WordCrowd – A Location-Based Application to Explore the City based on Geo-Social Media and Semantics

Kenzo Milleville[1,2,*], Dilawar Ali[1,2,*], Francisco Porras-Bernardez[3,*], Steven Verstockt[1], Nico Van de Weghe[2], Georg Gartner[3]

[1] IDLab, Ghent University-imec, Ghent, Belgium.
(kenzo.milleville@ugent.be, dilawar.ali@ugent.be, steven.verstockt@ugent.be)
[2] CartoGIS, Ghent University, Ghent, Belgium
nico.vandeweghe@ugent.be
[3] Research Division Cartography, TU Wien, Vienna, Austria.
(francisco.porras.bernardez@tuwien.ac.at, georg.gartner@tuwien.ac.at)
[*] These authors contributed equally to this work

Abstract. WordCrowd is a dynamic location-based service that visualizes and analyzes geolocated social media data. By spatially clustering the data, areas of interest and their descriptions can be extracted and compared on different geographical scales. When walking through the city, the application visualizes the nearest areas of interest and presents these in a word cloud. By aggregating the data based on the country of origin of the original poster, we discover differences and similarities in tourist interest between different countries. This work is part of the project Eureca: European Region Enrichment in City Archives and Collections of Ghent University (IDLab, CartoGIS), the Technical University of Vienna (Research Group Cartography) and several city and state archives from Ghent and Vienna.

Keywords. Geo-social Media, Location-based Service, Spatial Clustering, Word Clouds

LBS 2019
15th Conference on Location Based Services
Vienna, 11–13 November 2019

Published in "Adjunct Proceedings of the 15th International Conference on Location Based Services (LBS 2019)", edited by Georg Gartner and Haosheng Huang, LBS 2019, 11–13 November 2019, Vienna, Austria.

This contribution underwent double-blind peer review based on the paper. https://doi.org/10.34726/lbs2019.29 | © Authors 2019. CC BY 4.0 License.

1. Extraction and Visualization of Areas of Interest

A post on social media reflects the thoughts and feelings of the poster about a certain topic as a data point in space and time. By focusing on the location of the post instead of on the content, areas of interest (AOIs) can be extracted as areas with a higher post density. By spatially clustering these points, these AOIs are automatically extracted and most of the noise is filtered out. The dataset (Verstockt et al. 2019) used for this research consists of geolocated Flickr pictures and their associated tags. It covers continental Europe with metadata of all the images uploaded from 2004 to 2018. Nevertheless, our approach and application can work with any type of geolocated textual data.

The clustering technique is an essential part of this analysis and modifying it will impact the number of AOIs, their size, shape, and contents. In this research, HDBSCAN (Campello et al. 2013) is used as the main clustering algorithm. HDBSCAN is an extension of the popular density-based DBSCAN algorithm and performs better on datasets with varying densities. By changing the parameters of the algorithm, the clustering can be performed on different scales. This multi-scale clustering is necessary for an interactive LBS application, as the user might be overwhelmed with a large number of smaller clusters when he zooms out on the map. To ensure the application works in real time with a dynamic interface, we have preprocessed the data into multi-scale clusters and visualize only the nearby clusters instead of all the nearby points.

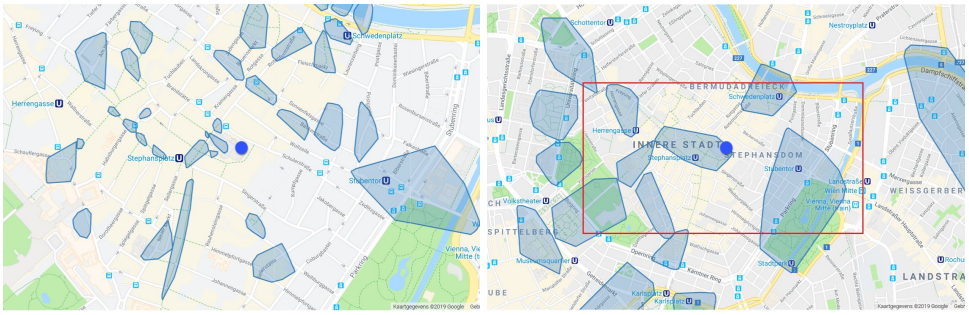

Figure 1. Visualization of the smallest clusters for a part of Vienna (left) and the larger clusters when the user zooms out (right). The user's current position is marked with a blue dot.

When the user zooms out, the application will fetch and visualize the larger nearby clusters from the database to reduce the network and memory load. *Figure 1* shows this functionality. Clicking an AOI displays its aggregated

tags in a word cloud. This provides an intuitive visualization of the tags contained in each AOI.

Initially, the points located in Austria (370,000 points) were clustered on three scales, resulting in 845, 79, and 8 clusters. For each cluster, we aggregate and preprocess all the related tags. The top 100 most frequently occurring tags and their frequency are then saved in our database. This gives us geolocated AOIs and their descriptions generated from the Flickr picture tags. Afterward, this process was repeated for all the points located in Belgium (430,000 points).

Because we are dealing with very noisy and multilingual picture tags, these need to be preprocessed to improve the generated word clouds. First, all the tags were translated into English to provide a common language for the following preprocessing steps. Next, irrelevant tags like brand names and stop words were removed. Afterward, NLP techniques (close/sequence matching and stemming/lemmatization) were used to group very similar words together. Finally, redundant multilingual place name tags were removed. Most pictures include a tag with the current place name, making that tag the most important one for that area. However, its inclusion in the word cloud is redundant, as the user already knows where he is or which area he is looking at on the map. These multilingual place names were filtered out with the use of Wikipedia and Wikidata.

These techniques made the resulting word clouds much clearer, but they still contain some errors. The most common errors are due to a bad translation or due to joined tags that are normally written with a space in between (e.g. *domkircheststephan)*. This is a common problem with social media tags. The emergence of tags relating to the name or company of the photographer is another problem that occurs within the word clouds of smaller clusters.

This spatial clustering of data shows us the AOIs for each region and its general description through the eyes of the crowd. The AOIs often coincide with landmarks and popular areas of each city. As a next step, we investigated if there is a difference in extracted AOIs when comparing people from different nationalities.

2. Tourism Interest Analysis

Only a fraction (32%) of the Flickr users provided information about their home location in their user profiles, limiting the available data for some countries of origin. To classify the other users, a home determination

method was developed based on Bojic et al. (2015). The method considers all the posts created by each user and the country in which he has the most pictures is considered as a potential country of residence. If the time span between the first and the last post is greater than 6 months, the user was classified as a resident of that country. This algorithm was validated on the fraction of users who supplied information about their country of residence. This information was first preprocessed with a gazetteer[1] to determine the English name of the city or country provided by the user. The developed algorithm achieved a precision of 0.87, a recall of 0.76 and an F_1 score of 0.81.

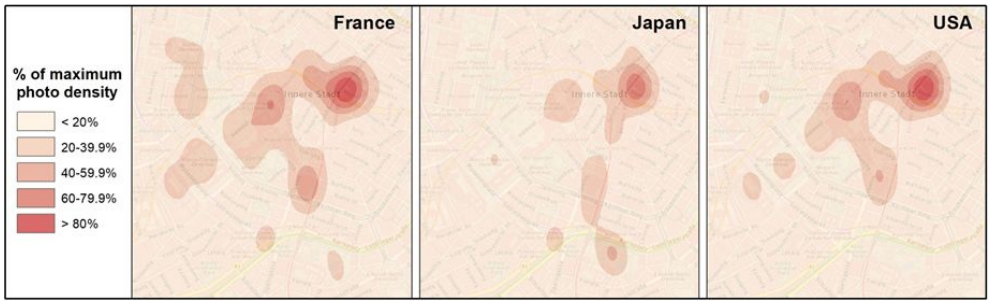

Figure 2. Footprints of visitors from France, Japan, and the USA in Vienna.

Kernel density estimation (KDE) (Grothe & Schaab, 2009) was selected as a visualization tool to generate continuous raster surfaces from the points. These rasters are heat maps representing areas with varying picture densities. KDE was chosen as most users are already familiar with the concept of heat maps and it is immediately clear where the hotspots are located. Each heat map shows the unique footprint of visitors from a certain country of origin. These heat maps can be compared for different countries, to analyze the differences in tourism interest. *Figure 2* shows the footprints of visitors from France, Japan, and the USA in Vienna. We see that the most popular areas of interest (the most popular tourist attractions) are shared and that larger differences occur in the less popular areas.

When looking at the generated tags for different nationalities, many tags are universal and widely used in the same locations. The most common tags are the more generic ones such as architecture, church, and travel. Between some nationalities, there can be major differences in the areas or topics of interest. *Figure 3* shows the word cloud for all points in Belgium from Dutch and English tourists. All of the points were included because the data is rather limited for specific regions of Belgium. It is clear that the interest of Dutch visitors, or at least those who posted on Flickr, is more focused on

[1] https://www.geonames.org/

leisure activities (*tomorrowland, motorsport*) whereas the English tourists are more focused on traditional tourism. Tags related to the First World War (*passchendaele, thegreatwar, memorial*) show up in the data for English tourists, as Belgium (especially Ypres) is often visited to commemorate the Great War and its casualties.

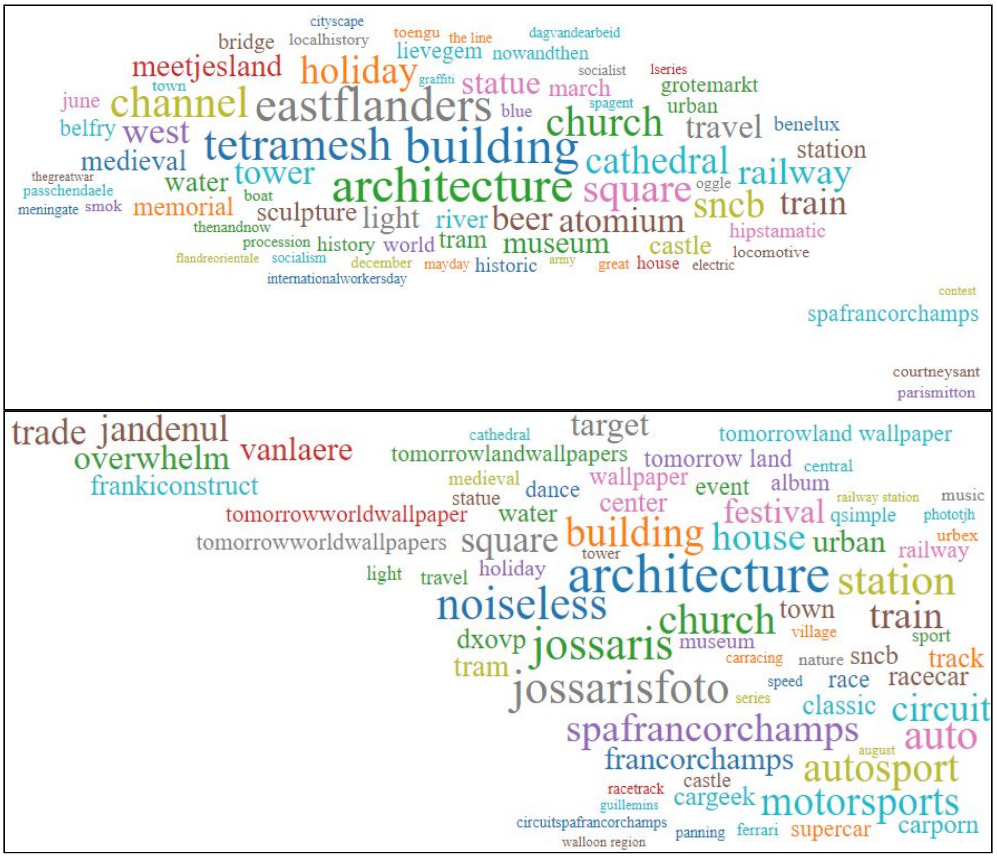

Figure 3. Word cloud of all points located in Belgium for English (top) and Dutch (bottom) tourists. Words are positioned relative to the user's location, which is Brussels for both word clouds.

3. Current State and Future Work

Currently, the AOIs situated in Belgium and Austria are extracted at three different scales and their tags have been preprocessed and clustered. These clusters are visualized in an interactive map, where the word cloud of each cluster is shown when it is clicked. The current prototype[2] is live and we

[2] https://tw06v072.ugent.be/wordcrowd/map/

have made the dataset[3] publicly available. As suggested by Tessem et al. (2015), the word cloud algorithm was adjusted based on the location of each tag. The positions of the words on the word cloud correspond to the location from where they were extracted, relative to the current user position. Both word clouds in *Figure 3* were constructed with Brussels as the user's location. The tag *Passchendaele* is located on the left side (west) and *Francorchamps* is grouped with motorsport-related tags on the bottom-right (southeast). This visualization offers the benefit that it often groups related tags from the same place together, at the cost of introducing additional whitespace.

As a next step, data aggregation and semantic clustering are the focus of this research. We aim to collect additional data sources, aggregate these based on mentioned place names via geoparsing and summarize the information with meaningful tags. Incorporating more advanced NLP techniques to hierarchically cluster semantically similar tags (e.g. replace *cathedral* with *church*) can also help reduce the noise in the generated word clouds. When a user shows interest in a specific tag of an AOI, he will be redirected to the original content. WordCrowd would then serve as an easy-to-use LBS tool to aggregate and explore the vast collection of public data available when visiting Europe.

References

Bojic, I., Massaro, E., Belyi, A., Sobolevsky, S., & Ratti, C. (2015). Choosing the right home location definition method for the given dataset. In International Conference on Social Informatics (pp. 194-208). Springer, Cham.

Campello, R. J., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In Pacific-Asia conference on knowledge discovery and data mining (pp. 160-172). Springer, Berlin, Heidelberg.

Grothe, C., & Schaab, J. (2009). Automated footprint generation from geotags with kernel density estimation and support vector machines. Spatial Cognition & Computation, 9(3), 195-211.

Tessem, B., Bjørnestad, S., Chen, W., & Nyre, L. (2015). Word cloud visualisation of locative information. Journal of location Based services, 9(4), 254-272.

Verstockt, S., Milleville, K., Ali, D., Porras-Bernárdez, F., Gartner, G., Van de Weghe, N. (2019). EURECA - EUropean Region Enrichment in City Archives and collections. 14th Conference on Digital Approaches to Cartographic Heritage (ICA DACH), Thessaloniki, Greece. 2019

[3] http://bit.ly/wordcrowd-dataset