Final Project - Is Yelp international? by Daniel Molnar

Intro

The US-based Yelp bought the European Qype. Did this make Yelp international? Based on the user generated contect metrics, not really.

Methods

Caveat. A bit of a market intel. I'm slightly familiar with the category. Some years ago I've written my thesis on the touristic effect of the location-based community activities in Hungary - it was a critical comparison geocaching and Foursquare.

(Molnár Dániel: Lokáció alapú közösségi tevékenységek turisztikai hatása Magyarországon (A geocaching és a Foursquare kritikai összehasonlítása.) Szakdolgozat, 2011, PDF (http://issuu.com/soobrosa/docs/soobrosa_thesis?e=0))

In the case of the current Yelp Data Challenge what's imminent that the activity differences how easily reveal the touristic spots.

In the exploration phase I came upon the following topics of interest:

1. Most likely residents and tourists differ in behaviour. Can you recognize them by activity? Can you cluster them - the ones who check in even on the toilet, the traveling elderly couple?
2. Do venue visits differ in cities? Does weekday noon, evening and every other time differ from weekend noon, evening and other time? Can you predict categories and visitors (resident versus tourist) from check-in time?
3. Reviews introduce all kinds of personal biases. Is it still possible to mine meaningful differentiators and attributes from reviews? Can you find all 'kneipe' (specific local category not explicitly covered by official categories) in Karlsruhe? Do reviews differ from tips in this sense? Do you have to do language detection for this?

I set out to solve this capstone in the command line.

I try to test the limits of Unix commands, pipes and some special, self-contained tools, three of them altogether slightly bigger than 1 Mbyte.

"*Not unlike jq (http://stedolan.github.io/jq/ (http://stedolan.github.io/jq/)) for JSON, Miller is written in modern C, and it has zero runtime dependencies. You can download or compile a single binary, scp it to a faraway machine, and expect it to work.*" **(Source (https://github.com/johnkerl/miller))**

Assessing the possibly performant and easy toolset I considered using the langdetect (https://pypi.python.org/pypi/langdetect) Python module for language detection, but I did not proceed with it.

As always first you have to shape the dataset explorable. For reproducability the Makefile (https://gist.github.com/soobrosa/4adf89ce197eb6299eb9) is the best.

```
...

# example

source_flattened: source.decompressed
  < source/yelp_academic_dataset_tip.json | json2csv -p=true -k business_id,date,likes,user_id > source_flattened/tip_no_text.csv
  < source/yelp_academic_dataset_review.json | json2csv -p=true -k business_id,date,review_id,stars,user_id > source_flattened/review_compact_no_text.csv
  < source/yelp_academic_dataset_business.json | jq -c '{business_id, category_main: .categories[0], category_sub: .categories[1], city, latitude, longitude, name, neighborhood: .neighborhoods[0], open, review_count, stars, state}' | json2csv -p=true -k business_id,category_main,category_sub,city,latitude,longitude,name,neighborhood,open,review_count,stars,state > source_flattened/business.csv
```

Now that we have all data in CSVs, let's query them in SQL.

I started exploring the three interests, but found no really promising patterns to hypothise on. Then I remembered how much I got used to use both Foursquare and Yelp living in Germany. The German Qype sold to Yelp in 2012. Did this acquisition make Yelp international? Can we tell from the data whether this deal was worth it for Yelp?

I counted the number of businesses per state and tried to get an estimate of their population.

```
#
# http://harelba.github.io/q/
#

q -H -d, "select state, count(*) from business.csv group by 1 having count(*) > 100 order by 2 desc"

AZ,25230 # Phoenix, AZ - capital, 1.5 m
NV,16485 # Las Vegas, NV - largest in state, 0.6 m
NC,4963  # Charlotte, NC - largest in state, 0.8 m
QC,3921  # Montreal - largest, 1.6 m
PA,3041  # Pittsburgh, PA - 0.3 m
EDH,2971 # Edinburgh, UK - capital - 0.5 m
WI,2307  # Madison, WI - 0.006 m
BW,934   # Karlsruhe - 0.3 m
IL,627   # Urbana-Champaign, IL - 0.003 m
ON,351   # Waterloo
SC,189   # Fort Mill == Charlotte, NC
MLN,123  # Edinburgh, UK
```

I have chosen a small, a midsized and a larger city from both the US and abroad with almost equal population. I calculated the reviewing and tipping activity trends in the three segments. As in production or on the call let's use a visual clue whether a linear will be enough to say something sure.

US:

- AZ,25230 # Phoenix, AZ - capital, 1.5 m
- NV,16485 # Las Vegas, NV - largest, 0.6 m
- PA,3041 # Pittsburgh, PA - 0.3 m

Control:

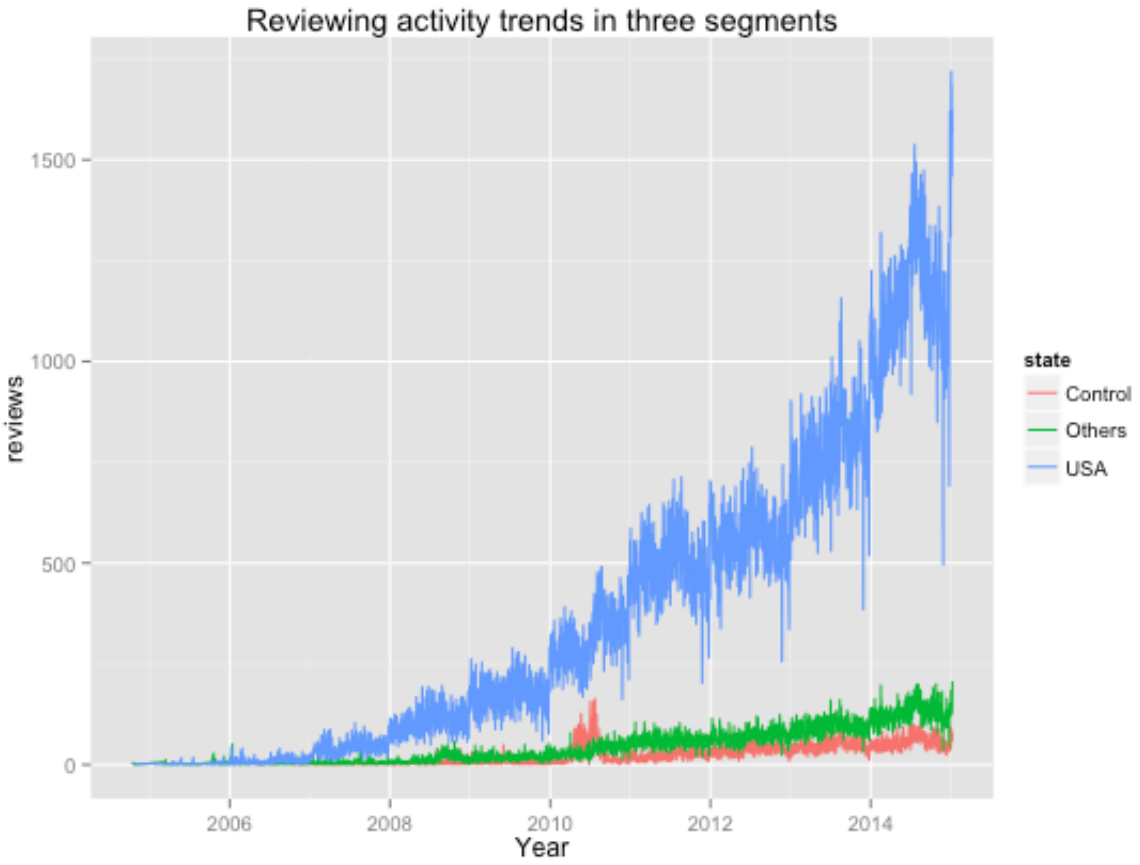
- QC,3921 # Montreal - largest, 1.6 m
- EDH,2971 # Edinburgh, UK - capital - 0.5 m
- MLN,123 # Edinburgh, UK
- BW,934 # Karlsruhe - 0.3 m

```
q -H -O -d, "SELECT business_id, date, count(*) AS reviews \
  FROM review_compact_no_text.csv GROUP BY 1, 2" > temp1.csv

q -H -O -d, " \
  SELECT t.date, \
  CASE WHEN b.state IN ('AZ', 'NV', 'PA') THEN 'USA' \
  WHEN b.state IN ('QC', 'EDH', 'MLN', 'BW') THEN 'Control' \
  ELSE 'Others' END AS state, \
  SUM(t.reviews) AS reviews \
  FROM temp1.csv t  \
  JOIN business.csv b \
  ON (t.business_id = b.business_id) \
  GROUP BY 1, 2 \
  ORDER BY 2, 1" > reviews_summed.csv

q -H -O -d, "SELECT business_id, date, count(*) AS tips \
  FROM tip_no_text.csv GROUP BY 1, 2" > temp2.csv

q -H -O -d, " \
  SELECT t.date, \
  CASE WHEN b.state IN ('AZ', 'NV', 'PA') THEN 'USA' \
  WHEN b.state IN ('QC', 'EDH', 'MLN', 'BW') THEN 'Control' \
  ELSE 'Others' END AS state, \
  sum(t.tips) AS tips \
  FROM temp2.csv t  \
  JOIN business.csv b \
  ON (t.business_id = b.business_id) \
  GROUP BY 1, 2 \
  ORDER BY 2, 1" > tips_summed.csv
```

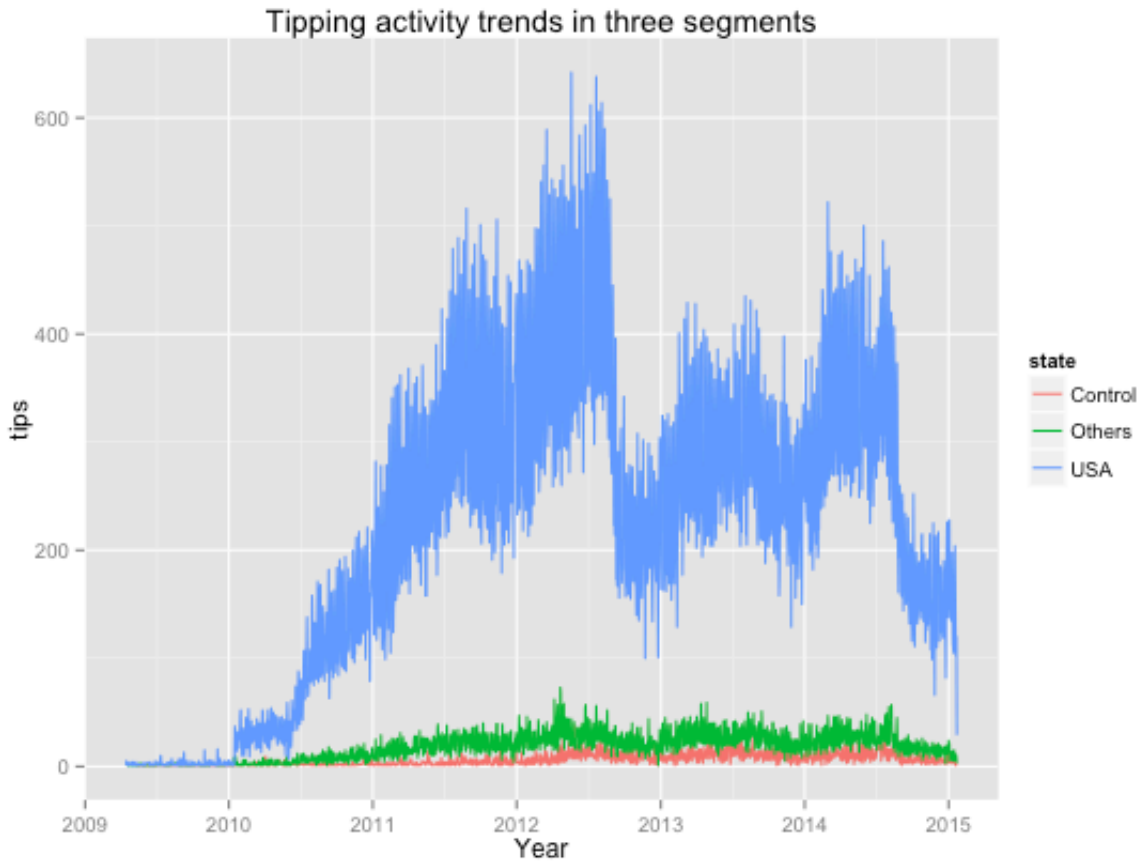
Results

Charting the change of user generated content does not show a rosy picture. The non-US control is way less productive than the US group.

```
< reviews_summed.csv ../Rio.sh -ge 'g+geom_line(aes(x=as.Date(date), '\
'y=reviews, group=state, color=state)) '\
'+ labs(x="Year", title="Reviewing activity trends in three segments")' | display
```




```
< tips_summed.csv ../Rio.sh -ge 'g+geom_line(aes(x=as.Date(date), '\
'y=tips, group=state, color=state)) '\
'+ labs(x="Year", title="Tipping activity trends in three segments")' | display
```



Discussion

Based on the reviews and tips timeline it's inevitable that the control grows much slower and no specific uptick happened at the acquisition. I tried hard to calculate the coefficients of the linear regression from the command line, but failed with this toolset.