TP 3-5 - Mini-project: geo-localised media

This mini-project is a modified version of a project created by Mehdi Kaytoue and Jean-François Boulicaut at INSA Lyon in 2017. The work is to be done in pairs and submitted in the form of a 6-page report presenting your solution and the results obtained. The file `.knwf` *of your workflow will also have to be submitted.*

1 Context : geo-localised media

In recent years, web, smartphone and tablet applications have flourished, providing a wide range of services (restaurant recommendations, hotel reservations, etc.). Some of these applications use the mass of information extracted from media such as social networks (Facebook, Twitter, Instagram, ...) to offer services where the geolocation of the media in question plays a crucial role, for example to suggest you to visit a tourist site that is very popular with your social network's contacts and that is close to where you are. These *distilleries of the social Web* filter the mass of messages to only keep their essence or added value.

Local authorities and governments are also interested in using these masses of information : we can monitor crowd movements in a city, track an epidemic of *dengue fever* in Brazil, discover events and influential users on social networks, etc. Companies also understand the importance of this data and are looking to exploit it to automatically evaluate the presence of their brand in the various social networks, to identify influential actors or even spontaneous unknown *hashtags*, etc. Thus applications based on this kind of data are being developed, such as job services connnecting employers and employees [1], event detection applications, creation of geo-localized galleries, etc. The possibilities of applications using geo-localised data from the media are limited only by our imagination.

2 Data : Flickr photos of Rennes

You recently responded to a public call for tenders by Rennes Métropole and won (congratulations !). In an effort to improve its public transportation and the lives of tourists visiting Rennes, Rennes Métropole is asking you to find the areas with a high density of tourists in a non-intrusive way and at small cost, by relying on photos posted on Flickr. This task is about discovering events in the broadest sense of the word : permanent events in a specific location, non-permanent events throughout the city, one-off events in a specific location, etc.

We imagine that there is an architecture to retrieve information from the Web (crawling, scraping, etc.) such as geo-localized photos. It is then necessary to automatically find points of interest, events, ... from a large collection of geo-localized photos. For example, 3000 photos taken around the Eiffel Tower correspond to a single point of interest. For this project, we assume that you have already made a collection of geo-localized media (you are very efficient !) using the Yahoo Flickr API [1]. You have more than 50 000 photos taken over the course of several years. Each photo is described as a tuple : $\langle id\_photo, id\_photographer, latitude, longitude, tags, title/description, dates\rangle$ Fig. 1 shows an extract of the dataset.

1. `https://www.flickr.com/services/api/`

Preview

Click column header to change column properties (* = name/type user settings)

Row ID	id_photogr...	lat	long	id_photo	tags	title	date_taken_time	date_t...	date_t...	date_t...	date_t...	date_t...	date_t...	date_t...
Row0_1	25782516@N05	48.11	-1.674	33,007,289,3...	rennes bretagne britany france...	Rennes la vilaine	2017-04-05T09:02:33	2017	2	4	5	4	95	14
Row0_2	25782516@N05	48.11	-1.686	33,007,288,9...	rennes bretagne britany france...	Les dames à q...	2017-04-05T09:02:32	2017	2	4	5	4	95	14
Row0_3	107075798@N...	48.115	-1.679	33,706,562,3...	instagramapp square squarefo...	Enfin... #Zelda...	2017-04-04T06:21:26	2017	2	4	4	3	94	14
Row0_4	25782516@N05	48.11	-1.686	33,823,411,0...	péniche boat constellationorion...	Péniches sous l...	2017-04-03T23:32:42	2017	2	4	3	2	93	14
Row0_5	7774173@N08	48.096	-1.74	33,667,020,0...	?	Etang d'Apigné...	2017-04-02T15:48:31	2017	2	4	2	1	92	13
Row0_6	13366873@N05	48.114	-1.677	33,438,108,9...	?	20170403_09...	2017-04-03T09:04:32	2017	2	4	3	2	93	14
Row0_7	13366873@N05	48.116	-1.678	33,823,239,4...	?	20170403_09...	2017-04-03T09:00:57	2017	2	4	3	2	93	14
Row0_8	13366873@N05	48.115	-1.679	33,823,234,0...	?	20170402_21...	2017-04-02T21:11:05	2017	2	4	2	1	92	13
Row0_9	69583894@N06	48.119	-1.603	33,774,247,1...	?	Appel à la pris...	2017-03-16T00:00:00	2017	1	3	16	5	75	11
Row0_10	27111862@N06	48.099	-1.669	32,952,063,6...	microscale effect standing man...	Microscale effe...	2015-11-08T11:59:01	2015	4	11	8	1	312	45
Row0_11	149180564@N...	48.11	-1.681	32,947,410,4...	?	IMG_0842	2017-04-02T04:00:37	2017	2	4	2	1	92	13
Row0_12	149180564@N...	48.114	-1.67	33,633,489,7...	parc thabor france rennes fran...	IMG_1985	2017-03-26T11:16:41	2017	1	3	26	1	85	12
Row0_13	149180564@N...	48.114	-1.67	32,976,923,9...	flower spring printemps pink r...	IMG_1986	2017-03-26T11:17:07	2017	1	3	26	1	85	12
Row0_14	149180564@N...	48.114	-1.67	33,405,278,4...	flower spring printemps pink r...	IMG_2001	2017-03-26T11:18:33	2017	1	3	26	1	85	12
Row0_15	149180564@N...	48.114	-1.67	33,405,265,5...	flower spring printemps pink r...	IMG_2002	2017-03-26T11:26:56	2017	1	3	26	1	85	12
Row0_16	149180564@N...	48.114	-1.67	33,789,862,2...	flower spring printemps pink r...	7CB60B22-D1...	2017-03-26T11:23:14	2017	1	3	26	1	85	12
Row0_17	149180564@N...	48.114	-1.67	32,976,868,6...	flower spring printemps pink r...	115F970C-8C...	2017-03-26T11:30:11	2017	1	3	26	1	85	12
Row0_18	149180564@N...	48.114	-1.67	32,976,858,0...	church eglise fleur pink flower ...	IMG_2015	2017-04-02T03:17:14	2017	2	4	2	1	92	13
Row0_19	21091233@N08	48.118	-1.644	33,620,405,7...	laurent duthion laurentduthion ...	Approximation...	2017-03-23T11:45:23	2017	1	3	23	5	82	12
Row0_20	107075798@N...	48.115	-1.679	32,961,507,9...	instagramapp square squarefo...	Trop bien le ca...	2017-04-01T08:03:51	2017	2	4	1	7	91	13
Row0_21	119588793@N...	48.113	-1.675	33,632,300,1...	?	Rennes	2017-03-31T13:37:42	2017	1	3	31	6	90	13
Row0_22	119588793@N...	48.114	-1.672	33,376,513,0...	?	Summit	2017-03-31T14:08:52	2017	1	3	31	6	90	13
Row0_23	119588793@N...	48.114	-1.67	33,604,246,7...	dog puppy chien park walk sit ...	Waiting	2017-03-31T13:47:12	2017	1	3	31	6	90	13
Row0_24	27111862@N06	48.103	-1.673	33,594,918,1...	rennes streets stairs escalier g...	Rennes streets...	2016-04-16T16:18:34	2016	2	4	16	7	107	15

FIGURE 1 – Raw sample of the dataset at your disposal.

3 Objective and methodology

Your mission is to automatically find points of interest in the city of Rennes, defined by the presence of a substantial amount of pictures. To do this, you will detail each step of the data mining process using the KNIME software. All useful files are available on Teams. The report of your work as well as the file generated by KNIME (extension `.knwf`, command `File > Export KNIME workflow...`) are to be sent by email (the deadline will be communicated to you).

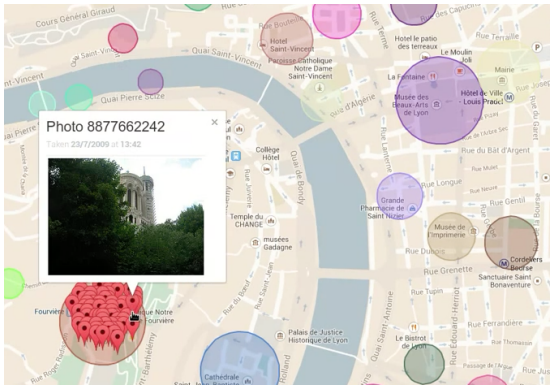

FIGURE 2 – Student project from INSA Lyon - 4IF

4 Data preparation and attribute selection

The data provided is raw, so it is important to "prepare" it before it can be used. The steps of this preparation are : understanding the data, cleaning, visualizing and editing statistics about the data. For example you must : check the consistency of the data w.r.t. your domain (such as dates and GPS positions) ; remove duplicates ; display the points on a map ; etc. In the KNIME tool, you can use, among others,

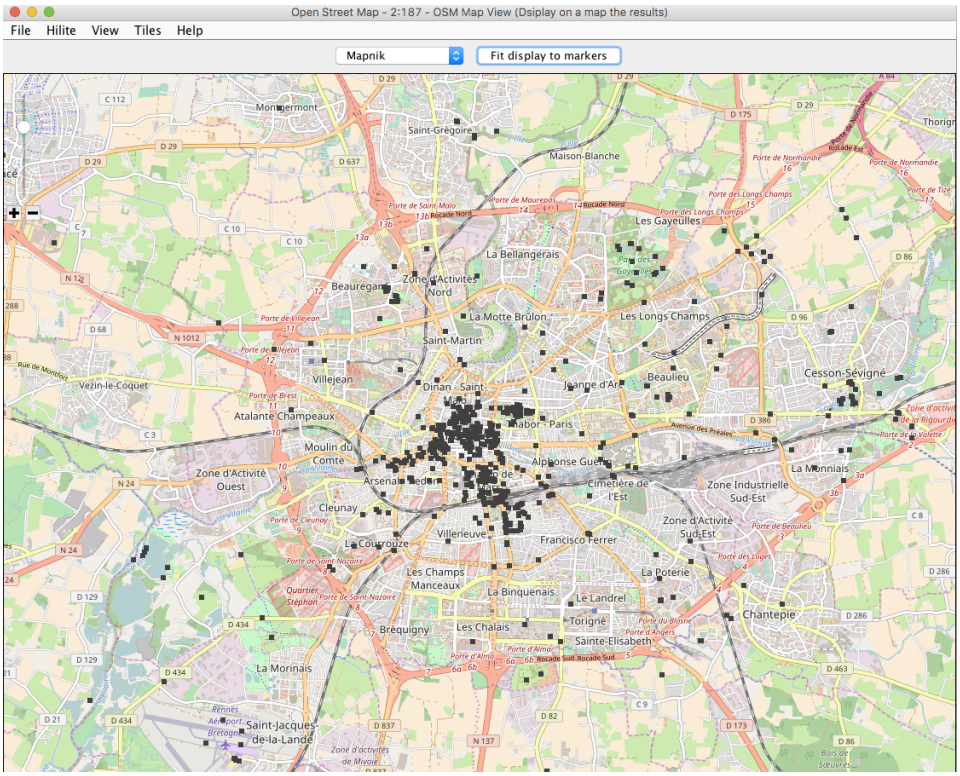

FIGURE 3 – Visualization of the cleaned data using the *OSM Map View* node.

the nodes : *File Reader, GroupBy, Row Filter, Geo-Coordinate Row Filter, Statistics, OSM Map View*. An example of data visualization using the *OSM Map View* node is shown in figure 3.

Note : if the *OSM Map View* node is not in the list of available nodes, you need to install the *KNIME Open Street Map Integration* using the `Help > Install New Software...` menu.

Another crucial step prior to data mining is the selection of attributes of interest for the current analysis. For this task, we recommend that you use the *Column Filter* node in KNIME to select the attributes relevant to the task at hand.

5 Discovering points of interest using clustering

The first task you have to perform is the discovery of points of interest. To do this, we suggest you use a clustering method. The work will be done in two steps :

1. Data mining using *clustering* : *k-means*
   We will use nodes *k-Means, Color Manager, OSM Map View*.
2. Evaluation, interpretation, visualization (on a map), discussion of results. How can your analysis help Rennes Métropole ? What knowledge does it provide to them ?

The last step is often neglected, but it is crucial. A data mining result is useless if it is not exploitable (*actionable*) : it must be useful, and the instructions for use must be given.

To go further : Depending on the time you have, you can try other clustering methods (e.g., *hierarchical clustering*) and compare these different methods. You will then need to use additional nodes in KNIME : *Hierarchical Clustering, Weka Cluster Assigner*.

6 Characterising points of interest using itemset mining

The previous step allowed us to extract candidate points of interest. However, a validation/understanding step is missing. For that, we will then try to describe the obtained clusters not in extension, but in intension. To do so, we propose to use a textual data processing pipeline to build a document/term matrix that can be made binary and then search for *frequent patterns* of terms for each cluster, or search for *discriminating patterns* if you want to go further.

For a start, we propose a simple text processing pipeline for the descriptions found in the tags and titles of the photos. The steps are the following : transform the tags into a "document" (which title and which author for the document ?) ; remove punctuation, remove words too short to be useful, remove numbers ; put everything in lower case ; remove "empty words" ; stemmatize (replace words by their root) ; create a "bag of words" representation for each document ; then a binary vector representation ; and apply a frequent patterns search algorithm. In KNIME you will need the following nodes : *Strings to Document, Punctuation Erasure, N Chars Filter, Number Filter, Case converter, Stop word Filter, Snowball Stemmer, Bag of Words Creator, Document Vector, Category to class, Create Bit Vector, Item Set Finder (Borgelt)*

Note : if the aforementioned nodes are not in the list of available nodes, you need to install the *KNIME Text Processing* and *KNIME Itemset Mining* extensions using the `Help > Install New Software...` menu.

To go further : Depending on how much time you have, you can use the KNIME tutorial(s) on text mining to improve your text processing pipeline.

7 Useful resources

— Retrieving data from the Web [2]
— Example of results on the dataset [4, 3]

Références

[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] Démo d'un excellent projet 4IF, INSA de Lyon. `https://www.youtube.com/watch?v=aM-zhxyVE54`.

[4] Démo d'un projet d'étudiants, UCBL, lyon. `http://liris.cnrs.fr/mehdi.kaytoue/sujets/ter-meanshift/demo1.html`.