Jointly analysis of geospatial and friendship of Gowalla data.

true

March 2017

Abstract

Gowalla is a location-based social networking website where users share their locations by checking-in. The friendship network is undirected and was collected using their public API, and consists of 196591 nodes and 950327 edges. It has a total of 6,442,890 check-ins of these users over the period of Feb. 2009 - Oct. 2010. The aim of the project is to use Data Science tools to analyse and build a recommendation system for both the geospatial checkins and the social network links in that database. As a part of the project I will tag a users as a locals or tourists (who mostly visit monuments, stay in hotels …) and adapt responses given that tag. I will use the 17400 Paris check-ins from 1366 users and I will join Gowalla location data with Google Places to separate the type of venues reported Gowalla's users.

Introduction - Motivation - Explanation of the context – Your contribution to the project

the authors [1] 2 3 4 5 6 7 8 9 10

Datasets description

Multiple sources of information were used to using in the developing of this project; the datasets includes the Stanford's SNAP Gowalla data [11], queries to Google Maps API and vectorial geodata from OpenStreetMaps.

The Gowalla is an anonymized and clean dataset collected from February 2009 to October 2010 from the Gowalla's startup to capture human mobility in a location-based social network (LBSN). In Gowalla LBSN, people were able to **check-in** at places or **spots** that they visited near to their local vecinity; Check-ins were

collected from a mobile application or through a mobile website, the incentive for users was to get advantages in the places they check-in[1]. SNAP's Gowalla data is splited in two datasets: a) a collection of more that 6.4 millions of individual check-ins[2] and b) an undirected graph[3] with 196591 nodes and 950327 edges [11]

For this project a subset of the check-in in Gowalla data was extracted; The subset correponds to data for Paris, selecting all datapoints in a cicle radious of 30km from the position 48.86°N, 2.35°E. The selected data consists in 17496 check-ins between September 2009 and October 2010 [see github]. Data from the undirected graph was not used in this Data Science pipeline given the constrains on the selected data as the low number of connections for Paris, searching time, etc.

The project's location data was enriched using information fetched from Google Places using their webservice API (function `nearbysearch`). I developped scripts to do multiple webscrapping and parsing Google's JSON geodata. Paris' Gowalla's data contains 4104 different check-ins locations, for each one of those point I searched all the POI (Points of interest) within 100m radius, and getting a total of 17373 individual POIs. Google's JSON POI contains many fields, for the DSSP pipeline I extracted a subset containing: ‘

- `geometry/location`: latitude and longitude, numerical values
- `id`: location google's hash, 10character string
- `name`: name of the place, a string
- `type`: a string with one of Google's supported types

Google classify their places in 130 types, the full list of supporte types are listed on the table +**???**. The location of Gowalla check-ins and Google Places are in the figure fig. 1

Exploration and Preprocessing

As mentioned before the raw data was given in multiples formats; Gowalla format consist on fixed width the columns, that contains: user identification number, the UTC time of the check-in, the location in i decimal latitude and logitude, the las column is the a unique spot id. For these datapoints the only needed modifications done on the data is on the UTC time, that time was corrected using a daysaving time function, this dataset was alredy clean for research by SNAP, then no extra modifications were done on this data.

[1] https://en.wikipedia.org/wiki/Gowalla

[2] the data's columns are 1-`user_id`, 2-`UTC_Time`, 3-`latitude`, 4-`longitude` and 5-`spot_id`; 1,3,4 are directly used on the DSSP's workflow; 2 is corrected for daysaving time, all variables are available on the dataframes.

[3] it contains in the columns `user_id` `friend_id`.

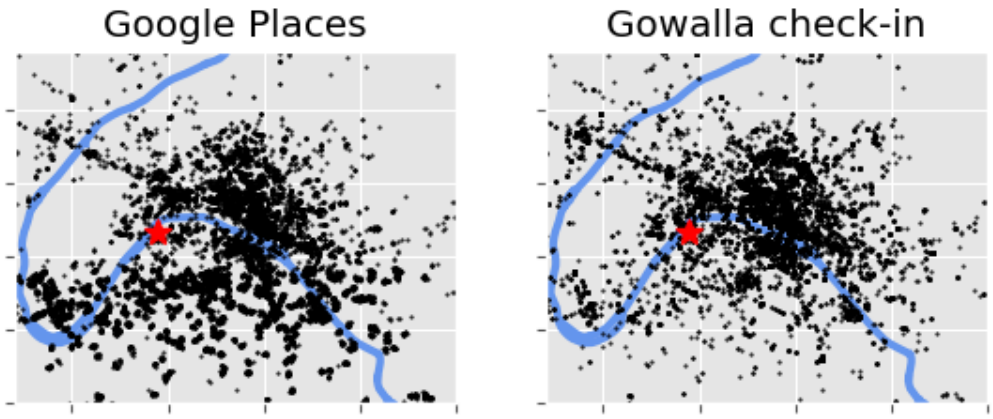

Figure 1: Map of selected datasets, left are the POI from Googles' Places database found, right are the Gowalla check-ins in Paris, the red start is the location of Eiffel Tower

To obtain Google's Places data, all unique locations from all check-ins in Paris from Gowalla dataset were downloaded in a radious of 100m to that point and selectin all type of sites (hotel, monuments, zoo, etc). The total the number check-ins in Gowalla's data is 17496, however that number is reduced to only 4178 when duplicated are removed. Duplicated data was searched from Google's site hash code.

Google Places webservice returns a JSON file for each query and a maximum of 2500 queries/day per developer key, I made 4178 queries in in two daily batches. Both batches were merged on a single
csv data file containing 30286 POI and 19089 POI after cleaning, the higher number of POI on Google Sites in comparison with Gowalla's check-ins is due to Gowalla locations are close one each other, and I am looking for an extenced searching radius in Gogle. Looking closely to figure fig. 1, it can be see that there are points in Gowalla that are not present on Google Sites and there are more point on Google's, this can happened because:

- The Gowalla spots doesn't exist any more, closed restaurants, shops etc.
- Check-ins in private properties (e.g. appartment, offices)
- Googles Sites were retrived around 100m of GoWalla's data, then there is a higher density of information.

Analysis of Gowalla data for Paris

From the analysis of 17496 Gowalla check-in locations, the places that are most visited are in +**???** table. The first place corresponds to is the CDG Airport and analysing the data it was found that check-in places at the airport correspond to car rental agencies, that could be due to promotional coupons or reductions when using the application and doing those check-ins. The other 9 major check-ins, corresponds to places near-to or in turistical places, that is around 10% of all the check-ins in Paris, meaning that a proportion of data is given by turists.

Table 1: Top 10 Places in Paris found from Google/Gowalla data. {#Table:top10}

#Check-ins	**Place**
402 198 194 171 267	CDG Airport Louvre Pont des Arts Eiffel
114 106 100 93 91	Tower BNF/François-Mitterrand Library
	Gare du Nord Notre Dame A place near to
	BNF Arc de Triomphe places near to porte de
	Montreuil

Raw data from Google Places are files stored from calls to Googles Webservce function nearbysearch

Google Places API returning data

returns a more complex and rich data

From fig. 1 it can

Feature Engineering/selection and decisions/assumptions you made

Methods used for learning/prediction

Analysis and interpretations of the results

difference can be explained on recurrent visit of same people to the same places or different people visiting the same places.

github repo

Faiss handles collections of vectors of a fixed dimensionality d, typically a few 10s to 100s. These collections can be stored in matrices. We assume row-major storage, ie. the j'th component of vector number i is stored in row i, column j of the matrix. Faiss uses only 32-bit floating point matrices.

Potential further work

Appendices

Table 2: List of returned types on google querey. {#Table:types_gogle}

account	ing cem	etery g	ym natur	al_feature rv_park
adminis	trative	_area_l	evel_1 c	hurch hair_care neighborhood school
adminis	trative	_area_l	evel_2 c	ity_hall hardware_store night_club shoe_store
adminis	trative	_area_l	evel_3 c	lothing_store health painter shopping_mall
adminis	trative	_area_l	evel_4 c	olloquial_area hindu_temple park spa
adminis	trative	_area_l	evel_5 c	onvenience_store home_goods_store parking stadium
airport	countr	y hospi	tal pet_	store storage
amuseme	nt_park	courth	ouse ins	urance_agency pharmacy store
aquariu	m denti	st inte	rsection	physiotherapist street_address
art_gal	lery de	partmen	t_store	jewelry_store place_of_worship street_number
atm doc	tor lau	ndry pl	umber su	blocality
bakery	electri	cian la	wyer poi	nt_of_interest sublocality_level_1
bank el	ectroni	cs_stor	e librar	y police sublocality_level_2
bar emb	assy li	ght_rai	l_statio	n political sublocality_level_3
beauty_	salon e	stablis	hment li	quor_store postal_code sublocality_level_4
bicycle	_store	finance	local_g	overnment_office postal_code_prefix sublocality_level_5
book_st	ore fir	e_stati	on local	ity postal_code_suffix subpremise
bowling	_alley	floor l	ocksmith	postal_town subway_station
bus_sta	tion fl	orist l	odging p	ost_box synagogue
cafe fo	od meal	_delive	ry post_	office taxi_stand
campgro	und fun	eral_ho	me meal_	takeaway premise train_station
car_dea	ler fur	niture_	store mo	sque real_estate_agency transit_station
car_ren	tal gas	_statio	n movie_	rental restaurant travel_agency
car_rep	air gen	eral_co	ntractor	movie_theater roofing_contractor university
car_was	h geoco	de movi	ng_compa	ny room veterinary_care
casino	grocery	_or_sup	ermarket	museum route zoo

References

[1] G. Ference, M. Ye, and W.-C. Lee, "Location recommendation for out-of-town users in location-based social networks," in *Proceedings of the 22Nd aCM international conference on information & knowledge management*, 2013, pp. 721–726 [Online]. Available: http://doi.acm.org/10.1145/2505515.2505637

[2] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann, "Time-aware point-of-interest recommendation," in *Proceedings of the 36th international aCM sIGIR conference on research and development in information retrieval*, 2013, pp. 363–372 [Online]. Available: http://doi.acm.org/10.1145/2484028.2484030

[3] H. Wang, Z. Li, and W.-C. Lee, "PGT: Measuring mobility relationship using personal, global and temporal factors," in *Data mining (iCDM), 2014 iEEE international conference on*, 2014, pp. 570–579.

[4] S. M. Rahimi and X. Wang, "Location recommendation based on periodicity of human activities and location categories," in *Advances in knowledge discovery and data mining: 17th pacific-asia conference, pAKDD 2013, gold coast, australia, april 14-17, 2013, proceedings, part iI*, J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 377–389 [Online]. Available: http://dx.doi.org/10.1007/978-3-642-37456-2_32

[5] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *Security and privacy, 2009 30th iEEE symposium on*, 2009, pp. 173–187.

[6] J.-D. Zhang and C.-Y. Chow, "GeoSoCa: Exploiting geographical, social and categorical correlations for point-of-interest recommendations," in *Proceedings of the 38th international aCM sIGIR conference on research and development in information retrieval*, 2015, pp. 443–452 [Online]. Available: http://doi.acm.org/10.1145/2766462.2767711

[7] Q. Yuan, W. Zhang, C. Zhang, X. Geng, G. Cong, and J. Han, "PRED: Periodic region detection for mobility modeling of social media users," 2016.

[8] L. Rossi and M. Musolesi, "It's the way you check-in: Identifying users in location-based social networks," in *Proceedings of the second aCM conference on online social networks*, 2014, pp. 215–226 [Online]. Available: http://doi.acm.org/10.1145/2660460.2660485

[9] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo, "A random walk around the city: New venue recommendation in location-based social networks," in *2012 international conference on privacy, security, risk and*

trust and 2012 international confernece on social computing, 2012, pp. 144–153.

[10] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gPUs," *CoRR*, vol. abs/1702.08734, 2017 [Online]. Available: http://arxiv.org/abs/1702.08734

[11] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proceedings of the 17th aCM sIGKDD international conference on knowledge discovery and data mining*, 2011, pp. 1082–1090 [Online]. Available: http://doi.acm.org/10.1145/2020408.2020579