Machine Learning Challenges for Global Media Monitoring

http://eventregistry.org/

Marko Grobelnik
Marko.Grobelnik@ijs.si
Artificial Intelligence Lab, Jozef Stefan Institute
Ljubljana, Slovenia

Contributions from Gregor Leban, Blaz Fortuna, Janez Brank, Jan Rupnik, Andrej Muhic, Aljaz Kosmerlj, Evgenia Belyaeva, Pat Moore

Workshop “Prioritise me! Side-effects of online content delivery”, June 12th 2017, UCL, London
What is Global Media Monitoring? (some general claims)

• Collecting and understanding data from as many as possible media across the world

• Media generally includes:
  • Main stream media – high quality, small size
  • Blogosphere – medium quality, medium size
  • Social-media (micro blogs) – low quality, big size

• Media is optimized mainly for human consumption and only slightly (like RSS standard) for machine consumption

• In general, we would like to see what is happening in society along many dimensions (to prevent filtered bubbles)
Where ML can help in media monitoring?

• ...on all levels – from data cleaning to causality modeling

• Some of the ML challenges in online media processing:
  • Transforming ugly HTML into a clean text document (+meta data)
  • Linguistic and Semantic annotation of textual content
  • Fact & Relation & Infobox extraction
  • Cross-lingual linking and categorization of news stories
  • Identification of events (micro-clusters of reports about the same physical events)
  • Modelling information propagation with barriers (language, culture, time-zone, ...)
  • Modelling news bias
    • dimensions like topic, language, geography, political orientation, source, sentiment, time, attention and some other contextual features
  • Construction of storylines (related event sequences)
  • Learning causal templates from event sequences & event prediction
  • Reasoning – extracting non-explicit facts about news
Outline

• Introduction
• Collecting Media Data
• Document Enrichment
• Cross-linguality
• News Reporting Bias
• Event Representation
• Event Visualization
• Event API
• Future/Ongoing Projects
Introduction
What questions we’ll try to answer?

• Where to get global media data?
• What is extractable from media documents?
• How to connect information across languages?
• What is an event?
• How to approach diversity in news reporting?
• How to visualize global event dynamics?
Systems/Demos used within the presentation

• NewsFeed (http://newsfeed.ijs.si/)
  • News and social media crawler
• Wikifier (http://wikifier.org/)
  • Language and Semantic annotation
• XLing (http://xling.ijs.si/)
  • Cross-lingual document linking and categorization
• Event Registry (http://eventregistry.org/)
  • Event detection and topic tracking
The overall goal

• The goal is to establish a real-time system
  • ...to collect data from global media in real-time
  • ...to identify events and track evolving topics
  • ...to assign stable identifiers to events
  • ...to identify events across languages
  • ...to detect diversity of reporting along several dimensions
  • ...to provide rich exploratory visualizations
  • ...to provide interoperable data export
Global Media Monitoring pipeline

Input data
- Main stream news
- Blogs

Pre-processing steps
- Article semantic annotation
- Extraction of date references
- Cross-lingual article matching
- Detection of article duplicates

Event construction
- Article clustering
- Cross-lingual cluster matching
- Event formation
- Event info. extraction
- Identifying related events

Event storage & maintenance
- Event registry

API Interface
- GUI/Visualizations

http://EventRegistry.org
Collecting Media Data

http://newsfeed.ijs.si/
Where to get references to news publishers?

- Good start is Wikipedia list of newspapers:

Lists of newspapers

From Wikipedia, the free encyclopedia

Below are lists of newspapers organized by continent.

<table>
<thead>
<tr>
<th>Contents [hide]</th>
<th>1 Africa</th>
<th>2 Asia</th>
<th>3 Europe</th>
<th>4 North America</th>
<th>5 Oceania</th>
<th>6 South America</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe [edit]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**List of newspapers in Greece**

1 Athens newspapers
   1.1 Daily political newspapers
   1.2 Daily financial newspapers
   1.3 Daily sports newspapers
   1.4 Weekly newspapers
   1.5 Semimonthly newspapers
   1.6 Monthly newspapers

2 Thessaloniki newspapers
   2.1 Daily political newspapers
   2.2 Daily financial newspapers
   2.3 Daily sports newspapers
   2.4 Weekly newspapers
   2.5 Monthly newspapers

3 Regional/local
   3.1 Main
   3.2 Sports
   3.3 Other

4 Albanian language

5 Macedonian language
From a newspaper home-page to an article

http://www.nytimes.com/

HTML

<link rel="alternate" type="application/rss+xml" title="RSS" href="http://www.nytimes.com/services/xml/rss/nyt/HomePage.xml" />

RSS Feed (list of articles)

Article to be retrieved
Example: news publishers from Rio de Janeiro

General News Media
- Agencia O Globo - agency
- Correio do Brasil - newspaper
- Destak - newspaper
- Extra Rio de Janeiro - newspaper
- Globo News - TV
- Ilha Noticias - newspaper
- Jornal Cidadania - newspaper
- Jornal da Barra - newspaper
- Jornal do Brasil - newspaper
- O Dia Rio de Janeiro - newspaper
- O Globo - newspaper
- Posto Seis - newspaper
- Povo do Rio - newspaper
- Tribuna da Imprensa - newspaper

Business Media
- Jornal do Commercio - newspaper
- Monitor Mercantil - newspaper

Entertainment Media
- Canal Brasil - TV
- Canal USA - TV
- Fox Life South America - TV
- Multishow - TV
- Piaui Magazine - magazine
- Rede Globo - TV
- TV Brasil - TV

Sports Media
- Diario Lance - newspaper
- Jornal dos Sports - newspaper
- Sportv - TV
- TV Esporte Interativo - TV

Society Media
- Folha Universal - newspaper
Collecting global media data

• Data collection service News-Feed
  • http://newsfeed.ijs.si/
  • ...crawling global main-stream and social media

• Monitoring
  • ~140k main-stream publishers (RSS feeds+special feeds)
  • ~250k most influential blogs (RSS feeds)
  • free Twitter feed

• Data volume: ~350k articles & blogs per day (+5M tweets)
• Languages: eng (50%), ger (10%), spa (8%), fra (5%)
Downloading the news stream (1/2)

• The stream is accessible at http://newsfeed.ijs.si/stream/

• To download the whole stream continuously, you can use the python script (http://newsfeed.ijs.si/http2fs.py)

• The script does the following:

```python
timestamp = [when you want to start downloading, e.g. now() minus 1 hour]
while True:
    fetch http://newsfeed.ijs.si/stream/?after=timestamp
    if [404 error]:
        # there is no new data
        pause 1 minute
    else:
        save data
        timestamp = [extract it from the Content-Disposition HTTP header]
```
Downloading the news stream (2/2)

• News Stream Contents and Format
  • The root element, `<article-set>`, contains zero or more articles in the following XML format:

• ...more details:
Document Enrichment

http://enrycher.ijs.si/
What can be extracted from a document?

- **Lexical level**
  - **Tokenization** – extracting tokens from a document (words, separators, ...)
  - **Sentence splitting** – set of sentences to be further processed

- **Linguistic level**
  - **Part-of-Speech** – assigning word types (nouns, verbs, adjectives, ...)
  - **Deep Parsing** – constructing parse trees from sentences
  - **Triple extraction** – subject-predicate-object triple extraction
  - **Name entity extraction** – identifying names of people, places, organizations

- **Semantic level**
  - **Co-reference resolution** – replacing pronouns with corresponding names; merging different surface forms of names into single entity
  - **Semantic labeling** – assigning semantic identifiers to names (e.g. LOD/DBpedia/Freebase) including disambiguation
  - **Topic classification** – assigning topic categories to a document (e.g. DMoz)
  - **Summarization** – assigning importance to parts of a document
  - **Fact extraction** – extracting relevant facts from a document
Enrycher (http://enrycher.ijs.si/)

Plain text

Enrycher is available as a web-service generating Semantic Graph, LOD links, Entities, Keywords, Categories, Text Summarization, Sentiment

Diego Maradona Semantics:
owl:sameAs: http://dbpedia.org/resource/Diego_Maradona
owl:sameAs: http://sw.opencyc.org/concept/Mx4rvofERZwpEbGdrcN5Y29ycA
rdf:type: http://dbpedia.org/class/yago/ArgentinaInternationalFootballers
rdf:type: http://dbpedia.org/class/yago/ArgentineExpatriatesInItaly
rdf:type: http://dbpedia.org/class/yago/ArgentineFootballManagers
rdf:type: http://dbpedia.org/class/yago/ArgentineFootballers

Robbie Keane Semantics:
owl:sameAs: http://dbpedia.org/resource/Robbie_Keane
rdf:type: http://dbpedia.org/class/yago/CoventryCityF.C.Players
rdf:type: http://dbpedia.org/class/yago/ExpatriateFootballPlayersInItaly
rdf:type: http://dbpedia.org/class/yago/F.C.InternazionaleMilanoPlayers

Enrycher is a web service consisting of a set of interlinked modules:

- covering lexical, linguistic and semantic annotations
- exporting data in XML or RDF
- To execute the service, one should send an HTTP POST request, with the raw text in the body:
  - `curl -d "Enrycher was developed at JSI, a research institute in Ljubljana. Ljubljana is the capital of Slovenia." http://enrycher.ijs.si/run`
Anaphora resolution (1)

• We want to link pronouns with their references
• We assume that pronouns refer only to named entities
• A difficult problem
  • Examples of difficult sentences:
    • Tom wrote a letter to Bill. He told him...
    • One passenger in King's car said they had been drinking liquor.
    • ...

  He refers to different person than him

• We link only 5 different pronouns: he (his, him, himself), she (her,...), I (me,...), they (them,...) and who
• Resolution procedure:
  • For each pronoun search backward (and forward) in text to find candidate name entities of correct type
  • Score each candidate name entity
    • Score is based on distance from pronoun, part of speech, other parser information (proper name, name of the county), ...
  • Pick a named entity with the best score
Anaphora resolution (2)

• Common mistakes:
  • Quoted speech: John said: “He is sick.”  Error: he == John
  • But: “I hope so," he replies after a pause.  he == I
  • Tom wrote a letter to Bill. He told him … Error: him == Tom
  • The relationship between active volcanoes and the communities that surround them is not always confrontational.  We don’t link them
  • Jordan's King Hussein and Yasser Arafat's open sympathy for Iraq has strained their relations with the U.S.  Can’t link their

• The most fatal case is when we wrongly resolve first occurrence of a pronoun and then follow many sentences using only the pronoun to refer to a person
Anaphora resolution evaluation

• We manually labeled 91 articles
  • Containing 1506 pronouns
  • 1024 (68%) pronouns are *he, she, I, they, who*
    • We try to link all of them
  • Other 482 (32%) pronouns are: *it, you, we, what, ...*
Anaphora resolution evaluation

<table>
<thead>
<tr>
<th>Pronoun</th>
<th>Frequency</th>
<th>Frequency [%]</th>
<th>Accuracy [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>He</td>
<td>681</td>
<td>45.22</td>
<td>86.9</td>
</tr>
<tr>
<td>They</td>
<td>244</td>
<td>16.20</td>
<td>67.2</td>
</tr>
<tr>
<td>It</td>
<td>204</td>
<td>13.55</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>64</td>
<td>4.25</td>
<td>82.8</td>
</tr>
<tr>
<td>You</td>
<td>50</td>
<td>3.32</td>
<td></td>
</tr>
<tr>
<td>We</td>
<td>44</td>
<td>2.92</td>
<td></td>
</tr>
<tr>
<td>That</td>
<td>44</td>
<td>2.92</td>
<td></td>
</tr>
<tr>
<td>What</td>
<td>27</td>
<td>1.79</td>
<td></td>
</tr>
<tr>
<td>She</td>
<td>24</td>
<td>1.59</td>
<td>62.5</td>
</tr>
<tr>
<td>This</td>
<td>22</td>
<td>1.46</td>
<td></td>
</tr>
<tr>
<td>Who</td>
<td>11</td>
<td>0.73</td>
<td>63.6</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1506</td>
<td>100</td>
<td>81.2</td>
</tr>
</tbody>
</table>

Accuracy on 5 selected **81.2%** (55.2% if counting all pronouns)
Cross-linguality

http://xling.ijs.si/
Cross-linguality
How to operate in many languages?

• Cross-linguality is a set of functions on how to transfer information across the languages
  • ...having this, we can track information independent of the language borders
  • Machine Translation is expensive and slow, so the goal is to avoid machine translation to gain speed and scale

• The key building block is the function for comparing and categorization of documents in different languages
  • http://XLing.ijs.si is an open web service to bridge information across 100 languages
Languages covered by XLing (top 100 Wikipedia languages)

Xling (http://XLing.ijs.si) service for comparing and categorization of documents across 100 languages.

Automatically Extracted Keywords

Similarity Between Two Documents

Chinese Text

Selection Of 100 Languages

English Text

 Automatically Extracted Keywords

Report

0.669834

Words that add the most to the similarity
China economy economic growth bank market Chinese rate that global

Summary of categories:
- business
- news
- services
- finance
- society
- events

World Bank cuts China and Thailand's growth forecasts

The World Bank has kicked off the week by cutting its growth forecasts for the Asian economy, including a sharp downgrade to Thailand following months of political unrest.

It warned that "there was a bumpy start to 2014, notably in China and the United States", as it lowered its forecast for growth across the East Asia and Pacific region.

It now expects GDP to rise by 7.1% in 2014 and 2015 across the area, down from the 7.2% previously forecast for both years.

And it admitted that this could be too optimistic, given the wider
Example: Cross-lingual News Recommendation

• What local media (e.g. German) is writing about the topic we are reading in English?

• Usual fear of publishers: are users we are sending away coming back?
  • ...evaluation shows they all come back
Example: Social Media Recommendation
(http://aidemo.ijs.si/xlike/hsdemo/)

• Usual problem publishers have what and when to publish to which social media channel?
  • ...the demo shows what Bloomberg articles should be the most relevant based on what is most ‘popular’ in a particular geographical area
News Reporting Bias

http://aidemo.ijs.si/diversinews/
UK soldiers cleared in Iraqi death – Seven British soldiers were acquitted on Thursday of charges of beating an innocent Iraqi teenager to death with rifle butts. A judge at a specially convened military court in eastern England ordered the adjudicating panel to return ’not guilty’ verdicts against the seven because he did not believe there was sufficient evidence against them, the Ministry of Defence said. . . .

British murderers in Iraq acquitted – The judge at a court-martial on Thursday dismissed murder charges against seven soldiers, from the 3rd Battalion, the Parachute Regiment, who’re accused of murdering Iraqi teenager; claiming there’s insufficient evidence to secure a conviction, The Associated Press reported Thursday. . . .
Detecting News Reporting Bias

• The task:
  • Given a news story, are we able to say from which news source it came?

• We compared CNN and Aljazeera reports about the same events from the war in Iraq
  • ...300 aligned articles describing the same story from both sources

• The same topics are expressed in both sources with the following keywords:
  • CNN with:
    • Insurgents, Troops, Baghdad, Iran, Militant, Police, Suicide, Terrorist, United, National, Hussein, Alleged, Israeli, Syria, Terrorism...
  • Aljazeera with:
    • Attacks, Claims, Rebels, Withdrawing, Report, Fighters, President, Resistance, Occupation, Injured, Army, Demanded, Hit, Muslim, ...
DiversiNews iPad App (1/2)

- DiversiNews iPad App is using newsfeed.ijs.si and enrycher.ijs.si services
- ...in its initial screen is shows list of current hot topics and current trending events
DiversiNews iPad App (2/2)

- DiversiNews “diversity search” screen allows dynamic reranking of articles describing an event along three dimensions:
  - **Geography** – where is a content being published from
  - **Subtopics** – what are subtopics of an event
  - **Sentiment** – what are good and what are bad news
- For each query it provides
  - Automatically generated **summary**
  - List of corresponding **articles**
New Bias Experiment

• 30 news publishers selected
• All articles from Dec 15th 2014 – Aug 15 2014
• Bussiness related: Business Insider, Boston Business Journal, Economic Times, Financial Times, WSJ Blogs
• Tech related: Gizmodo, GigaOM, The Next Web
• Foreign: ABC.es, El Mundo, Die Welt, Stern Magazine
Dimensions of news bias to be observed

- Analysis of article length differences
- Analysis of grammatical differences
- Readability differences
- Newswire citations
- Geographical bias
- Topic (category) bias
- Speed of reporting
- Predicting article news source
- Similarity between news publishers
Article length differences
Number of tokens in articles per topic
Number of tokens in article headlines
Grammatical differences (% of adjectives in articles)
Grammatical differences (% of adverbs in articles)
Readability differences

• Dale-Chall readability formula.

\[
0,1579 \left( \frac{\text{difficult words}}{\text{words}} \times 100 \right) + 0,0496 \left( \frac{\text{words}}{\text{sentences}} \right)
\]

• Flesch-Kincaid Reading Ease

  • This measure is a standard way for many US agencies for evaluating technical documents.

\[
206,835 - 1,015 \left( \frac{\text{total words}}{\text{total sentences}} \right) - 84,6 \left( \frac{\text{total syllables}}{\text{total words}} \right)
\]
Readability in articles (Flesch-Kincaid test)
Readability in articles
(Flesch-Kincaid test on Washington Post topics)
Publishers citing news agencies

- USA Today
- The Moscow Times
- FOX News
- Washington Post
- ABC News
- Stern Magazine
- Huffington Post
- The Next Web
- Sydney Morning Herald
- Time
- Die Welt
- Wall Street Journal
- Globe and Mail
- The Hindu
- Economic Times
- Daily Telegraph
- Business Insider
- WSJ Blogs
- DNA India
- CNN Europe
- The Independent
- The Guardian
- BBC
- GigaOM
- Boston Business Journal
- ABC.es
- Daily Mail
- Financial Times
- Gizmodo
- El Mundo

Percentage of articles that cite at least one news agency [%]
Agencies

Citations

- Associated Press
- Reuters
- Agence France-Presse
- Deutsche Presse-Agentur
- Interfax
- Xinhua News Agency
- Press Association
- RIA Novosti
- American Press Agency
- Syrian Arab News Agency
- IRNA
- Dow Jones Newswires
- MENA
- Agenzia Nazional Stampa Associata
- ITAR-TASS
- China News Service
- Algemeen Nederlands Persbureau
- United Press International
- ISNA
- Slovenian Press Agency
- All Headline News
- Prensa Latina
- FARS
- Agencia Brasil
- Mehr News Agency
- Khaama Press
- Qatar News Agency
- BNO News
- Ukrainian News Agency
- Algeria Press Service
Geographical bias

• What events are being reported by news publishers?
USA Today geographical coverage
ABC.es geographical coverage
“Die Welt” geographical coverage
Separability between news sources

• How well we can determine which news publisher is the author of the article
• Tested on pairs of news publishers
• Used 10,000 articles from each news publisher
• Removed a custom-built set of tokens (like huffpost, mailonline, ...)
• Used keywords as features, used 1,000 top ranking features
• Linear SVM
• Used 70% articles for training, 30% for testing
• Use accuracy as a distance measure $\rightarrow$ use MDS for plotting
<table>
<thead>
<tr>
<th>News source</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>selfie, uber, rapper, minister, researchers, Hillary, sex, party, relationships</td>
</tr>
<tr>
<td>Washington post</td>
<td>Maryland, redskins, Georgetown, Montgomery, reelection, capital, Arlington, officials, capitol</td>
</tr>
<tr>
<td>FOX News</td>
<td>Jerusalem, Christians, minister, dynasty, officials, prosecutors</td>
</tr>
<tr>
<td>Daily mail</td>
<td>Mr, uk, centre, behavior, football, Scotland, British, premier, Cameron</td>
</tr>
<tr>
<td>BBC</td>
<td>Correspondents, Belfast, council, Scotland, UK, Wales, Ireland</td>
</tr>
<tr>
<td>CNN</td>
<td>Editor, commentary, opinions, affiliate, according, London, official, Pakistan</td>
</tr>
<tr>
<td>Sydney Morning Herald</td>
<td>Australian, Australia, Melbourne, Fairfax, Canberra, rugby, Queensland, Brisbane,</td>
</tr>
<tr>
<td>Moscow times</td>
<td>Russian, Russia, Interfax, soviet, Barack, Petersburg, Siberia, Sergei, Kremlin, Duma, Yury, Oleg</td>
</tr>
<tr>
<td>DNA India</td>
<td>Mumbai, indian, singh, Bollywood, Kapoor, cricket, crore, Bangalore, lakh, Pune, Kumar, Tata, Gandhi</td>
</tr>
</tbody>
</table>
Separability on reports about Ukraine

• Analyze articles reporting about Russia-Ukraine incident
• Selected articles
  • Mention Ukraine and Russia
  • between Jan 15th, 2014 and June 1st, 2014
  • Categorized into appropriate Dmoz category
• Selected publishers that reported about 500 articles
<table>
<thead>
<tr>
<th>News source</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Russian sources</td>
<td>troops, allies, invasion, aggression, excellency, insurgents, investigation, defense, ethnic, seized</td>
</tr>
<tr>
<td>The Moscow Times</td>
<td>overthrow, authorities, western, election, khodorkovsky, detained, annexation, pressure, seizure, fanning, imposed, domestic</td>
</tr>
</tbody>
</table>
Similarity between news publishers regarding the events they cover

• For each article we know the corresponding event in the Event Registry
• Compute a list of events covered by each news publisher
• Compute the pairwise similarity between the lists
• We used cosine similarity as the similarity measure
Event Representation

http://eventregistry.org/
What is an event?
(abstract description)

• ...more practical question: what definition of is computationally feasible?

• In general, an event is something which “sticks out” from the rest of information in some kind of (high dimensional) data space
  • ...could be interpreted as an “anomaly”
  • ...densification of data points (e.g. many similar documents)
  • ...significant change of distribution (e.g. a trend on Twitter)

• In practice, the event could be:
  • A cluster of documents / change of a distribution in data
    • Detected in an unsupervised way
  • A fit to a pre-built model
    • Detected in a supervised way
How to represent an event?

• Baseline data for a news event is usually a cluster of documents
  • ...with some preprocessing we extract linguistic and semantic annotations
  • ...semantic annotations are linked to ontologies providing possibility for multiresolution annotations

• Three levels of event representation:
  • Feature vector event representation:
    • ...light weight representation that can be easily represented as a set of feature vectors augmented with external ontologies – suitable for scalable ML analysis
  • Structured event representation:
    • Infobox representation (slots filling) using open schema or event taxonomy
  • Deep event representation
    • Semantic representation linked to a world-model (e.g. CycKB common sense knowledge) – suitable for reasoning and diagnostics
Feature vector event representation

• Feature vectors easily extractable from news documents:
  • **Topical dimension** – what is being talked about? (keywords)
  • **Social dimension** – which entities are mentioned? (named entities)
  • **Temporal aspect** – what is the time of an event? (temporal distribution)
  • **Geographical aspect** – where an event is taking place? (location)
  • **Publisher aspect** – who is reporting? (publisher identifiers)
  • **Sentiment/bias aspect** – emotional signals (numeric estimates)

• Scalable Machine Learning techniques can easily deal with such representation
  • ...in “Event Registry” system we use this representation to describe events
Example of “feature vector” event representation: Event Registry “Chicago” related events

Query: “Chicago”

Where? (geography)

When? (temporal distribution)

Who? (named entities)

What? (keyword/topics)
Structured event representation

• Structured event representation describes an event by its “Event Type” and corresponding information slots to be filled

• Event Types should be taken from “Event Taxonomy”

• ...at this stage of development this level of representation still requires human intervention to achieve high accuracy (Precision/Recall) extraction

• Example on the right – Wikipedia event infobox:
  • 2011 Tōhoku earthquake and tsunami
“Event Taxonomy” – preview to the current development
Prototype for event InfoBox extraction: XLike annotation service

- The goal is to build a system for economically viable extraction of event infoboxes
  - ...using crowd-sourcing
  - ...aiming at high Precision & Recall for a small cost
Event sequences & Hierarchical events

• Once having events identifies and represented we can connect events into “event sequences” (also called story-lines)
• “Event sequences” include events which are supposedly related and constitute a larger story

• Collection of interrelated events can be also organized in hierarchies (e.g. World Cup event consists from a series of smaller events)
An example event: Microsoft Windows 9

Windows 9 release date, price, features: why Microsoft can't wait to launch Windows 9 in 2015

WHEN: April 23, 2014

CATEGORIES: Shopping → Home and Garden → Windows, Home → Home Improvement → Windows and Doors, Computers → Software → Freeware

ENTITIES: Microsoft, Twitter, Google, Apple Inc., Computerworld, Gartner, Facebook, Universal Studios, Nokia, United States

KEYWORDS: Microsoft Windows, History of Microsoft Windows, Patch (computing), Software release life cycle, Operating system, Desktop computer, Electronics, Company, NeXT, Software build

ARTICLES: 239 (239 eng, 0 ger, 0 spa, 0 chi, 0 slo)

We outline what we know about Windows 9 release date, Windows 9 price and Windows 9 features. It's fair to say that Windows 8 has not been a universal success. In many ways a necessary step not terribly well executed, Windows 8's attempt to stitch together desktop and mobile has been way too radical to encourage new users. And as the longer life of existing PCs and laptops combines with the growth of smartphones and tablets to continually slow PC sales, Window 8 has come to be seen as a dud....
Similar events example: similar events to Microsoft Windows 9 event

<table>
<thead>
<tr>
<th>83%</th>
<th>Windows 8.1 Update: Five features we love (2014/04/04)</th>
</tr>
</thead>
<tbody>
<tr>
<td>82%</td>
<td>Windows 8.2 Rumors: New Start Menu, Apps on Desktop, More (2013/12/18)</td>
</tr>
<tr>
<td>81%</td>
<td>Windows 9 release date, price, features: new Start menu and no Charms bar (2014/08/13)</td>
</tr>
<tr>
<td>80%</td>
<td>Microsoft's new, new Windows cadence: Once a year is not enough (2014/04/26)</td>
</tr>
<tr>
<td>80%</td>
<td>Microsoft making Windows free on devices with screens under nine inches</td>
</tr>
<tr>
<td>77%</td>
<td>Here are the latest details on Windows 9 and Microsoft's other major Windows updates (2014/06/25)</td>
</tr>
<tr>
<td>77%</td>
<td>Mouse and keyboard-friendly Windows 8.1 update arrives April 8 (2014/04/02)</td>
</tr>
<tr>
<td>77%</td>
<td>Five Places to Learn How to Use Windows 8.1 Better (2014/01/18)</td>
</tr>
<tr>
<td>77%</td>
<td>Microsoft's 'Project Siena' lets anyone create Windows 8.1 apps (2013/12/19)</td>
</tr>
<tr>
<td>76%</td>
<td>Microsoft strips some Windows 7 users of IE11 patch privileges (2014/07/11)</td>
</tr>
<tr>
<td>76%</td>
<td>No Microsoft Start Menu for Windows 8 until 2015 (2014/06/02)</td>
</tr>
<tr>
<td>76%</td>
<td>Intel plans a CES coup: Android and Windows in the same computer (2014/01/03)</td>
</tr>
<tr>
<td>76%</td>
<td>New Windows 'Threshold' rumor claims virtual desktop support is being tested (2014/08/06)</td>
</tr>
<tr>
<td>75%</td>
<td>Microsoft Hopes Windows 9 Will Win Over Desktop Users (2014/07/03)</td>
</tr>
<tr>
<td>75%</td>
<td>Microsoft showcases latest updates to Windows, opportunities for developers (2014/04/03)</td>
</tr>
<tr>
<td>75%</td>
<td>Windows 8.1 update sweeps 'Metro' under the rug, boots to desktop by default (2014/02/02)</td>
</tr>
<tr>
<td>75%</td>
<td>In Six Months, Microsoft Will Pull The Plug On Some Parts Of Windows 7 Support (MSFT) (2014/07/09)</td>
</tr>
<tr>
<td>74%</td>
<td>Microsoft talks Windows 8.1 Update 1 features, April 8 release, teases future Start menu (2014/04/05)</td>
</tr>
</tbody>
</table>
Event sequence identification
Hierarchy of events
Example
Microsoft hierarchy of events
Zoom-in Example
Microsoft hierarchy of events
Event Visualization

http://eventregistry.org/
Live Event tracking with [http://EventRegistry.org/](http://EventRegistry.org/)
Event description through entities and Semantic keywords

**ENTITIES**

- Barack Obama
- National Security Agency
- United States
- Washington D.C.
- Josiah Bartlet
- President of the United States
- United States Foreign Intelligence
- Angela Merkel
- Presidency of George W. Bush
- Chancellor of Germany (Federal...)
- Executive Office of the President...
- Yakov Smirnoff
- Supreme Court of the United States
- Republican Party (United States)
- Romeo Miller
- Jay Carney
- United States Department of justice
- Larry Klayman
- European Parliament

**KEYWORDS**

- Surveillance
- White House
- Privacy
- Espionage
- Review
- Telephone
- Meeting
- Leadership
- Official
- Legislator
- United States Congress
- Debate
- Board of directors
- Foreign policy
- Report
- Data
- Intelligence
- National security
- Document
Collection of events described through Entity relatedness
Collection of events described through trending concepts.
Collection of events described through three level categorization
Obama ponders limiting NSA access to phone records

Washington D.C., United States

Washington -- President Barack Obama is expected to rein in spying on foreign leaders and is considering restricting National Security Agency access to the government's surveillance program, according to people familiar with the government's review. Obama could unveil his highly anticipated decisions as early as next week. On Thursday, the president is expected to discuss his review with congressional lawmakers, while his top lawyer plans to meet with privacy groups.

Representatives from tech companies are meeting with White House staff on Friday.

The White House...

Nr. of articles: 105 (89 eng, 5 ger, 11 spa, 0 chi, 0 slo)

Articles

Obama ponders limiting

Washington -- President Barack Obama is expected to restrict National Security Agency access to Americans' phone records and rein in spying on foreign leaders, according to people familiar with the government's surveillance programs.

Obama could unveil his highly anticipated decisions as early as next week. On Thursday, the president is expected to discuss his review with congressional lawmakers, while his top lawyer plans to meet with privacy groups. Representatives from tech companies are meeting with White House staff on Friday.

Obama likely to adjust US spying on foreign leaders, nearing decision on intelligence changes

WASHINGTON - President Barack Obama is expected to restrict National Security Agency access to Americans' phone records and rein in spying on foreign leaders, according to people familiar with a White House review of the government's surveillance programs.

Obama could unveil his highly anticipated decisions as early as next week. On Thursday, the president is expected to discuss his review with congressional lawmakers, while his top lawyer plans to meet with privacy groups. Representatives...

Obama might limit how much the NSA can spy on you

President Barack Obama speaks at a press conference on Dec. 20, 2013. (Pablo Martinez Monsivais/AP)
Collection of events described through Reporting dynamics
Collection of events described through a story-line of related events
Event Registry API
Python code to access “Event Registry”

- https://github.com/gregorleban/event-registry-python
Searching for events using Python

from EventRegistry import *
er = EventRegistry()
q = QueryEvents()
# get events related to Barack Obama
q.addConcept(er.getConceptUri("Obama"))
# and are related to issues in society
q.addCategory(er.getCategoryUri("society issues"))
# and have been reported by the BBC
q.addNewsSource(er.getNewsSourceUri("bbc"))
# return event details for first 30 events
q.addRequestedResult(RequestEventsInfo(page = 0, count = 30))
# execute query and obtain results
res = er.execQuery(q)
Result of the query

```json
'events': { 'resultCount': 122,
    'results': [
        {'articleCounts': {'eng': 54.0, 'total': 54.0},
            'categories': [{}],
            'concepts': [{}],
            'eventDate': '2014-08-29',
            'eventDateEnd': '',
            'multiLingInfo': {'eng': {
                'title': ..., 'summary': ...}},
            'uri': '1211229', 'wgt': 9.0}
    ], ...
}
```
Getting info for a particular event

# get information about event with ID 123
>>> q = QueryEvent("123");

# return concept labels in 3 languages
>>> q.addRequestedResult(RequestEventInfo(["eng", "spa", "slv"]))

# get 10 most central articles
>>> q.addRequestedResult(RequestEventArticles(0, 10))

# get information how articles about the event were trending
>>> q.addRequestedResult(RequestEventArticleTrend())

# get top keywords
>>> q.addRequestedResult (RequestEventKeywordAggr())

>>> eventRes = er.executeQuery(q);
Searching for articles

```python
>>> q = QueryArticles();
# articles should be from a particular time period
>>> q.setDateLimit(datetime.date(2014, 4, 16), datetime.date(2014, 4, 28))
# they should mention apple
>>> q.addKeyword("apple")
# they should also mention iphone
>>> q.addKeyword("iphone")
# get top 30 articles that match criteria
>>> q.addRequestedResult(RequestArticlesInfo(page=0, count = 30));
>>> res = er.execQuery(q)
```
Event Registry exports event data through API and RDF/Storyline ontology

- API to search and export event information
  - Export of all the system data in JSON

- Event data is exported in a structured form
  - BBC Storyline ontology

- SPARQL endpoint:
  - [http://eventregistry.org/rdf/search](http://eventregistry.org/rdf/search)
  - [http://eventregistry.org/rdf/event/{eventID}](http://eventregistry.org/rdf/event/{eventID})
  - [http://eventregistry.org/rdf/article/{articleID}](http://eventregistry.org/rdf/article/{articleID})
  - [http://eventregistry.org/rdf/storyline/{storylineID}](http://eventregistry.org/rdf/storyline/{storylineID})
  - Example: [http://eventregistry.org/rdf/event/1234](http://eventregistry.org/rdf/event/1234)