linshi. Yale University.

Round Two Challenge Winner

From the completed entries we received, a team of our data mining engineers selected the following as a grand prize winner:

- "Valence Constrains the Information Density of Messages." David W. Vinson, Rick Dale. University of California, Merced.

Round One Challenge Winners

From the completed entries we received, a team of our data mining engineers selected four entries as grand prize winners (in alphabetical order by entry name):

- "Clustered Layout Word Cloud for User Generated Review." Ji Wang, Jian Zhao, Sheng Guo, Chris North. Virginia Tech and University of Toronto. Presented at Graphics Interface 2014 Montreal
- "Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text." Julian McAuley, Jure Leskovec. Stanford University. Published in ACM RecSys '13 Proceedings
- "Improving Restaurants by Extracting Subtopics from Yelp Reviews." James Huang, Stephanie Rogers, Eunkwang Joo. University of California, Berkeley. Presented at iConference 2014 Berlin
- "Inferring Future Business Attention." Bryan Hood, Victor Hwang, Jennifer King. Carnegie Mellon University.

Notes on the Dataset

Each file is composed of a single object type, one json-object per-line.
Take a look at some examples to get you started: https://github.com/Yelp/dataset-examples.

yelp_academic_dataset_business.json

```
{
    "business_id":"encrypted business id",
    "name":"business name",
    "neighborhood":"hood name",
    "address":"full address",
    "city":"city",
    "state":"state -- if applicable --",
    "postal code":"postal code",
    "latitude":latitude,
    "longitude":longitude,
    "stars":star rating, rounded to half-stars,
    "review_count":number of reviews,
    "is_open":0/1 (closed/open),
    "attributes":["an array of strings: each array element is an attribute"],
    "categories":["an array of strings of business categories"],
```

```
    "hours":["an array of strings of business hours"],
    "type": "business"
}
```

yelp_academic_dataset_review.json

```
{
    "review_id":"encrypted review id",
    "user_id":"encrypted user id",
    "business_id":"encrypted business id",
    "stars":star rating, rounded to half-stars,
    "date":"date formatted like 2009-12-19",
    "text":"review text",
    "useful":number of useful votes received,
    "funny":number of funny votes received,
    "cool": number of cool review votes received,
    "type": "review"
}
```

yelp_academic_dataset_user.json

```
{
    "user_id":"encrypted user id",
    "name":"first name",
    "review_count":number of reviews,
    "yelping_since": date formatted like "2009-12-19",
    "friends":["an array of encrypted ids of friends"],
    "useful":"number of useful votes sent by the user",
    "funny":"number of funny votes sent by the user",
    "cool":"number of cool votes sent by the user",
    "fans":"number of fans the user has",
    "elite":["an array of years the user was elite"],
    "average_stars":floating point average like 4.31,
    "compliment_hot":number of hot compliments received by the user,
    "compliment_more":number of more compliments received by the user,
    "compliment_profile": number of profile compliments received by the user,
    "compliment_cute": number of cute compliments received by the user,
    "compliment_list": number of list compliments received by the user,
    "compliment_note": number of note compliments received by the user,
    "compliment_plain": number of plain compliments received by the user,
    "compliment_cool": number of cool compliments received by the user,
    "compliment_funny": number of funny compliments received by the user,
    "compliment_writer": number of writer compliments received by the user,
    "compliment_photos": number of photo compliments received by the user,
    "type":"user"
}
```

yelp_academic_dataset_checkin.json

```
{
    "time":["an array of check ins with the format day-hour:number of check ins from hour to hour+1"],
    "business_id":"encrypted business id",
    "type":"checkin"
}
```

yelp_academic_dataset_tip.json

```
{
    "text":"text of the tip",
    "date":"date formatted like 2009-12-19",
    "likes":compliment count,
    "business_id":"encrypted business id",
    "user_id":"encrypted user id",
    "type":"tip"
}
```

photos (from the photos auxiliary file)

This file is formatted as a JSON list of objects.

```
[
    {
        "photo_id": (encrypted photo id),
        "business_id" : (encrypted business id),
        "caption" : (the photo caption, if any),
        "label" : (the category the photo belongs to, if any)
    },
    {...}
]
```

Frequently Asked Questions

1. I am a college professor - can I use and distribute the dataset for a class assignment?

Yes! As long as the assignment is academic in nature and everyone who will be accessing and using the data agrees and signs our Terms of Use, feel free to use the dataset! Please note, however, that only students are eligible to participate in the Yelp Dataset Challenge.

2. Can I use the dataset without participating in the challenge?

Absolutely - as long as it's for academic use! For access to data for non-academic purposes, check out Yelp's Fusion API.

3. Can I present and/or publish my research that was based off of your dataset?

You may present and/or publish your research, but you should not disclose the contents of the dataset. Also, the purpose must be academic in nature.

4. Can I enter the Dataset Challenge as a group?

Of course, as long as you are all students and otherwise eligible to enter the contest! Just keep in mind that if your group wins, the prize will only go to one submission - the winning team would have to divide the prize themselves. Additionally, all contributors must be named on the submission as co-authors.

5. Can I get copies of old datasets or get additional data?

Unfortunately, we do not provide old datasets or additional data outside of the current dataset at this juncture.

6. Do you have to be a current student to participate in the Dataset Challenge?

Yes, you do!

7. I have trouble downloading the files - can you email them to me?

As much as we'd love to help, we cannot send you the dataset directly; the dataset can only be downloaded from our Dataset Challenge site. Some students have experienced success by downloading the files again or using a different connection.

8. How do I withdraw from the Dataset Challenge?

We are sad to see you withdraw, but certainly understand things come up and plans change. If you would like to withdraw from the Dataset Challenge, please shoot an email to dataset@yelp.com with your name and email address. No questions asked!

9. Why is the user review count different than the actual number of reviews returned for that user?

The review count represents the total number of reviews a user had posted at the time of data collection, whether Yelp recommended them or not. As for the reviews, only the reviews that were recommended at the time of data collection are included. Also, we only include businesses that have had at least 3 reviews older than 14 days. So the review count number may differ from the number of actual reviews for any given user.

10. Which reviews are included in the reviews files?

Only the reviews that Yelp recommended at the time of data collection are included.