D4.2 PRELIMINARY VERSION OF THE CONTENT EXTRACTION AND ANALYSIS MODULE FOR ONLINE SOCIAL PLATFORMS

Grant Agreement nr	611057
Project acronym	EUMSSI
Start date of project (dur.)	December 1st 2013 (36 months)
Document due Date :	30/11/2014
Actual date of delivery	23/12/2014
Leader	UPF
Reply to	jens.grivolla@upf.edu
Document status	Submitted

Project co-funded by ICT-7th Framework Programme from the European Commission

Project ref. no.	611057
Project acronym	EUMSSI
Project full title	Event Understanding through Multimodal Social Stream Interpretation
Document name	EUMSSI_D4.2_Preliminary version of the content extraction and analysis module for online social platforms_20141223
Security (distribution level)	PU – Public
Contractual date of delivery	30/11/2014
Actual date of delivery	23/12/2014
Deliverable name	Preliminary version of the content extraction and analysis module for online social platforms
Type	P – Prototype
Status	Submitted
Version number	1
Number of pages	25
WP / Task responsible	GFAI / UPF
Author(s)	Jens Grivolla
Other contributors	Maite Melero
EC Project Officer	Mrs. Aleksandra WESOLOWSKA Aleksandra.WESOLOWSKA@ec.europa.eu
Abstract	D4.2 documents a preliminary version of the module that extracts publications from online social networks, blog/news providers and social news websites, together with a module that analyses the influence and impact of the extracted contents.
Keywords	Social Media
Circulated to partners	No
Peer review completed	No
Peer-reviewed by	-
Coordinator approval	Yes

TABLE OF CONTENTS

1 Background **4**

2 Introduction **5**

3 Social Media Sites **6**
3.1 Twitter 6
3.2 YouTube 6

4 Social Media Data Collection **7**
4.1 Data crawlers 7
4.2 Crawling of Twitter data 7
4.3 Crawling of Youtube data 8

5 Social Media Analysis **9**
5.1 Statistics 9
5.2 Social Media Algorithms 9
5.3 Temporal dynamics 15

6 Conclusions and future work **16**
6.1 Current state and ongoing work 16
6.2 Next steps and future work 16

7 Appendix – Statistics (IPython) **18**
7.1 Setup 18
7.2 Analysis 18
7.2.1 number of content items 18
7.2.2 number of tweets 18
7.2.3 number of Youtube videos 18
7.2.4 top tags 18
7.2.5 languages 24

1 BACKGROUND

D4.2 is the first deliverable of Tasks 4.4. and 4.5. It is a prototype deliverable consisting of a preliminary version of the module that extracts publications from online social networks, blog/news providers and social news websites, together with a module that analyses the influence and impact of the extracted contents. The present report is intended to document the implementation of this module. The module will be progressively enhanced and will yield two more deliverables in Months 24 and 32.

2 INTRODUCTION

The term *Social Media Analysis* covers two distinct aspects. On one hand it refers to the analysis of User Generated Content (UGC) and Social Media content (e.g. Twitter, Youtube, Facebook, ...), with the purpose of extracting information, such as entities, relations or opinion, from this content. On the other hand, it denotes the use of techniques and algorithms that model interactions or social dynamics, such as graph-based analysis, or the analysis of social / temporal dynamics, usually based on the associated metadata.

Task 4.4 deals with the extraction of content from online social platform so that it can be processed by the EUMSSI multimodal platform, whereas task 4.5 covers the analysis of said content, in terms of social and temporal dynamics. The work carried out by M12 of the EUMSSI project has focused mainly on the content extraction, with some preliminary social analysis. Further analysis will be facilitated in the short term, facilitated by having an efficient and flexible access to all gathered content through the EUMSSI platform, which now is starting to be functional.

3 Social Media Sites

3.1 Twitter

Twitter is an online social networking and microblogging service that enables users to send and read short 140-character text messages, called "tweets". It has emerged as one of the premier social media analytics channels, with over 3 billion tweets and 15 billion API calls generated daily [DuVander, 2012]. From its inception, tweets were set to a largely constrictive 140-character limit for compatibility with SMS messaging, introducing the shorthand notation and slang commonly used in SMS messages. The 140-character limit has also increased the usage of URL shortening services. Since June 2011, Twitter has used its own t.co domain for automatic shortening of all URLs posted on its website.

Twitter has a history of both using and releasing open source software. The service's application programming interface (API) allows other web services and applications to integrate with Twitter. Individual tweets are registered under unique IDs using software called snowflake and geolocation data is added using 'Rockdove'. The URL shortner t.co then checks for a spam link and shortens the URL. Tweets are stored in a MySQL database using Gizzard and acknowledged to users as having been sent.

Thanks to its metadata, it has become a discovery engine for finding out what is happening right now, "what are the trending topics". In addition, there are numerous tools for adding content, monitoring content and conversations. In https://dev.twitter.com/docs/platform-objects/tweets there is a comprehensive list with the all possible metadata fields, although not all may appear in all contexts. In EUMSSI, we are collecting the following information from tweets: geo-coordinates, creation time and date, language, number of retweets, user, and , the text itself.

Apart from the metadata, the text itself is an important source of information, from which we extract entities, relations or opinion, notwithstanding the difficulties associated to the analysis of tweet text, and UGC in general.

3.2 YouTube

YouTube is a video-sharing website, currently owned by Google, on which users can upload, view and share videos. The most relevant metadata associated to the videos, which we plan to collect, includes the number of likes, number of dislikes, number of views, number of favorites, author, publication time and date and, finally, the list of comments made by users on each particular video.

Our purpose is to enrich the original YouTube metadata by aggregating the information resulting from the analysis of the content itself, both the actual video and the list of user generated comments.

4 Social Media Data Collection

4.1 Data crawlers

As detailed in D2.3 and D5.3, the EUMSSI platform incorporates a variety of input sources (or crawlers) providing constant updates to the data that is stored and managed by the platform. This includes in particular the real-time collection of social media data, such as Twitter and Youtube.

4.2 Crawling of Twitter data

Twitter corpus collection is facilitated by the API itself. However its exploitation and reuse is seriously restricted by the Twitter terms of service, which do not allow the sharing of aggregated resources of tweets. A common workaround to this problem is the distribution of only lists of tweet IDs, as is done for example in the TREC microblog shared task (http://trec.nist.gov/data/tweets/).

In EUMSSI, Twitter content is retrieved by following specific relevant hashtags, keywords and users, using the Twitter Streaming API. The list of hashtags was manually built starting with the 'seed' hashtag #fracking and then iteratively looking for the "related hashtags" provided by the site #hashtags.org. This gave us a list of 20 different, multilingual hashtags, some of which are very productive. At the same time, a list of the most prolific users for each hashtag was collected.

Later on, the initial list has been iteratively expanded by adding further relevant terms appearing more frequently in the crawled tweets, particularly for languages different than English. This tag list expansion process is currently done manually, but later in the project it will become automatized so as to be able to dynamically respond to the emerging trends in the live Twitter flow.

In order to classify tweets by language, we currently look at the metadata autodetected by Twitter itself (i.e. the language attribute), but at some point we may need to use an external language identification module, such as LangID (Lui and Baldwin, 2012) or Google language detector, which is part of Google Translate (McCandless, 2011)).

An extract of the initial list, along with the users most actively using each tag, is provided in table 1.

The platform currently contains a continuously running crawler component that integrates tweets into the EUMSSI platform database in real time. Those tweets are immediately available for analysis and show up in the Solr indexes used by the end applications or demonstrators, within seconds of having been posted on Twitter.

The crawler, as all project code, can be found on GitHub[1], and is written in Python[2] using the Twython[3] library to access the Twitter Streaming API[4].

Having collected over two million tweets, unfortunately all data was lost in a major incident with the project's server (hosted on Azure[5]) in November 2014, shortly before

[1] https://github.com/EUMSSI/EUMSSI-platform/tree/master/crawlers/twitter
[2] https://www.python.org/
[3] https://github.com/ryanmcgrath/twython
[4] https://dev.twitter.com/streaming/overview
[5] http://azure.microsoft.com/

#fracking	@frackoff_	updates: 87,073 followers: 2,491
	@johnlundin	updates: 160,206 followers: 9,407
	@thetruelorax	updates: 2,632 followers: 281
	@marcellus_SWPA	updates: 35,271 followers: 2,684
	@LAGOPUEBLA	updates: 6,415 followers: 174
	@DWBerkley	
#oilandgas	@APTOilgastrans	updates: 3,127 followers: 32
	@Guilly2P	updates: 1,722 followers: 430
	@gorman_mary	updates: 8,794 followers: 1,691
	@OilFinity	updates: 3,056 followers: 22,316
	@SAPOilandGas	updates: 857 followers: 1,755
#frackoff	@SkyzaLimitPro	updates: 3,017 followers: 297
	@RomaniaRising	updates: 5,337 followers: 833
	@ecoforumorg	updates: 10,085 followers: 2,011
	@BarbaraQuigley1	updates: 12,994 followers: 478
#NOFRACKING	@OElika	updates: 39,834 followers: 1,084
	@thecynth	updates: 5,108 followers: 450
	@JoBurgess11	updates: 26,229 followers: 629
	@fuller_derek	updates: 147,671 followers: 5,875
#shale	@TheEarthNetwork	updates: 170,428 followers: 5,092
#shalegas	@ShaleNOW	updates: 3,589 followers: 2,700
	@overges	updates: 6,147 followers: 1,044
	@gardencatlady	updates: 140,222 followers: 23,387

Figure 1: Hashtags and most active users

this report was being compiled. At this point, $252\,559$ tweets have again been collected since the incident, in addition to $192\,703$ tweets from an earlier collection outside the EUMSSI platform, getting presently close to half a million tweets.

4.3 Crawling of Youtube data

Youtube videos are also being collected, using a crawler written in Java[6], with additional code in Python to import the results into the platform. A full integration into the platform with live updates is planned for the near future.

This collection contains videos from Deutsche Welle's Youtube channels[7], and from The Guardian[8] channel, as well as others that correspond to relevant keywords for the project's thematic scope.

This collection will be expanded to contain additional videos that are referenced from other media (such as Twitter), as well as to include user comments referring to the collected videos.

[6] https://java.com/

[7] https://www.youtube.com/channels?q=deutsche+welle

[8] https://www.youtube.com/user/TheGuardian

5 Social Media Analysis

Analysing the content collected from Social Media can be approached in many ways, be it to gain a deeper understanding of the data, extracting actionable insight, or extracting content and visualizations for the end user. Social analysis is centered around the combination of three key aspects: user activity, content, and temporality.

5.1 Statistics

There is currently an IPython Notebook[9] that can be viewed on nbviewer[10] with a variety of statistics, including the number of available items per language, the most prominent tags per language, etc. A current snapshot of those statistics is provided in the appendix at the end of this document.

The statistics are currently generated directly from the Mongo database, but the Solr indexes will make it easier and more efficient to extract this and similar data in the future.

At this point in time, beyond giving an overview of the content collection process, the results allow us to improve the tag list and detect e.g. the current imbalance between languages which suggests that the tag list is too focussed on tags used in English tweets. As hinted above, the list of frequent tags by language can help improve the tag lists in order to reduce this imbalance, and could in the future help develop automatic mechanism for expanding and improving the data collection process.

5.2 Social Media Algorithms

An important part of of analyzing social media is what is commonly referred to as Social Network Analysis:

> Social network analysis (SNA) is the use of network theory to analyse social networks. Social network analysis views social relationships in terms of network theory, consisting of nodes, representing individual actors within the network, and ties which represent relationships between the individuals, such as friendship, kinship, organizations and sexual relationships.[Pinheiro, 2011][Abraham et al., 2009] These networks are often depicted in a social network diagram, where nodes are represented as points and ties are represented as lines. *[Wikipedia, 2014]*

Social Network Analysis is used often to identify influential sources, using a variety of metrics. Some of the most common are indegree, outdegree, and centrality. Similarly, metrics can be used to characterize the whole network, showing its structure and identifying interesting subgraphs (cliques). A good introduction can be found in [Hanneman and Riddle, 2005].

Another use is finding trends by observing the temporal evolution, or the propagation of concepts, terms or hashtags throughout the network.

[9] `http://ipython.org/notebook.html`

[10] `http://nbviewer.ipython.org/github/EUMSSI/EUMSSI-tools/blob/master/scripts/eumssi-social.ipynb`

Related algorithms are also used to illustrate and visualize complex interactions, plot interactions and relations, or for visual clustering of concepts or users, based on their interactions in the graph. A very popular tool that integrates a wide range of algorithms, both for numeric analysis and for visualization is Gephi [Bastian et al., 2009], which was used to produce the plots shown below.

Figure 2 shows the relations between Twitter users in our *fracking* related collection. The graph is constructed by linking users that mention other users (most frequently in the form of retweets[11]). The size of a user's name reflects how many times this user mentioned other users in their tweets, whereas the intensity of color corresponds to the frequency with which a user was mentioned by others. Figure 3 shows the same graph, inverting outgoins and incoming mentions.

A very clear conclusion from seeing these two mirrored views of the Twitter activity is that there is a strong distinction between "users who mention others" and "users who are mentioned by others". It becomes apparent that some users act as *aggregators*, propagating the content, whereas other are *content creators* who are then cited by others. In fact, many of the aggregators (such as ShaleMarkets) appear to be automatic systems that retweet content from a number of sources, or even any content matching certain keywords.

[11] https://support.twitter.com/articles/77606-faqs-about-retweets-rt

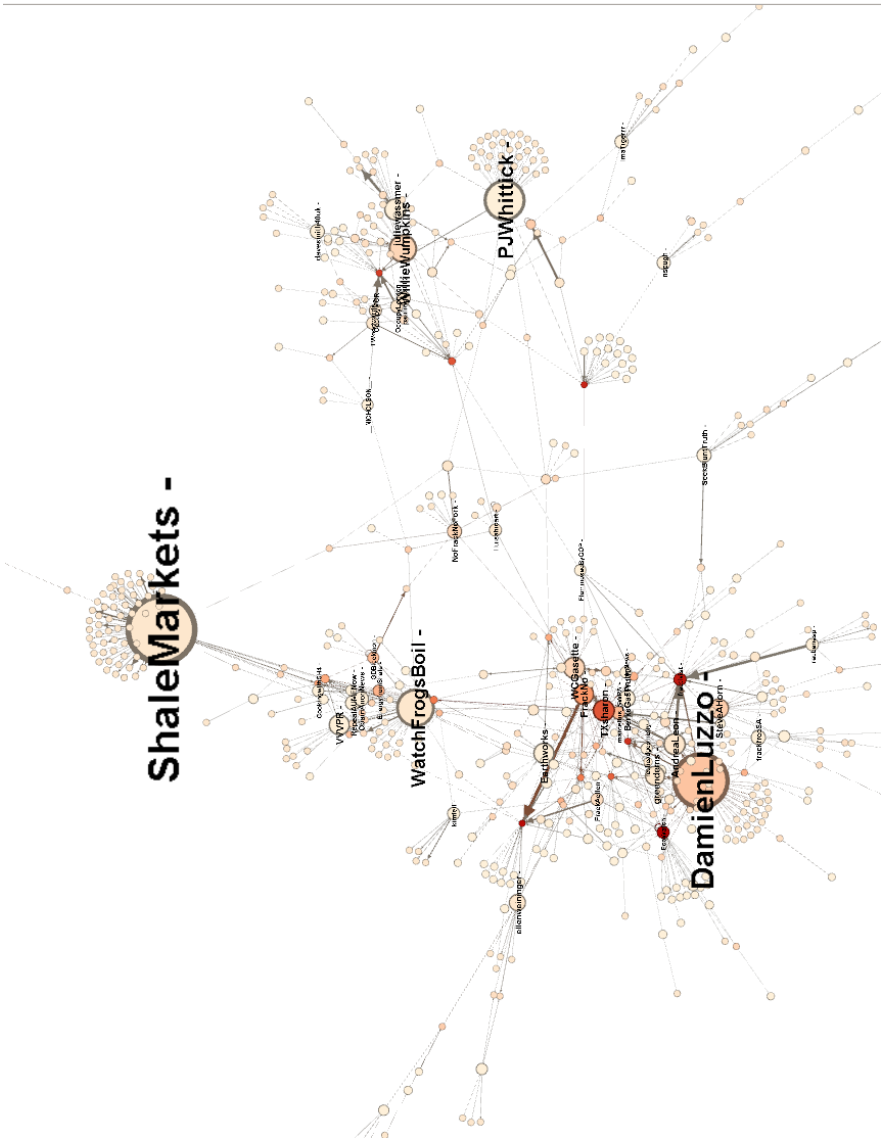

Figure 2: Users by out-degree

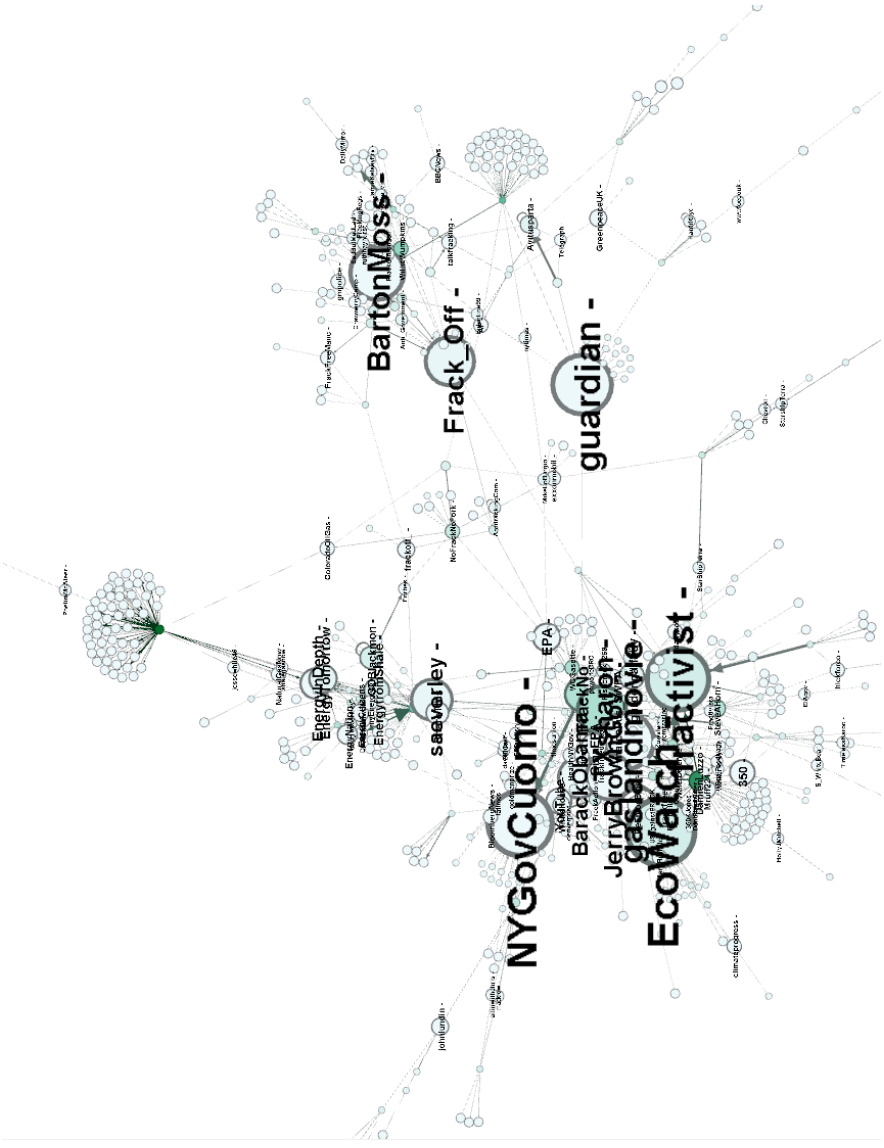

Figure 3: Users by in-degree

In figure 4 we see the hashtags, plotted according to which users employ them (the users are represented by clear circles without names), and weighted by their in-degree, i.e. how many times they appear in tweets. Some hashtags form groups, if they are only used by one or a few users, but the main result is that the hashtag *fracking* dominates the graph so much that not much of the remaining structure is visible. This may well be a reflection of the content collection process and suggests that it could be beneficial to balance the retrieval, getting away from the narrow focus on one specific tag.

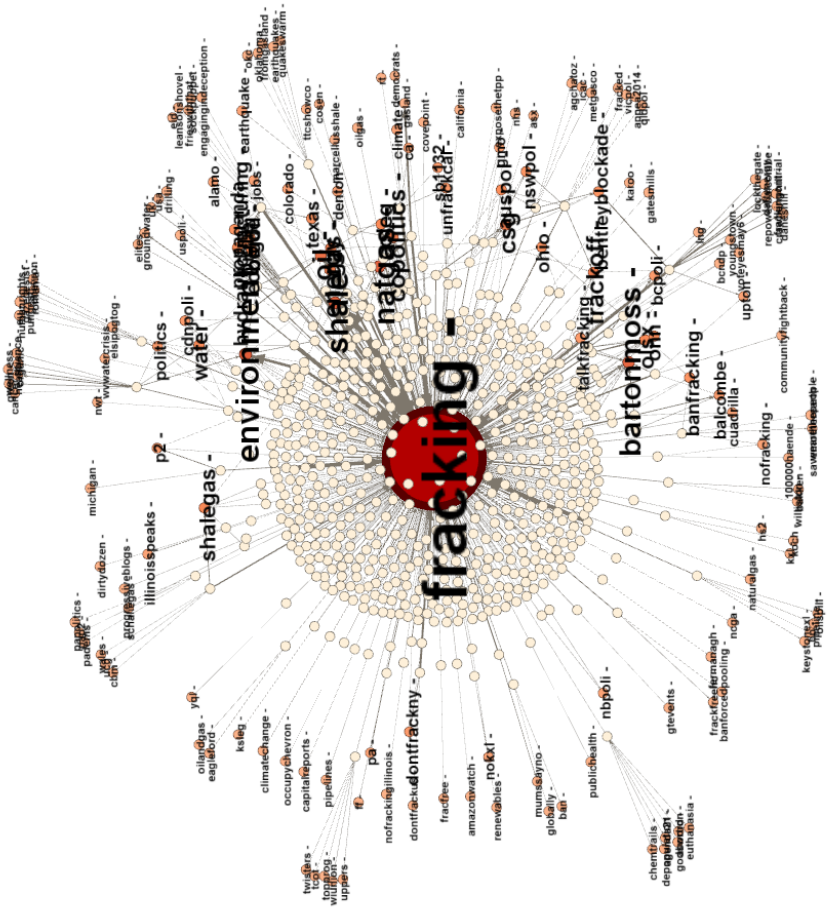

Figure 4: Hashtags by in-degree

5.3 Temporal dynamics

While temporal analysis of social media activity is a major aspect of interest, no analysis of the temporal dynamics (trends, message propagation, activity fluctuations, etc.) has been conducted so far, for a variety of reasons.

Firstly, the data collection, in particular with the recent data loss mentioned above, does not yet provide a continuous collection that would allow us to analyse activity over longer timespans (several months). Secondly, the Solr indexes, which have become available only recently, are fundamental in extracting temporal activity trough *date faceting*[12][13]. And lastly, work on the demonstrators, in close collaboration with end users, will undoubtedly guide the analysis, helping determine what kinds of insights can be useful for the project.

It is therefore expected that temporal aspects will be an important part of the upcoming work on the Social Media analysis task, which will be reflected on D4.5.

[12] https://cwiki.apache.org/confluence/display/solr/Faceting#Faceting-DateFacetingParameters

[13] https://wiki.apache.org/solr/SimpleFacetParameters#Date_Faceting_Parameters

6 Conclusions and Future Work

6.1 Current state and ongoing work

At this point in time, a working crawler for Twitter content is fully integrated in the platform and provides real-time updates that can make twitter data available to the demonstrators within seconds of being posted. We are currently investigating some stability problems that are likely due to intermittent network problems, which integrates with the more general task of improving the overall platform stability, as well as facilitate monitoring and deployment of components.

Youtube videos are also being collected and available through the EUMSSI platform, albeit without real-time updates of the data.

The statistics and graph based analysis are scripted and reproducible, however updates are not yet fully automated. General statistics are published trough *GitHub* and can be easily viewed through *nbviewer*.

6.2 Next steps and future work

Some of the next steps include improving the collection criteria (tags, users, ...), including the use of trend detection to automatically add new tags to track. Other sources will also be added to the system, in particular the comments linked to Youtube videos. Statistics and visualization will be fully automated, ideally making them available to the demonstrators *on-demand*, adapting to filter criteria based on user interaction.

Having all data available through the EUMSSI platform, which now is functional, will greatly facilitate further work on analysing and using the social media content that is being gathered. The MongoDB backend allows for complex and advanced queries of the content, including costly statistical analysis using the aggregation framework, if necessary distributing calculations through map-reduce. The Solr indexes, on the other hand, allow us to efficiently query the content collections in real-time, allowing for interactive analysis of the data, even in direct response to user interactions.

REFERENCES

[Abraham et al., 2009] Abraham, A., Hassanien, A.-E., and Snášel, V. (2009). *Computational Social Network Analysis: Trends, Tools and Research Advances*. Springer Science & Business Media.

[Bastian et al., 2009] Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks.

[DuVander, 2012] DuVander, A. (2012). Which APIs are handling billions of requests per day? [Online; accessed 22-December-2014].

[Hanneman and Riddle, 2005] Hanneman, R. A. and Riddle, M. (2005). Introduction to social network methods, published in digital form at `http://faculty.ucr.edu/~hanneman`. *Riverside CA: University of California*.

[Pinheiro, 2011] Pinheiro, C. A. R. (2011). *Social Network Analysis in Telecommunications*. Wiley.

[Wikipedia, 2014] Wikipedia (2014). Social network analysis — wikipedia, the free encyclopedia. [Online; accessed 22-December-2014].

7 Appendix – Statistics (IPython)

7.1 Setup

```
In [1]: import pymongo
        client=pymongo.MongoClient()
        db=client['eumssi_db']
        col=db['content_items']
```

7.2 Analysis

7.2.1 number of content items

```
In [2]: col.count()

Out[2]: 1217390
```

7.2.2 number of tweets

```
In [3]: col.find({'source':{'$in':['Twitter','Twitter-DW']}}).count()

Out[3]: 457224
```

7.2.3 number of Youtube videos

```
In [4]: col.find({'source':{'$in':['Youtube-video-GeneralChannel',
                                   'Youtube-video-dwEnglishChannel',
                                   'Youtube-video-theguardianChannel'
                ]}}).count()

Out[4]: 7826
```

7.2.4 top tags

```
In [5]: top_tags = col.aggregate([
        {'$match' : {'source' : {'$in':['Twitter','Twitter-DW']}}}, # only count tw
        {'$project' : {'meta.original.entities.hashtags.text':1}}, # only keep hash
        {'$group' :{ '_id' : "$meta.original.entities.hashtags.text",'groupCount' :
        {'$unwind':"$_id"}, # split hashtag groups
        {'$group' :{ '_id' : {'$toLower':"$_id"},'tagCount' : {'$sum':'$groupCount'
        {'$sort':{'tagCount':-1}} # top hashtags first
        ])['result']

In [6]: print '\n'.join(['\t'.join((str(x['tagCount']),x['_id'])) for x in top_tags

246129          fracking
101423          climate
49214           environment
31814           sustainability
```

30710	nuclear
28989	cop20
12971	shale
12410	energy
10230	climatechange
9660	oil
9289	oilandgas
8313	auspol
8306	green
7936	natgas
7007	thorium
6727	cdnpoli
5553	water
5437	iran
5124	gas
4982	health
4271	frackoff
4270	shalegas
4244	ttip
3832	solar
3352	globalwarming
3197	bartonmoss
3167	texas
3156	csr
3120	csg
3009	usa
2978	lima
2959	science
2890	p2
2814	earth
2479	coal
2458	gentechnik
2437	acta
2414	eco
2384	economy
2301	tarsands
2296	nature
2128	pollution
2121	ukraine
2119	uk
2074	us
2031	copolitics
2021	ford
2012	cmax
1953	uranium
1944	tpp

top tags by language

```
In [7]: for lang in ('en','es','de','fr'):
            top_tags = col.aggregate([
        {'$match' : {'source' : {'$in':['Twitter','Twitter-DW']},'meta.source.inLan
        {'$project' : {'meta.original.entities.hashtags.text':1}}, # only keep hash
        {'$group' :{ '_id' : "$meta.original.entities.hashtags.text",'groupCount' :
        {'$unwind':"$_id"}, # split hashtag groups
        {'$group' :{ '_id' : {'$toLower':"$_id"},'tagCount' : {'$sum':'$groupCount'
        {'$sort':{'tagCount':-1}} # top hashtags first
        ])['result']
            print '==   '+lang+'   =='
            print '\n'.join(['\t'.join((str(x['tagCount']),x['_id'])) for x in top_
            print
```

```
==   en   ==
211930        fracking
98805        climate
46940        environment
30734        sustainability
28552        cop20
27705        nuclear
12309        shale
12163        energy
9993        climatechange
9373        oil
8529        oilandgas
7910        green
7861        auspol
7794        natgas
6741        thorium
6607        cdnpoli
5399        water
5336        iran
4871        health
4751        gas
4050        frackoff
3648        shalegas
3614        solar
3140        bartonmoss
3078        globalwarming
3073        csr
3059        csg
3037        texas
2909        lima
2852        p2
2754        science
```

```
2723        earth
2484        usa
2435        coal
2345        economy
2240        tarsands
2099        nature
2062        uk
2057        pollution
2020        ford
2011        cmax
2009        us
2005        copolitics
1932        uranium
1908        eco
1871        planet
1861        irantalks
1856        tpp
1835        renewables
1816        ukraine

==   es   ==
17309        fracking
1610        nuclear
709        medioambiente
648        frackingno
446        ucrania
347        climate
319        cantabria
302        shale
299        shapoporose
280        méxico
269        environment
250        reformaenergética
247        eeuu
241        Últimahoratve
237        burgos
223        shalegas
209        oilandgas
206        renovables
202        españa
201        marcaespaña
199        mexico
197        tamaulipas
185        sustainability
168        coahuila
165        reformaenergetica
```

```
145         agua
137         gas
133         nl
133         cop20
129         unasur
129         integración
129         eeuusanciones
124         nofracking
123         merindades
120         science
119         pemex
118         argentina
114         vacamuerta
113         falso
111         ttip
110         bbc
108         auspol
107         thorium
100         prospecciones
98        frackingez
98        energía
89        pp
89        mitosdelfracking
88        nuevolaredo
87        petróleo

==   de   ==
12572         fracking
2920        ttip
2457        gentechnik
2409        acta
477       wm2014
471       schiefergas
381       erdgas
332       nofracking
286       eu
285       100000haende
259       usa
232       energiewende
231       ewendemo
218       climate
185       umwelt
179       piraten
171       spd
150       gas
150       oilandgas
```

```
149         cdu
147         gasbohren
146         groko
143         ukraine
140         energie
120         nrw
119         shalegas
114         shale
108         gabriel
93         russland
91         deutschland
88         bigoil
87         engagingindeception
85         nato
83         eid
72         energy
72         natgas
70         oil
68         ceta
68         grüne
68         exxon
63         sockpuppet
63         propaganda
62         nokxl
58         kohle
58         nuclear
57         co2
57         environment
56         atom
55         oettinger
53         auspol

==   fr   ==
959         fracking
369         climate
360         gazdeschiste
355         environment
129         nuclear
114         cop20
107         schiste
89         sustainability
70         oilandgas
48         jobs
47         cofrentes17
47         solaridad
42         climat
```

```
40        pétrole
37        gaz
35        australia
34        metals
34        occupychevron
34        sweden
34        shalegas
30        shale
28        pollution
28        chevron
27        toxic
26        svpol
26        ericgarner
26        icantbreathe
26        investors
25        canada
25        agriculture
25        québec
25        gettheffout
24        green
24        holyfieldholywar
23        environnement
22        climatechange
22        tafta
22        polcan
21        polqc
21        change
21        bartonmoss
21        france
21        usa
20        pungesti
20        ukraine
20        europe
19        rechauffementclimatique
19        romania
18        texas
17        total
```

7.2.5 languages

```
In [8]: langs = col.aggregate([
        {'$match' : {'source' : {'$in':['Twitter','Twitter-DW']}}}, # only count tw
        {'$project' : {'meta.source.inLanguage':1}}, # only keep language field
        {'$group' :{ '_id' : "$meta.source.inLanguage",'langCount' : {'$sum':1} } },
        {'$sort':{'langCount':-1}} # top languages first
        ])['result']
```

```
In [9]: print '\n'.join(['\t'.join((str(x['langCount']),str(x['_id']))) for x in la

413189          en
20204          es
12862          de
2563          und
1956          fr
1178          it
885          nl
809          ja
604          pt
361          in
320          sk
296          ro
214          tl
180          pl
167          da
165          ar
165          et
147          sv
121          sl
81          cy
79          no
79          tr
77          ht
63          fi
59          vi
58          bs
42          ru
41          el
33          fa
32          lt
29          hi
26          hu
26          id
19          hr
18          is
15          bg
13          zh
12          lv
10          None
9          uk
8          ta
6          th
4          ko
2          iw
2          bn
```

```
1       ne
1       ur
1       sr
```