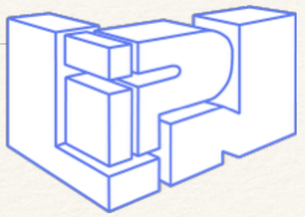
Geographic information, texts and social media

Davide Buscaldi
LIPN, Université Paris 13

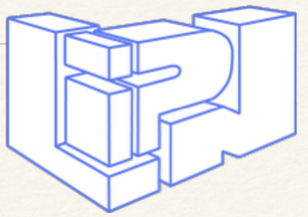
EXtraction de Connaissances à partir de donnéEs Spatialisées (EXCES) - SAGEO 2017
Rouen, 6/11/2017

Plan of the talk



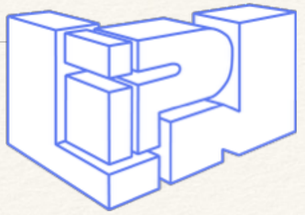
- Social Media and Geographic Information
- Extracting Geographic Information from Tweets
- An application in the Disaster Management domain

Social Media



- Key aspects:
  - **Wide coverage** and continuously growing
  - Temporally and **spatially** aware
  - Allow for mining social behaviour, opinions, tendencies
- Metaphor: users as “active sensors” in an online environment
  - Ubiquitous crowdsourcing, participatory sensor networks (Silva et al., 2013, Zhao et al., 2007)

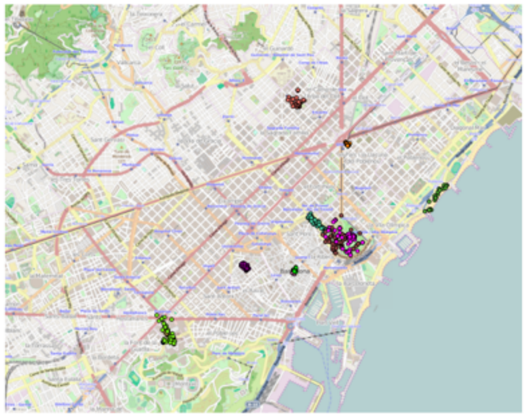
Social Media as Sensor Networks



- Social Networks can therefore be viewed as sources of data about the physical world, with millions of users reporting about “what’s happening” around them
  - Something that is happening = **event**
- Main application: <u>event detection or discovery</u>
  - Planned events such as cultural or sport events;
  - But also and especially: unplanned events such as demonstrations, crimes, accidents, natural disasters, etc.

Examples

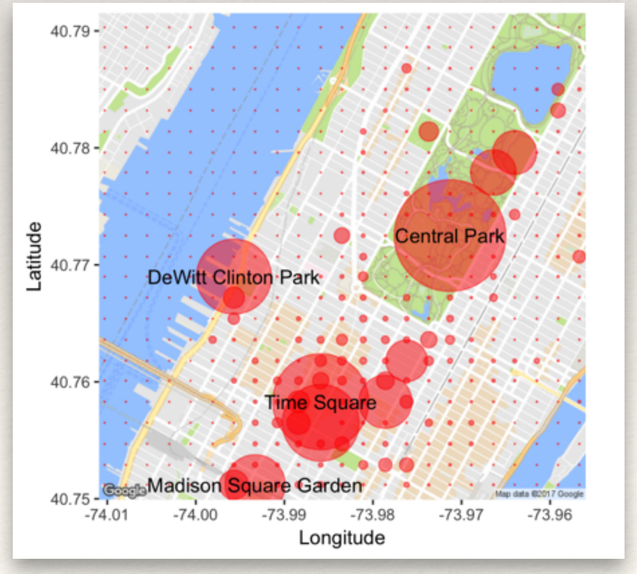
- Tweet-SCAN (Capdevila et al., 2016) uses Twitter to discover events in the city of Barcelona
- (Oostdijk et al., 2016) used Twitter to detect traffic incidences in the Netherlands
- (Sakaki et al., 2010) used Twitter data to estimate earthquakes location
- (Santos et al., 2017) used Instagram data to detect events in Manhattan



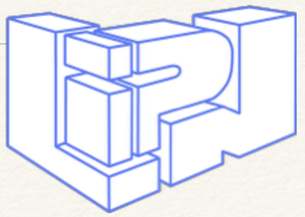
(a) Tagged events over Barcelona

Event	Number of tweets	Most common hashtags
CaixaForum conference	21	#mlearningcat
ATypl2014 conference	99	#ATypl2014
Bogatell Beach concerts	61	#Merce2014 #Txarango #mediterràniament
Fàbrica Damm concerts	89	#AntigaFàbrica, #BAM14 #Merce2014
MACBA concerts	131	#macba, #BAM14 #Merce2014
Maria Cristina concerts	30	#40merce,#Merce2014
Referendum law	167	#lleiconsultes, #9N, #Parlament
Human towers	47	#castellers, #humantowers, #Merce2014
Fireworks	59	#piromusical, #tricentenari, #Merce2014
Mapping Barcekholm	80	#Barcekholm, #projecció, #Merce2014
Van market	281	#VanVanMarket, #ParcdelaCiutadella, #food
Wine Tasting	119	#vins, #mostracat, #Merce2014

(b) Events overview

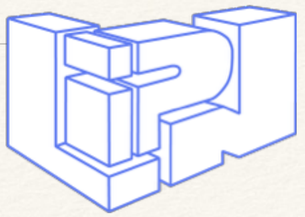


Social Media and Geographic Information



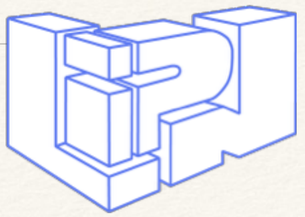
- Localizing sensors (i.e. users) is critical
- How people tell their position?
  - **Exact location** (GPS coordinates) linked to the posted content (image, text)
  - **Place mention** (toponym) - only if text is admitted as content
  - **Origin** of user (profile information) - can be both an exact location or a place mention

Exact location



- Pro: best geolocalization
- Con:
  - Not all Social Media support it
    - At least Foursquare, Instagram and Twitter have support for geo-tagged content
  - GPS may be disabled on mobile devices for various reasons:
    - Privacy issues
    - Battery duration
  - Bad or no GPS signal; location not updated (resulting in an incorrect or imprecise geo-tag)
  - Geo-tagged content is less than 1% (Gonzalez et al., 2012) of all the published content

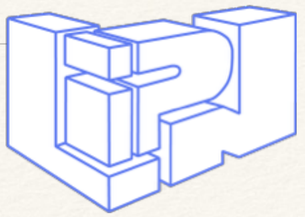
Place mention



- Pro: widely used
  - (Hoang and Mothe, 2017): between 8.8% and 17.6% in various Twitter datasets
  - Microposts NEEL Challenge[1]: 20.79% of tweets had a Location marker
- Con:
  - Potentially ambiguous
    - Toponym vs. non-toponym: ex. la Gave (de Pau) vs. gave (past form of the English verb to give)
    - Strictly geographic ambiguity: ex. Cambridge, UK vs. Cambridge, USA
    - Street names
  - Imprecise
  - Non-standard formatting: abbreviations, hashtags, etc. (BCN, #paris, #nuitblancheparis)

1 http://microposts2016.seas.upenn.edu/challenge.html

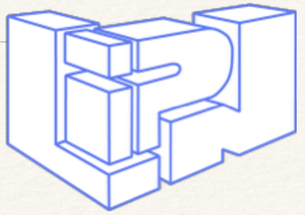
User origin



- Pro: even more widely used
- Con:
  - Highly imprecise
    - Users are moving
    - Users may talk about an event that they're not involved in (but this may occur also with automated geo-tagging using GPS coordinates)
  - If specified in a textual way, same problems with toponyms

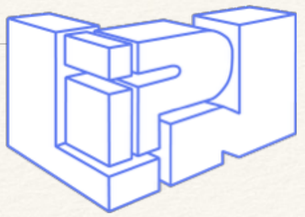
Extracting Geographic Information from Tweets

Geo-tagging Tweets



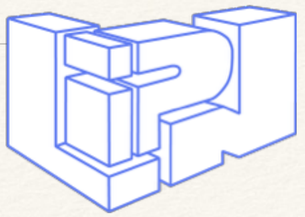
- Twitter is one of most popular social media, not only from the point of view of users but also from researchers
  - Rich context, relatively easy to analyze (text)
- But (as said before) geo-tagged content is extremely rare in Twitter
- In order to improve the effectiveness of any geographically-aware application based on Twitter, we need to geo-tag more data

Detecting placename mentions



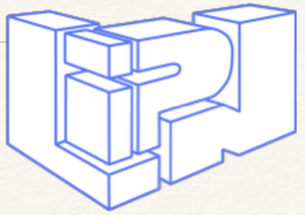
- Sub-task of Named Entity Recognition task
- Wide-coverage NER tools: LingPipe, Stanford NER, GATE, etc. (Co-NLL accuracy ~90%)
  - However: not as accurate as on other categories of names (accuracy on GeoCLEF ~55%)
- Problem: Inter-class ambiguity
  - Washington, president or place? (PER vs. LOC)
- In Twitter: additional problem (hashtag, abbreviations, noisy text)

Detecting placename mentions



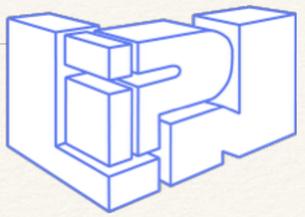
- Standard NER tools fail on Tweets
  - (Gelernter, 2011) studied the performance of the Stanford NER on a set of tweets related to the Christchurch earthquake, finding an accuracy of 34.4%
- (Ritter et al., 2011) introduced a NER tool specifically designed for tweets
  - https://github.com/aritter/twitter_nlp
- Most solutions try to normalize hashtags and abbreviations, exploiting clues in the text or external knowledge
  - For instance: #ParisFashionWeek -> Paris fashion week
  - CDMX -> Ciudad de México

Resolving place names

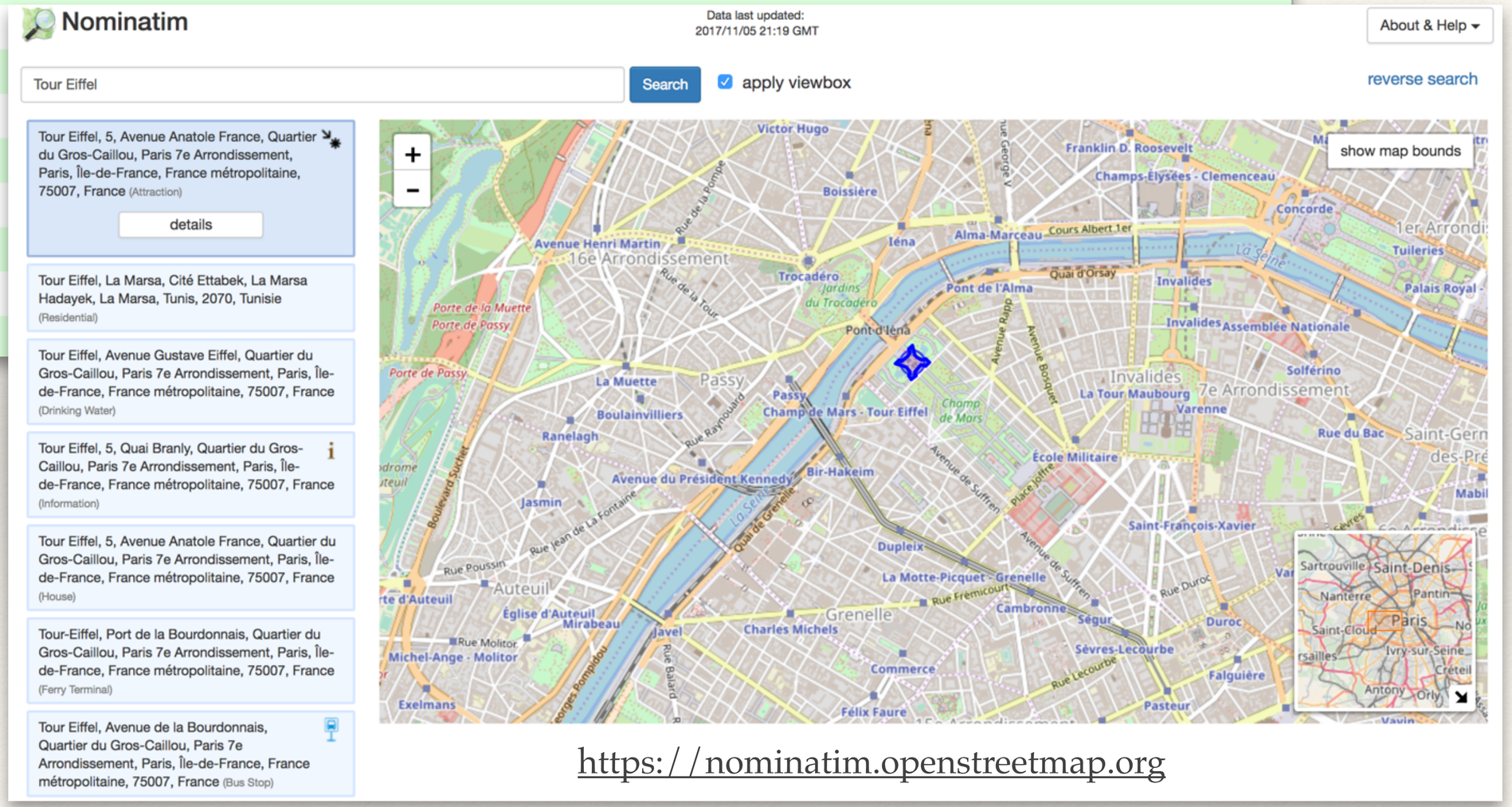


- Being able to detect mentions of places is not enough
- The toponyms may be ambiguous
- Resolving a toponym: to assign to a toponym the right referent, selected among a set of places with the same name
  - This allows to assign geographical coordinates
  - Key for the NLP – GIS bridge
- The ambiguity of a toponym depends from the world knowledge that a system has

Dealing with ambiguity

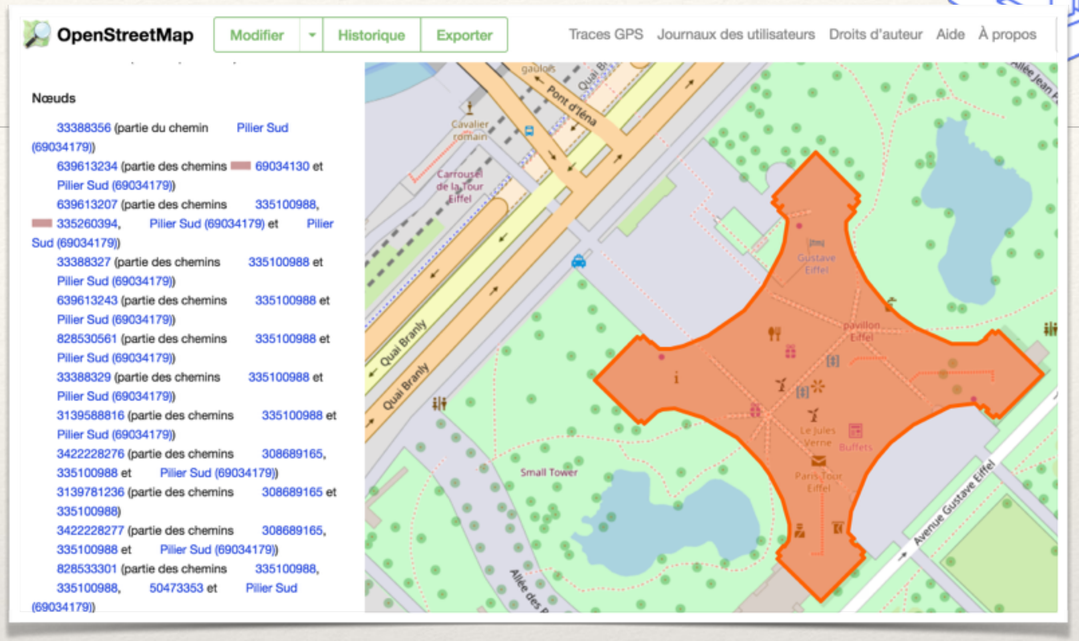


- Factors that modify the risk of ambiguity:
  - Geographical scope of an application: if the application is monitoring a small region, the less the risk of ambiguity
  - Detail of the resource used as inventory of placenames:
    - OpenStreetMap > GeoNames > (Wiki | DB)pedia

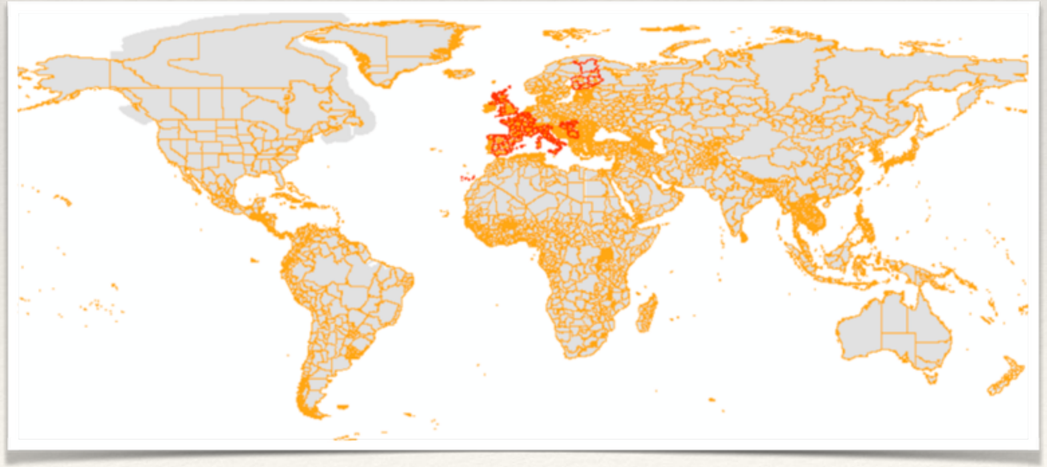


https://nominatim.openstreetmap.org

Talking resources…



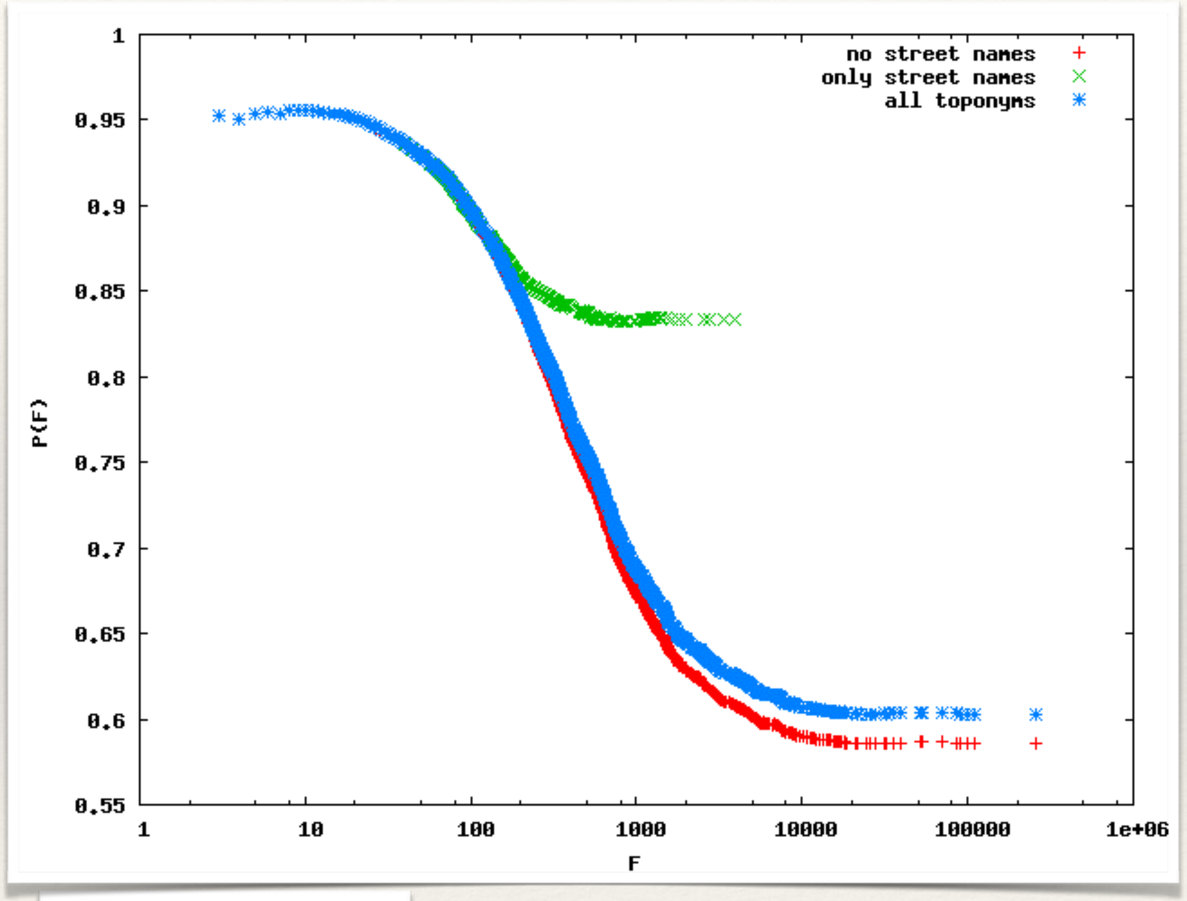
- OpenStreetMap has also polygons
- GeoNames provides only the center coordinates
  - Quattroshapes has boundaries for various administrative regions, with GeoNames IDs



- http://quattroshapes.com/#preview

Ambiguity and Frequency

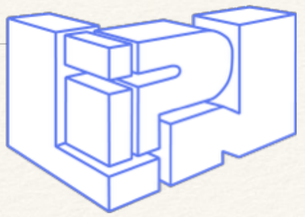
- Most frequent toponyms tend to be less ambiguous than those rarely used
- Street names are particularly ambiguous



$$P(F) = \frac{|T_{amb_F}|}{|T_F|}$$

Probability of finding an ambiguous toponym at Frequency F (Buscaldi and Magnini, 2010)

Resolving toponym ambiguity (from GIR)

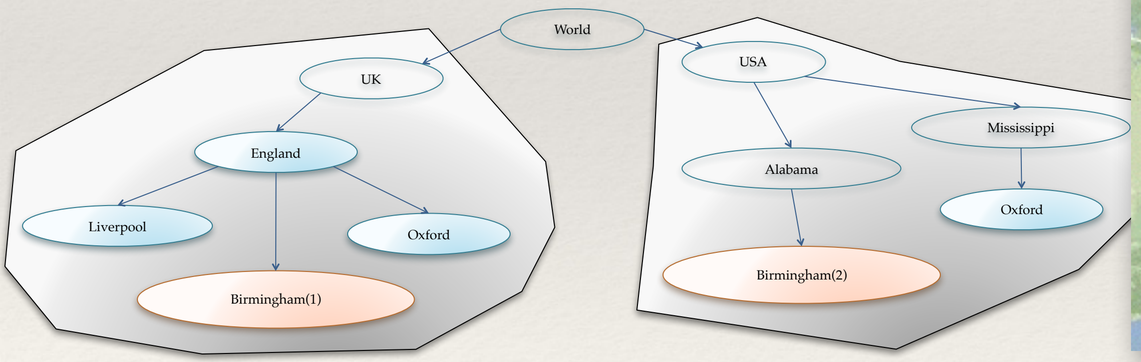


- Existing methods can be subdivided into three broad categories:
  - **Map-based**
    - They need geographical coordinates
  - **Knowledge-based**
    - They need resources providing clues for the disambiguation
  - **Data-driven** or supervised
    - They need a large enough set of labelled data
    - Many names occurring only once (impossible to estimate their probabilities)

Exploiting geographical context

- The right referent is the one with minimum average distance (geographical or conceptual) from the context locations

*"One hundred years ago there existed in **England** the Association for the Promotion of the Unity of Christendom. ... A **Birmingham** newspaper printed in a column for children an article entitled "The True Story of Guy Fawkes," ... An Anglican clergyman in **Oxford** sadly but frankly acknowledged to me that this is true. ... A notable example of this was the discussion of Christian unity by the Catholic Archbishop of **Liverpool,** …"*

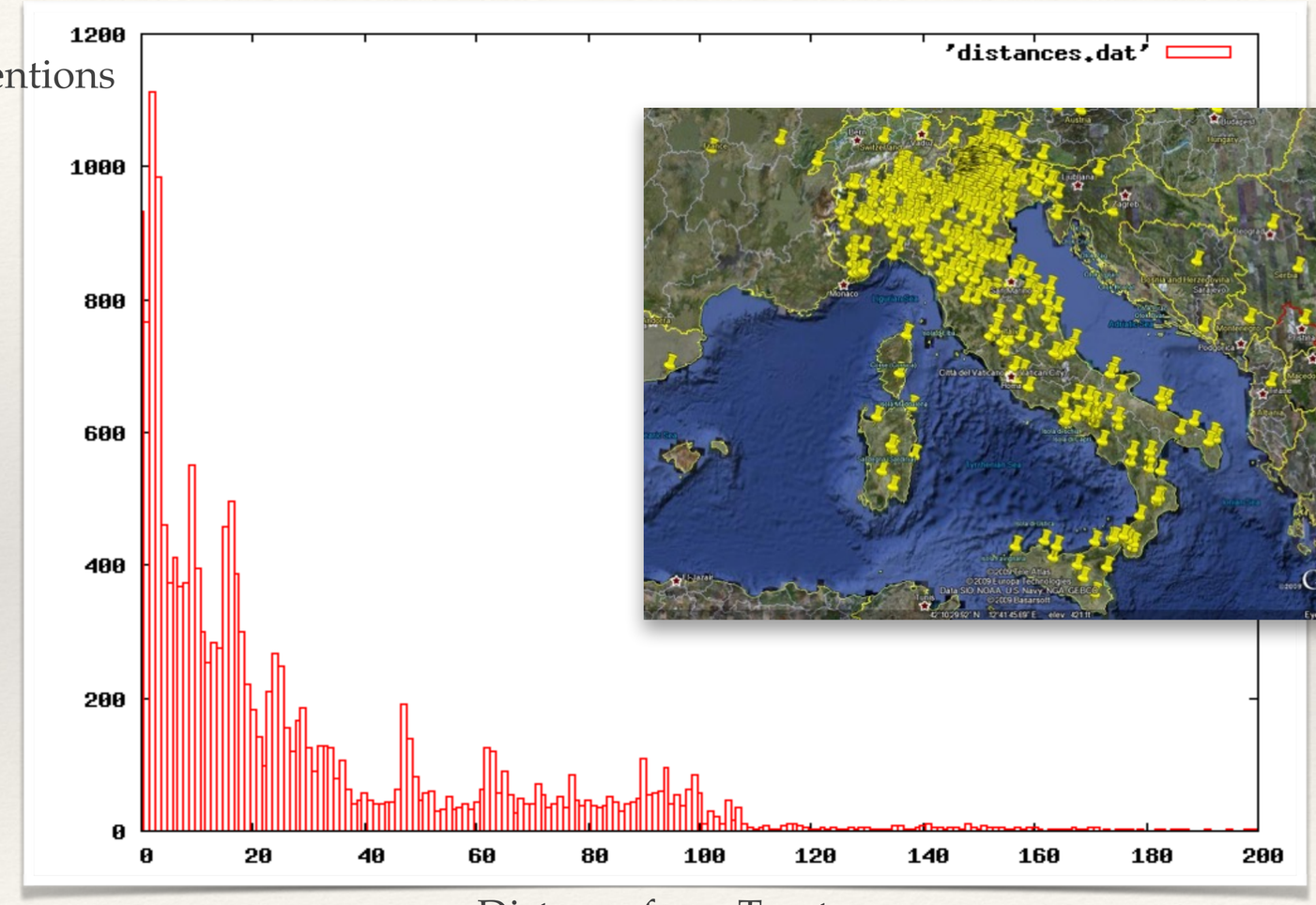


The “Steinberg hypothesis”

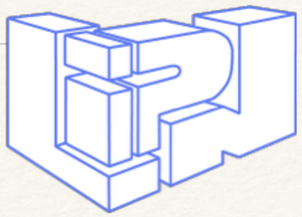
- These methods cannot be applied in Twitter because of the **lack of context toponyms**
  - Among all the tweets containing a toponym in Microposts 2016 collection, 68.8% contained just only one toponym
- (Overell, 2009) formulated the Steinberg hypothesis (based on the famous New Yorker drawing by Steinberg):

Testing the Steinberg hypothesis

- Number of mentions wrt distances from Trento of places in the Trento newspaper "L'Adige" (Buscaldi and Magnini, 2010)

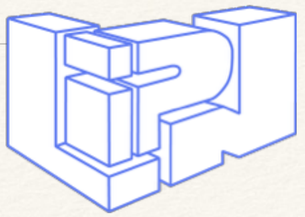


Resolving toponym ambiguity in Tweets



- The “Steinberg hypothesis” may help in disambiguating toponyms in Tweets
- Need to take into account additional context:
  - Mention of places in the history of the user
  - User origin
  - Size of places
- (Zhang and Gelernter, 2014) found the following best features:
  - population of place in gazetteer entry.
  - number of alternative names within an entry, and among matching entries.
  - other location expressions mentioned in the tweet

Geo-tagging without place names



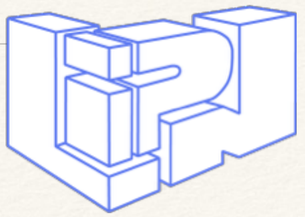
- Technique introduced by (Paraskevopoulos and Palpanas, 2015)
- Idea: use tweets that are geo-tagged with precise coordinates to build a signature corresponding to an event taking place at the same coordinates
- Tweets that are not geo-tagged but have words contained in a signature derived from geo-tagged data are tagged with the same location
- Ideal for narrow, focused scope

INPUT: A training set of timestamped and geotagged tweets, a timestamped query-tweet (Q_t) that is not geotagged.
OUTPUT: The most eligible candidate location.

1: **for all** $i \in$ {candidate geolocations: Geolocs} **do** ▷ process training dataset, for all locations
2: **for all** $t \in$ {time intervals} **do** ▷ and for all time intervals
3: $Doc_{i_t} \leftarrow$ all tweets in location i at time interval t
4: $kwVector_{i_t} \leftarrow$ create vector of Doc_{i_t} keywords and their weights
5: $kwVector_{Q_t} \leftarrow$ create vector of Q_t keywords and their weights ▷ process non-geotagged tweet Q_t
6: $location \leftarrow argmax_{i \in Geolocs}$ {similarity between $kwVector_{i_t}$ and $kwVector_{Q_t}$} ▷ identify location of tweet Q_t
7: **return** $location$

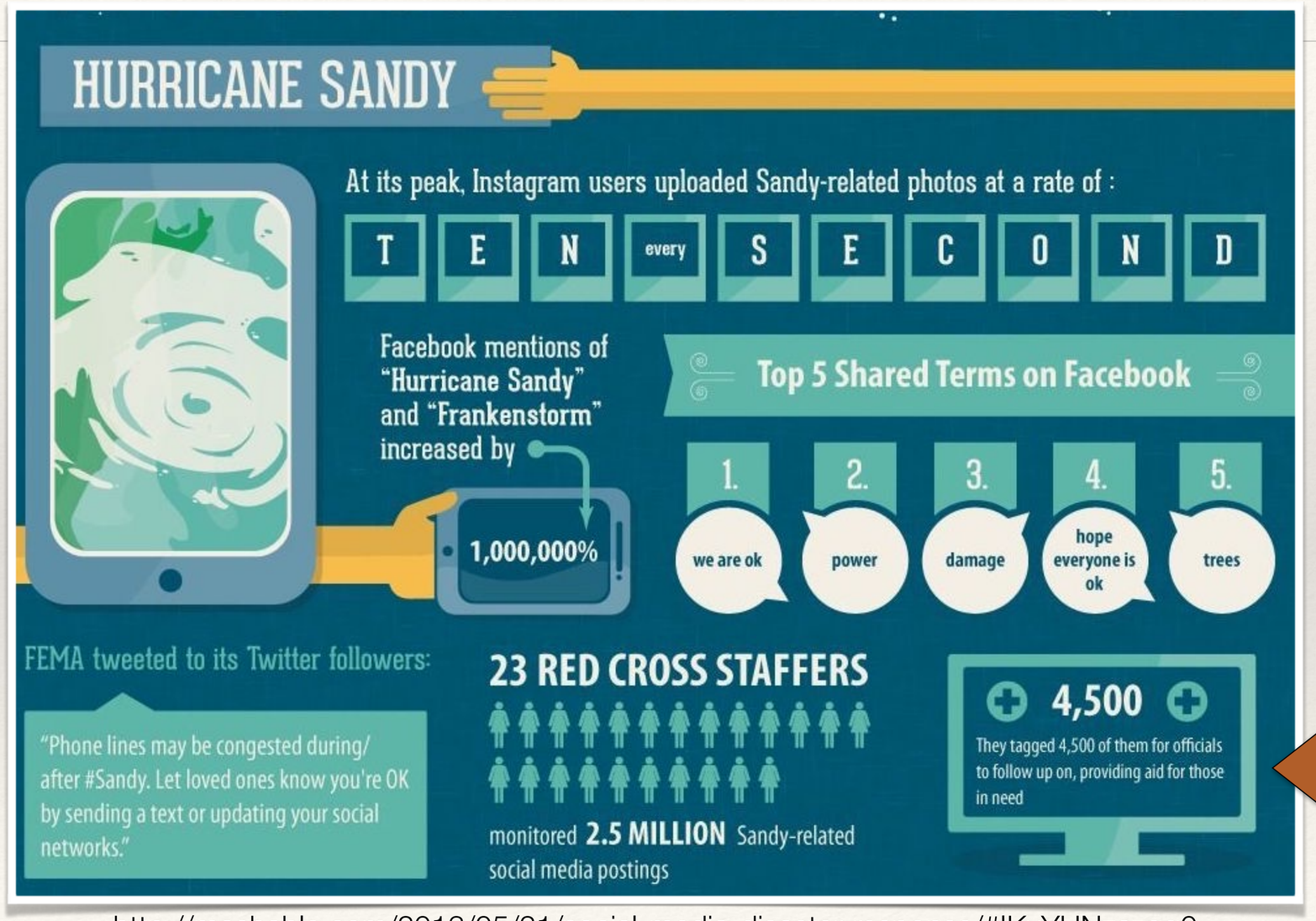
An application in the disaster management domain

Not only event detection



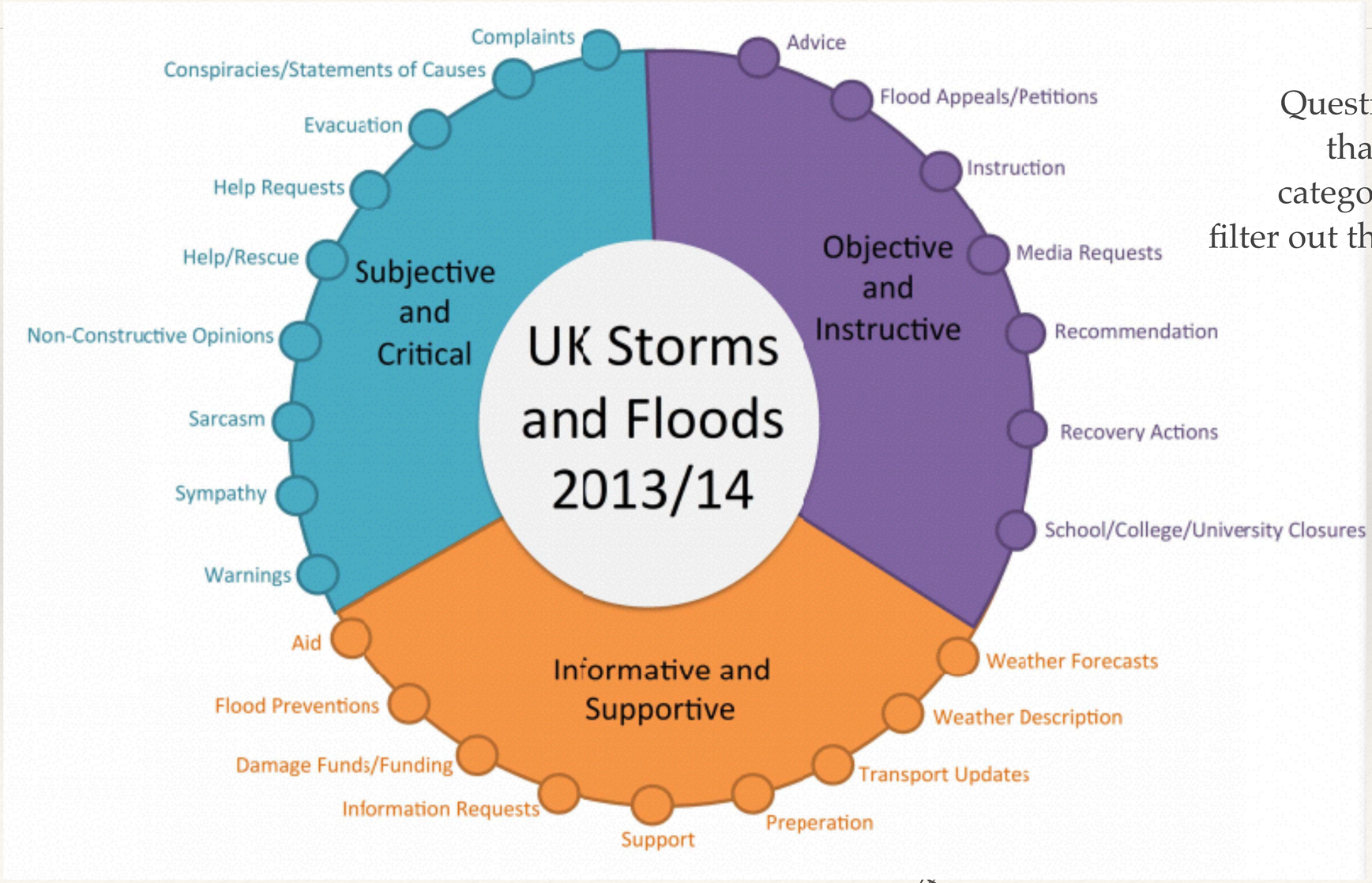
- Many applications that exploit the spatial data are focused exclusively on the detection of ongoing events
- Tweets provide more context that is worth to be analyzed
  - Going beyond counting the number of tweets in a certain place
- Applying NLP techniques such as sentiment analysis to geo-tagged data has many potential interesting applications

Social Media and Disaster Management



http://mashable.com/2013/05/21/social-media-disaster-response/#IKxYUNzazgq0

Social Media and Disaster Management



Question: are we able to select tweets that are related to "interesting" categories (such as help requests) and filter out the "noisy" categories (sarcasm, etc.)?

From: Parsons et al., Thematically Analysing Social Network Content During Disasters Through the Lens of the Disaster Management Lifecycle

Case Study: the 2014 Genoa flood

Total annotations:

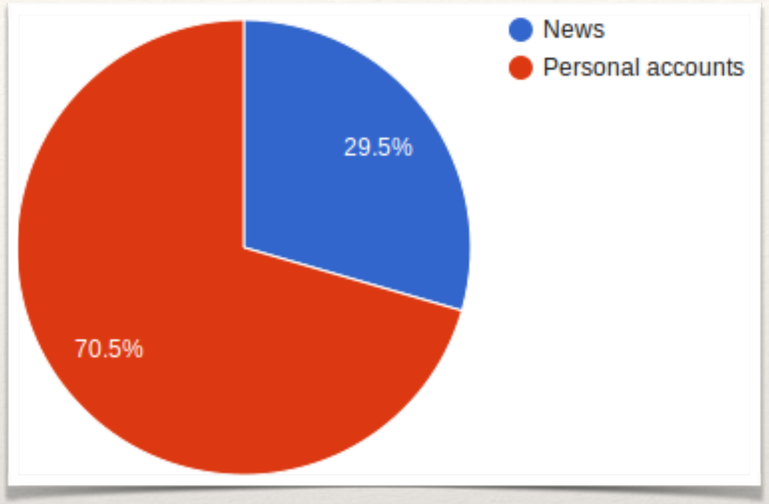
8922 subj
499 pos
3519 neg
1019 iro

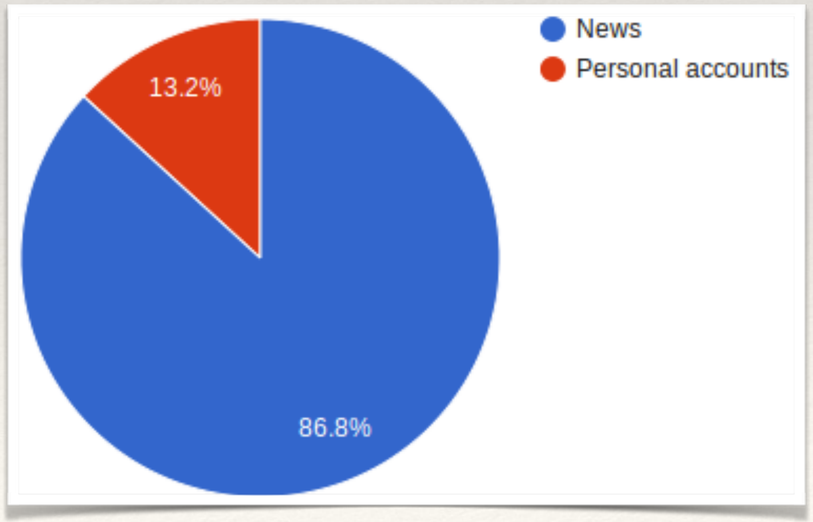
6033 users

Problem: evaluation

Evaluating subjectivity accuracy

- Manual annotation of 75 sources
- **News** (59.2%)
  - @ScoopSquareGE, @MeteoWeb_eu, @infoitinterno, @ITnewsGE, @TopTrendIT…
- **Personal accounts** (40.8%)
  - @Miti_Vigliero, @mauneobux, @LaRouge_DOC, …





Indirect evaluation - relevant items

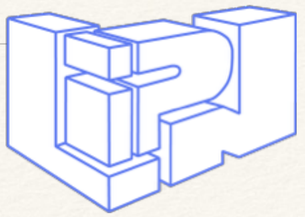
- Subset of relevant **hashtags**, **toponyms** and **topics** :
  - #allertameteo, #protezionecivile, #alluvionege, ...
  - Montoggio, Sturla, Fereggiano, ...
  - “Ondata di piena”, “Invaso dal fango”, “esonda …”…

Indirect evaluation - relevant items

* acqua , deraglia un freccia
* allagamenti e paura
* allerta massima
* allerta meteo
* alluvione in corso
* croce rossa
* crolla muraglione
* danni enormi
* danni ingenti
* deraglia un treno
* emanata allerta
* esce dai binari
* esonda il bisagno
* esonda lo scrivia
* esondano i fiumi
* esondano il bisagno
* esondano i torrenti
* esonda un torrente
* evacuate 16 famiglie
* fa paura
* frana improvvisa
* ho paura
* invaso dal fango
* livello di allerta
* mancato allarme
* meteo da allerta
* momenti di ansia
* morto e dispersi
* morto e gravi
* morto foto
* nuova alluvione
* nuovo alluvione
* ondata di piena
* piani alti
* prestare massima attenzione
* protezione civile
* rubano nei negozi
* sciacalli in azione
* situazione critica
* situazione drammatica
* tanta paura
* trascinate da acqua
* trascinate dall' acqua
* travolge auto
* trovato il corpo
* usate l' auto
* uscire di casa
* uscite da casa
* #allerta2
* #allertameteo
* #allertameteoge
* #allertameteogenova
* #allertameteolg
* #allertameteolig
* #alluvione
* #alluvionege
* #angelidelfango
* #arpal
* #bisagno
* #busalla
* #campoligure
* #chiaravagna
* #crocerossa
* #emergenza24
* #emergenze

#entella
#fereggiano
#forzagenovarisorgidalfango
#genovaalluvione
#iononrischio
#maltempo
#molassana
#montoggio
#noncefangochetenga
#ovada
#polcevera
#protezionecivile
#stura
#voltri

Detecting trending items



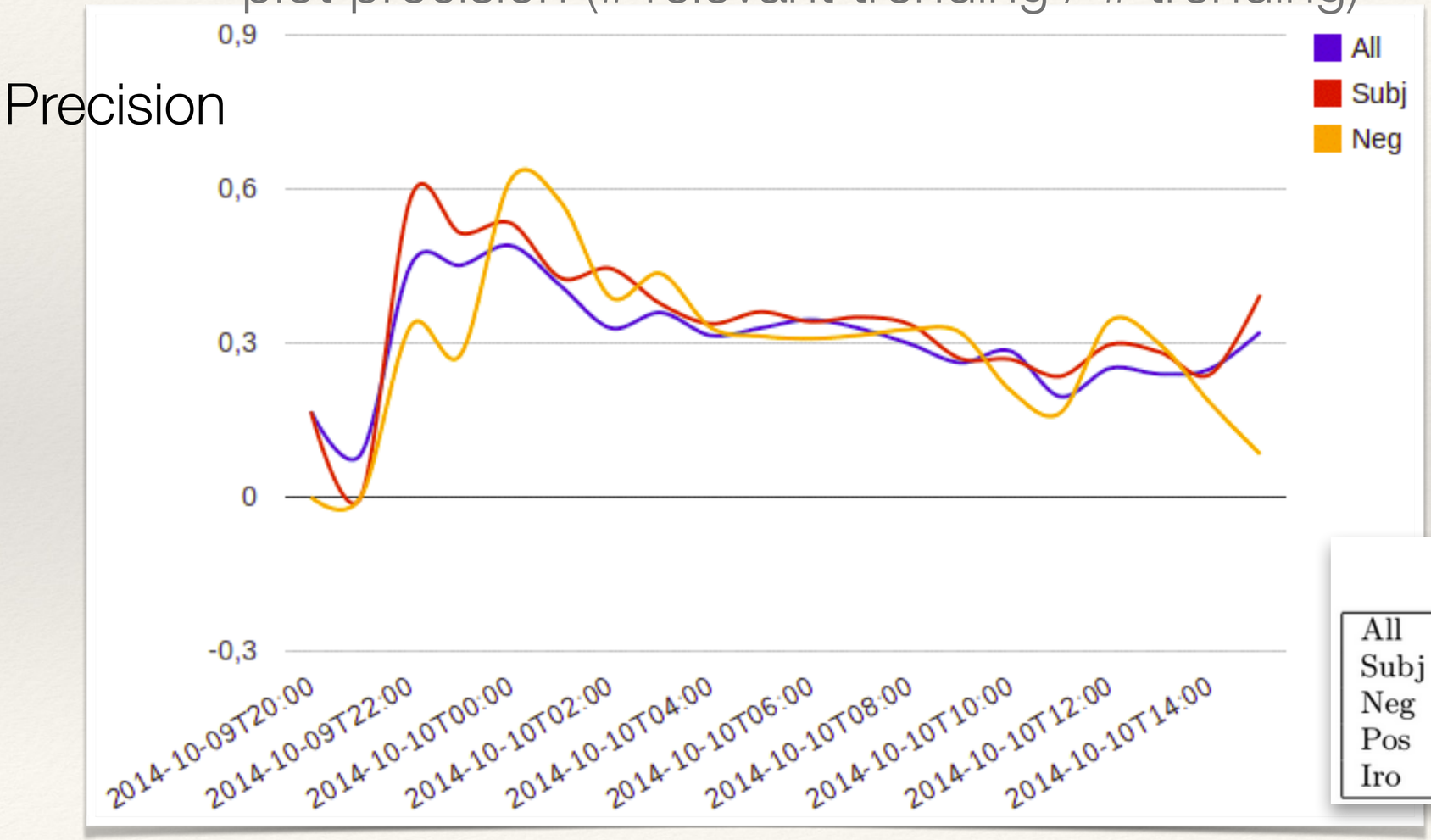
- Extraction of trending* (hashtags | toponyms | topics) for each hour

$$Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} > \epsilon$$

- The key is to model λ to capture the expected frequency of an item
- With enough data, it is possible to model the frequency over the same period (week, hour, etc.) of the year
- In our test, the threshold ε was set to 0
- λ was set to the estimated frequency during the previous hour

Indirect evaluation

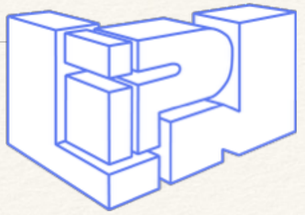
- plot precision (# relevant trending / # trending)



	Hashtags		Toponyms		Topics	
	Acc.	Cov.	Acc.	Cov.	Acc.	Cov.
All	0.160	1.000	0.660	0.875	0.104	0.980
Subj	0.197	0.875	0.661	0.844	0.156	0.804
Neg	0.254	0.625	0.482	0.594	0.137	0.451
Pos	0.208	0.188	0.155	0.156	0.164	0.157
Iro	0.322	0.313	0.362	0.344	0.118	0.098

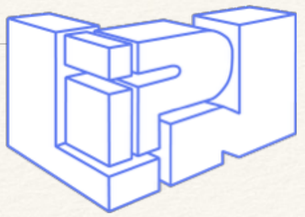
Mean Average Precision

Qualitative analysis



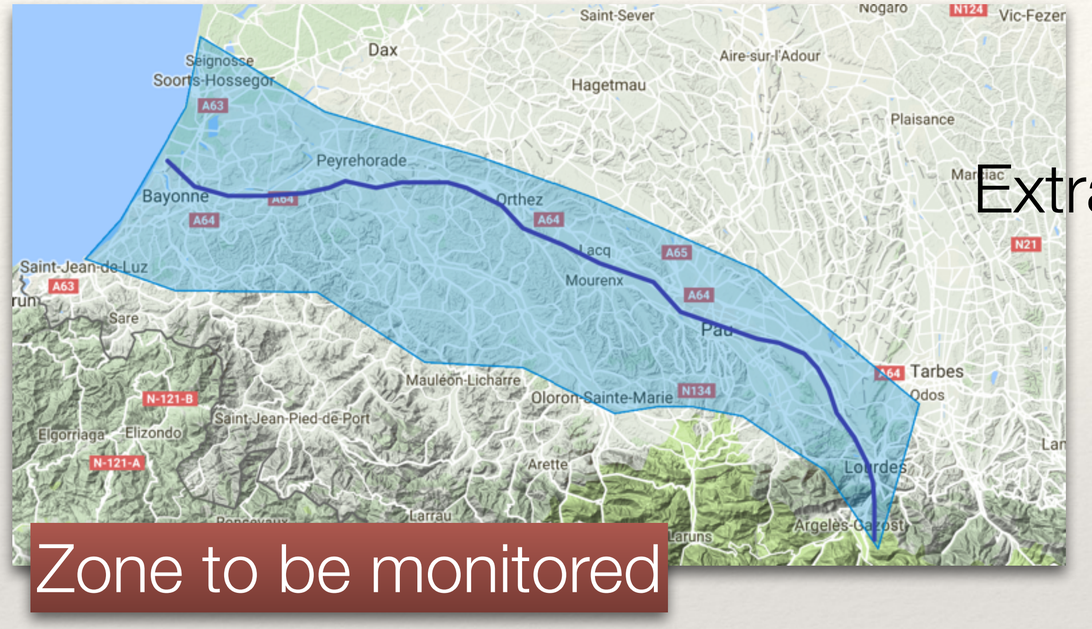
- Most of the positive tweets (82 out of 493 - ~16%) came from the same account (spam?):
  - “#News e #Astronomia Genova alluvionata, esondano i fiumi: un morto... -... http://bit.ly/110yfvW Buona visione ! :))”
- Many false positives for irony but also some good answers: “Ragazzi tranquilli #Renzi ha detto che non ci lascerà soli”
- Negative tweets may convey **different emotions**:
  - Fear: “Basta guardare fuori per avere paura"
  - Rage: "In tre anni non è stato fatto nulla, amministrazioni delinquenti”
  - Sadness: "Mi viene da piangere”
- Some features may **change polarity** (meaning) depending on the context:
  - “Coraggio Genova!!!” vs. “Salite ai piani alti!!!”
  - “Speriamo bene” vs. “Stiamo bene”

Geographical aspects



- The narrow scope of the study limited also the ambiguity of the involved toponyms
- Only 13 toponyms in the collection, all with only one possible referent
- In most cases they were written correctly, with the only problems related to the use of hashtags and incorrect capitalization
- Less than 190 tweets were geo-tagged (GPS coordinates)

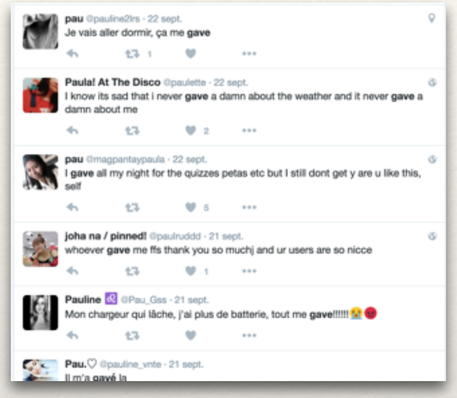
Future plans: project NaDiA



Zone to be monitored

Meteorological alert

Extracting related flux

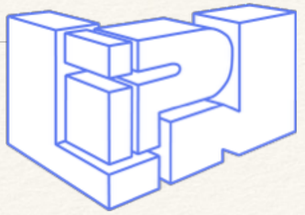


Signalling relevant situations

Seine dataset

- We collected 45,831 tweets related to the spring 2016 flood in France (annotating…)
  - <tweet id="736542320611446784" authorid="@rdvavecdamenat" time="2016-May-28 14:59" loc="**Mirebeau, France**">ça va être costaud les orages cette après-midi sur le **Poitou** fort risque d'inondations locales à prévoir #**Vienne86**</tweet>
  - <tweet id="735868225506947072" authorid="@AuCoeurMeteo" time="2016-May-26 18:20">#Inondations locales à #**StMalo** sous un orage diluvien...https://twitter.com/Meteovilles/status/735867818781069312 …</tweet>
  - <tweet id="735850798316441600" authorid="@KeraunosObs" time="2016-May-26 17:11">Inondations locales en #**Bretagne** à **Plougasnou** dans le #**Finistère** lors des #orages de l'après-midi via @LeTelegrammepic.twitter.com/HZtrRK5sQg</tweet>

Conclusions



- Social Media are a rich source of information with spatial and temporal markers
- Most of the geographic information is coded using text (toponyms)
- Resolving toponyms may be a major problem - or not, depending on the scale of the application
  - GIR-derived Algorithms and techniques often lack context to be effective at the same level
  - Need to exploit the network to enrich context (where possible)
- NER algorithms need to adapt to the writing style of social media

Selected Bibliography

- Thi Bich Ngoc Hoang, Josiane Mothe: Location Extraction from Tweets, Information Processing and Management, in Press (2017)
- Pavlos Paraskevopoulos, Themis Palpanas: Fine-Grained Geolocalisation of Non-Geotagged Tweets, Advances in Social Networks Analysis and Mining (ASONAM), 2015
- Judith Gelertner, Nikolai Mushegian: Geo-parsing Messages from Microtext, ransactions in GIS, 2011, 15(6): 753–773
- Zhang, Wei, and Judith Gelernter. "Geocoding location expressions in Twitter messages: A preference learning method." Journal of Spatial Information Science 2014.9 (2014): 37-70.
- Capdevila, Joan, et al. "Tweet-scan: An event discovery technique for geo-located tweets." Pattern Recognition Letters 93 (2017): 58-68.
- Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. "Earthquake shakes Twitter users: real-time event detection by social sensors." Proceedings of the 19th international conference on World wide web. ACM, 2010.
- Buscaldi, Davide, and Bernardo Magnini. "Grounding toponyms in an Italian local news corpus." Proceedings of the 6th Workshop on Geographic Information Retrieval. ACM, 2010.
- Nelleke Oostdijk, Ali Hürriyetoglu, Marco Puts, Piet Daas, Antal van den Bosch: Information extraction from social media: A linguistically motivated approach. TALN 2016, Paris