Graph Visualization Tool for Twittersphere users based on a high-scalable Extract, Transform and Load System

Pablo Aragón, Íñigo García and Antonio García

May, 27th 2011

CIERZO DEVELOPMENT

smmart
social media marketing
analysis & reporting tool

WIMS'11

INDEX

INTRODUCTION

Cierzo Development and SMMART
Structure of Twitter
Volume of Twitter
Detection of influencers

DISTRIBUTED COMPUTATION

Hadoop
Amazon EC2

PIPELINE DESIGN

Crawling Module
Metadata Extraction Module
Indexing Module
Graph Visualization Module

RESULTS

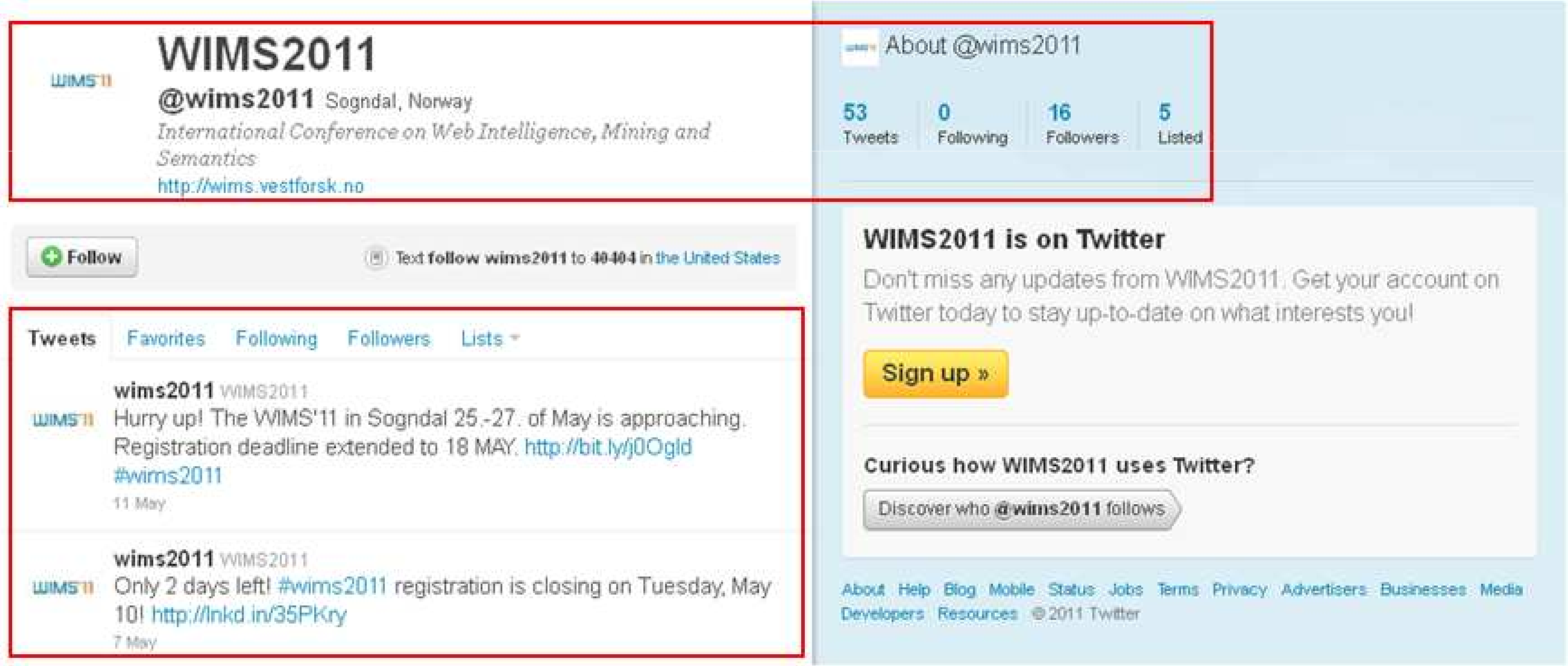
Western Sahara Conflict
Patxi López
Conclusions
Future work

INTRODUCTION: CIERZO DEVELOPMENT AND SMMART

SMMART (Social Media Marketing Analysis and Reporting Tool) is the system developed by Cierzo Development for:

- ✓ Corporate social reputation
- ✓ Measuring effectiveness of marketing campaigns
- ✓ Detection of new trends

INTRODUCTION: STRUCTURE OF TWITTER



WIMS2011

@wims2011 Sogndal, Norway

International Conference on Web Intelligence, Mining and Semantics

http://wims.vestforsk.no

About @wims2011

53 Tweets | 0 Following | 16 Followers | 5 Listed

Follow

Text follow wims2011 to 40404 in the United States

Tweets | Favorites | Following | Followers | Lists

wims2011 WIMS2011
Hurry up! The WIMS'11 in Sogndal 25.-27. of May is approaching. Registration deadline extended to 18 MAY. http://bit.ly/j0OgId #wims2011
11 May

wims2011 WIMS2011
Only 2 days left! #wims2011 registration is closing on Tuesday, May 10! http://lnkd.in/35PKry
7 May

WIMS2011 is on Twitter

Don't miss any updates from WIMS2011. Get your account on Twitter today to stay up-to-date on what interests you!

Sign up »

Curious how WIMS2011 uses Twitter?

Discover who @wims2011 follows

About Help Blog Mobile Status Jobs Terms Privacy Advertisers Businesses Media Developers Resources © 2011 Twitter

Structure of a profile

INTRODUCTION: STRUCTURE OF TWITTER

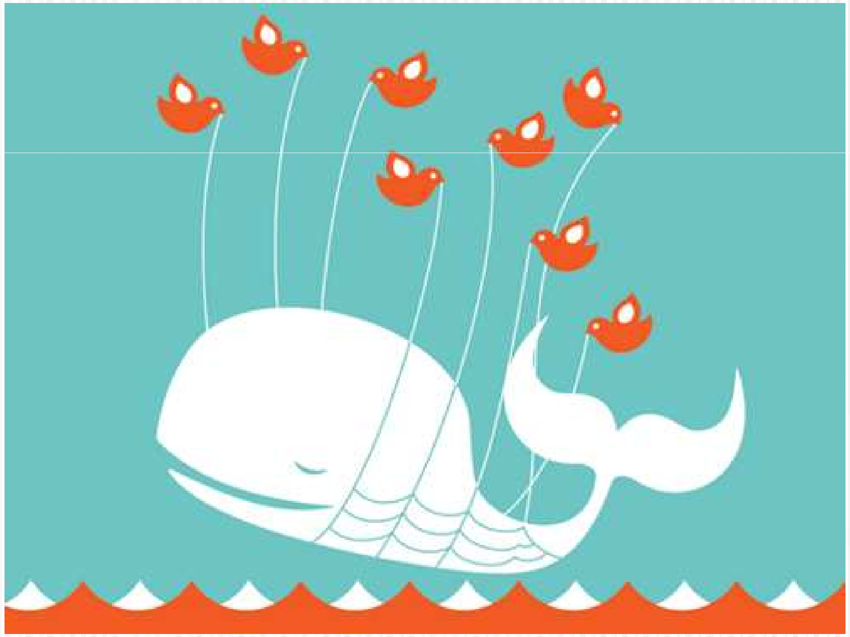
A user can set a relationship with another user by:

❑ Reply: Update that begins with @username

❑ Mention: Update that contains @username in the body of the tweet

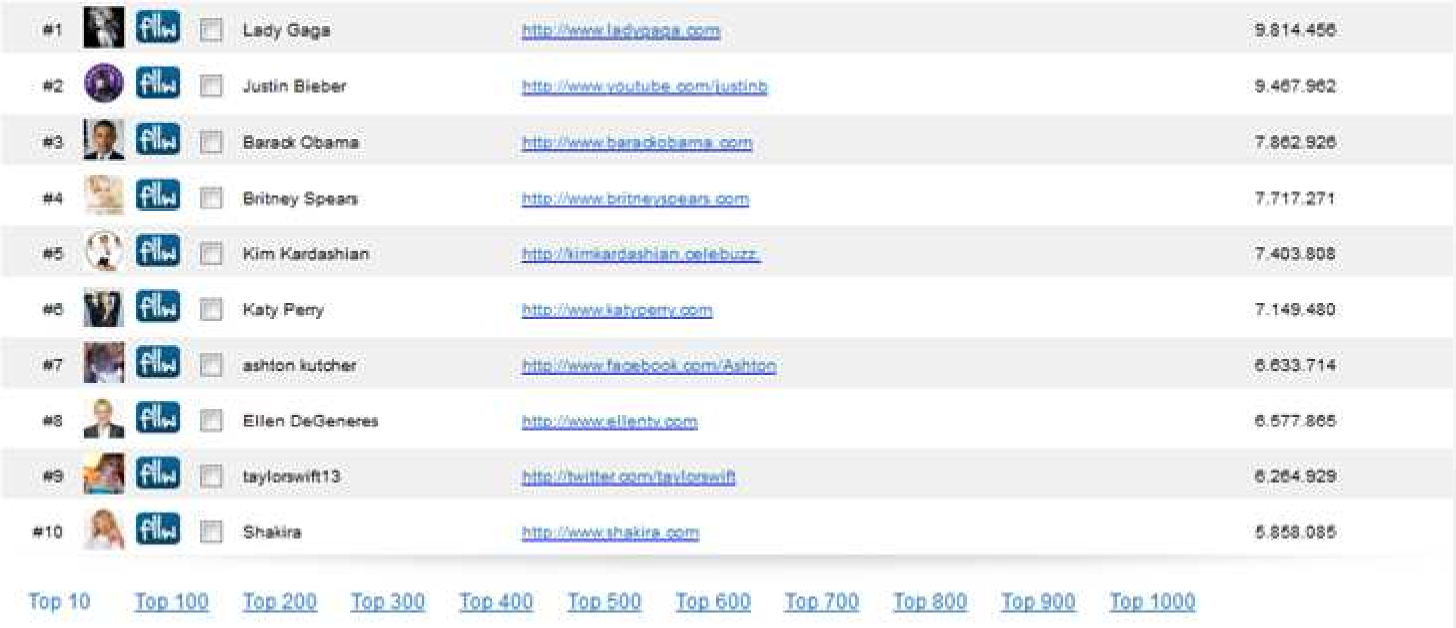
❑ Retweet: Update that contains the body of another user tweet by specifying the original author

INTRODUCTION: VOLUME OF THE TWITTER



More than 200M users publishing millions of tweets per day

INTRODUCTION: DETECTION OF INFLUENCERS



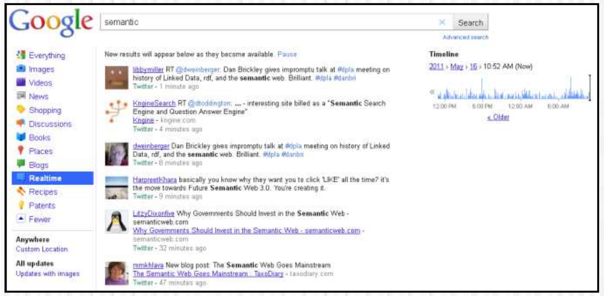
#1	flw	□	Lady Gaga	http://www.ladygaga.com	9.814.456
#2	flw	□	Justin Bieber	http://www.youtube.com/justinb	9.467.962
#3	flw	□	Barack Obama	http://www.barackobama.com	7.862.926
#4	flw	□	Britney Spears	http://www.britneyspears.com	7.717.271
#5	flw	□	Kim Kardashian	http://kimkardashian.celebuzz.	7.403.808
#6	flw	□	Katy Perry	http://www.katyperry.com	7.149.480
#7	flw	□	ashton kutcher	http://www.facebook.com/Ashton	6.633.714
#8	flw	□	Ellen DeGeneres	http://www.ellentv.com	6.577.865
#9	flw	□	taylorswift13	http://twitter.com/taylorswift	6.264.929
#10	flw	□	Shakira	http://www.shakira.com	5.858.085

Top 10 Top 100 Top 200 Top 300 Top 400 Top 500 Top 600 Top 700 Top 800 Top 900 Top 1000

Old metrics based on data as:

- ❑ Absolute info: Number of followers
- ❑ Relative info: Quotient of following users and followers

INTRODUCTION: DETECTION OF INFLUENCERS



New results will appear below as they become available. Pause

libbymiller RT @dweinberger: Dan Brickley gives impromptu talk at #dpla meeting on history of Linked Data, rdf, and the **semantic** web. Brilliant. #dpla #danbri
Twitter - 1 minute ago

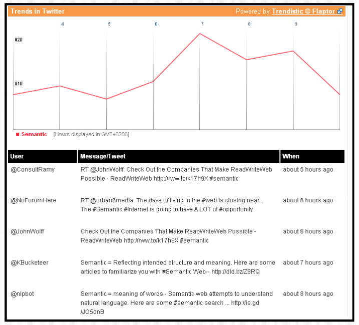
KngineSearch RT @dtoddington: ... - interesting site billed as a "**Semantic** Search Engine and Question Answer Engine"
Kngine - kngine.com
Twitter - 4 minutes ago

dweinberger Dan Brickley gives impromptu talk at #dpla meeting on history of Linked Data, rdf, and the **semantic** web. Brilliant. #dpla #danbri
Twitter - 8 minutes ago

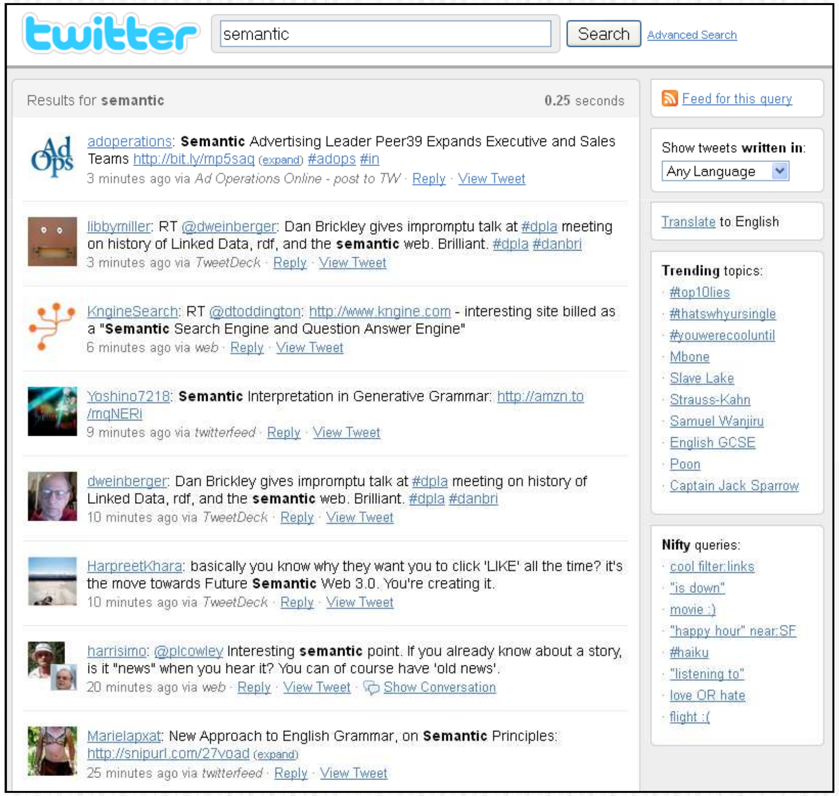
HarpreetKhara basically you know why they want you to click 'LIKE' all the time? it's the move towards Future **Semantic** Web 3.0. You're creating it.
Twitter - 9 minutes ago

LitzyDixonfive Why Governments Should Invest in the **Semantic** Web - semanticweb.com
Why Governments Should Invest in the Semantic Web - semanticweb.com - semanticweb.com
Twitter - 32 minutes ago

mmkhlava New blog post: The **Semantic** Web Goes Mainstream
The Semantic Web Goes Mainstream : TaxoDiary - taxodiary.com
Twitter - 47 minutes ago


Trends in Twitter
Powered by Trendistic © Flaptor
4
5
6
7
8
9
#20
#10
Semantic
[Hours displayed in GMT+0200]

User	Message/Tweet	When
@ConsultRamy	RT @JohnWolff: Check Out the Companies That Make ReadWriteWeb Possible - ReadWriteWeb http://rww.to/k17h9X #semantic	about 5 hours ago
@NoForumHere	RT @urban6media: The days of living in the #web is closing near... The #Semantic #Internet is going to have A LOT of #opportunity	about 6 hours ago
@JohnWolff	Check Out the Companies That Make ReadWriteWeb Possible - ReadWriteWeb http://rww.to/k17h9X #semantic	about 6 hours ago
@KBucketeer	Semantic = Reflecting intended structure and meaning. Here are some articles to familiarize you with #Semantic Web-- http://dld.bz/Z8RQ	about 7 hours ago
@nlpbot	Semantic = meaning of words - Semantic web attempts to understand natural language. Here are some #semantic search ... http://is.gd/JO5onB	about 8 hours ago



Results for **semantic** 0.25 seconds

adoperations: **Semantic** Advertising Leader Peer39 Expands Executive and Sales Teams http://bit.ly/mp5saq (expand) #adops #in
3 minutes ago via *Ad Operations Online - post to TW* · Reply · View Tweet

libbymiller: RT @dweinberger: Dan Brickley gives impromptu talk at #dpla meeting on history of Linked Data, rdf, and the **semantic** web. Brilliant. #dpla #danbri
3 minutes ago via *TweetDeck* · Reply · View Tweet

KngineSearch: RT @dtoddington: http://www.kngine.com - interesting site billed as a "**Semantic** Search Engine and Question Answer Engine"
6 minutes ago via *web* · Reply · View Tweet

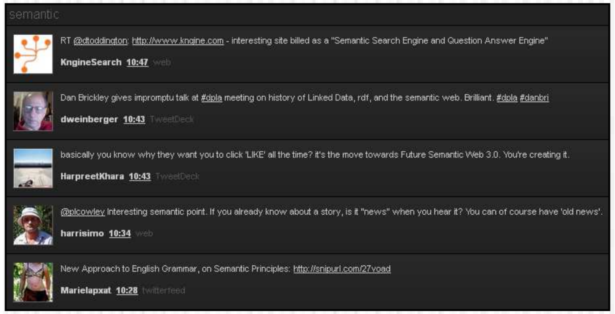
Yoshino7218: **Semantic** Interpretation in Generative Grammar: http://amzn.to/mqNERi
9 minutes ago via *twitterfeed* · Reply · View Tweet

dweinberger: Dan Brickley gives impromptu talk at #dpla meeting on history of Linked Data, rdf, and the **semantic** web. Brilliant. #dpla #danbri
10 minutes ago via *TweetDeck* · Reply · View Tweet

HarpreetKhara: basically you know why they want you to click 'LIKE' all the time? it's the move towards Future **Semantic** Web 3.0. You're creating it.
10 minutes ago via *TweetDeck* · Reply · View Tweet

harrisimo: @plcowley Interesting **semantic** point. If you already know about a story, is it "news" when you hear it? You can of course have 'old news'.
20 minutes ago via *web* · Reply · View Tweet · Show Conversation

Marielapxat: New Approach to English Grammar, on **Semantic** Principles: http://snipurl.com/27voad (expand)
25 minutes ago via *twitterfeed* · Reply · View Tweet



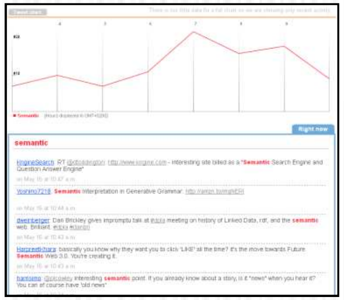
RT @dtoddington: http://www.kngine.com - interesting site billed as a "Semantic Search Engine and Question Answer Engine"
KngineSearch 10:47 web

Dan Brickley gives impromptu talk at #dpla meeting on history of Linked Data, rdf, and the semantic web. Brilliant. #dpla #danbri
dweinberger 10:43 TweetDeck

basically you know why they want you to click 'LIKE' all the time? it's the move towards Future Semantic Web 3.0. You're creating it.
HarpreetKhara 10:43 TweetDeck

@plcowley Interesting semantic point. If you already know about a story, is it "news" when you hear it? You can of course have 'old news'.
harrisimo 10:34 web

New Approach to English Grammar, on Semantic Principles: http://snipurl.com/27voad
Marielapxat 10:28 twitterfeed



KngineSearch: RT @dtoddington: http://www.kngine.com - interesting site billed as a "**Semantic** Search Engine and Question Answer Engine"
on May 16 at 10:47 a.m.

Yoshino7218: **Semantic** Interpretation in Generative Grammar: http://amzn.to/mqNERi
on May 16 at 10:44 a.m.

dweinberger: Dan Brickley gives impromptu talk at #dpla meeting on history of Linked Data, rdf, and the **semantic** web. Brilliant. #dpla #danbri
on May 16 at 10:43 a.m.

HarpreetKhara: basically you know why they want you to click 'LIKE' all the time? it's the move towards Future **Semantic** Web 3.0. You're creating it.
on May 16 at 10:43 a.m.

harrisimo: @plcowley Interesting **semantic** point. If you already know about a story, is it "news" when you hear it? You can of course have 'old news'.

Available search engines track Twitter and list results,

but they do not set a value to the users from the response.

News / World news / Spain

Spanish voters head to the polls, as city square protests continue

Millions turn out to elect municipal councils and regional governments, despite protests over politics-as-usual

Giles Tremlett in Madrid
guardian.co.uk, Sunday 22 May 2011 19.46 BST
Article history

Protesters react a
vote in regional el

#spanishrevolution
#yeswecamp
#15m

Opinion

'Yes we camp' activists hit Spanish streets

When Spanish officials banned a protest camp in central Madrid ahead of elections, thousands defied the ruling.

Leila Nachawati Last Modified: 22 May 2011 10:34

Email Print Share Send Feedback Tweet 258 Like 2K

Listen

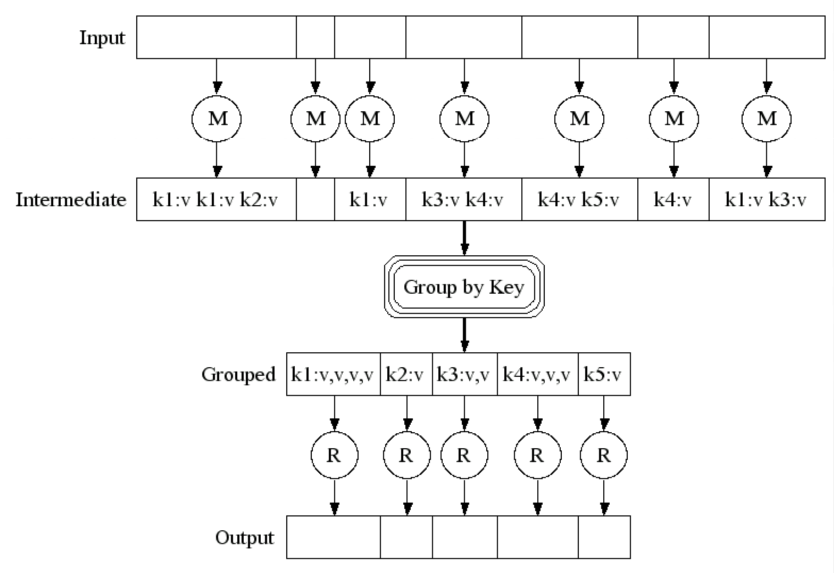
ement that protests against the ongoing financial crisis, politicians and bankers sit on the roof a
bway station as they camp at Madrid's Puerta del Sol May 17, 2011 [Reuters]

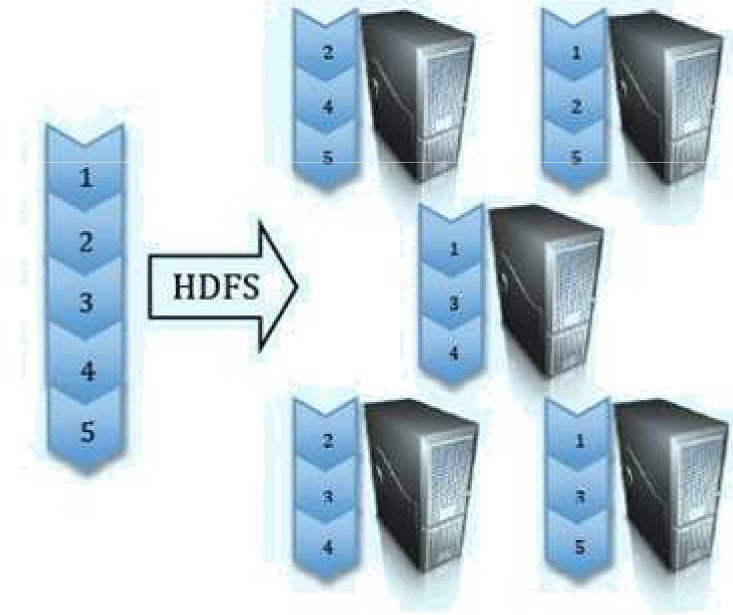
DISTRIBUTED COMPUTATION

❑ Management of large volumes at the lowest cost

❑ Automatic adjustment to the daily growth of users and the oscillations in the frequency of publication

DISTRIBUTED COMPUTATION: HADOOP





Map Reduce

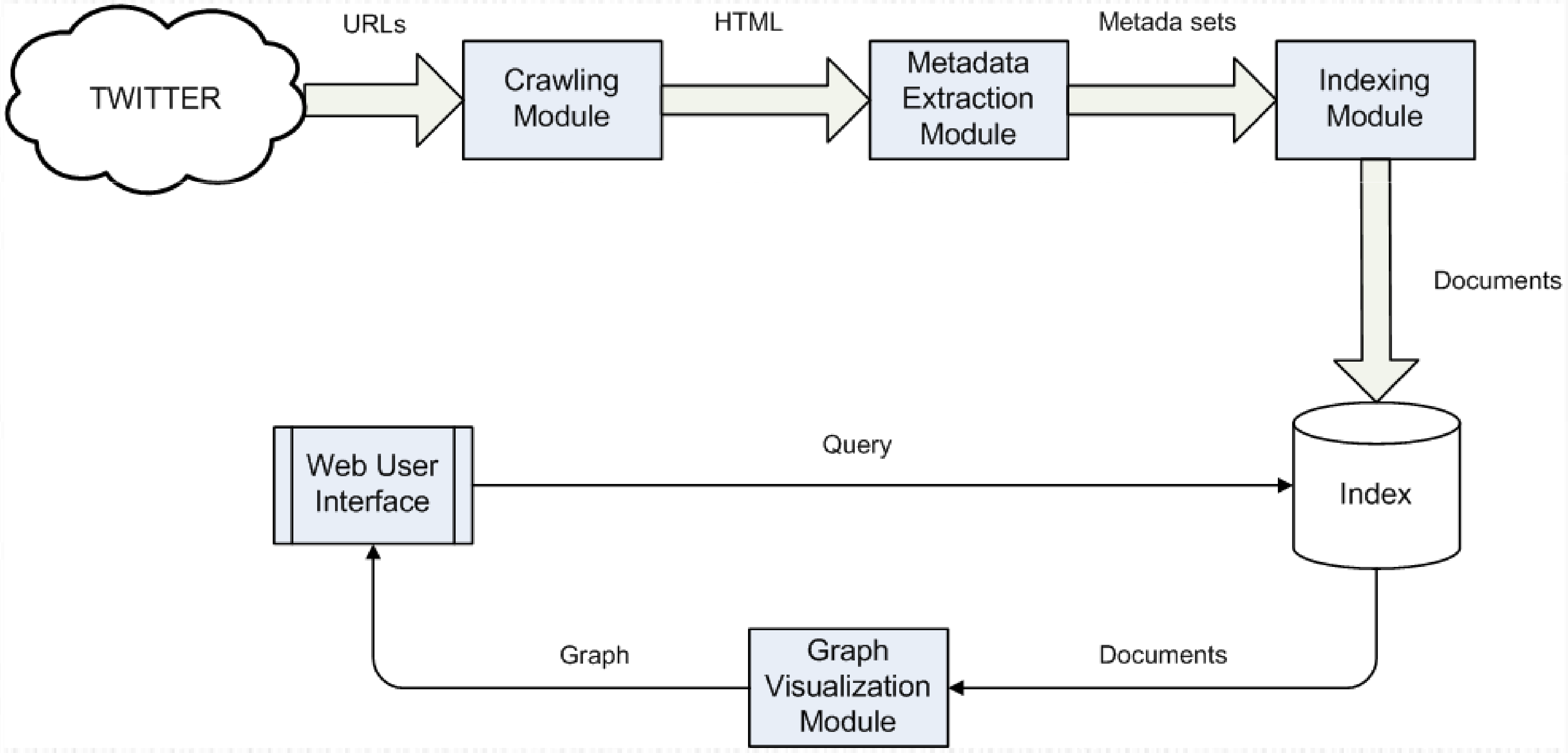
Distributed File System

DISTRIBUTED COMPUTATION: AMAZON EC2

Definition of a Hadoop node as a machine image in Amazon Elastic Compute Cloud.

The system balancing mechanism adds and removes Hadoop nodes in real time on demand.

PIPELINE DESIGN

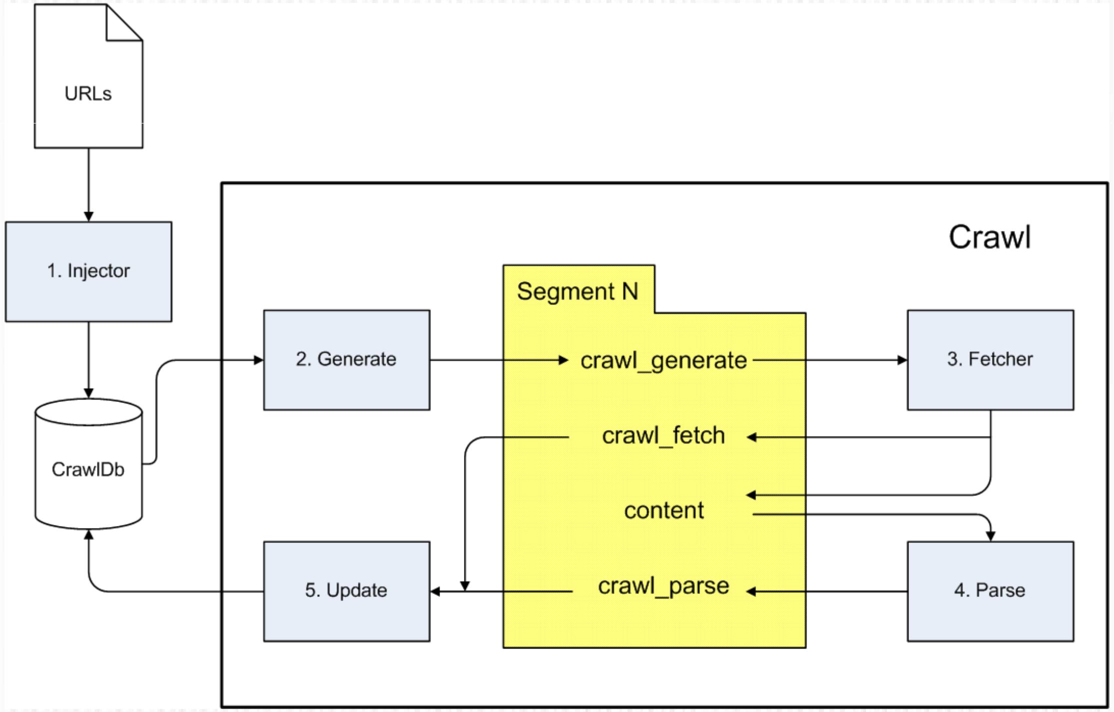


PIPELINE DESIGN: CRAWLING MODULE

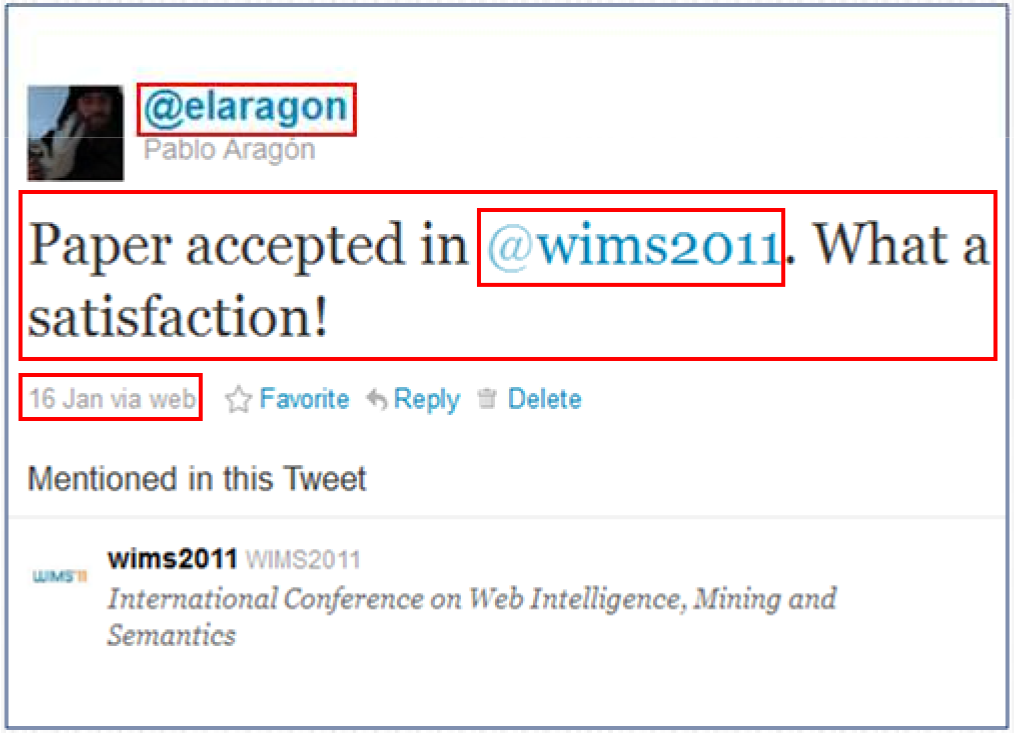
Based on Nutch

1. Crawl the Twitter profiles stored in a DB

2. Extract outlinks to new profiles



PIPELINE DESIGN: METADATA EXTRACTION MODULE



@elaragon
Pablo Aragón

Paper accepted in @wims2011. What a satisfaction!

16 Jan via web

Mentioned in this Tweet

wims2011 WIMS2011
International Conference on Web Intelligence, Mining and Semantics

The portion of HTML of a tweet contains a set of metadata:

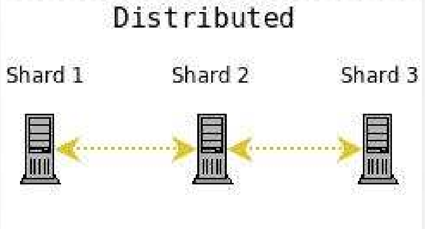
- □ Textual content
- □ Publication date
- □ Author
- □ Mention to other users

PIPELINE DESIGN: INDEXING MODULE

Apache Solr (enterprise search server based on Lucene)

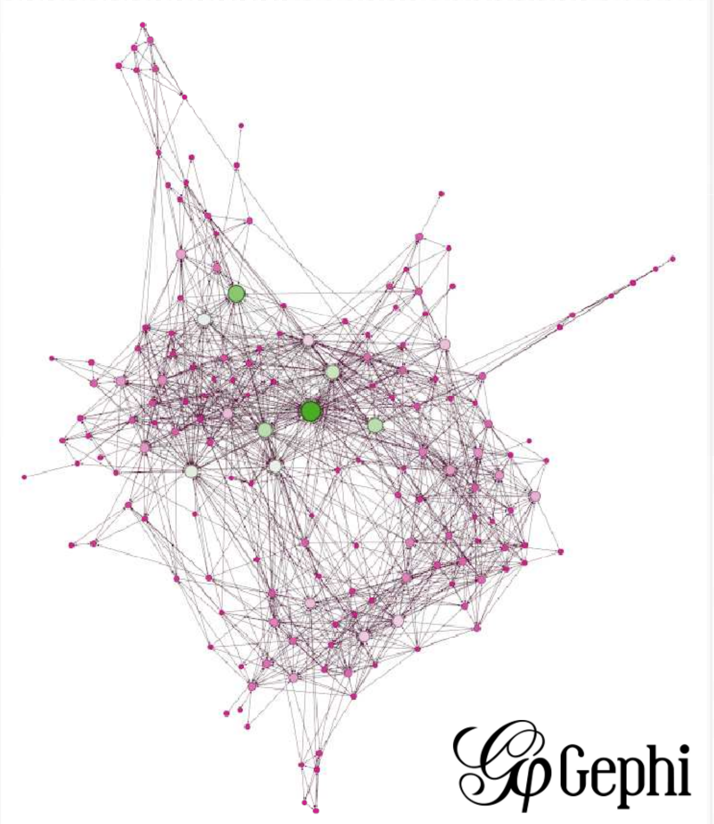
- ✓ Sorting algorithms
- ✓ Stemming
- ✓ Stopwords filters
- ✓ Faceted searchs

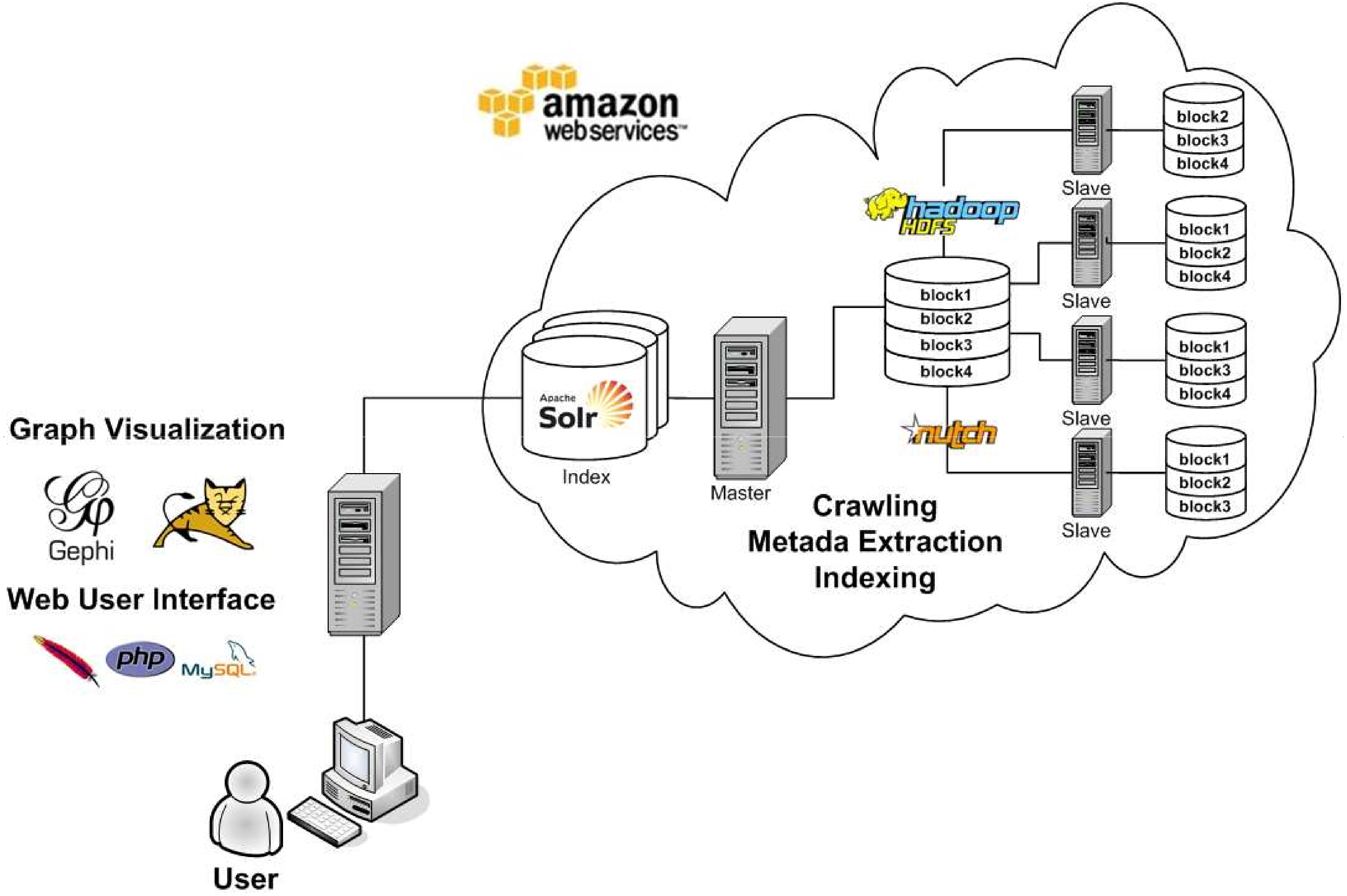
Multicore architecture sharding by publication date.



PIPELINE DESIGN: GRAPH VISUALIZATION MODULE

The Graph Visualization module transforms the responses from the index into a graph by the force-based multilevel layout Yifan Hu's algorithm provided in Gephi Toolkit.



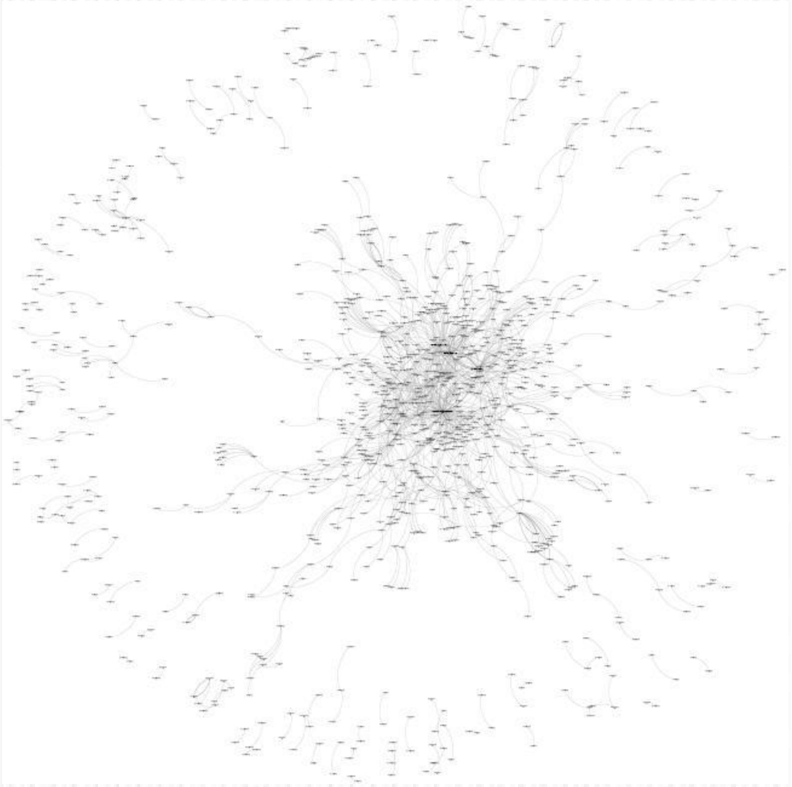

amazon webservices
Slave
block2
block3
block4
hadoop HDFS
Slave
block1
block2
block4
block1
block2
block3
block4
Slave
block1
block3
block4
Apache Solr
Index
Master
nutch
Slave
block1
block2
block3
Graph Visualization
Gephi
Crawling
Metada Extraction
Indexing
Web User Interface
php
MySQL
User

RESULTS: WESTERN SAHARA CONFLICT

www.theodora.com/maps
Canary Islands
(SPAIN)
MOROCCO
North
Atlantic
Ocean
Laayoune
(El Aaiún)
Semara
Bu
Craa
Cabo Bojador
ALGERIA
WESTERN
Guelta
Zemmur
SAHARA
MAURITANIA
Ad Dakhla
Bir Gandús
0 100 200 km
0 100 200 mi
AFRICA

In November 2010, Moroccan security forces involved in a camp in Western Sahara. This action was criticized by part of the Spanish society.

RESULTS: WESTERN SAHARA CONFLICT



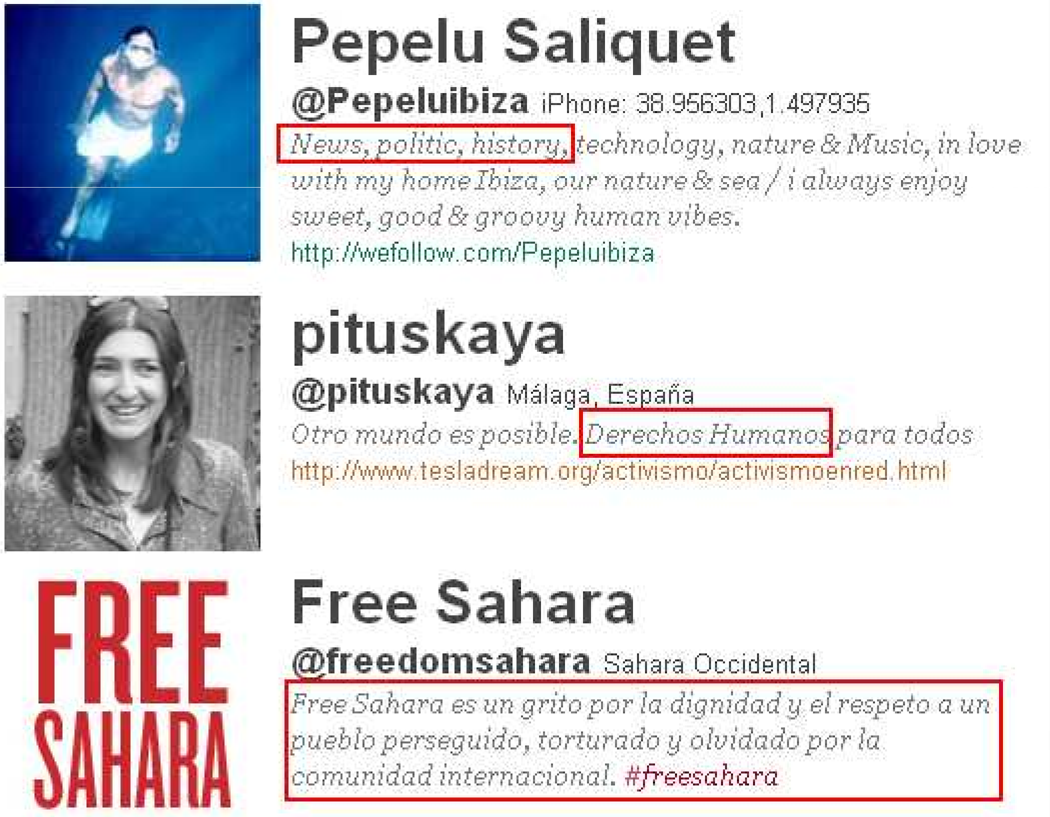
Search

- ❑ content:‘sahara’
- ❑ language:’es’
- ❑ date:[2010-11-10 TO 2010-11-18]

Results

- ✓ 1721 users
- ✓ 3925 tweets
- ✓ 707 mentions

RESULTS: WESTERN SAHARA CONFLICT



Pepelu Saliquet

@Pepeluibiza iPhone: 38.956303,1.497935

News, politic, history, technology, nature & Music, in love with my home Ibiza, our nature & sea / i always enjoy sweet, good & groovy human vibes.

http://wefollow.com/Pepeluibiza

pituskaya

@pituskaya Málaga, España

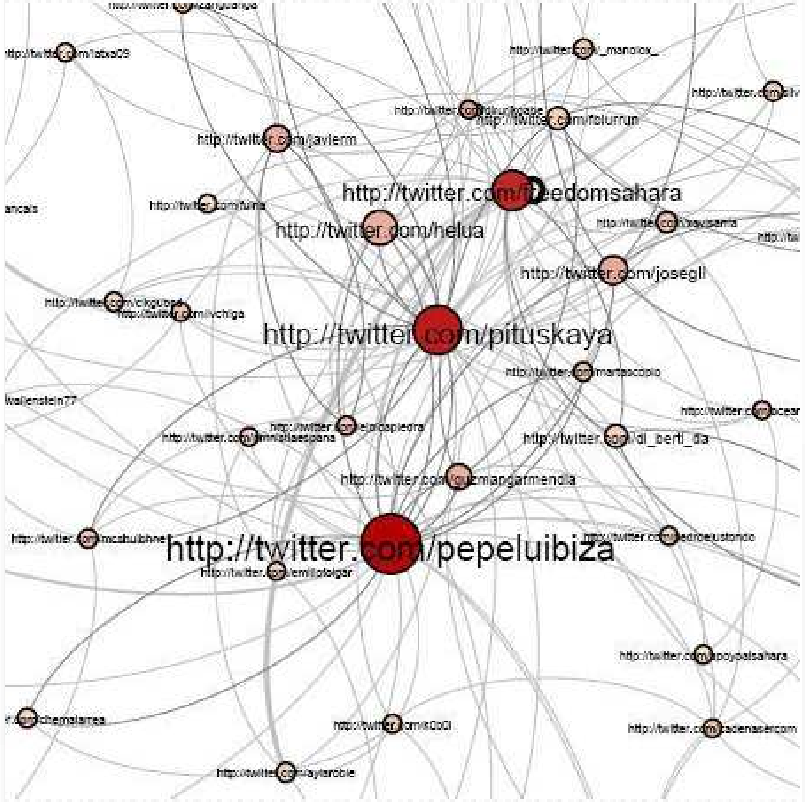
Otro mundo es posible. Derechos Humanos para todos

http://www.tesladream.org/activismo/activismoenred.html

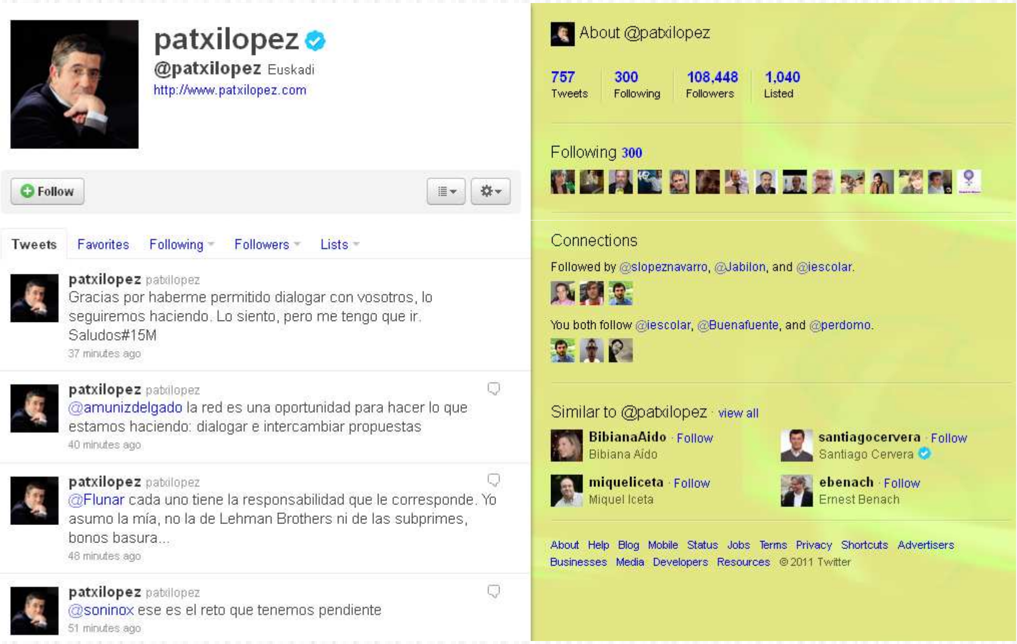
Free Sahara

@freedomsahara Sahara Occidental

Free Sahara es un grito por la dignidad y el respeto a un pueblo perseguido, torturado y olvidado por la comunidad internacional. #freesahara


http://twitter.com/freedomsahara
http://twitter.com/helua
http://twitter.com/josegll
http://twitter.com/pituskaya
http://twitter.com/pepeluibiza
http://twitter.com/javierm
http://twitter.com/fslurrun
http://twitter.com/_manolox_
http://twitter.com/iatxa09
http://twitter.com/fuina
http://twitter.com/cikgubes
http://twitter.com/vchiga
http://twitter.com/martascopio
http://twitter.com/dl_berri_da
http://twitter.com/amnistiaespana
http://twitter.com/elpiedra
http://twitter.com/guemangarmendia
http://twitter.com/mcabulbhne
http://twitter.com/emiliololgar
http://twitter.com/pedroeustondo
http://twitter.com/apoyoalsahara
http://twitter.com/chemalarrea
http://twitter.com/s0b0l
http://twitter.com/cadenasercom
http://twitter.com/aylarobie
wallenstein77
ancais

RESULTS: PATXI LÓPEZ



patxilopez
@patxilopez Euskadi
http://www.patxilopez.com

patxilopez patxilopez
Gracias por haberme permitido dialogar con vosotros, lo seguiremos haciendo. Lo siento, pero me tengo que ir. Saludos#15M
37 minutes ago

patxilopez patxilopez
@amunizdelgado la red es una oportunidad para hacer lo que estamos haciendo: dialogar e intercambiar propuestas
40 minutes ago

patxilopez patxilopez
@Flunar cada uno tiene la responsabilidad que le corresponde. Yo asumo la mía, no la de Lehman Brothers ni de las subprimes, bonos basura...
48 minutes ago

patxilopez patxilopez
@soninox ese es el reto que tenemos pendiente
51 minutes ago

About @patxilopez

757	300	108,448	1,040
Tweets	Following	Followers	Listed

Following 300

Connections

Followed by @slopeznavarro, @Jabilon, and @iescolar.

You both follow @iescolar, @Buenafuente, and @perdomo.

Similar to @patxilopez · view all

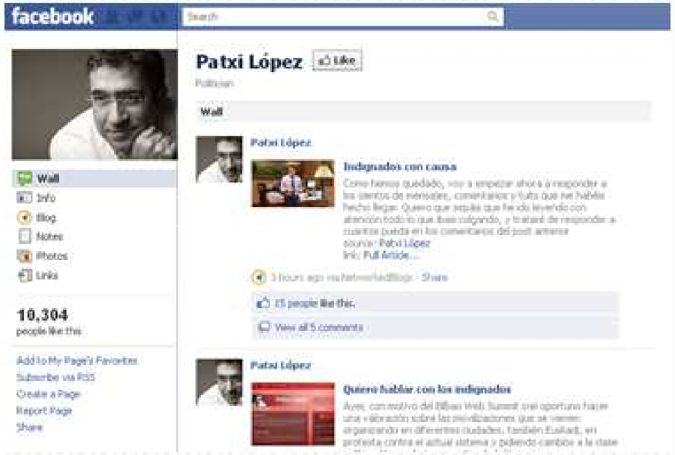
BibianaAido · Follow
Bibiana Aído

santiagocervera · Follow
Santiago Cervera

miqueliceta · Follow
Miquel Iceta

ebenach · Follow
Ernest Benach

About Help Blog Mobile Status Jobs Terms Privacy Shortcuts Advertisers
Businesses Media Developers Resources © 2011 Twitter



Patxi López
Politician

Like
Wall
Wall
Info
Blog
Notes
Photos
Links

10,304
people like this

Add to My Page's Favorites
Subscribe via RSS
Create a Page
Report Page
Share

Patxi López

Indignados con causa
Como hemos quedado, voy a empezar ahora a responder a los cientos de mensajes, comentarios y tuits que me habéis hecho llegar. Quiero que sepáis que he ido leyendo con atención todo lo que ibais colgando, y trataré de responder a cuantos pueda en los comentarios del post anterior
source: Patxi López
link: Full Article...

3 hours ago via NetworkedBlogs · Share

15 people like this.

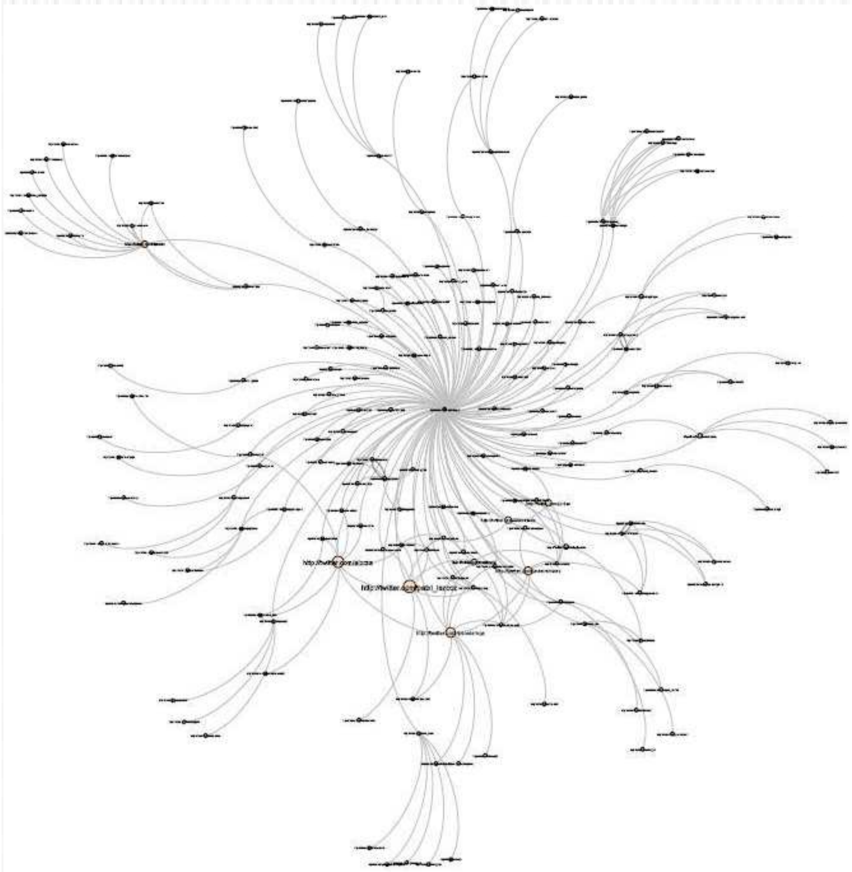
View all 5 comments

Patxi López

Quiero hablar con los indignados
Ayer, con motivo del Bilbao Web Summit creí oportuno hacer una valoración sobre las movilizaciones que se vienen organizando en diferentes ciudades, también Euskadi, en protesta contra el actual sistema y pidiendo cambios a la clase

Patxi López holds the position of the President of the Basque Country Government. His campaign included strategies in social networks.

RESULTS: PATXI LÓPEZ



Search

□ mention:‘patxi_lopez’

□ language:’es’

□ date:[2010-11-10 TO 2010-11-18]

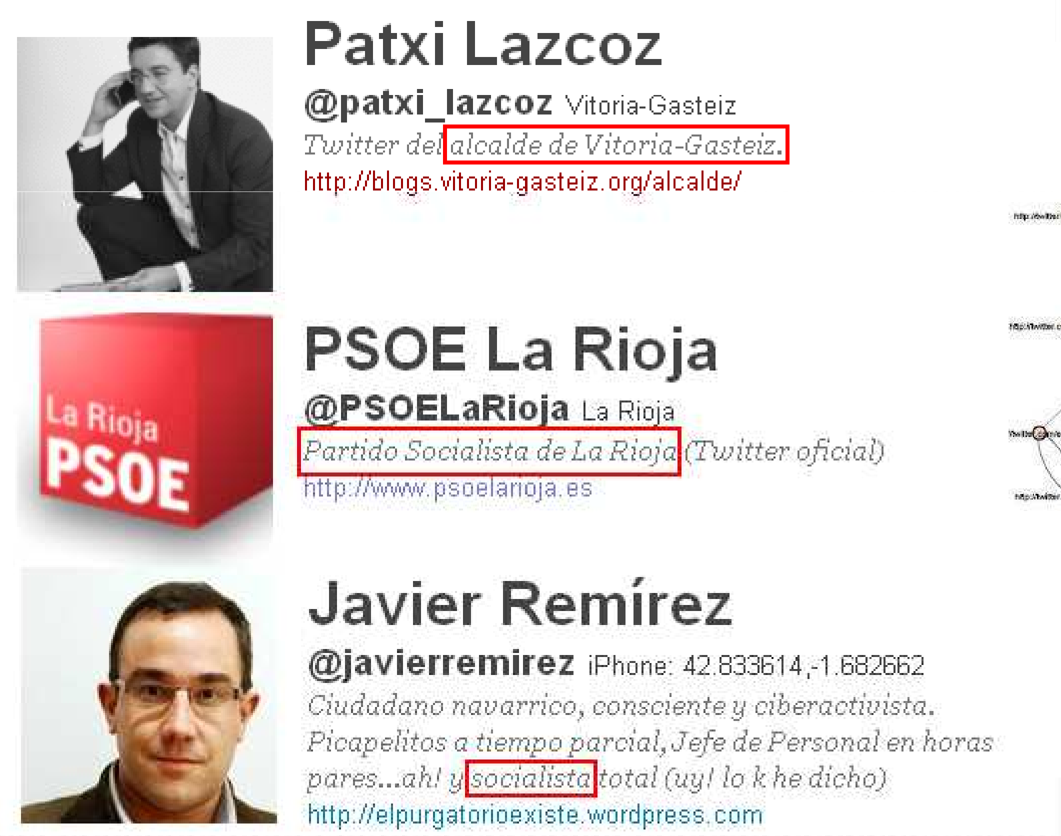
Results

✓ 186 users

✓ 196 tweets

✓ 366 mentions

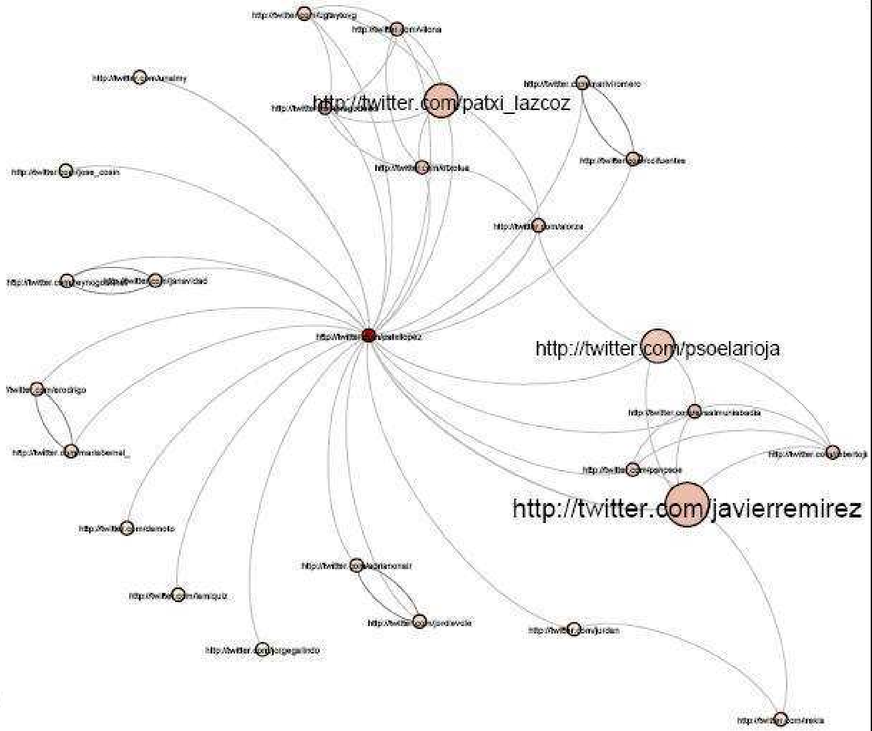
RESULTS: PATXI LÓPEZ



Patxi Lazcoz
@patxi_lazcoz Vitoria-Gasteiz
Twitter del alcalde de Vitoria-Gasteiz.
http://blogs.vitoria-gasteiz.org/alcalde/

PSOE La Rioja
@PSOELaRioja La Rioja
Partido Socialista de La Rioja (Twitter oficial)
http://www.psoelarioja.es

Javier Remírez
@javierremirez iPhone: 42.833614,-1.682662
Ciudadano navarrico, consciente y ciberactivista. Picapelitos a tiempo parcial, Jefe de Personal en horas pares...ah! y socialista total (uy! lo k he dicho)
http://elpurgatorioexiste.wordpress.com


http://twitter.com/patxi_lazcoz
http://twitter.com/psoelarioja
http://twitter.com/javierremirez

RESULTS: CONCLUSIONS

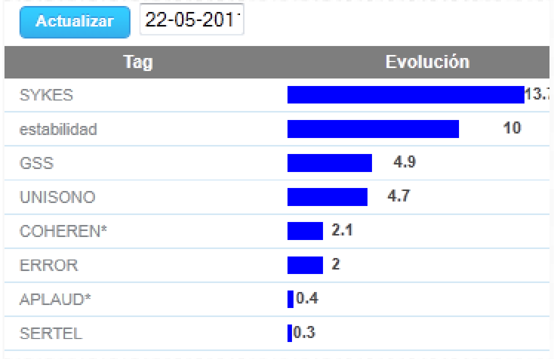
✓ The implemented tool identifies main influencers in a specific topic or around a concrete user

✓ The high-scalable design adapts to a large social network as Twitter

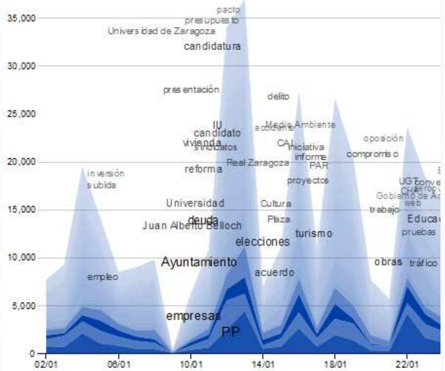
✓ Enterprises can deploy social media monitoring systems using exclusively open source technologies

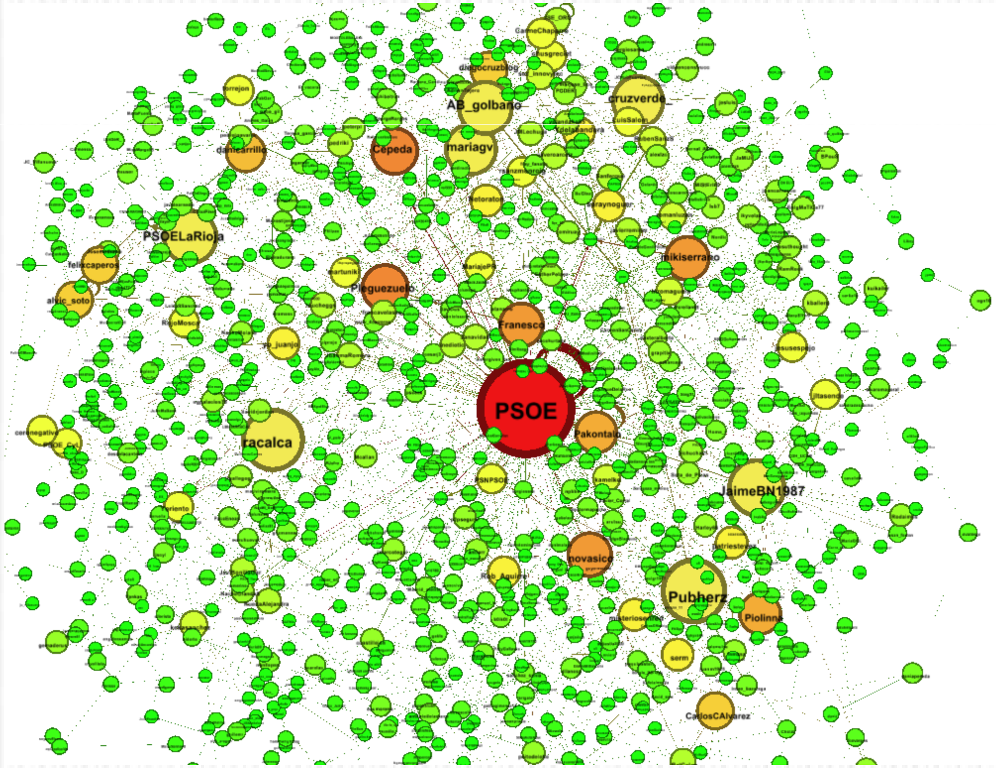
✓ The tool provides information for performing crisis management

RESISTENCIA A LA DICTADURA ECONOMICA CIENTIFICA
ERROR404 DEMOCRACY NOT FOUND
NO ESTAMOS DE BOTELLÓN



Tag	Evolución
SYKES	13.7
estabilidad	10
GSS	4.9
UNISONO	4.7
COHEREN*	2.1
ERROR	2
APLAUD*	0.4
SERTEL	0.3





RESULTS: FUTURE WORK

✓ New versions for more social media sources

✓ Real-time results

✓ New data mining applications

  - ❑ Predictive models

Thanks for your attention