UNIVERSITÄT PASSAU

Semantic enrichment and data filtering in social networks for subject centered collection.

Student : Anthony FARAUT
Supervisor 1 : Prof. Dr. Michael GRANITZER (Passau)
Supervisor 2 : Dr. Habil. Elöd EGYED-ZSIGMOND (Lyon)
Chair : Prof. Dr. Harald KOSCH

http://www.hustisford.lib.wi.us/wp-content/uploads/2014/05/Robotreading.jpg

_ Motivations

- Social networks have become an important source of information, connecting people all around the world in almost real-time

- Demands for extracting meaningful and interesting information from them have dramatically increased

- Social networks can be queried through their API (Application programming interface)

_ Research questions

- How to deal with heterogeneity of the data ?

  -> Textual data cleaning

- How to deal with the short context of (hollow) social network posts ?

  -> Textual data enrichment

- What is the best numerical representation of the textual data ?

  -> Word2vec, Doc2vec, TF-IDF ?

- What is the best way to group tweets together ?

  -> Classification (SVM), Clustering ?

- How to keep a bag of relevant words over the time ?

_ Problem statement

- The main goal of this master thesis was to :

"*Collect the most information on an event described beforehand as a set of words while being robust (i.e. eliminating noise) in real time.*"

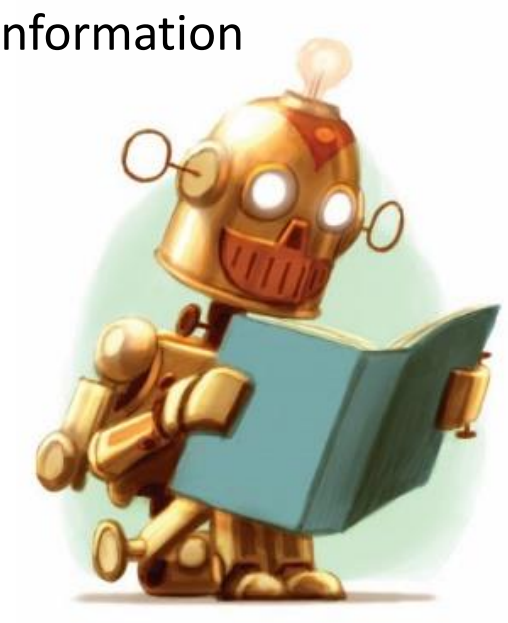
_ Shall I continue the presentation?

- Social networks have become an important source of information

However,

  - Heterogeneity of the data ?
  - Numerical representation of the textual data ?
  - Short context of social network posts ?

"*Collect the most information on an event described beforehand as a set of words while being robust (i.e. eliminating noise) in real time.*"

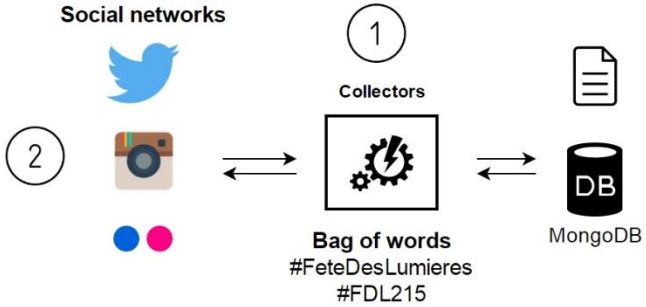


_ Agenda

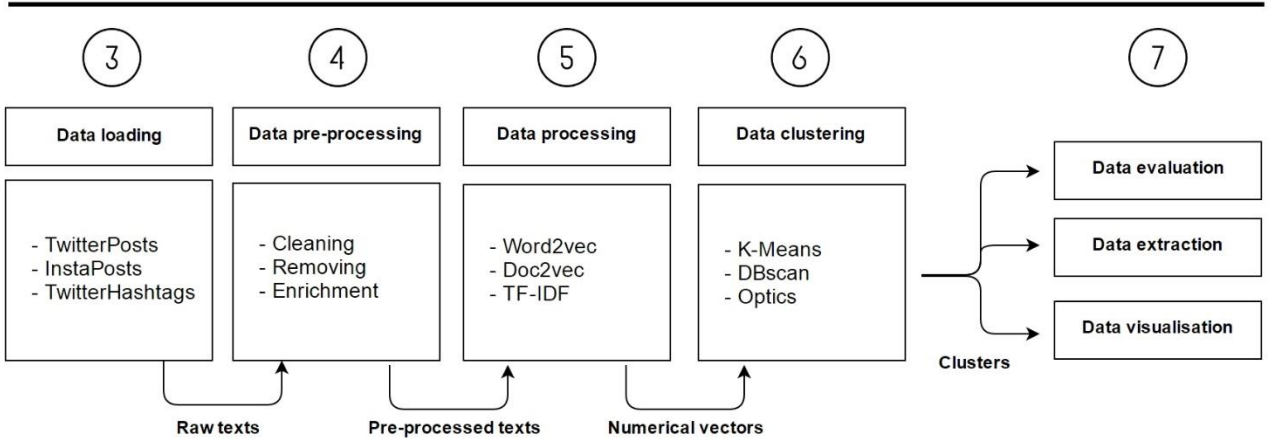
- Overall overview
- Understanding the data
- Approach
- Experimentation
- Evaluation
- Results
- Perspectives
- Conclusion

_ Overall overview

A



B



- Corpus
- Sample examples
- Facts about the data
- Handmade clustering

UNDERSTANDING THE DATA

_ Understanding the data – Corpus

- Focus on the “Fête des lumières 2015”
- Initial request’s inputs:

#Lyon #Candle
#FDL2015
#FeteDesLumieres2015

Geographical coordinates

_ Understanding the data – Sample of tweets

- #Lyon #8decembre #hommageauxvictimes https://t.co/eWFVmChqU8

- Fêtes des Lumières #8decembre #lyon #parisattacks #werenotafraid #forabetterworld #PrayForParis. . . https://t.co/t0tLug7XhM

- #FF @berniezinck for a good music

- Sie sind #endlich wieder da ! ???? @phillaude @derTC @oguz @Y_Titty @PatrickBuenning

_ Understanding the data – Facts about the data

- ~ 31 000 tweets;
- 5% of tweets with a specific geolocation;
- 12% of tweets with at least one media (photo/video);
- 13% of tweets with at least one link;
- 21% of tweets with at least one #hashtag;
- 51% of tweets with at least one user mention;
- 38 languages are represented.

_ Understanding the data – Handmade clustering

- A handmade clustering have been made by Mrs. Oriane PIQUER-LOUIS (PhD student working on the IDENUM project)

- (3%) - 1048 tweets talking about the "Fête des lumières" (97%) - 29958 noise tweets

- The tweets were labeled as related to "Fête des lumières"

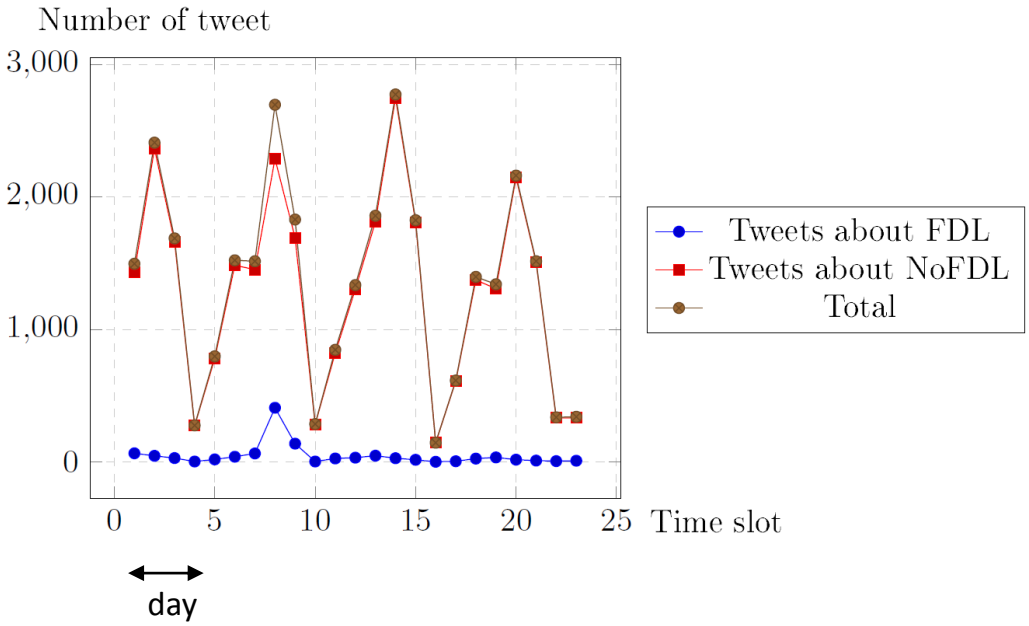
_ Understanding the data – Handmade clustering

	Language	Number of tweets in this language
Language 1	French	771
Language 2	English	122
Language 3	Undefined	118
Language 4	Spanish	13
Language 5	Japanese	5
Language 6	Norwegian	2
Language 7	Russian	2

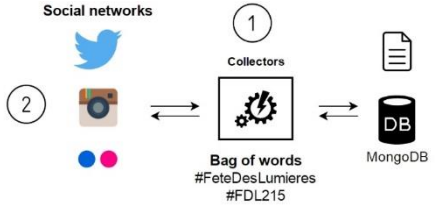
Seems to have a correlation between the language used and the event

Most of the French population does not speak English ☺

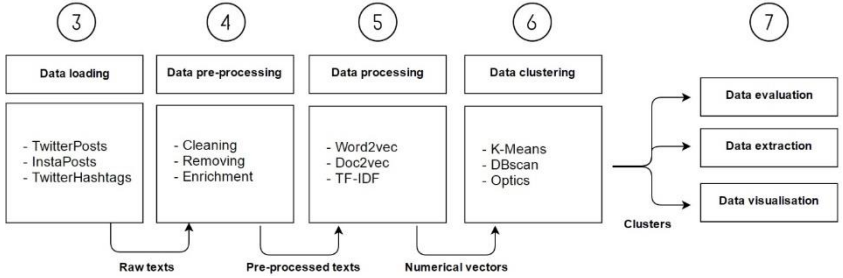
Step	Dates	FDL	NoFDL	Percent	Total
1	07 dec 13:00 - 07 dec 17:00	64	1,432	4.5%	1,496
2	07 dec 17:00 - 07 dec 21:00	45	2,366	1.9%	2,411
3	07 dec 21:00 - 08 dec 01:00	28	1,660	1.7%	1,688
4	08 dec 01:00 - 08 dec 05:00	2	273	0.7%	275
5	08 dec 05:00 - 08 dec 09:00	18	779	2.3%	797
6	08 dec 09:00 - 08 dec 13:00	38	1,484	2.6%	1,522
7	08 dec 13:00 - 08 dec 17:00	63	1,452	4.3%	1,515
8	08 dec 17:00 - 08 dec 21:00	408	2,289	17.8%	2,697
9	08 dec 21:00 - 09 dec 01:00	137	1,693	8.1%	1,830
10	09 dec 01:00 - 09 dec 05:00	2	283	0.7%	285
11	09 dec 05:00 - 09 dec 09:00	25	821	3.0%	846
12	09 dec 09:00 - 09 dec 13:00	31	1,304	2.4%	1,335
13	09 dec 13:00 - 09 dec 17:00	46	1,813	2.5%	1,859
14	09 dec 17:00 - 09 dec 21:00	27	2,748	1.0%	2,775
15	09 dec 21:00 - 10 dec 01:00	15	1,810	0.8%	1,825
16	10 dec 01:00 - 10 dec 05:00	0	144	0%	144
17	10 dec 05:00 - 10 dec 09:00	4	610	0.7%	614
18	10 dec 09:00 - 10 dec 13:00	24	1,373	1.7%	1,397
19	10 dec 13:00 - 10 dec 17:00	33	1,308	2.5%	1,341
20	10 dec 17:00 - 10 dec 21:00	16	2,146	0.7%	2,162
21	10 dec 21:00 - 11 dec 01:00	9	1,506	0.6%	1,515
22	11 dec 01:00 - 11 dec 05:00	5	331	1.5%	336
23	11 dec 05:00 - 11 dec 09:00	8	332	2.4%	340



A



B



APPROACH

- Data collection & storage

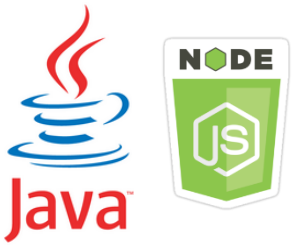
- Data loading
- Data pre-processing
- Data processing
- Data clustering
- Data extraction
- Data visualization

_ Data collection – Tools developed (Collectors)

Part A

REST API,
Streaming API

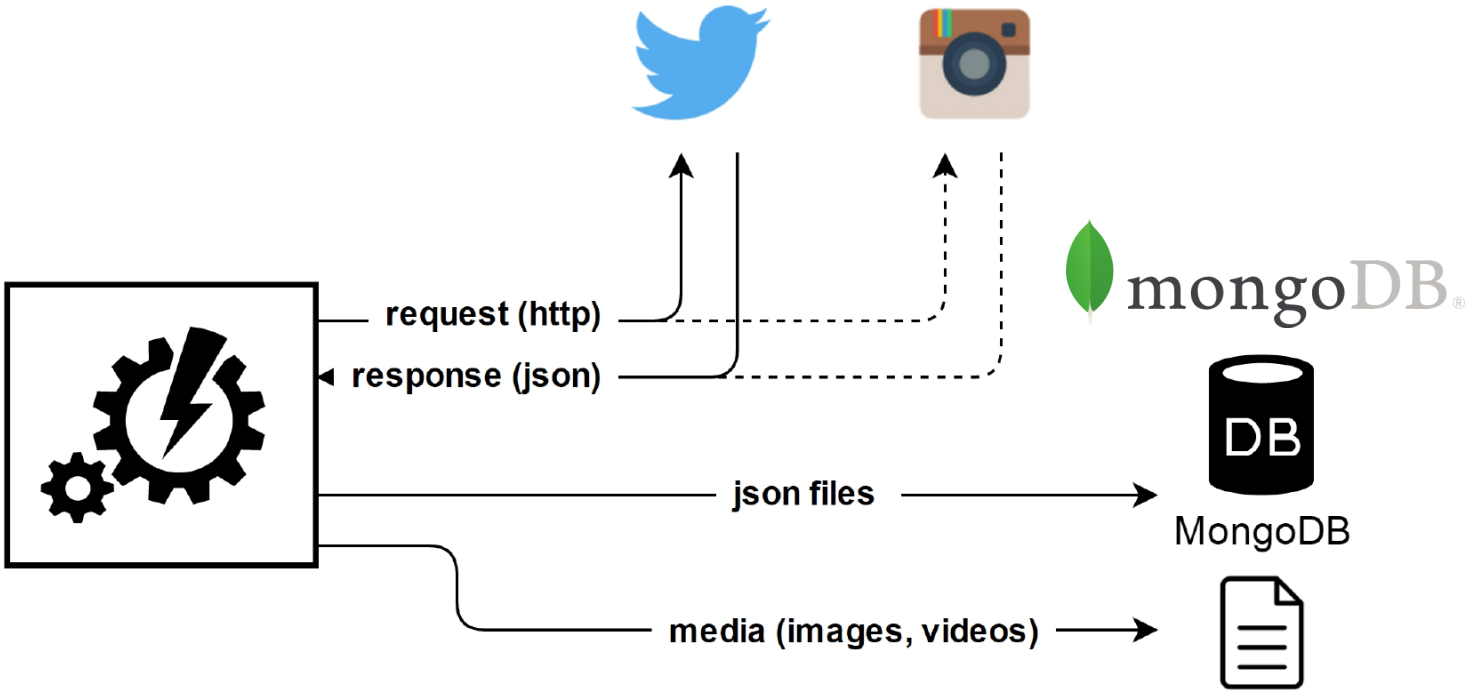
REST API,
Streaming API



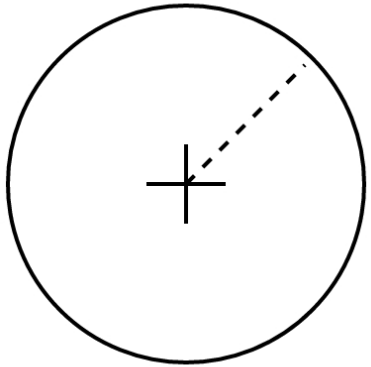
REST API

https://github.com/afaraut

_ Data collection – Tools developed (Collectors)



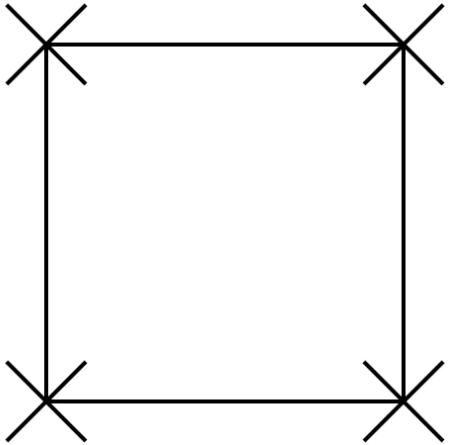
_ Data collection – Querying tools



Point (x,y) + radius

#FDL2015
#Lyon #Lights
#Lumignons
#lumieres #8decembre

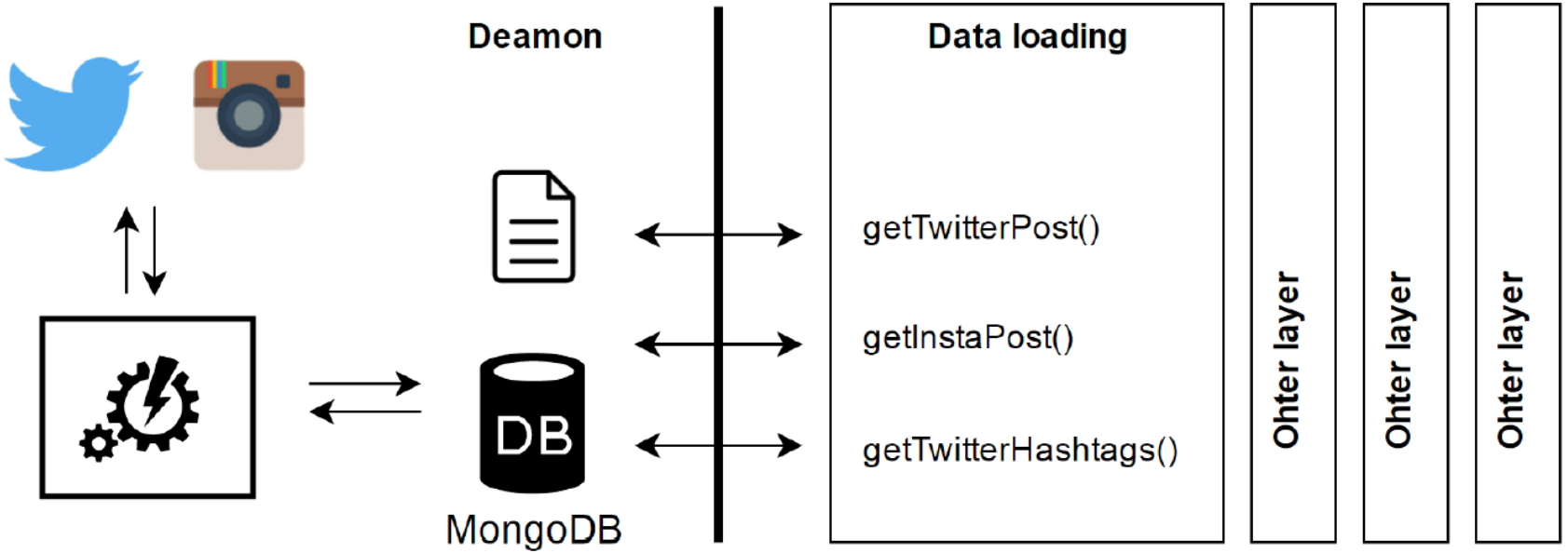
Keywords



Zones (x,y) x4

_ Data loading

Part B



abstraction

_ Data pre-processing

The data is heterogeneous

- Removing stage

  Will lose information (that is not very useful for the project)

- Cleansing stage

  Will clean the tokens in order to improve the further token connections

- Enrichment stage

  Will enrich the data in order to improve the relevance of the entire corpus

_ Data pre-processing – Removing stage

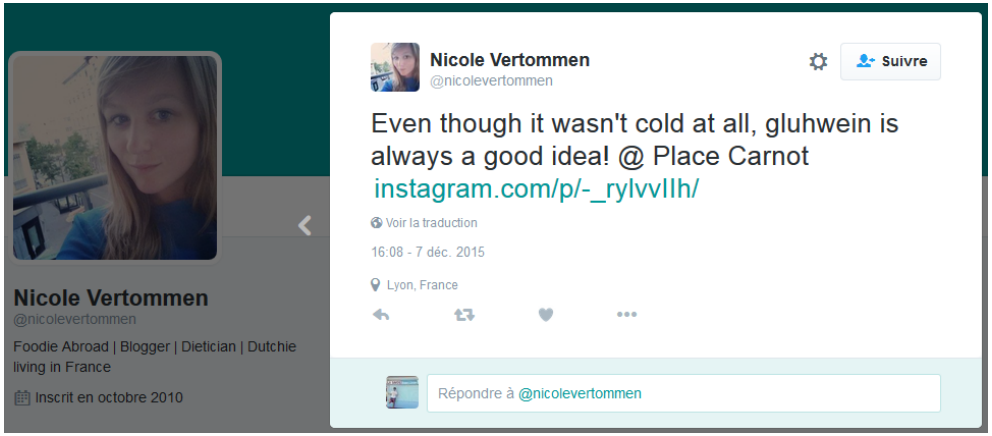
- Removing the line breaks
- Removing the usernames (user-mentions)
- Removing the links
- Removing accents

_ Data pre-processing – Cleansing stage

- Clean the following points “?????”-> “?”
- Clean space between punctuations “hello,” -> “hello ,”
- Lowercase

_ Data pre-processing – Enrichment stage

- Enrich raw post with hashtag from at least 2 users



- Even though it wasn't cold at all, gluhwein is always a good idea! @Place Carnot https://t.co/MFIpjfphA0
- even though #it wasn't cold at a l , gluhwein is always a good idea ! @place carnot

- Mal gut, dass es draufsteht... @ Confluence https://t.co/qMmHetjiOj
- mal gut , dass es draufsteht . @ #confluence

_ Data Processing – Word2Vec vectors

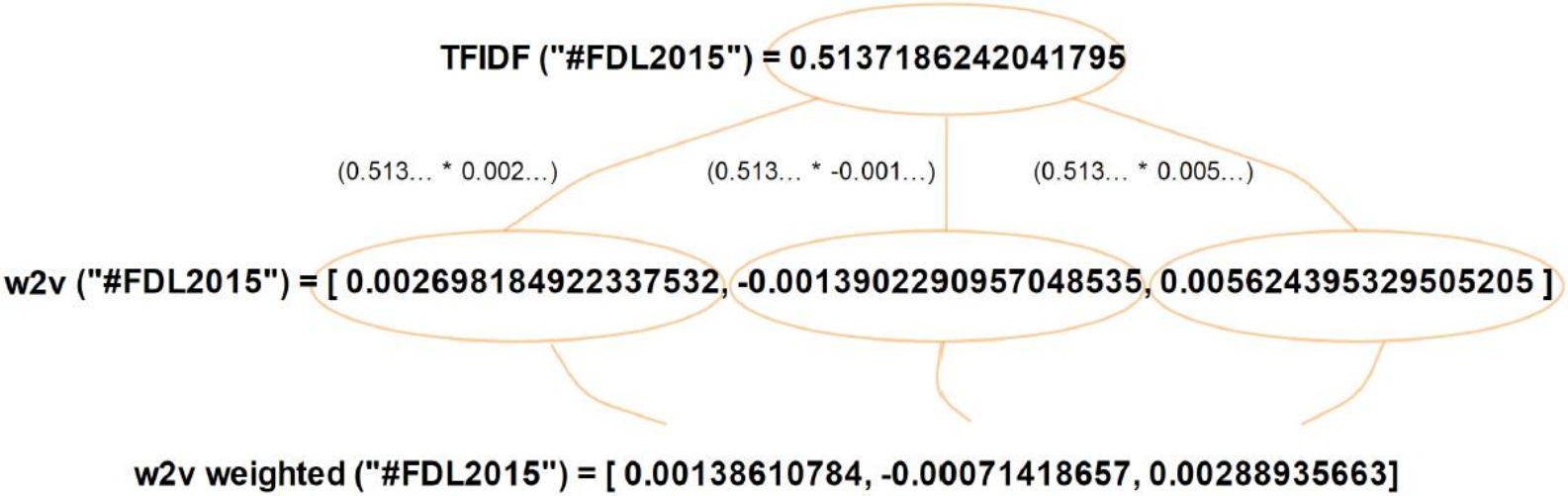
- Vector representations of words;
- Groups vectors of similar words together in vector space;
- Allows to detect similarities mathematically.

_ Data Processing – TF-IDF

- TF (term frequency): The number of times that a term T occurs in document D;
- DF (Document frequency): The number of times a term T occurs in all the entire corpus;

  (IDF means : corpus size / df)
- Weighting words from Word2Vec thanks to TFIDF formula.

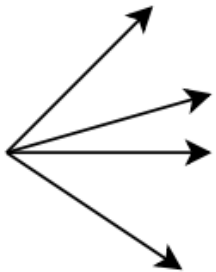
_ Data Processing – Word2Vec + TF-IDF



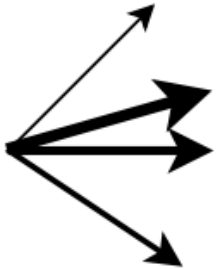
_ Data Processing – Word2Vec + TF-IDF

Need vectors corresponding to tweets -> combination of the word vectors.

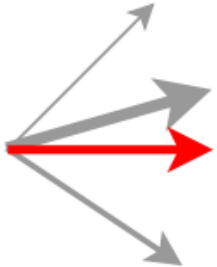
Tweet : "I have to go"



4 words = 4 vectors
Word2Vec



Weighted vectors
with TFIDF



The vector corresponding
to the tweet

_ Data Processing – Doc2vec vectors

- Vector representations of documents
- An extension of word2vec that learns to correlate documents with other documents, rather than words with other words
- Here, a document is a tweet

_ Data Processing – TF-IDF vectors

- Value close to 0 -> common to the overall corpus (stop word, or a very used word).
- Value close to 1 -> means that the word is specific to a given document

The length of the vectors is the number of unique words in the entire corpus (hollow vectors, considerable problem in practice)

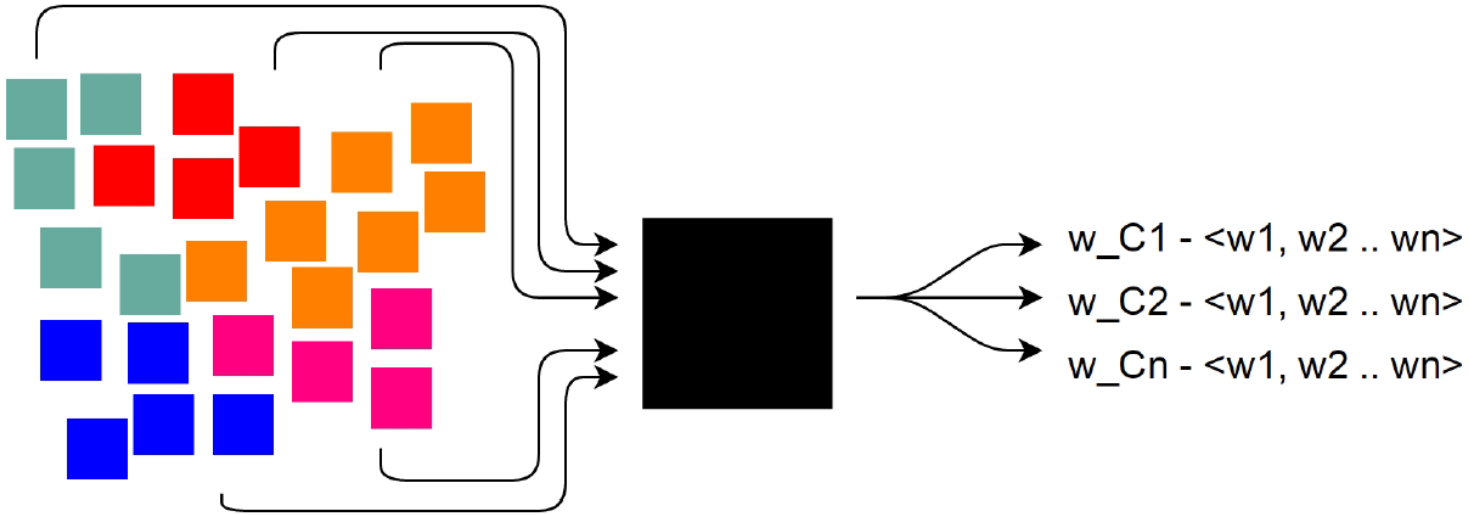
_ Data clustering

- For the process, the Kmeans algorithm were tested in order to get exactly the number of cluster wanted

  (2) FDL – Not FDL

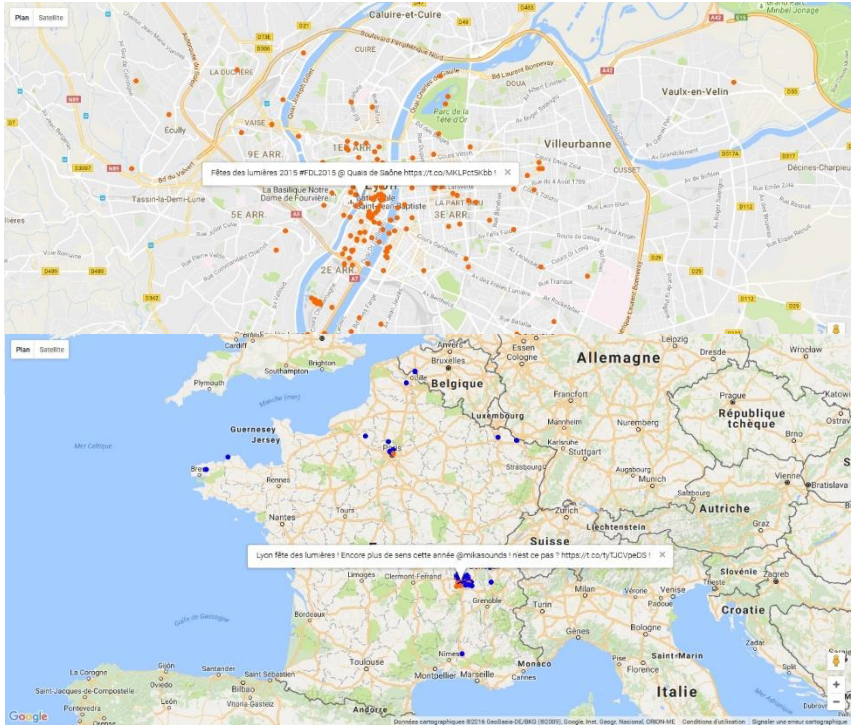
- DBScan algorithm seems to be a better algorithm in order to evolve over the time

_ Data extraction

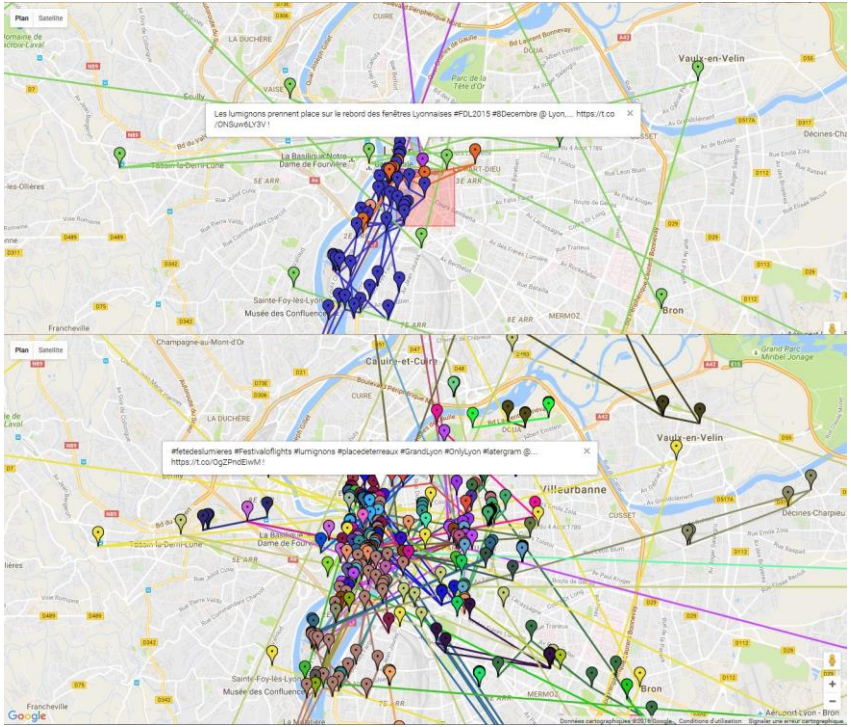


_ Data visualization

Points of interest



Movements of the users



- SVM
- Backtracking

EXPERIMENTATION

_ Experimentation – SVM

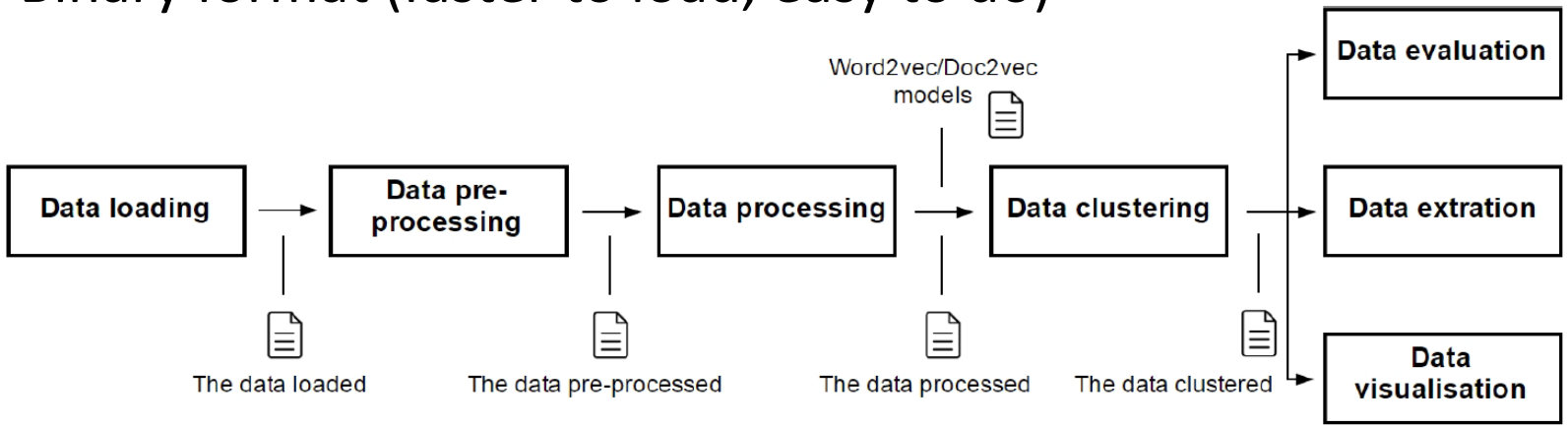
Know whether the clustering stage can have good results or not

- Linear kernel works well then K-means might work well too
- RBF kernel works well then density based clustering might work well too (DBScan)

_ Experimentation – Backtracking

Time consuming …

- Store for each step the current result in a serialization format
- Binary format (faster to load, easy to do)



EVALUATION

- Models generation
- Measures

_ Evaluation – Models generation

- ~700 Doc2vec and ~700 Word2vec models were generated
- On the entire corpus in order to improve the precision

-> To find the best representation of a tweet.

_ Evaluation – Measures

- Precision, Recall, F1

		Classified as: Positive	Classified as: Negative
Reality (Ground truth)	Positive	**True positive**	**False negative**
	Negative	**False positive**	**True negative**

Precision: How many selected items are relevant?

Recall: How many relevant items are selected?

F1: A measure that combines precision and recall

_ Evaluation – Measures

- Rand index

Look if the clustering is good without worrying about labels

- Normalized Mutual Information (NMI)

Measure the mutual dependence between two random variables

RESULTS

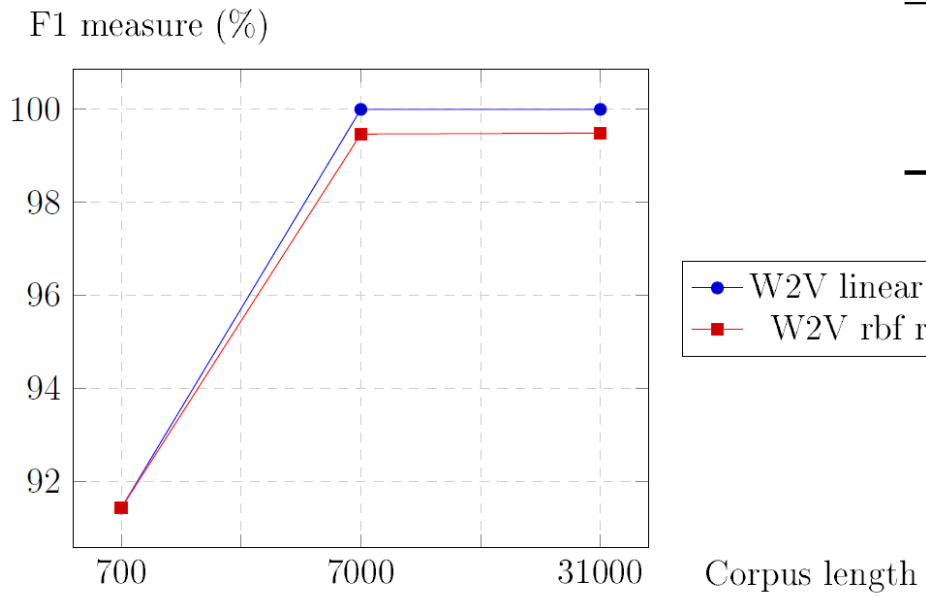
- Classification
- Clustering

_ Results – Classification

Table 7.8: Confusion matrix on 31000 tweets.

D2Vs + TF lin		False (GT)	True (GT)	
	False (predicted)	6003 (TN)	0 (FN)	6003
	True (predicted)	1 (FP)	198 (TP)	199
		6004	198	6202
W2V linear + TF rbf		False (GT)	True (GT)	
	False (predicted)	6004 (TN)	0 (FN)	6004
	True (predicted)	0 (FP)	198 (TP)	198
		6004	198	6202
W2V rbf		False (GT)	True (GT)	
	False (predicted)	6004 (TN)	2 (FN)	6006
	True (predicted)	0 (FP)	196 (TP)	196
		6004	198	6202

_ Results – Classification



Corpus	W2V linear	W2V rbf
700	91.43	91.43
7,000	100	99.47
31,000	100	99.49

_ Results – Clustering

700 tweets

Model	Precision	Recall	F1	Rand index	nmi
D2V worst one	20,36%	64,75%	30,98%	49,93%	0,66%
D2V best one	16,36%	86,07%	27,49%	66,94%	1,1%
W2V worst one	19,46%	95,08%	32,31%	57,49%	1,57%
W2V best one	88,24%	61,48%	72,46%	85,02%	30,19%
TF-IDF	17,45%	100%	29,72%	70,99%	0,12%

7000 tweets

Model	Precision	Recall	F1	Rand index	nmi
D2V worst one	17,13%	55,43%	26,17%	50,19%	0,31%
D2V best one	15,27%	96,76%	26,37%	69,07%	0,23%
W2V worst one	14,98%	82,82%	25,37%	60,55%	0%
W2V best one	15%	99,71%	26,08%	73,96%	0,04%
TF-IDF	14,97%	100%	26,05%	74,52%	0,01%

31000 tweets

Model	Precision	Recall	F1	Rand index	nmi
D2V worst one	3,81%	56,20%	7,13%	50%	0,04%
D2V best one	3,47%	94,94%	6,69%	81,15%	0,05%
W2V worst one	3,56%	91,79%	6,85%	73,72%	0,07%
W2V best one	3,39%	100%	6,55%	93,03%	0,08%
TF-IDF	3,38%	100%	6,55%	93,23%	0,05%

UNIVERSITÄT PASSAU INSA

_ Results – Clustering

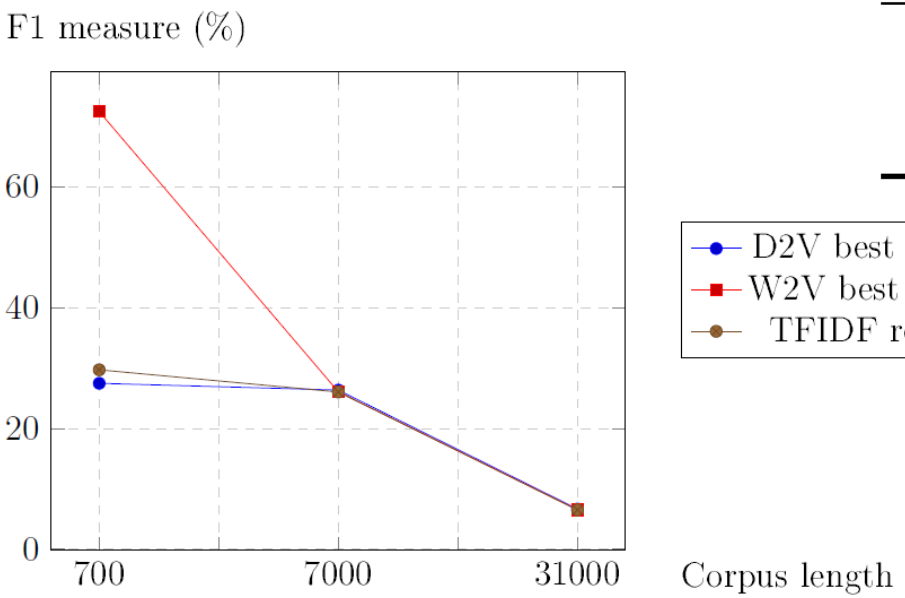
Table 7.12: Confusion matrix on 700 tweets.

D2V worst one		False (GT)	True (GT)	
	False (predicted)	269 (TN)	43 (FN)	312
	True (predicted)	309 (FP)	79 (TP)	388
		578	122	700
D2V best one		False (GT)	True (GT)	
	False (predicted)	41 (TN)	17 (FN)	58
	True (predicted)	537 (FP)	105 (TP)	642
		578	122	700
W2V worst one		False (GT)	True (GT)	
	False (predicted)	98 (TN)	6 (FN)	104
	True (predicted)	480 (FP)	116 (TP)	596
		578	122	700
W2V best one		False (GT)	True (GT)	
	False (predicted)	568 (TN)	47 (FN)	615
	True (predicted)	10 (FP)	75 (TP)	85
		578	122	700
TF-IDF		False (GT)	True (GT)	
	False (predicted)	1 (TN)	0 (FN)	1
	True (predicted)	577 (FP)	122 (TP)	699
		578	122	700

Table 7.15: Confusion matrix on 7000 tweets.

D2V worst one		False (GT)	True (GT)	
	False (predicted)	3141 (TN)	467 (FN)	3608
	True (predicted)	2811 (FP)	581 (TP)	3392
		5952	1048	7000
D2V best one		False (GT)	True (GT)	
	False (predicted)	324 (TN)	34 (FN)	358
	True (predicted)	5628 (FP)	1014 (TP)	6642
		5952	1048	7000
W2V worst one		False (GT)	True (GT)	
	False (predicted)	1024 (TN)	180 (FN)	1204
	True (predicted)	4928 (FP)	868 (TP)	5796
		5952	1048	7000
W2V best one		False (GT)	True (GT)	
	False (predicted)	32 (TN)	3 (FN)	35
	True (predicted)	5920 (FP)	1045 (TP)	6965
		5952	1048	7000
TF-IDF		False (GT)	True (GT)	
	False (predicted)	1 (TN)	0 (FN)	1
	True (predicted)	5951 (FP)	1048 (TP)	6999
		5592	1048	7000

_ Results – Clustering



Corpus	D2V	W2V	TFIDF
700	27.49	72.46	29.72
7,000	26.37	26.08	26.05
31,000	6.69	6.55	6.55

PERSPECTIVES

- Enrichment
- Clustering
- Location restrictions

_ Perspectives

Enrichment :

- Use metadata as (timestamp, geo-location, language …)
- Content-based image retrieval (get images which are close to the tweet's image and extract the #hashtags they have)
- Retrieve the tags "title" and "meta keywords" from links
- Event website - Get texts and try to get out keywords in order to use them as a base keywords for the beforehand bag of words

_ Perspectives

Clustering:

- DBScan (seems to be a good clustering algorithm in order to evolve over time -> not re-compute all the data, when some new data come.)

Location restrictions:

- Messages near of a border -> Enlarge the geographical limitation

_ Conclusion

- The (Word2vec + weighting) representation seems to be the best one

- Classification task is better than the clustering with Kmeans
  (machine learning techniques tend to give better results)

- Clusterize messages from social networks is not an easy task

_ Acknowledgements

- Prof. Dr. Harald KOSCH and Prof. Dr. Lionel BRUNIE (double master initiative, commitment to the german-french collaboration)

- Prof. Dr. Michael GRANITZER and Dr. Habil. Elöd EGYED-ZSIGMOND (the time they spent helping me, the meetings we made and the precious advice)

- Mrs Morwenna JOUBIN and all the people of the chair of Prof. Dr. Harald KOSCH (for all Franco-German courses she taught us and they good humor and kindness)

THANK YOU FOR YOUR ATTENTION