Data Mining project: Discover and describe areas of interest and events from geo-located data

Diana Nurbakova (diana.nurbakova@insa-lyon.fr)
Killian Barrere (killian.barrere@insa-lyon.fr)

2025 – 2026

Project details

The project is to be carried out in **groups of 3 students**[1], during 3 practical sessions lasting 4h each.

You will be implementing code using Python, which is commonly used for most data mining tasks thanks to many packages: scikit-learn (sklearn) that provide many learning algorithms and clustering algorithms, as well as Natural Language Toolkit (nltk) to tackle text processing tasks.

The project is intentionally provided with little guidance. We encourage you to try new things (new ideas, clustering algorithms not seen in classes, further work, etc.) and we will take this into account during your evaluation.

You may use generative AI to assist you in coding, or developing ideas, but you MUST use it in a responsible way: you are accountable for the usage of the AI, the produced code (you must be able to explain every line of your code and justify why you use it) and your learning. We advise you to use AI to discover new things and to get a better understanding. Lastly, remember that, as professors, we are conversational agents trained over many years. Feel free to ask us any questions or for advice. We (might) provide you with better answers and feedback, even with poor prompting!

Context

You have submitted your proposal to a public call for tenders from Grand Lyon and won it (congratulations!). In order to improve public transport and the lives of tourists visiting Lyon, Grand Lyon asks you to find areas with high densities of tourists using a cost-effective and non-intrusive way.

We can then think about a solution capable of retrieving information from the Web (using crawling/scraping), such as geolocated pictures. The aim is to **automatically find areas of interest, events, etc., by grouping together data coming from a large database of geolocated pictures**. For instance, 3,000 pictures of the Eiffel Tower are expected to match together with a single area of interest.

Pedagogical aims of the project

- Implement and use techniques for handling large collections of data.

[1] and 1–2 group(s) of 2 students in case the number of students is not divisible by 3. We will adapt the evaluation criteria.

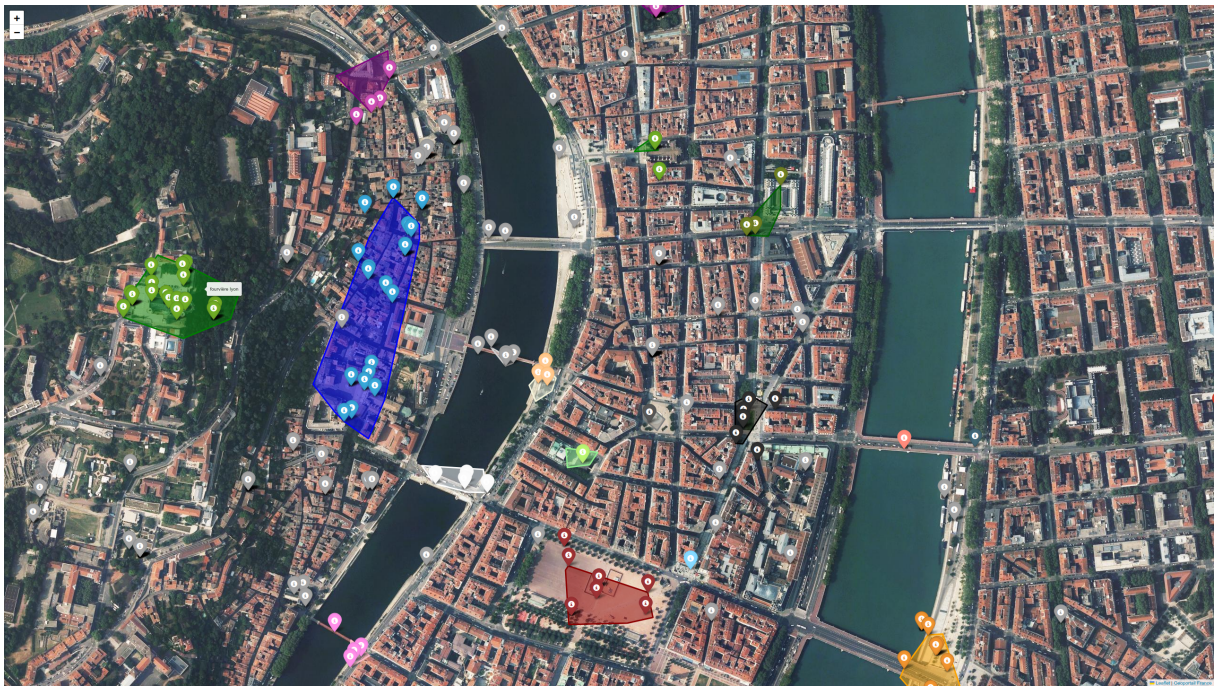

Figure 1: An example of a possible output.

- Experiment with various clustering algorithms (k-means, hierarchical clustering, DBSCAN, etc.) explore and report about their inputs and outputs, the meaning and influence of their parameters, algorithms' complexity, as well as their pros and cons.
- Discover and apply text processing algorithms.
- Demonstrate scientific methodology and rigor in your choices: questioning, hypotheses and justifications.

Data

Your team already collected geo-located data using the Flickr's API (your Web scrapper deserves a raise!). The dataset contains more than 400,000 rows of data describing photos. Each picture is described using the following format:
$\langle id\_photo, id\_photographe, latitude, longitude, tags, description, dates \rangle$.

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	user	lat	long	tags	title	date_taken_minute	date_taken_hour	date_taken_day	date_taken_month	date_taken_year	date_upload_minute	date_upload_hour	date_upload_day	date_upload_month	date_upload_year
2	77161041@N00	45.768121	4.801776	square,sierra,squareformat,iphone	Enfin. #instabeer #beer #chimay	46	18	24	11	2015	46	18	24	11	2015
3	11328031@N03	45.7597	4.8422	square,squareformat,iphoneogr	https://www.facebook.com/Pasi	3	17	24	11	2015	3	17	24	11	2015
4	132999708@N08	46.027633	4.699855		compagnons_dev_arnas20 (1)	0	15	7	11	2015	50	15	24	11	2015
5	132999708@N08	46.027572	4.699966		compagnons_dev_arnas20 (3)	1	15	7	11	2015	50	15	24	11	2015
6	133835212@N03	45.699105	4.474932	sunset,sky,cloud,sun,soleil,cou	Un soir dans les Monts du Lyon	20	20	31	8	2015	50	13	24	11	2015
7	129394312@N07	45.763429	4.849875	france,architecture,lyon,offices,h	hiCity, Lyon, France, 2015	11	16	7	9	2015	21	9	24	11	2015
8	19710808@N05	45.739289	4.814243	orange,building,architecture,lyon	edificio arquitectura arangoiope	29	12	25	6	2015	12	9	24	11	2015
9	35210768@N02	45.762661	4.8269	square,squareformat,iphoneogr	@Bidule_officiel C'est à la Ren	2	23	23	11	2015	2	8	24	11	2015
10	124810342@N04	45.586944	4.774444	france,animaux,fr,oiseau,rhone	Grand Cormoran (Phalacrocor	55	13	3	10	2015	7	7	24	11	2015
11	124810342@N04	45.586944	4.774444	france,animaux,fr,oiseau,rhone	Grand Cormoran (Phalacrocor	54	13	3	10	2015	7	7	24	11	2015
12	124810342@N04	45.586944	4.774444	france,animaux,fr,oiseau,rhone	Martin-pêcheur d'Europe (Alce	39	13	3	10	2015	7	7	24	11	2015
13	124810342@N04	45.586944	4.774444	france,animaux,fr,oiseau,rhone	Martin-pêcheur d'Europe (Alce	39	13	3	10	2015	7	7	24	11	2015
14	124810342@N04	45.586944	4.774444	france,animaux,fr,oiseau,rhone	Martin-pêcheur d'Europe (Alce	38	13	3	10	2015	7	7	24	11	2015
15	124810342@N04	45.586944	4.774444	france,animaux,fr,oiseau,rhone	Grand Cormoran (Phalacrocor	33	13	3	10	2015	7	7	24	11	2015
16	124810342@N04	45.586944	4.774444	france,animaux,fr,oiseau,rhone	Grand Cormoran (Phalacrocor	33	13	3	10	2015	6	7	24	11	2015
17	124810342@N04	45.586944	4.774444	france,animaux,fr,oiseau,rhone	Grand Cormoran (Phalacrocor	33	13	3	10	2015	6	7	24	11	2015
18	124810342@N04	45.586944	4.774444	france,animaux,fr,oiseau,rhone	Grand Cormoran (Phalacrocor	33	13	3	10	2015	6	7	24	11	2015
19	124810342@N04	45.586944	4.774444	france,animaux,fr,oiseau,rhone	Grand Cormoran (Phalacrocor	33	13	3	10	2015	6	7	24	11	2015
20	124810342@N04	45.586944	4.774444	france,animaux,fr,oiseau,rhone	Grand Cormoran (Phalacrocor	33	13	3	10	2015	6	7	24	11	2015
21	124810342@N04	45.586944	4.774444	france,animaux,fr,oiseau,rhone	Grand Cormoran (Phalacrocor	33	13	3	10	2015	6	7	24	11	2015
22	124810342@N04	45.586944	4.774444	france,animaux,fr,oiseau,rhone	Grand Cormoran (Phalacrocor	33	13	3	10	2015	6	7	24	11	2015
23	124810342@N04	45.586944	4.774444	france,animaux,fr,oiseau,rhone	Grand Cormoran (Phalacrocor	32	13	3	10	2015	6	7	24	11	2015
24	124810342@N04	45.586944	4.774444	france,animaux,fr,oiseau,rhone	Martin-pêcheur d'Europe (Alce	32	13	3	10	2015	6	7	24	11	2015
25	124810342@N04	45.586944	4.774444	france,animaux,fr,oiseau,rhone	Martin-pêcheur d'Europe (Alce	32	13	3	10	2015	6	7	24	11	2015
26	124810342@N04	45.586944	4.774444	france,animaux,fr,oiseau,rhone	Grand Cormoran (Phalacrocor	32	13	3	10	2015	5	7	24	11	2015

Figure 2: Illustration of some extracted data.

It is possible to access a photo on `https://www.flickr.com/photos/<user>/<id>`, where `<user>` is the user identifier and `<id>` is the photo identifier, e.g. `https://www.flickr.com/photos/95450872@N03/45122361361`.

1 Discovering areas of interests using clustering

The first, and most important objective that your team has to achieve is to **automatically find areas of interest** in the city of Lyon. An area of interest is defined as a localized area of varying size with strong photo-taking activity. You may follow the following steps of knowledge discovery in databases:

- Understanding of the data, data clearing and preparation, visualizations as well as useful statistics. In particular, it is expected that you check data coherency (dates and GPS positions); remove duplicates; visualize geolocated points on a map (e.g. using Folium[2]) ; etc.
- Selection of relevant attributes to perform analysis of data.
- Data mining using clustering: k-means, hierarchical clustering, as well as DBSCAN (you might consider using Scikit-Learn). Comparing and discussing these different approaches are of important matter.
- Evaluation, interpretation, visualization (mostly using a map) and discussing of the results. How is your analysis might help Grand Lyon, and what knowledge did you extracted?

The last step is often overlooked, but is nevertheless crucial. A data mining output is of no interest if it cannot be acted upon: it must be used for something, and the instructions for use must be provided.

Please notice that **you are expected to detail and discuss the main steps** in your presentations.

2 Description of areas of interest using text pattern mining

The first objective of the project enabled your team to extract and discover candidates' area of interest. However, a second step of validation and understanding of the different areas is missing.

The second objective aims to use textual data (title and tags of the publication) to describe and understand each area of interest. Your understanding might enable you to improve the first steps of the project (feel free to tell us about it!).

As this second objective extends further than the scope of the course, we provide a small tutorial with some hints to help you in the process. Yet, you are also expected to learn by yourself these useful notions. Consider reading additional resources online. To implement the code, you might use either scikit-learn and/or Natural Language Toolkit Python libraries.

2.1 Preprocessing

As with other types of data, data preprocessing plays a major role with textual data.

- Removing stop words (words that are used a lot while not bringing meaningful information, such as “is”, “the”, “a”, etc. or their french equivalent “est”, “le”, “un”, etc.).
- Similarly, it will be interesting to remove frequent words in the dataset that are not a stop word or meaningful (e.g. “picture”). You might consider visualizing the data with a word cloud.

[2] https://python-visualization.github.io/folium/latest/

- A common processing technique for text processing is to tokenize the text (split a sentence/text into smaller units). A basic solution will be to split a sentence into words, while a more advanced technique consist in splitting a sentence in lexical units, enabling to retrieve the root of a word (e.g. "drinking" will be split into "drink" and "-ing"). These techniques are generally available in libraries through tokenizers.

- Depending on the approach chosen, it might also be meaningful to create binary features representing whether a word exists in a sentence.

2.2 Term frequency and inverse document frequency (TF-IDF)

A first approach to find words describing an area of interest is to study word frequencies.

You might want to have a look to term frequencies (TF) (how frequent a word appears in a text), and describe a cluster using the words with the higher frequencies. However, a problem that might arise is that a given term might have a high frequency for many areas of interest, therefore being unmeaningful to identify a single area of interest.

Hence, it is important to compare the TF to the document frequency (DF) (how frequent is the word in all documents/texts). The Term frequency and inverse document frequency (TF-IDF) metric provides a score that shows how meaningful each word in a sentence is.

2.3 Association rules

A second approach will be to use association rules. The goal is then to find a set of items (words or lexical units) that best describe an area of interest.

3 Events: study of dense areas through time and space

As a third objective of this project, **you are expected to study whether the extracted areas of interest are located in time**. Indeed, each area of interest can be a one-time event, a recurrent event (such as the Fête des Lumières), or not correlated with any specific event.

It might be necessary to adapt the data preparation, clustering algorithm, and pattern mining steps. You will describe and discuss your different choices.

For this third objective, you have more freedom to explore, study and discuss one thematic of your choice as long as you explore something related to temporal axis.

4 Further work?

If you really liked this project (you can first tell us so!) and want to explore further ideas, below are some ideas that you might consider exploring. Of course, you might explore any idea from your imagination.

- You might consider using and extracting further data from other sources (e.g. Instagram). With this, you are free to explore and experiment deeper with more data extraction techniques.

- You might want to improve the text processing, with further algorithms to describe the areas of interest. Why not using some generative AI to automatically describe the clusters?

- It might not be useful for learning useful things related to your studies, but you might apply your methods to other of your favorite areas around the world. Who does not want to know why people are so eager to visit La Creuse?

5 Evaluation details

At the end of the project, you will be evaluated during a presentation of your works in a demonstration format. **The final presentation should last no more than 10'** (plus 5' of questions). The final presentation will take place between February 2 and February 6 during the launch break. You should discuss your **methodology**, the **algorithms used**, your **experimentations** and your **main results** (what information did you find in the data and how could it be useful for Grand Lyon?). Your presentation shall show your comprehension, scientific rigor and methodology. You are encouraged to highlight a functionality of your choice (a part of your project implementation you are most proud of).

In addition, you are expected to reach milestones at the end of each practical session. Your work and progress will be evaluated and validated at the beginning of the next practical session (5' + questions). The last milestones will be evaluated during the final presentation.

Milestones for the 1st session (5/20 points):

- Explore the data and discover the most important problems;
- Propose and implement a data cleaning (your are expected to cover most important problems);
- Implement a working visualization application with a working map, display some data points on the map;
- Show a first (working) clustering algorithm of your choice (not yet expected to appear on the map), even a very early work that does not yet provid a meaningful result.

Milestones for the 2nd session (5/20 points):

- Complete the remaining parts of the data cleaning;
- Try 3 clustering algorithms and optimize the parameters to get meaningful clusters, show a visualization of the clusters on the application;
- Implement a first text pattern mining algorithm to find words describing a given cluster.

Milestones for the 3rd session (10/20 points):

- Complete the optimization of the clustering algorithms, compare and discuss the results obtained and decide which algorithm you recommend;
- Finalize the implementation of 2 text mining algorithms to automatically name clusters, make it appear on the application;
- Explore the data through the scope of time, show the results of your exploration;
- Prepare the final demo to avoid any bug during the demonstration.

Please note that we encourage you to try new things and take this into account during your evaluation. Feel free to do so and ask us for advice.