



Creating Hashtags from Yelp Reviews to Explore Business Relationships and Attributes

By Terin Ambat, Tony Ta, and Zafeer Syed



Problem

- Difficult to summarize what people think about businesses
 - Yelp has a reviews system, but without reading the reviews individually, it is hard to gauge what words are most important in the review
- Hard to recommend businesses for users
 - Finding restaurants that are similar to a given restaurant isn't readily available to users
- Not easy to find keywords that relate most to a business

Recommended Reviews




Yelp Sort Date Rating Elites English 16



Jenn P.
San Francisco, CA
👤 1 friend
★ 22 reviews

★★★★★ 10/17/2013

Absolutely Outstanding! The Grounds at Grace Vineyards are stunning...there are SO many photo ops. I must give 5 stars for Steve the owner he is simply wonderful. He was so organized, flexible and prompt I never was stressed. The food was great and the vino was delicious! If your looking for a beautiful venue with many things included this is the place.



Ben M.
Woodside, NY
👤 1 friend
★ 3 reviews

★★★★★ 6/22/2016

From the initial scheduling and explanation of service from Chelsea to the service itself, provided by Peter, my experience with Green Earth Pest Control was top-notch. From a customer service vantage point, everyone was friendly, helpful, courteous, and honest and from an actual service vantage point they did everything they were supposed to do completely, thoroughly, and without issue. Will definitely use again (though hopefully not for a while).

Solution

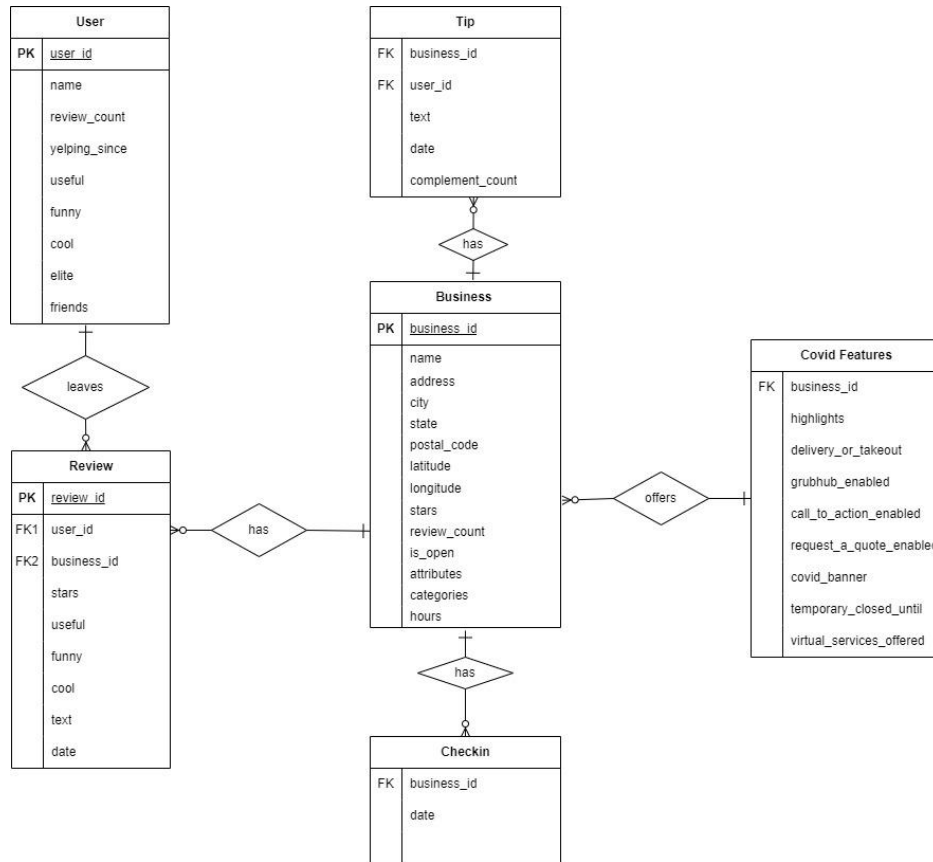


- Make hashtags from Yelp reviews to summarize reviews and businesses with just a few keywords
- Create applications using the hashtags to find similar businesses, the most common hashtags for businesses, and patterns in the hashtags



Methodology

1. Use TF-IDF on Yelp review data to find the most relevant words in each review
2. Get the 5 most relevant words (hashtags) for each review
3. Create a graph that has nodes for each business and each hashtag using Neo4j
 - a. Hashtag to Hashtag relationship if two tags are found in the same review (with attribute for number of occurrences)
 - b. Business to Hashtag relationship if the business had a review that contained that hashtag
4. Make seven applications that allow the user to query from the graph to gain insights about similar businesses, relevant words for each business, most common hashtags in their area, and more
 - a. Implemented a Redis cache to store queries



E/R Diagram



Data Processing

- Yelp Business JSON file (113 MB) - contains information on over 150,000 businesses in the U.S.
 - Key Attributes: business_id, name, address, city, state, postal_code, latitude, longitude, stars
- Yelp Review JSON file (4.97 GB) - contains nearly 7 million Yelp reviews for businesses across the United States
 - Key Attributes: review_id, user_id, business_id, stars, text, date
- Merged Review and Business Datasets in order to add city and state attributes for easier querying
- Limited Dataset to 30,000 businesses in California to reduce Neo4j data loading time



Generating Hashtags

- Tools Used: Regex, Sklearn TFIDF-Vectorizer and NLTK.
- First, removed punctuation and stop words from each review in the data set, and tokenized them into a list of tokens
- Performed tf-idf on the reviews, with the entire set of reviews being the corpus
- Extracted 5 hashtags with highest tf-idf score from each review
- **Limitation:** Generated some unhelpful hashtags with this approach

Creating the Graph Data

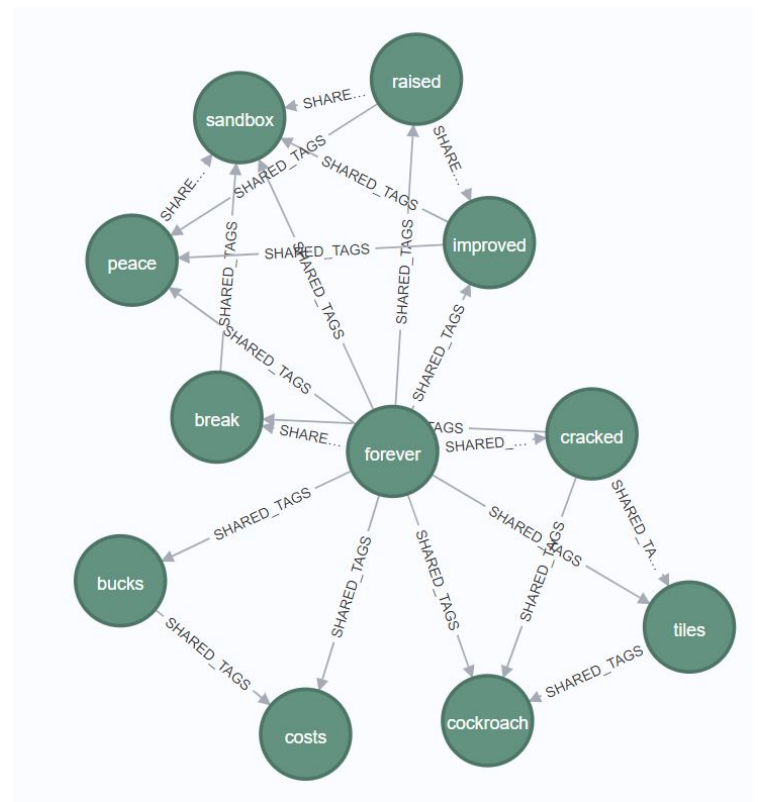
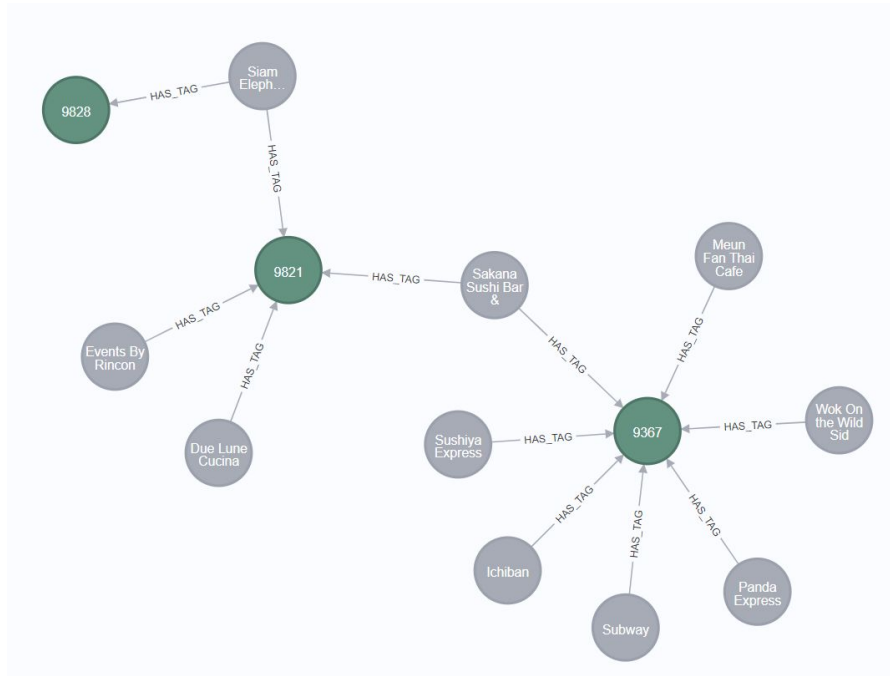
- 2 resulting CSV files:

Businesses to Tags

	business_id	hashtag	count	name	address	city	state	latitude	longitude	stars	categories
0	B5XSoSG3SfvQGtKEGQ1tSQ	dead	1	Los Padres National Forest	San Marcos Pass Rd	Santa Barbara	CA	34.597239	-119.510772	4.5	Parks, Active Life
1	B5XSoSG3SfvQGtKEGQ1tSQ	national	5	Los Padres National Forest	San Marcos Pass Rd	Santa Barbara	CA	34.597239	-119.510772	4.5	Parks, Active Life
2	B5XSoSG3SfvQGtKEGQ1tSQ	grass	1	Los Padres National Forest	San Marcos Pass Rd	Santa Barbara	CA	34.597239	-119.510772	4.5	Parks, Active Life
3	B5XSoSG3SfvQGtKEGQ1tSQ	forest	2	Los Padres National Forest	San Marcos Pass Rd	Santa Barbara	CA	34.597239	-119.510772	4.5	Parks, Active Life
4	B5XSoSG3SfvQGtKEGQ1tSQ	rid	1	Los Padres National Forest	San Marcos Pass Rd	Santa Barbara	CA	34.597239	-119.510772	4.5	Parks, Active Life

Tags to Tags

	tag1	tag2	sharedOccurrences
0	dead	national	1
1	dead	grass	1
2	dead	forest	1
3	dead	rid	1
4	dead	poisoning	1



Visualization of how the graph looks in Neo4j



Applications

1. Find geographically close businesses that share the same hashtags as a given business
2. Finding businesses in a given city that are most similar to a given business
3. Find the highest rated businesses in a city that contains at least one hashtag from a list of given hashtags
4. Find the most common hashtags associated with a given business
5. Find the most common hashtags that occur in the same reviews as a given hashtag
6. Find the top hashtags for a given category
7. Find the most common hashtags in a given city



Demo

- Using subset of 30,000 reviews in California to minimize processing time, data loading time, and querying time
- Over 9,000 unique hashtags and 250 businesses
- Over 120,000 total relationships