Université de Rennes

Université de Rennes 1

ISTIC - Informatique et Électronique

Data Mining (FSY)

ISTIC - Informatique et Électronique

Authors:

José Antonio Ruiz Heredia

Pietro Manni

Teacher:

Bertrand Couasno

Date:

October 30, 2025

Contents

1 Introduction

Geo-localized data from social media and online platforms offer valuable insights for both commercial and public applications, such as monitoring crowds or detecting events. For this reason, Rennes Métropole seeks to identify tourism hotspots efficiently while using non-intrusive techniques. In this context, this project analyzes Flickr's geo-localized photos to automatically detect points of interest across the city of Rennes.

The selected dataset comprises over 50,000 images, each including descriptive metadata such as the author, date, or geographic coordinates, and also textual annotations such as titles and tags. This set of information enables detailed analysis of spatial and temporal patterns, which is essential for identifying locations with high tourist activity or recurring events.

The objective of this work is to develop an automated workflow, implemented in *KNIME* [5], to process and analyze the dataset. Key steps include data cleaning, attribute selection, visualization, and detection of locations with a significant density of photos, which serve as indicators of events or popular sites. This report presents the methodology applied and the results obtained in identifying these points of interest.

2 Data Preparation

The raw dataset provided consisted of 54,800 geo-localized photos, each described by several attributes including photo ID, author, date, location coordinates, and textual annotations. To ensure data quality and relevance for analysis, we have developed a series of preprocessing steps using *KNIME* desktop software.

Firstly, the dataset was imported into *KNIME* using the `CSV Reader` node. Then, we removed all the duplicate entries with the `Duplicate Row Filter` node, considering only the `photo_id` column to identify duplicated photos. This step significantly reduced the dataset from 54,800 to 4,195 unique pictures.

In addition, we deleted all the irrelevant or redundant attributes using the `Column Filter` node. Specifically, the columns `row_id` (a duplicate of the existing index), `date_taken_time`, `date_taken_year` `date_taken_quarter`, `date_taken_month`, `date_taken_day_of_month`, `date_taken_day_of_week`, `date_taken_day_of_year`, and `date_taken_week_of_year` were excluded to simplify the analysis.

Moreover, we observed that the values in the `tags` column were neither relevant nor useful for the analysis. Many entries contained nonspecific or unrelated information, such as "*voigtlander f095 prime lens fixed focal*" or "*lysistrata paleblueskin idealcrash lecartelloco*". Since this column did not provide meaningful insights and the data was too noisy to process effectively, we also removed it using the `Column Filter` node.

Secondly, to focus the analysis on Rennes, the dataset was spatially delimited using the `Geo-Coordinate Row Filter` node, filtering the latitude and longitude to the city boundaries. As visualized in Figure 1, the remaining photos are located within the city of Rennes.

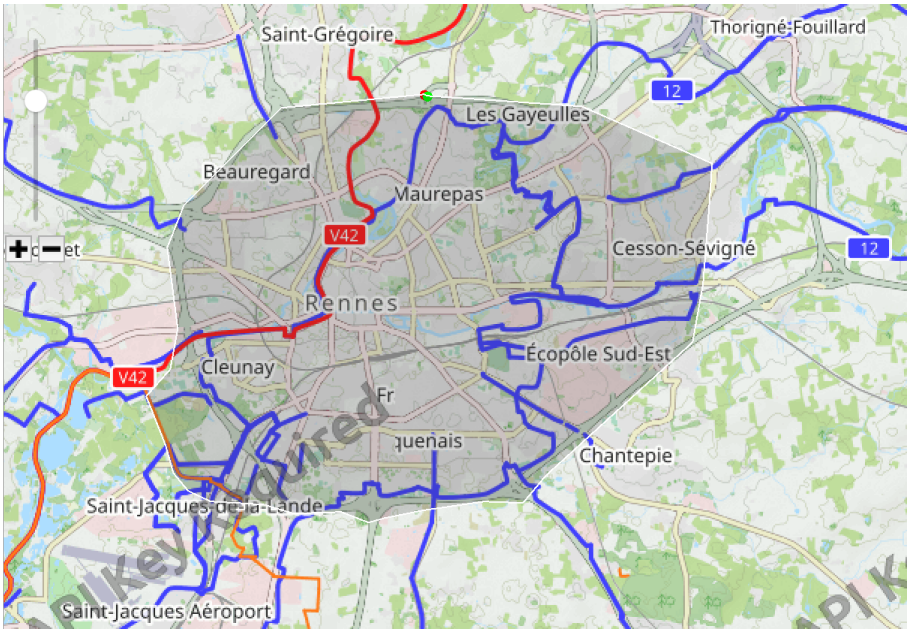


Figure 1: Geo-Coordinate row filter of the city of Rennes

Thirdly, we aimed to extract meaningful keywords that describe each cluster from the `title` column. To ensure cleaner text for analysis, we added a `String Manipulation` node to remove special characters and numbers, resulting in a standardized and noise-free string.

Additionally, we used the `Cell Splitter` node to separate the `title` column into a list of individual words, which were then expanded into separate rows using the `Ungroup` node. After splitting the titles into distinct words and rows, we grouped them by word using the `GroupBy` node to count their occurrences.

The resulting rows were then sorted by count with the `Sorter` node. This distribution is illustrated in Figure 2, where the most frequently occurring keywords are *rennais*, *atana*, *studio*, and *rennes*.

Rows: 1476 | Columns: 4

#	RowID	title_SplitResultList (String)	Count*(id_photo) (Number (Integer))	Mean(lat) (Number (Float))	Mean(long) (Number (Float))
1	Row11	rennaise	1560	48.139	-1.641
2	Row70	atana	1159	48.104	-1.672
3	Row12	studio	1159	48.104	-1.672
4	Row11	rennes	1086	48.108	-1.675
5	Row12	streets	565	48.106	-1.671
6	Row75	lego	139	48.099	-1.671
7	Row78	automne	85	48.103	-1.672
8	Row78	loft	71	48.099	-1.671
9	Row51	fog	69	48.105	-1.677
10	Row54	fun	69	48.105	-1.676

Figure 2: Frequency of Title Words Sorted by Number of Appearances

3 Discovering Points of Interest

3.1 First look at the clusters

First, Figure 3 provides an overview of the available points within the city of Rennes. From this visualization, we can conclude that the majority of points in the dataset are concentrated in the city center, particularly around key points of interest, while peripheral areas are less represented.

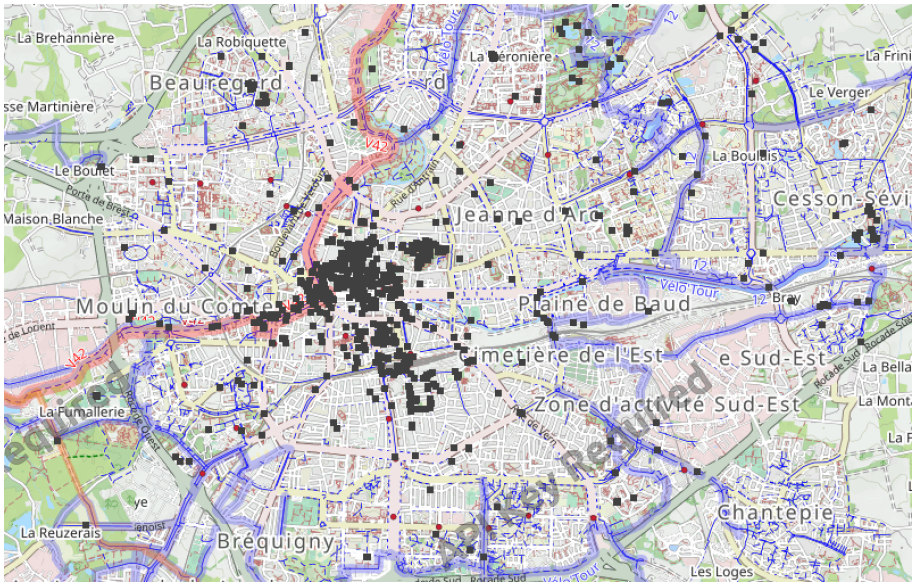


Figure 3: Map of Rennes with points of interest

After analyzing the available points, Figure 4 provides a first view of the clusters obtained based on `latitude` and `longitude`. We used an initial configuration of 50 clusters generated with the `k-Means` node and displayed with the `OSM Map View` node.

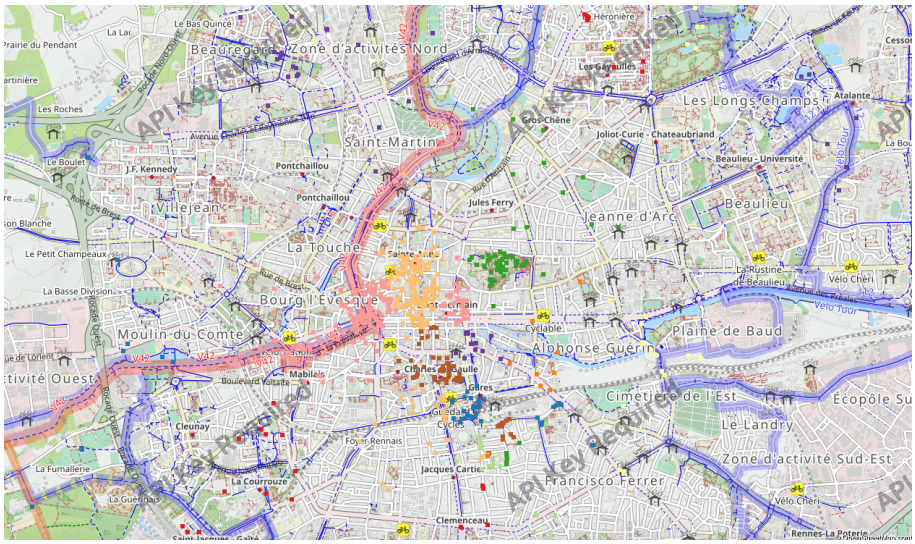


Figure 4: OSM map view of clusters

In Figure 5, we have a closer look at the city center where most of the points are concentrated. We can observe that some clusters correspond to well-known neighborhoods or points of interest, such as *Gares*, *Saint-Germain*, *Sainte-Anne*, and *Parc du Thabor*. However, there are other clusters that do not align with any specific or widely recognized locations, and some areas are divided into multiple clusters. This suggests that the current clustering configuration may not capture the spatial patterns of the city center. To improve the results, it may be necessary to adjust the number of clusters and perform additional data cleaning to extract more meaningful and actionable insights.

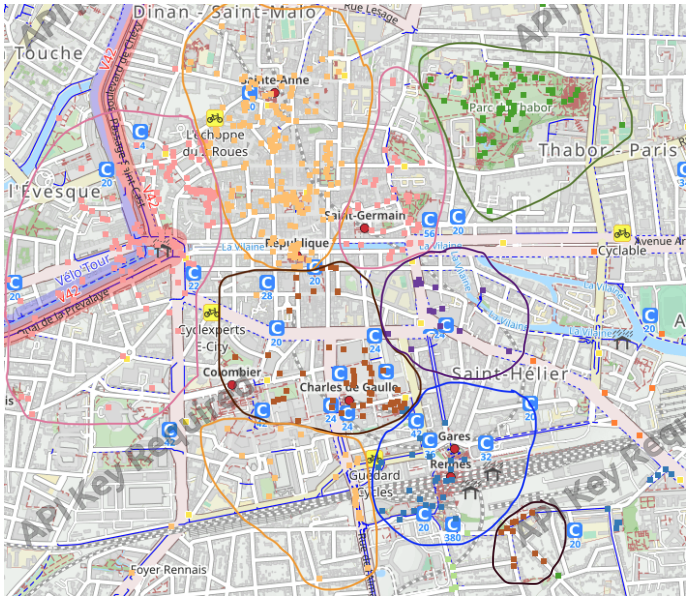


Figure 5: Clusters delimited around the center of Rennes

In addition, we increased the number of clusters in the configuration to 200 to gain more detailed insights into the segmentation of the city and its points of interest. In Figure 6, we can observe that some areas that were previously represented by a single cluster are now subdivided into multiple clusters.

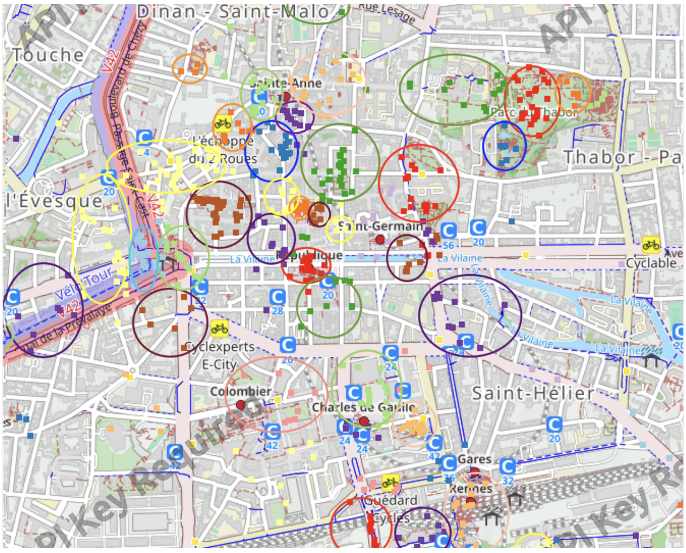


Figure 6: Number of clusters increased around the center of Rennes

This new segmentation was intended to reveal additional meaningful points of interest that might have been obscured in the broader clustering. However, while it did increase the granularity of the analysis, it also resulted in an oversaturation of clusters, many of which could not be clearly associated with recognizable locations or landmarks. This suggests that simply increasing the number of clusters does not necessarily improve the interpretability of the results, and highlights the importance of balancing cluster resolution with meaningful spatial patterns.

3.2 Reduction of points with title words segmentation

For these new experiments, we applied the filtering procedure described in the Data Preparation section. Specifically, we focused on transforming the `title` column to retain only the main keywords while removing repeated or irrelevant entries. By applying this process, the dataset was reduced from the initial 54,800 photos (raw dataset) to around 1,500 meaningful points (final filtered and normalized dataset), ensuring that only the most informative points and distinct titles were retained for analysis.

The final distribution of these points is visualized in Figure 7, which shows the clustering result with a configuration of 50 clusters. This filtering step was crucial to improve the quality and interpretability of the clustering outcomes, as it minimizes noise and emphasizes useful data.

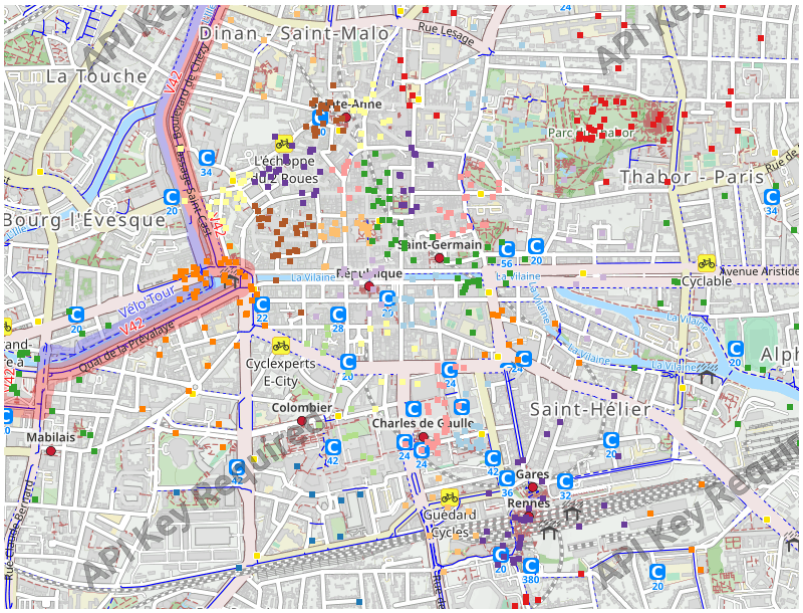


Figure 7: Clusters with a reduction in the points displayed by title words

In Figure 8, to better illustrate our filtering and segmentation of title keywords, we can observe, for instance, *cluster_39*, which corresponds to the area of La République. In this cluster, the keywords *biennale* and *contemporain* are predominant, each appearing 38 times, alongside other terms such as *edit*, *center*, *day*, and *alignments*.

At first, this cluster may seem uninformative. However, focusing on the most descriptive keywords, such as `park`, `architecture`, `people`, `retail`, and `outdoors`, reveals meaningful insights. These titles reveal streets with notable buildings, human activity, open public space with parks and vegetation, and shops or markets with retail activity. This provides a meaningful description that can be visually associated with the cluster corresponding to *Place de la République* in Rennes on the map.

RowID	title_SplitResultList String	Count*(id_photo) Number (Integer)	Mean(lat) Number (Float)	Mean(long) Number (Float)	Cluster String
Row13	biennale	38	48.111	-1.678	cluster_39
Row31	contemporain	38	48.111	-1.678	cluster_39
Row42	edit	13	48.11	-1.68	cluster_39
Row23	centre	8	48.109	-1.679	cluster_39
Row35	day	8	48.11	-1.677	cluster_39
Row19	alignements	7	48.108	-1.678	cluster_39
Row13	tout	7	48.109	-1.678	cluster_39
Row18	bus	6	48.11	-1.68	cluster_39
Row60	guénaël	6	48.11	-1.679	cluster_39
Row10	pour	5	48.109	-1.678	cluster_39
Row11	rues	5	48.109	-1.677	cluster_39
Row66	imag	4	48.109	-1.678	cluster_39
Row55	architecture	3	48.108	-1.677	cluster_39
Row10	people	3	48.11	-1.679	cluster_39
Row77	auto	2	48.109	-1.68	cluster_39
Row32	couleurs	2	48.109	-1.679	cluster_39
Row76	let	2	48.11	-1.679	cluster_39
Row99	park	2	48.109	-1.677	cluster_39
Row10	person	2	48.11	-1.679	cluster_39
Row13	vasselot	2	48.109	-1.678	cluster_39
Row60	arrêt	1	48.109	-1.678	cluster_39
Row18	built	1	48.11	-1.679	cluster_39
Row21	cageots	1	48.109	-1.677	cluster_39
Row45	exterior	1	48.11	-1.679	cluster_39
Row54	full	1	48.11	-1.679	cluster_39
Row59	group	1	48.11	-1.679	cluster_39
Row61	halles	1	48.108	-1.68	cluster_39
Row64	homme	1	48.109	-1.677	cluster_39
Row67	indoors	1	48.11	-1.679	cluster_39
Row75	length	1	48.11	-1.679	cluster_39

Figure 8: Table of the different keyword titles in a same cluster

In Figure 8 we can visualize the spatial distribution of the points belonging to *cluster_39* (highlighted in purple). This indicates a clear concentration around *Place de la République*, confirming the relevance of this area as a point of interest.

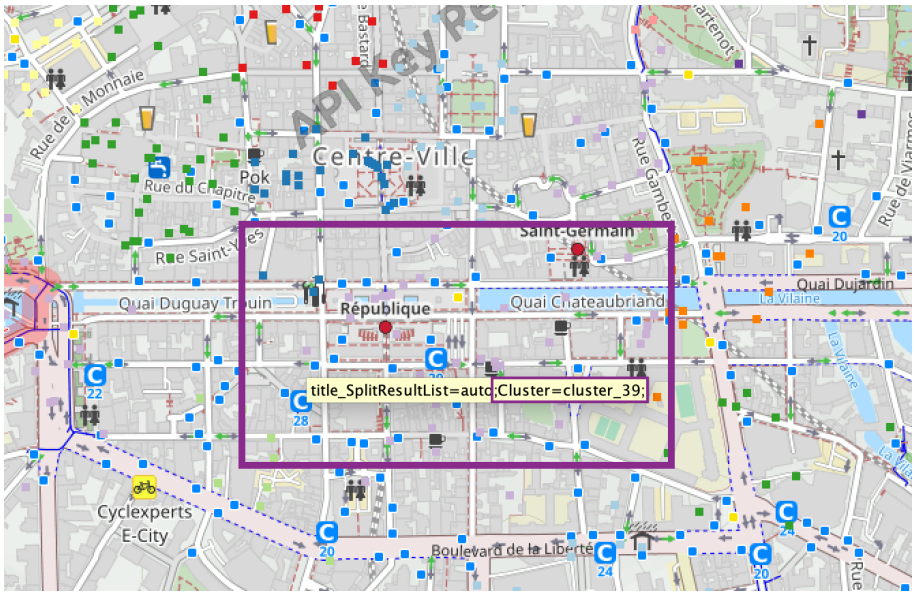
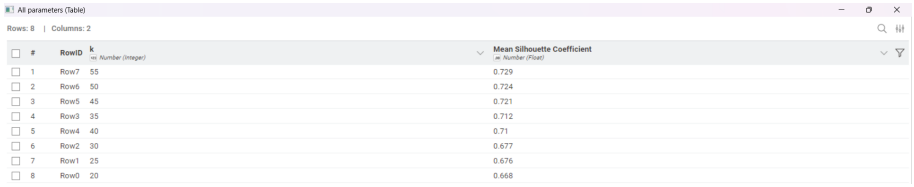


Figure 9: Cluster_39 of *La République*

3.3 Best model selection

We used the `Optimized K-Means (Silhouette Coefficient)` node to select the number of clusters k from geo-normalized coordinates (latitude/longitude). The component runs K-Means across a range of k values and chooses the model with the highest average silhouette (in $[-1, 1]$, higher is better; it reflects compactness within clusters and separation between clusters). To avoid overly fragmented solutions which can maximize silhouette but yield tiny, uninterpretable POIs, we restricted the search to a practical range (e.g., $k \in [20, 55]$ with step 5). Overall, this procedure balances numerical quality (high silhouette) with practical interpretability (well-populated, meaningful POIs).

In Figure 10 we can observe that we tested several values of k and looked at the average silhouette. $k = 55$ gives 0.729 and $k = 50$ gives 0.724, which is almost the same. We chose $k = 50$ because, on the map, it splits a few areas a bit more clearly into meaningful POIs, while keeping the same clustering quality.



All parameters (Table)

Rows: 8 | Columns: 2

#	RowID	k (Number (Integer))	Mean Silhouette Coefficient (Number (Float))
1	Row7	55	0.729
2	Row6	50	0.724
3	Row5	45	0.721
4	Row3	35	0.712
5	Row4	40	0.71
6	Row2	30	0.677
7	Row1	25	0.676
8	Row0	20	0.668

Figure 10: Table showing the best K

In Figure 11 we can visualize the final map of the center of Rennes with the optimized number of clusters set to 50.

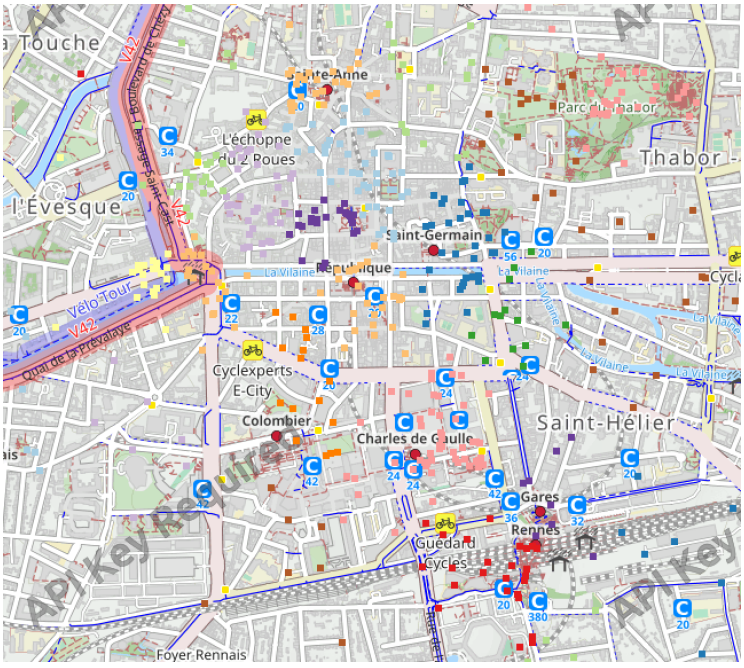


Figure 11: K-means optimized to 50 clusters

4 Characterizing Points of Interest

To understand and validate the clusters of candidate points of interest, we developed a text processing and itemset mining pipeline in *KNIME*. The goal of this process was to study the textual patterns that describe the items more deeply.

The workflow begins by aggregating the textual information available for each photo (`title` and `tags`) into a single text document representing that image. Each document is then transformed into a structured representation suitable for text mining using the `Strings to Document` node. The preprocessing steps include:

1. **Text cleaning:** removal of punctuation, numeric values, and tokens shorter than three characters to eliminate irrelevant noise.
2. **Normalization:** conversion of all words to lowercase using the `Case Converter` node.
3. **Stop word filtering:** removal of very common words such as artciles or other meaningless words for the study.
4. **Stemming:** application of the `Snowball Stemmer` node to reduce words to their linguistic roots.

After cleaning, each document is converted into a *bag-of-words (BoW)* representation, where each term corresponds to a feature. This representation is further transformed into a binary vector that represents the appearance of each term in a document.

To analyse the textual content of each cluster, the pipeline applies the `Itemset Finder` node (*Borgelt's algorithm*) separately to the subset of documents belonging to each cluster. This algorithm identifies frequent itemsets, combinations of terms that appear in at least a specified number of documents within a cluster. Each found itemset can be interpreted as a descriptive pattern that summarises the main semantic content of that cluster.

In Figure 12, we can visualize the final table of the association rules for *cluster_18* corresponding to *Parc du Thabor*.

Rows: 29 | Columns: 4

#	RowID	ItemSet (Collection (Set))	ItemSetSize (Number (Integer))	ItemSetSupport (Number (Integer))	RelativeItemSetSupport% (Number (Float))
28	Row27	[parc,thabor]	2	31	37.349
29	Row28	[ren,thabor]	2	26	31.325
26	Row25	[parc,ren,thabor]	3	20	24.096
27	Row26	[parc,ren]	2	20	24.096
22	Row22	[fleur,thabor]	2	14	16.867
25	Row24	[jardin,thabor]	2	13	15.663
16	Row15	[fleur,parc]	2	12	14.458
19	Row18	[fleur,parc,thabor]	3	11	13.253
1	Row0	[roserai,fleur,parc,...]	4	9	10.843
2	Row1	[roserai,fleur,parc,...]	4	9	10.843
3	Row2	[roserai,fleur,parc]	3	9	10.843
4	Row3	[roserai,fleur,ren,...]	4	9	10.843
5	Row4	[roserai,fleur,ren]	3	9	10.843
6	Row5	[roserai,fleur,thabor]	3	9	10.843
7	Row6	[roserai,fleur]	2	9	10.843
8	Row7	[roserai,parc,ren,...]	4	9	10.843
9	Row8	[roserai,parc,ren]	3	9	10.843
10	Row9	[roserai,parc,thabor]	3	9	10.843
11	Row10	[roserai,parc]	2	9	10.843
12	Row11	[roserai,ren,thabor]	3	9	10.843
13	Row12	[roserai,ren]	2	9	10.843
14	Row13	[roserai,thabor]	2	9	10.843
15	Row14	[ros,thabor]	2	9	10.843
17	Row16	[fleur,parc,ren,...]	4	9	10.843
18	Row17	[fleur,parc,ren]	3	9	10.843
20	Row19	[fleur,ren,thabor]	3	9	10.843
21	Row20	[fleur,ren]	2	9	10.843
23	Row22	[jardin,ren,thabor]	3	9	10.843

Figure 12: Table of Association Rules of Cluster_18

5 Conclusions

This study demonstrates the potential of geo-localized social media metadata for identifying points of interest within an urban area. By analyzing over 50,000 geo-tagged photos of the city of Rennes from Flickr, we developed a robust workflow in *KNIME* that integrates data cleaning, filtering, spatial clustering, and text mining to extract meaningful insights from both geographic and textual information.

The preprocessing steps improved the quality of the dataset by implementing different cleaning and normalization techniques, reducing the dataset to a subset of 1,385 informative points. This filtering enabled a more precise and interpretable clustering of locations, highlighting the central areas of Rennes and key neighborhoods such as *Place de la République* or *Parc du Thabor*.

The application of K-Means clustering, optimized with the silhouette coefficient, revealed the optimal amount of clusters to group the data corresponding to well-known points of interest. While increasing the number of clusters improved granularity, it result in a negative impact on the interpretability of the clusters.

And finally, the text processing and itemset mining pipeline characterizes the semantic content of each cluster, complementing the spatial analysis. By aggregating and cleaning the textual information from photo titles and tags, we were able to extract meaningful patterns that describe the main activities, landmarks, or features within each point of interest. The resulting association rules, as illustrated for *cluster_18*11 (*Parc du Thabor*), provideing interpretability of the identified points of interest.

References

[1] Article de le monde. http://www.lemonde.fr/economie/article/2015/02/25/votrejob-quand-twitter-s-aventure-sur-le-terrain-de-pole-emploi_4582863_3234.html

[2] Data publica : Crawling et au scraping (livre blanc). http://www.data-publica.com/content/2013/09/le-livre-blanc-de-data-publica-consacre-au-crawling-et-au-scraping/

[3] Demo d'un excellent projet 4IF, INSA de Lyon. http://www.data-publica.com/content/2013/09/le-livre-blanc-de-data-publica-consacre-au-crawling-et-au-scraping/

[4] Demo d'un projet d'´etudiants, UCBL, Lyon. http://liris.cnrs.fr/mehdi.kaytoue/sujets/ter-meanshift/demo1.html

[5] KNIME Analytics Platform, KNIME. https://www.knime.com