Emotional rating of events based on tweets

Master 1 SIC Synthesis project

20/06/2018

Team Members:
Hervé-Madelein Attolou
Loïc Bachelot
David de Castilla

Reporter:
Philippe Gaussier

Technical tutors :
Pierre ANDRY
Dimitris KOTZINOS

Project Management Supervisor :
Tianxiao LIU

Summary

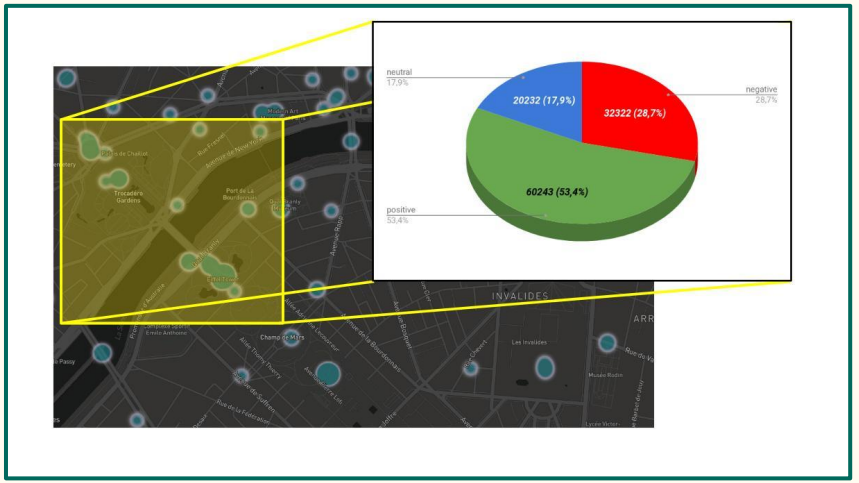
The Project

From a real-time stream of tweets, our project :

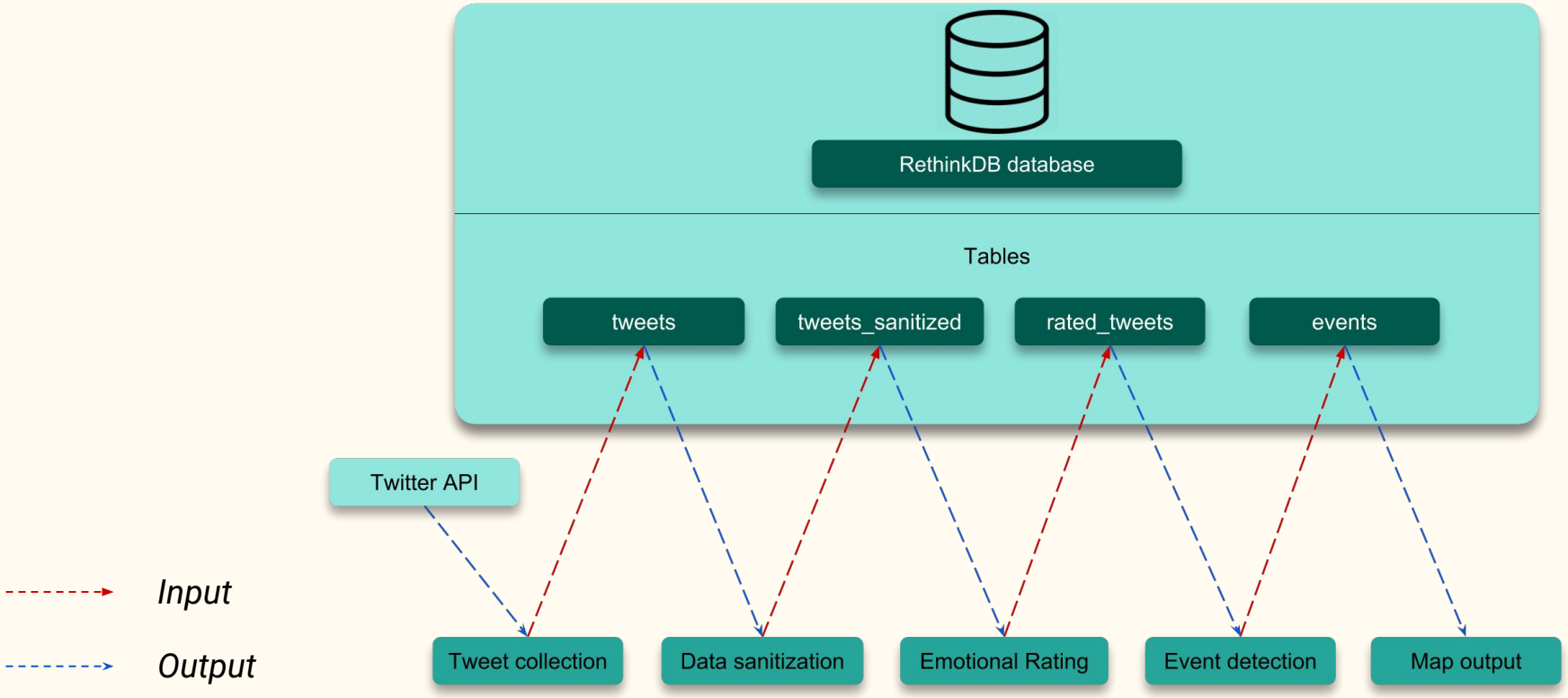
- Gives a rate to each tweet depending on its positivity or negativity
- Detects spatio-temporal events (examples: concerts, sport …)
- Gives a rate to each event depending on its positivity or negativity
- Displays the events on a map, each event will have a color associated to its grade

Use cases : Event analytics

- Online satisfaction service
- Prediction of the success of a venue



A modular conception



Work environment

Libraries and Platforms:

- RethinkDB : database management system
- Tweepy: a library to use the Twitter API
- Snowball (pystemmer) : Stemming library
- Scikit-learn : Machine learning library
- Tensorflow : Machine Learning library
- Mapbox : An open source mapping platform for custom designed maps

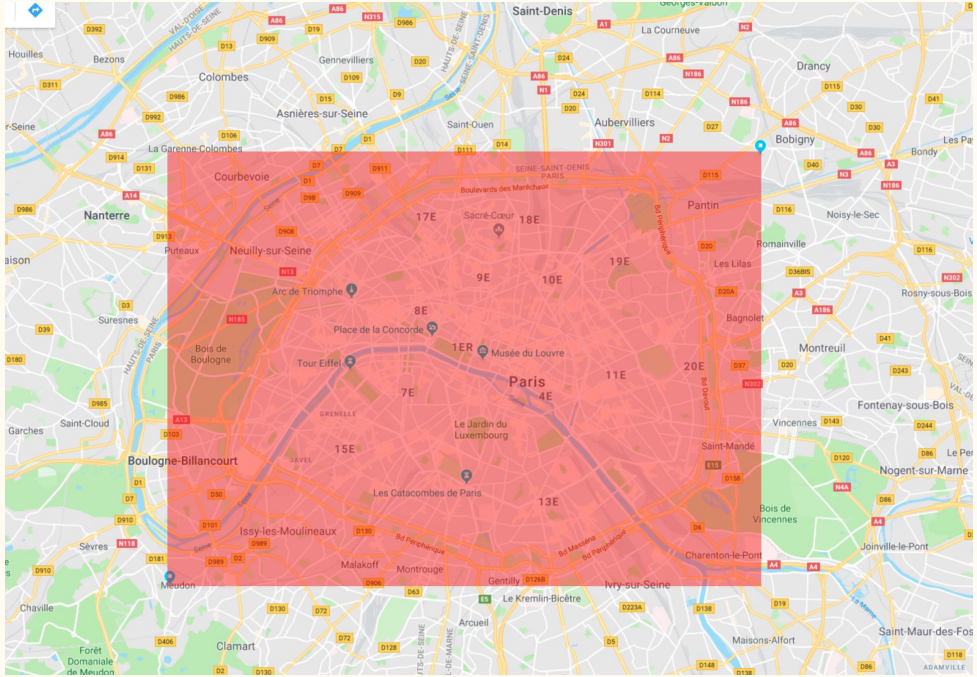
Virtual Private Server :

VPS Cloud 2 from OVH

- 2 vCore(s) 3,1 GHz
- RAM : 4 Go + SWAP 4 Go
- Data storage : 50 Go
- Operating System : Ubuntu 16.04.04 LTS

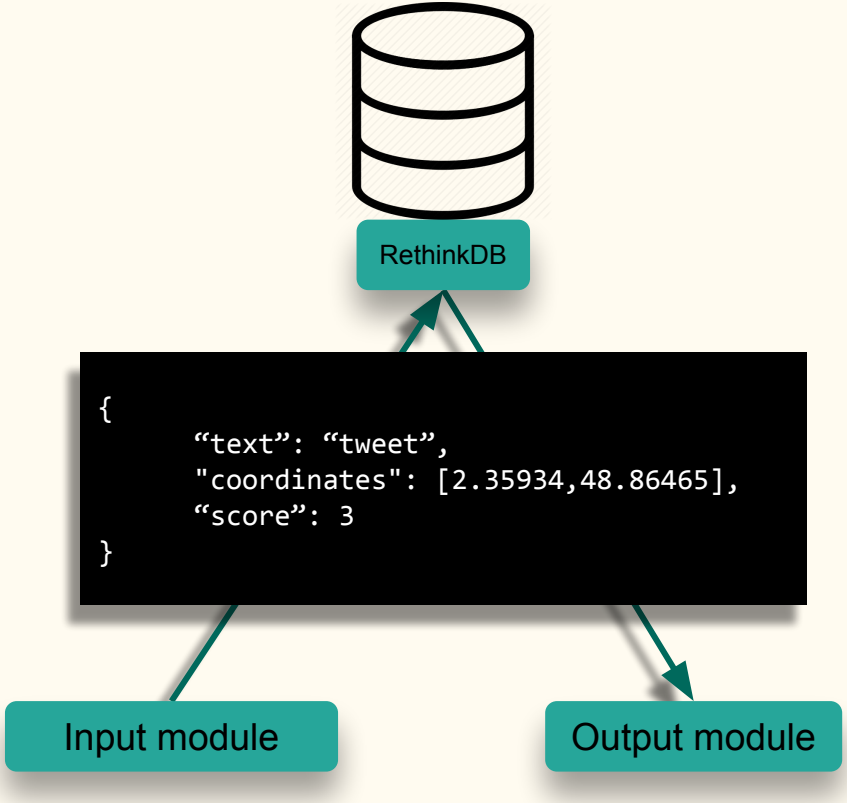
Tweet collection

- A python program that :
  - Gets a stream of tweets from the Twitter API
  - Applies a location filter based on coordinates (Paris Area)
- It uses :
  - Tweepy library to use the Twitter API in python
  - API credentials to connect to the Twitter API



Paris Area

Tweets Storage



- Real-time NoSQL Database RethinkDB :
  - SDK in various languages (JavaScript, Python, Ruby and Java)
  - JSON

We are keeping every tweets we are receiving from the Twitter API

Data sanitization

ME: Joe, about halfway through the speech, I'm gonna wish you a happy birth--
BIDEN: IT'S MY BIRTHDAY!
ME: Joe.

Happy birthday to @JoeBiden, my brother and the best vice president anybody could have. pic.twitter.com/sKbXjNiEj

ME: Joe, about halfway through the speech, I'm gonna wish you a happy birth--
BIDEN: IT'S MY BIRTHDAY!
ME: Joe.

*Happy birthday to **@JoeBiden**, my brother and the best vice president anybody could have. pic.twitter.com/sKbXjNiEj*

Joe, halfway speech wish happy birth--
BIDEN: BIRTHDAY!
Joe.

Happy birthday brother best vice president anybody.

- Remove URL
- Remove @usernames
- Remove common words
- (Semi) Stemming
  - Take off all the suffix of the word

Benefits : less computation for the same or even better results.

Emotional rating : dataset creation

Tweet

AFINN List

🔥 → 0.2
❤️ → 0.7
⚽ → 0.6

Average Score Output = 0.42

Distance Score Output = 0.7

Unicode Emoji 😀 →1 😐 →0 ☹️ →-1

- Use data from the tweet as unigrams (each word is an item that we analyse individually)
- Use the AFINN (emotional dictionary) list to get start point of our emotional dictionary (with emojis)
- Use an average (or distance) computation to get the emotional score of a tweet between 0 and 1

Emotional rating : Machine Learning

We are using three different datasets to train our machine learning algorithms:

- **Emoji Unicode**
- **Emotional Emoji (average)**
- **Emotional Emoji (distance)**

As **Emoji Unicode** is classifying tweet as positive (1), neutral (0) or negative (-1), we are using a classifier with this dataset.
As both **Emotional Emoji** datasets gives scores to tweets varying between 0 and 1, we are using a regressor with these datasets.

Dataset name	SVM Classifier	Deep neural network Regressor
Emoji Unicode	✔	
Emotional Emoji (average)		✔
Emotional Emoji (distance)		✔

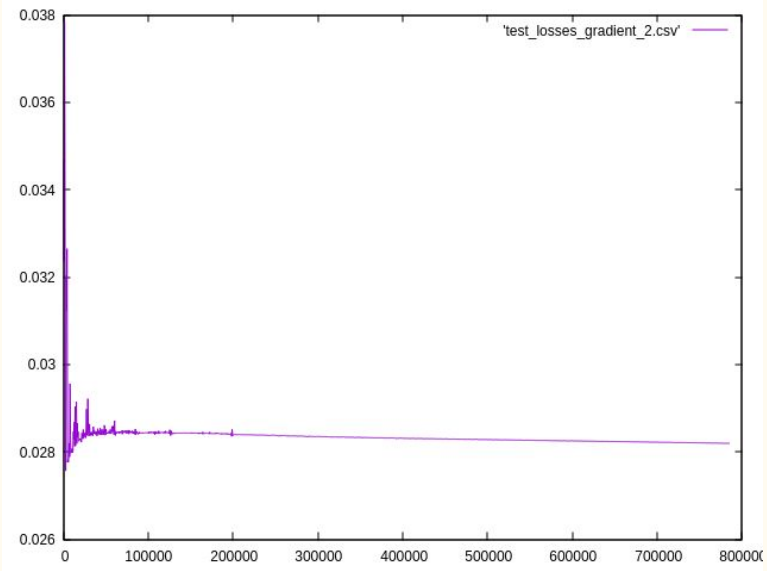
Emotional rating : Certification

SVM Classifier:

- Dataset : Unicode Emoji
- Mean accuracy on training dataset: 0.7527
- Mean accuracy on test dataset: 0.6112

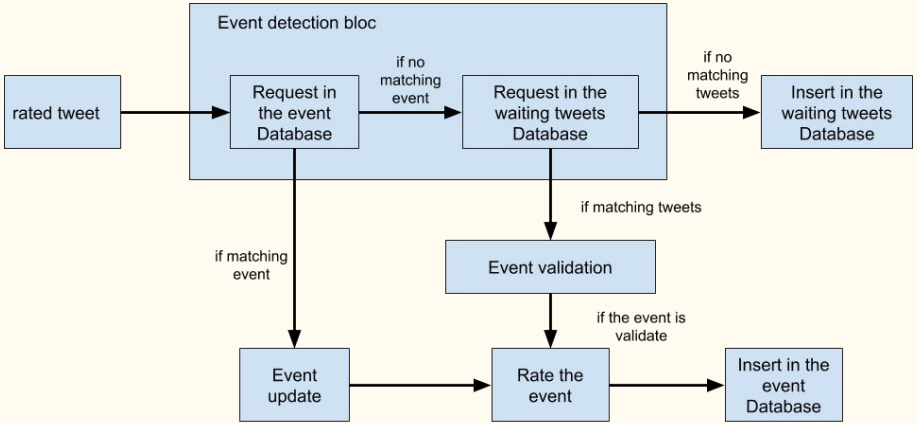
Deep Neural Network (DNN) Regressor:

- Dataset : Emotional Emoji (Average)
- Hidden Layers : 3
- Numbers of neurons : 1000, 100, 10
- Input layer size : 8417
- Learning rate : 0.1
- Optimizer : Gradient Descent
- Number of steps: 784500
- Final loss : 0.028195387



Loss of the DNN on the training dataset depending on the training step

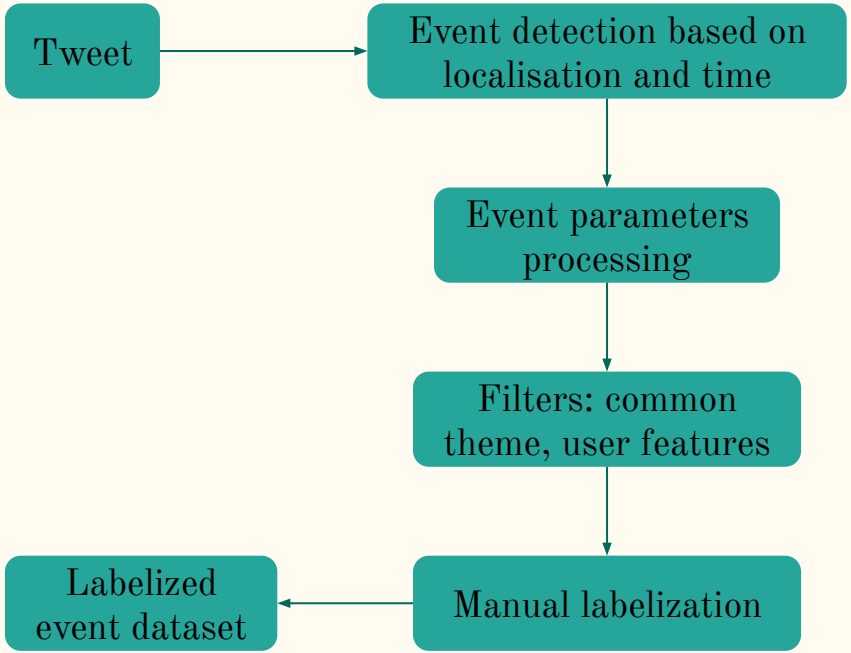
Events detection



Data from tweets we use to detect events:

- commontheme
- numberOfHashtag
- frequencyOnHashtags
- frequencyOnTweets (for the most used hashtag)
- numberOfMention
- frequencyOnMentions
- frequencyOnTweets (for the most used mention)
- numberOfTweets
- numberOfUser
- ratioMaxTweetPerUser

Dataset creation



Dataset statistics:

Dataset	Real events	False events	Total events
Dataset 1	22	179	201
Dataset 2	14	211	225
Dataset 3	36	390	426
Dataset 4	22	22	44

Events detection : Certifications

Results of the validation Multi-layer perceptrons:
Tests on dataset number 2

Hidden Layers	Training	Test
20,10	93,3%	93,3%
15,5	94%	93,3%
18	93,3%	93,3%
20,10,5	93,3%	93,3%

AI trained on dataset 3

Usability:
1 real event detected on 1 detected

Hidden Layers	Training	Test
20,10	64,6%	53,3%
15,5	38%	28%
18	31,3%	36%
20,10,5	42%	52%

AI trained on dataset 4

Usability:
12 real event detected on 97 detected

Improvements

Problem: Low number of tweets

Solution: Larger spatio-temporal zone

Problem: Tweets related to the event are sent before and after the event, without the right location.

Solution: Detection based on textual features like hashtags, mentions, and text.

DataBase statistics

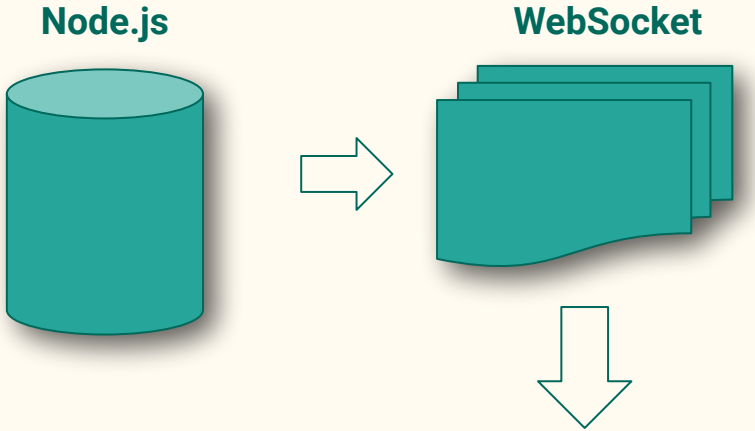
Overall statistics

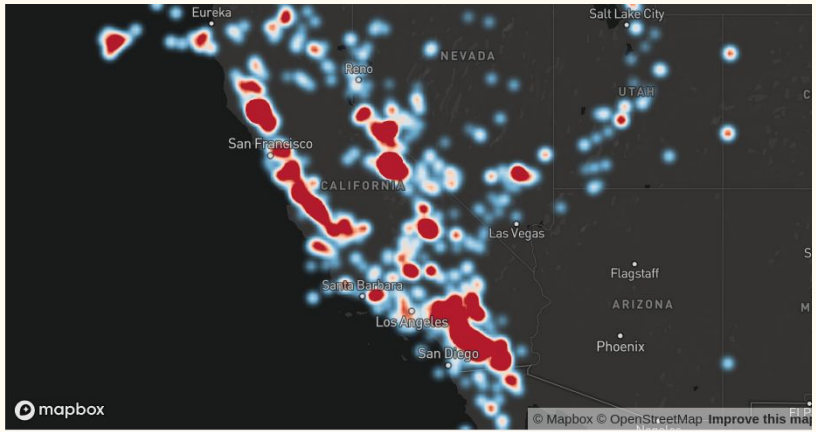
Number of tweets	325 621
Localisation area	Paris
Tweets in French	208 788 / 64%
Localisation and in French	5811 / 1.78%
Included in events	~500 / 0.14%

Statistics by day:

Number of tweets/day	~11000
Number of French tweets/day	~7040
Number of usable tweets a day for events	~200

Map output





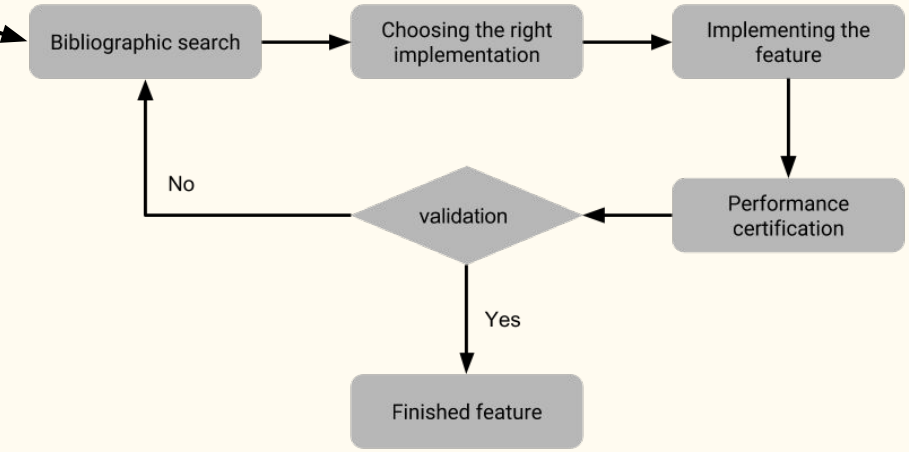
Web Page

- Use a node.js listener to get real time information from the database
- Use a websocket
- Update a map (MapBox) to display a heat map on a webpage

Benefits : constant information is displayed no need to update the page to get new information from the database

Project Management

Life cycle and task assignment



This diagram represents the process of developing a feature in our project

Task assignment

Task Completed	Hervé-M	Loïc	David
Tweet collection Script API Twitter Statistiques		✓	✓ ✓
Data sanitization Stemming Remove entities	✓ ✓		✓
Emotional rating Dataset creation: - Emoji Unicode - Emoji List average - Emoji List distance Machine learning	✓ ✓		✓ ✓
Event detection		✓	
Map output	✓		✓

Project Management

Parallel working

The modules *Emotional rating* and *Event detection* are fully independent

Release Planning

Task	Planned date	Completion date
Tweet collection	31/01/2018	31/01/2018
Data sanitization	15/03/2018	30/03/2018
Emotional rating	31/05/2018	14/06/2018
Event detection	31/05/2018	14/06/2018
Map output	10/06/2018	13/06/2018

Conclusion

Twitter API	Removing entities Stemming words	Dataset Creation Machine learning algorithms	Dataset creation Machine learning algorithms	NodeJS Mapbox

Conclusion

Unsolved Issues:

- The quality of our datasets
- Our machine learning algorithms need more experimentation

Possible solutions:

- Find an already big dataset from Internet
- Use different Machine Learning algorithm to provide the event detection and the emotional rating

Possible extensions:

- Improve the map by adding a time slider to navigate in the dataset
- Use a machine learning to predict the events and their emotional rate
- Use a bigger analysis area
- Use more social networks as data sources

Questions ?

Annexe

AFINN

AFINN is a list of English words rated for valence with an integer between minus five (negative) and plus five (positive). The words have been manually labeled by Finn Årup Nielsen in 2009-2011.

We use that list with French translation and some useful additions as emoticons for example


```
{
  "verdict": "POSITIVE",
  "score": 3,
  "comparative": 0.3,
  "positive": [
    "best"
  ],
  "negative": []
}
```

:)	2
:(	-2
:|	-1
:]	2
:[	-2
:}	2
:{	-2
:/	-2
:\	-2
:*	2
:-)	2
:-(	-2
:-|	-1
:-]	2
:-[	-2
:-}	2
:-?	-1
:->	2

beloved	3
benefactor	2
benefit	2
benefitting	2
benevolent	3
bereave	-2
bereaving	-2
best	**3**
best damn	4
betray	-3
betrayal	-3
betrays	-3
better	2
bias	-1
biased	-2
big	1
bitch	-5
bitches	-5
bitter	-2

Twitter API characteristics

Stream

- Users randomly choose
- Stream $<$ 1% total twitter stream

Filters

- 5000 follow users
- 25 0.1-360 degree location boxes filter
- 400 keywords filter

SVN and other uses of ML

In this case, tweets are simply classified as positive or negative. The main goal, would be to check the feedback of a product by a query on some concerning tweets. The marketers could also use the tool to search public opinion and analyse customer satisfaction.
This project could replace all the star-rated reviews for some shopping websites (Amazon, BestBuy, Boulanger, Fnac, ...) or even for some flat renting platforms (AirBnB, Booking, …)

We can see that the presented results are mostly over 80% which seems like precise for our usage.

Table 1: Example Tweets

Sentiment	Query	Tweet
Positive	jquery	dcostalis: Jquery is my new best friend.
Neutral	San Francisco	schuyler: just landed at San Francisco
Negative	exam	jvici0us: History exam studying ugh.

Table 6: Classifier Accuracy

Features	Keyword	Naive Bayes	MaxEnt	SVM
Unigram	65.2	81.3	80.5	82.2
Bigram	N/A	81.6	79.1	78.8
Unigram + Bigram	N/A	82.7	83.0	81.6
Unigram + POS	N/A	79.9	79.9	81.9

POMS-ex, a psychometric instrument

In this case, the goal is to analyse link tweets emotion and big social events in the world.
The emotion classification is made using a psychometric instrument POMS (the Profile of Mood States). This tool measured six individual dimensions of the mood including :

- Tension
- Depression
- Anger
- Vigour
- Fatigue
- Confusion

November 4 On U.S. election day, Tension skyrockets to over +2 standard deviations. The day after Vigour jumps from baseline levels to +3 standard deviations, while fatigue steadily drops to -2 standard deviations.

November 27 On U.S. Thanksgiving day, Vigour notably records a sharp increase from baseline levels to +4 standard deviations.

$$\mathcal{P}(t) \rightarrow m \in \mathbb{R}^6 = [||w \cap p_1||, ||w \cap p_2||, \cdots, w \cap p_6||]$$

This version of the instrument is not intended to be administered as a questionnaire to human subjects, but rather to be applicable to large textual corpora. POMS-ex extends the original set of 65 POMS mood adjectives to 793 terms, including synonyms and related word constructs, thus augmenting the possibility of matching terms in large data, such as online textual corpora.

RETHINKDB

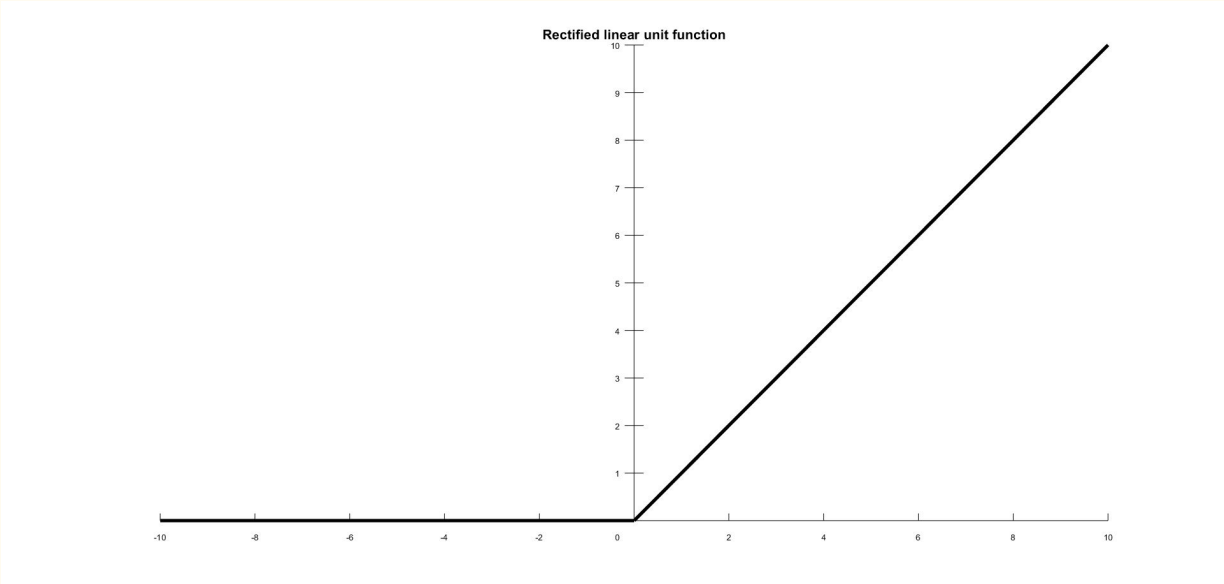
- Open-source JSON database
- Made for real-time web

Perfect to handle the tweet stream:

- Notification on any change on the database

Events MLP

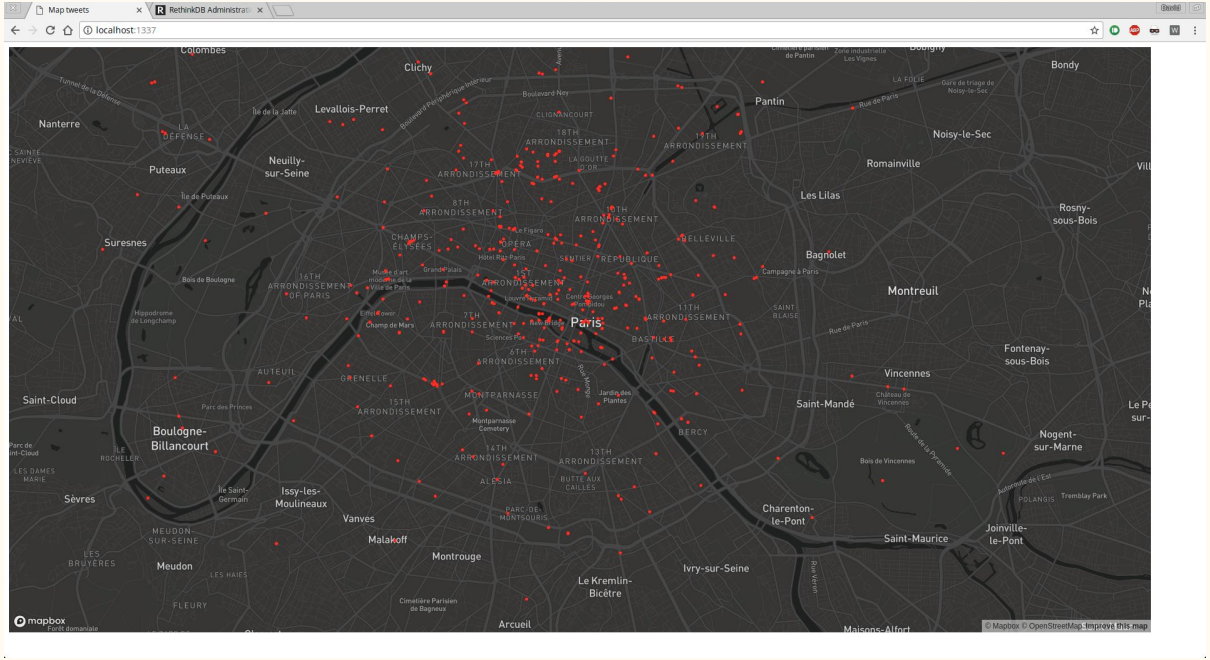
- Activation function: Rectified Linear units
- Batch size: Minimum between 200 and number of samples in the dataset
- Learning rate: 0.001



MAPBOX

Mapbox is a location data platform for mobile and web applications. It provide us building blocks to add location features like a map.

Here is a screenshot of our app with the map and the tweet displayed on it.



Bibliography

- Walther, M., & Kaisser, M. (2013, March). Geo-spatial event detection in the twitter stream. In European conference on information retrieval (pp. 356-367). Springer, Berlin, Heidelberg.
- Bertrand, K. Z., Bialik, M., Virdee, K., Gros, A., & Bar-Yam, Y. (2013). Sentiment in new york city: A high resolution spatial and temporal view. arXiv preprint arXiv:1308.5010.
- Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(12).