Yelp Dataset Challenge - Capstone Project

Hardeep Arora

17 November 2015

Title

This is the report for my data science capstone project. In this project we were given the data from Yelp Dataset Challenge 6 and asked to formulate a question and answer it using this data.

The data is in streaming JSON format and has information on entities like business, review, user, check-in and tip. For this project I framed an Inferential question based on the dataset and tried to explore the findings to formulate an answer.

Introduction

Problem

Since YELP dataset is primarily reviews, I have a pretty simple question that I want to explore:

Analysis on the text of review to figure out what makes a business good/bad, basically what specific feature people value in a business? Do people living in different cities value same/different things in the business?

Rationale

This question is of interest to me, as it would help uncover what people really care about when they write a review about a business. It can be good or a bad review, what matters is what is that they are reviewing, getting this information can be of great help to
1. Entrepreneurs planning to start a new business
2. Existing business owners

Also it would be interesting to see how these discovered features vary across cities.

Methods and Data

Data

```
The Challenge Dataset:

* 1.6M reviews and 500K tips by 366K users for 61K businesses
* 481K business attributes, e.g., hours, parking availability, ambience.
* Social network of 366K users for a total of 2.9M social edges.
* Aggregated check-ins over time for each of the 61K businesses
```

This data is in streaming *JSON* format. It order to begin any form of analysis on this dataset we first needed to cleanse and format it.

Data Cleansing

I read this data in using the package jsonlite::steam_in and then flattened it using jsonlite::flatten

```
## Sample Code
library(jsonlite)
business <- stream_in(file("yelp_academic_dataset_business.json"))
businessF <- flatten(business)
```

Since stream_in and flatten is a costly operation I stored this data as intermediate RData files.Using this flattened dataset I conducted exploratory analysis on the data and came up with the problem statement as discussed above.

Exploratory Analysis

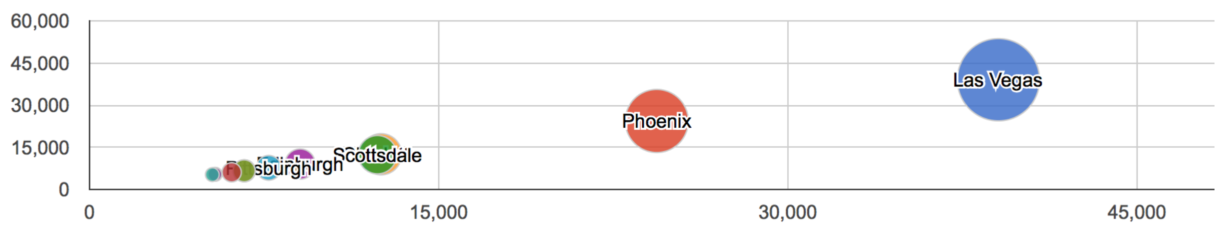
Shortlisting the business

In order to answer the question I needed to shortlist a business category to explore first.

Based on the above exploration, I decided to pick up Food and Restaurants as a business to explore.

Shortlisting the city

For choosing a city, I conducted a similar exploration and shortlisted on following cities from 3 different regions (Phoenix US, Montreal Canada and Edinburgh, UK)



Method

Since I needed to focus on different cities for the businesses, I have to first mark/cluster the business around the different regions

Clustering data as per region

The data for the yelp review is given for 10 different regions, one of the first challenges was to how to successfully cluster the businesses correctly around these regions. I used ggmap::geocode to get the latitude and longitude data for the 10 locations and then used kmeans clustering to cluster businesses around these regions

```
city.centres<-geocode(cities)
geo.cluster<-kmeans(business[,c('longitude','latitude')],city.centres)
```

Topic/Feature discovery using NLP

This is the most challenging part of this project, where I needed to pick up the reviews for Food and Restaurants in Edinburgh and then use some NLP techniques to extract the topic/features in a unsupervised way.

To accomplish this task I decided to use **Latent Dirichlet allocation (LDA)**. As per wikipedia *"In natural language processing, Latent Dirichlet allocation (LDA) is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics."*

Click here to read LDA explained in plain english.

LDA Steps

In order to carry out this analysis I used the following steps

1. First I created a word corpus using all the reviews, I used package tm::Corpus function
2. Then I cleansed corpus to retain meaningful words
   - Remove stemming words
   - Remove stop words
   - Remove numbers
   - Remove punctuations
   - Remove words shorter than 3 letters
3. Then I created a Document Term Matrix out of the cleansed corpus, DTM is basically a matrix having each review as a row and each word the review as a column. This matrix is very sparse. Used tm::DocumentTermMatrix function for the same.
4. Then I re-weighted the terms using a concept of Term Frequency Inverse Document Frequency (TF-IDF). This technique helps to put a penalty on the common words, so that they don't come up prominently in when we use LDA to discover topics.
5. Finally I applied the LDA algo on the above cleansed DTM to discover the topics. I used package topicmodels & RKEA to generate cluster of 30 topics.
6. Then I extracted the top 10 words in each of the 30 topics and tried to manually put labels on each category of words.
7. These labels are the features/topics that the LDA algo has helped us to learn automatically in an unsupervised manner.
8. Order the topics as per the probabilities (gamma) and get list of features customers care about the most.

Results

The results of the analysis is a list of topics discovered by the LDA algo, below are the partial screen shots of topics discovered from LDA (Can't accommodate all 30 in this space).

Edinburgh

	coffee	topic2	thai	mexican	chinese	drinks	british	japanese	indian
1	scone	pie	thai	mexican	chines	whiski	mash	sushi	curri
2	starbuck	haggi	pad	burrito	noodl	selection	sausag	japanes	indian
3	browni	hotel	haddock	nacho	flavor	danc	british	tuna	ale
4	cappuccino	castl	surprise	frozen	wing	broughton	nom	tofu	naan
5	yard	thorough	ton	taco	chop	whisky	oyster	miso	cupcak
6	peter	tatti	shrimp	margarita	wev	scotch	gravi	tempura	dosa
7	bakeri	specials	tonight	nachos	efficient	isnt	sausages	sashimi	tikka
8	starbucks	forth	fring	asparagus	wings	son	costa	soy	india
9	scones	east	average	illeg	buffalo	relaxed	gravy	ramen	pakora
10	muffin	cheddar	satay	los	overcook	con	character	sake	masala

Phoenix

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
1	late	tea	sandwich	burger	drive	kid	cake	often	pizza
2	disappointed	theyr	italian	wing	slow	ladi	wish	buffet	crust
3	twice	chain	pasta	fries	location	rude	thank	varieti	garlic
4	pleasant	kinda	turkey	five	money	poor	valley	prices	pie
5	mushroom	options	oil	joint	charg	son	treat	averag	thin
6	simpl	vegan	homemad	chili	card	phone	birthday	rate	deliv
7	stick	boy	oliv	bun	worst	young	bake	standard	deliveri
8	remind	singl	stuf	soda	car	horribl	cute	low	fire
9	spinach	damn	sandwiches	wings	complain	groupon	box	coupon	mozzarella
10	origin	seriously	deli	greasi	donut	state	daughter	abov	oven

Montreal

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
1	amaz	service	coffe	husband	shop	owner	breakfast	tasti	burger
2	fish	excel	cafe	stars	ice	okay	egg	awesome	poutin
3	fun	vegetarian	coffee	boyfriend	for	charg	brunch	taco	poutine
4	favourit	famili	cup	custom	store	rib	toast	averag	ramen
5	uniqu	wow	thai	man	buy	rude	sunday	yum	joint
6	solid	twice	neighborhood	birthday	market	deal	benedict	stuff	gravi
7	chip	incred	espresso	worst	die	tapa	sausag	mexican	bun
8	dark	ton	milk	poor	park	bother	morn	girl	classic
9	try	middl	latt	simpli	sell	complain	ham	too	greasi
10	vibe	yummy	wifi	anyon	card	dollar	complaint	burrito	dog

Discussion

Based on the topic discovery as shown above, I tried to label the topic categories and came up with the list of following features and how they differ among various places

Restaurants & Food

The below table lists the top 3 features customers talk about in the reviews in each region

Sno	Edinburgh	Montreal	Phoenix
1	Stores	Value	Mexican Food
2	Pizza	Coffee	Atmosphere
3	Breakfast	Desserts	Desserts

Also another it was very interesting to see kind of words/language people used between different cities e.g. there were lot of french usage in Montreal.

Next Steps

Validations and Fine Tuning

Next steps for me would be do conduct some extensive validations of these findings using some alternate clustering algo's or manual verification, so that I can fine tune the hyper-parameters for LDA and get more better results.

Go across other dimensions

Compare features across various categories of businesses.