Migrations Research Initial Report

(Initial Twitter Scrapping)

Jeong Hyun Lee

March 2021

The objective of this preliminary analysis was to explore the validity and utility of incorporating Twitter data to identify immigrant users and their related public space usage through identifying the language of the Tweets posted and querying specific keywords that reference public spaces in the area of Santa Coloma de Gramenet, Spain.

1 Identifying and Collecting Users

The initial pool of users were collected by searching for Tweets that contained references to the public spaces within the month of July, 2019. The reason for selecting a very specific time interval was to minimize the time it took to query data as the API library being used took time to scrap. Hence, I chose a pre-pandemic year and chose the month during the summer that would incorporate a lot of immigrants. Thus, within this time period, I searched for all Tweets that mention the below keywords that correspond to the public space (Keywords include both Spanish and Catalan descriptions). Table 1 describes the all keywords used.

Table 1: Keywords for user collection

Name	Relevant Keywords
Santa Coloma de Gramenet	Santa Coloma de Gramenet, Santa Coloma
Parque Fluvial del Besos	Parque Fluvial del Besos, Parc Fluvial del Besos, Fluvial del Besos
Parque Molinet	Parque Molinet, Parc Molinet, Molinet
Plaza del Rellotge	Plaza del Rellotge, Plaça del Rellotge
Rambla San Sebastian	Rambla San Sebastian

Table 1: Keywords for user collection

Name	Relevant Keywords
Parque Can Zam	Parque Can Zam, Parc Can Zam, Can Zam
Instituto Can Peixauet	Instituto Can Peixauet, Institut Can Peixauet, Can Peixauet
Parque Gran Sol	Parque Gran Sol, Parc Gran Sol, Gran Sol
Escuela Tanit	Escuela Tanit, Escola Tanit
Instituto Terra Roja	Instituto Terra Roja, Institut Terra Roja, Terra Roja
Instituto Gassol	Instituto Gassol, Institut Gassol
CAP Santa Rosa	CAP Santa Rosa
Cinto Verdaguer	Cinto Verdaguer
Mercado del Fondo	Mercado del Fondo, Mercat del Fondo, del Fondo
Nus de la Trinitat	Nus de la Trinitat

In addition, in case any Tweets incorporated geotags (where users can tag specific geographical locations), I queried for additional Tweets based on specific location so that it covers roughly the municipal area of Santa Coloma de Gramenet.

Location 1 and radius: (41.46287400801948, 2.2028934732857177), 1km
Location 2 and radius: (41.45039468429977, 2.212764002746006), 0.75km

Once I collected the Tweets with the above keywords and location tags, I compiled the users who created the Tweets and then, I divided the users into native users (i.e. those who have continuously lived in the area) and foreign users (i.e. those who are more likely to be immigrants) based on the language of the Tweet. That is, if the Tweet was categorized as a Spanish or Catalan Tweet (categorization was done by Twitter's own language identification system), the user would be categorized as a native user. Any users who's Tweet's that had no language tag (mostly spam Tweets with advertisement urls) were ignored. Table 2 shows the distributions users as well as total user counts.

Table 2: User Distribution

Native user count	14570
Foreign user count	808
Total user count	15378

2 Collecting Public Space Mentions Data

In order to get a crude estimate of how these users interact with the various public spaces in Santa Coloma, I decided to use the frequencies of mentions and the date/time of the Tweet. Hence, once I had groups of users, I would iterate through all of their 2019 Tweets and search again for specific public spaces. This time, however, since there is already a presumption made that these users are already related to the general region, I was able to use more generic keywords that would still be specific to certain location. For example, in searching for Tweets that mention 'Parque de Fluvial Besos', I could not search for simple terms 'rio' added with 'Besos', therefore increasing the accuracy of the search. In addition, it allowed me to search for more public spaces that inherently had a generic name and hence couldn't have been used in the initial collection stage. Table 3 describes these additional search terms and keywords.

Table 3: Additional Search Keywords

Name	Relevant Keywords
Parque Fluvial del Besos	(Any combination of 'rio' and 'Besos' in Tweet)
Macanet str	Maçanet str
Iglesia Evangelica	Iglesia Evangelica, Iglesia Esglesia

3 Preliminary Results

Below, I list for each public space, its frequency of mentions and user count, language distribution, as well as the distribution of the Tweets throughout the week.

3.1 Native Users Results

Table 4: Native User Tweet mention statistics

Name	Statistics	Language Distribution	Time of week
Santa Coloma de Gramenet	Frequency: 7874 User counts: 1088 Frequency / User: 7.24	CA: 62.24% ES: 36.41% IT: 0.43% EN: 0.28%	Weekends: 23.16% Weekday: 76.84%

Table 4: Native User Tweet mention statistics

Name	Statistics	Language Distribution	Time of week
Santa Coloma de Gramenet	Frequency: 7874 User counts: 1088 Frequency / User: 7.24	FR: 0.15% RO: 0.13% PT: 0.11% FI: 0.08% IN: 0.03% DA: 0.01% CY: 0.01% ET: 0.01%	Weekends: 23.16% Weekday: 76.84%
Parque Fluvial del Besos	Frequency: 1608 User counts: 626 Frequency / User: 2.57	ES: 96.89% IN: 2.11% CA: 0.44% LT: 0.31% PT: 0.12% JA: 0.06% EN: 0.06%	Weekends: 27.74% Weekdays: 72.26%
Parque Molinet	Frequency: 741 User counts: 340 Frequency / User: 2.18	ES: 70.85% CA: 20.24% IT: 6.75% EN: 0.54% FI: 0.27% ET: 0.13% PT: 0.13% EU: 0.13% RO: 0.13% FR: 0.13% CY: 0.13%	Weekends: 22.67% Weekdays: 77.33%

Table 4: Native User Tweet mention statistics

Name	Statistics	Language Distribution	Time of week
Plaza del Rellotge	Frequency: 43 User counts: 10 Frequency / User: 4.3	CA: 100.00%	Weekends: 25.58% Weekdays: 74.42%
Rambla San Sebastion	N/A	N/A	N/A
Parque Can Zam	Frequency: 356 User counts: 99 Frequency / User: 3.60	CA: 55.62% ES: 40.45% RO: 1.69% EN: 0.56% IT: 0.56% PT: 0.56% TR: 0.28%	Weekends: 27.81% Weekdays: 72.19%
Instituto Can Peixauet	Frequency: 201 User counts: 45 Frequency / User: 4.47	CA: 78.61% ES: 20.90% EN: 0.50%	Weekends: 15.92% Weekdays: 84.08%
Parque Gran Sol	Frequency: 1308 User counts: 588 Frequency / User: 2.22	ES: 96.18% CA: 2.75% EN: 0.69% IN: 0.15% HT: 0.08% EU: 0.08% TR: 0.08%	Weekends: 24.85% Weekdays: 75.15%
Escuela Tanit	Frequency: 1 User counts: 1 Frequency / User: 1	ES: 100.00%	Weekends: 0% Weekdays: 100.00%
Instituto Terra Roja	Frequency: 14 User count: 6 Frequency / User: 2.33	CA: 85.71% ES: 14.29%	Weekends: 7.14% Weekdays: 92.86%
Instituto Gassol	N/A	N/A	N/A

Table 4: Native User Tweet mention statistics

Name	Statistics	Language Distribution	Time of week
CAP Santa Rosa	N/A	N/A	N/A
Cinto Verdaguer	Frequency: 56 User count: 42 Frequency / User: 1.33	CA: 58.93% ES: 39.29% PT: 1.79%	Weekends: 14.29% Weekdays: 85.71%
Mercado del Fondo	Frequency: 29737 User counts: 10558 Frequency / User: 2.82	ES: 99.03% CA: 0.62% IT: 0.28% EN: 0.04% PT: 0.0067% IN: 0.0067% ET: 0.0034% HT: 0.0034%	Weekends: 20.38% Weekdays: 79.62%
Nus de la Trinitat	Frequency: 531 User counts: 22 Frequency / User: 24.14	CA: 100.00%	Weekends: 13.75% Weekdays: 86.25%
Macanet Str	N/A	N/A	N/A
Iglesia Evangelica	Frequency: 36 User counts: 25 Frequency / User: 1.44	ES: 97.22% CA: 2.78%	Weekends: 27.78% Weekdays: 72.22%

3.2 Foreign Users Results

Table 5: Foreign User Tweet mention statistics

Name	Statistics	Language Distribution	Time of week
Santa Coloma de Gramenet	Frequency: 599 User counts: 137 Frequency / User: 4.37	EN: 26.54% JA: 22.37% PT: 9.68% CA: 8.37%	Weekends: 22.87% Weekday: 77.13%

Table 5: Foreign User Tweet mention statistics

Name	Statistics	Language Distribution	Time of week
Santa Coloma de Gramenet	Frequency: 599 User counts: 137 Frequency / User: 4.37	ES: 8.01% PL: 4.00% RO: 3.84% IT: 2.84% FR: 2.84% IN: 2.17% FI: 1.67% TR: 1.50% ET: 0.83% SV: 0.83% NO: 0.33% SL: 0.33% VI: 0.33% DA: 0.33% LV: 0.17% DE: 0.17% TL: 0.17% HU: 0.17%	Weekends: 22.87% Weekday: 77.13%
Parque Fluvial del Besos	Frequency: 16 User counts: 6 Frequency / User: 2.67	ES: 100.00%	Weekends: 25% Weekdays: 75%
Parque Molinet	Frequency: 113 User counts: 42 Frequency / User: 2.69	IT: 59.29% EN: 34.51% IN: 0.88% CA: 0.88% FI: 0.88% ES: 0.88%	Weekends: 24.78% Weekdays: 75.22%

Table 5: Foreign User Tweet mention statistics

Name	Statistics	Language Distribution	Time of week
Plaza del Rellotge	Frequency: 2 User counts: 2 Frequency / User: 2	CA: 100.00%	Weekends: 0% Weekdays: 100.00%
Rambla San Sebastion	N/A	N/A	N/A
Parque Can Zam	Frequency: 32 User counts: 7 Frequency / User: 4.57	EN: 68.75% CA: 12.50% ES: 6.25% TR: 6.25% FI: 3.13% RO: 3.13%	Weekends: 28.13% Weekdays: 71.88%
Instituto Can Peixauet	Frequency: 3 User counts: 2 Frequency / User: 1.5	ES: 33.33% CA: 33.33% PT: 33.33%	Weekends: 0% Weekdays: 100.00%
Parque Gran Sol	Frequency: 44 User counts: 37 Frequency / User: 1.19	EN: 36.36% ES: 31.82% IT: 22.73% PT: 2.73% CA: 2.27% JA: 2.27% FR: 2.27%	Weekends: 45.45% Weekdays: 54.55%
Escuela Tanit	N/A	N/A	N/A
Instituto Terra Roja	Frequency: 1 User count: 1 Frequency / User: 1	EN: 100%	Weekends: 0% Weekdays: 100.00%
Instituto Gassol	N/A	N/A	N/A
CAP Santa Rosa	N/A	N/A	N/A

Table 5: Foreign User Tweet mention statistics

Name	Statistics	Language Distribution	Time of week
Cinto Verdaguer	Frequency: 1 User count: 1 Frequency / User: 1	PT: 100%	Weekends: 0% Weekdays: 100.00%
Mercado del Fondo	Frequency: 1798 User counts: 592 Frequency / User: 3.04	IT: 96.50% ES: 1.72% PT: 0.50% EN: 0.50% HT: 0.22% JA: 0.22% CA: 0.11% CY: 0.05% IN: 0.05% NL: 0.05% TR: 0.05%	Weekends: 14.68% Weekdays: 85.32%
Nus de la Trinitat	N/A	N/A	N/A
Macanet Str	N/A	N/A	N/A
Iglesia Evangelica	N/A	N/A	N/A

3.3 Location-based statistics

In addition to querying based on the user, I did a search purely on Tweets that included the location names of the public places throughout 2019 to serve as a standard to which we can compare the above two groups.

Table 6: Location Based Tweet mention statistics

Name	Statistics	Language Distribution	Time of week
Santa Coloma de Gramenet	Frequency: 25895 User counts: 10032 Frequency / User: 2.58	CA: 50.18% ES: 43.70% EN: 1.77% PT: 0.85%	Weekends: 27.92% Weekday: 72.08%

Table 6: Location Based Tweet mention statistics

Name	Statistics	Language Distribution	Time of week
Santa Coloma de Gramenet	Frequency: 25895 User counts: 10032 Frequency / User: 2.58	IT: 0.65% JA: 0.63% FR: 0.42% RO: 0.29% IN: 0.26% PL: 0.26% TR: 0.16% FI: 0.10% ET: 0.05% NO: 0.05% DA: 0.05% SV: 0.04% VI: 0.03% DE: 0.03% HU: 0.02% SL: 0.02% CY: 0.02% TL: 0.01% LV: 0.0077% etc...	Weekends: 27.92% Weekday: 72.08%
Parque Fluvial del Besos	Frequency: 274 User counts: 145 Frequency / User: 1.89	CA: 77.74% ES: 21.90% IT: 0.36%	Weekends: 20.80% Weekdays: 79.20%
Parque Molinet	Frequency: 2508 User counts: 722 Frequency / User: 3.47	FR: 32.74% CA: 22.69% EN: 16.63% ES: 11.60% FI: 0.24%	Weekends: 30.94% Weekdays: 69.06%

Table 6: Location Based Tweet mention statistics

Name	Statistics	Language Distribution	Time of week
Parque Molinet	Frequency: 2508 User counts: 722 Frequency / User: 3.47	IT: 0.24% PT: 0.16% EU: 0.16% CY: 0.12% ET: 0.04% IN: 0.04% RO: 0.04% etc...	Weekends: 30.94% Weekdays: 69.06%
Plaza del Rellotge	Frequency: 107 User counts: 47 Frequency / User: 2.27	CA: 100.00%	Weekends: 21.50% Weekdays: 78.50%
Rambla San Sebastion	N/A	N/A	N/A
Parque Can Zam	Frequency: 813 User counts: 361 Frequency / User: 2.25	CA: 54.00% ES: 34.93% EN: 6.27% IN: 0.86% TR: 0.86% RO: 0.86% HT: 0.49% PT: 0.37% IT: 0.25% PL: 0.12% NL: 0.12% FI: 0.12%	Weekends: 28.29% Weekdays: 71.71%
Instituto Can Peixauet	Frequency: 525 User counts: 170 Frequency / User: 3.09	CA: 72.76% ES: 14.29% PT: 12.19% EN: 0.38%	Weekends: 38.29% Weekdays: 61.71%

Table 6: Location Based Tweet mention statistics

Name	Statistics	Language Distribution	Time of week
		CS: 0.19% EU: 0.19%	
Parque Gran Sol	Frequency: 3694 User counts: 2289 Frequency / User: 1.61	ES: 89.58% CA: 4.30% EN: 3.55% PT: 0.54% IT: 0.43% HT: 0.38% FR: 0.32% JA: 0.24% EU: 0.16% IN: 0.05% etc...	Weekends: 28.21% Weekdays: 71.79%
Escuela Tanit	Frequency: 15 User counts: 13 Frequency / User: 1.15	CA: 80.00% ES: 20.00%	Weekends: 13.33% Weekdays: 86.67%
Instituto Terra Roja	Frequency: 106 User count: 62 Frequency / User: 1.71	CA: 67.92% ES: 30.19% PT: 0.94% EN: 0.94%	Weekends: 17.92% Weekdays: 82.08%
Instituto Gassol	N/A	N/A	N/A
CAP Santa Rosa	Frequency: 5 User counts: 4 Frequency / User: 1.25	ES: 60.00% IT: 20.00% CA: 20.00%	Weekends: 0% Weekdays: 100.00%
Cinto Verdaguer	Frequency: 347 User count: 256 Frequency / User: 1.36	CA: 80.40% ES: 16.43% PT: 1.73% EN: 0.58%	Weekends: 24.78% Weekdays: 75.22%

Table 6: Location Based Tweet mention statistics

Name	Statistics	Language Distribution	Time of week
		NO: 0.29% FR: 0.29% NL: 0.29%	
Mercado del Fondo	Frequency: 29737 User counts: 10558 Frequency / User: 2.82	ES: 95.17% IT: 4.34% CA: 0.31% EN: 0.07% PT: 0.04% IN: 0.02% JA: 0.0074% HT: 0.0064% TR: 0.0044% CY: 0.0039% ET: 0.0015% NL: 0.00099%	Weekends: 23.04% Weekdays: 76.96%
Nus de la Trinitat	Frequency: 659 User counts: 132 Frequency / User: 4.99	CA: 96.81% ES: 3.03% FR: 0.15%	Weekends: 15.48% Weekdays: 84.52%

3.4 Remarks

In comparing the location based statistics with the two user groups, the groups with additional keywords are not fit for comparison as in searching for just location, we could not make the same assumptions about the search and hence couldn't utilize a more specific range of keywords. Hence, for example, in 'Parque Fluvial del Besos,' there will be a discrepancy in the variety of languages as well as frequency of mentions between the user groups and location-based data due to this difference.

In the location-based data's langauage description, if the list was long, only the languages that were also present in the public space description of the user group data's were presented, and the rest excluded under 'etc...' Also, there was one other category in languages 'und' that classified any url based Tweets and non-identifiable languages (mostly url) which was excluded in the above description of the statistics.

The process of collecting the frequency data took a very long time as the more users I had to iterate through, there were more tweets to process. In addition, real time suspension and deletion of accounts would result in numerous errors that prompted a restart of the query and hence more time needed to complete the search. Therefore, the search method was revised so that I would only search 800 users at a time, therefore minimizing the additional time of search in case of deleted users (For reference, it takes about $10 \sim 13$ hours to run 800 users). The language code table for the language codes mentioned above is presented below.

Table 7: Language Code Table

Code	Language
AR	Arabic
CA	Catalan
CS	Czech
CY	Welsh
DA	Danish
DE	German
EN	English
ES	ES
ET	Estonian
EU	Basque
FI	Finnish
FR	French
HT	Haitian Creole
HU	Hungarian
IT	Italian
JA	Japanese
KO	Korean
LT	Lithuanian
LV	Latvian
NL	Dutch
NO	Norwegian
PL	Polish
PT	Portugese
RO	Romanian
RU	Russian
SL	Slovenian
SV	Swedish
TL	Tagalog
TR	Turkish
VI	Vietnamese
ZH	Chinese