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

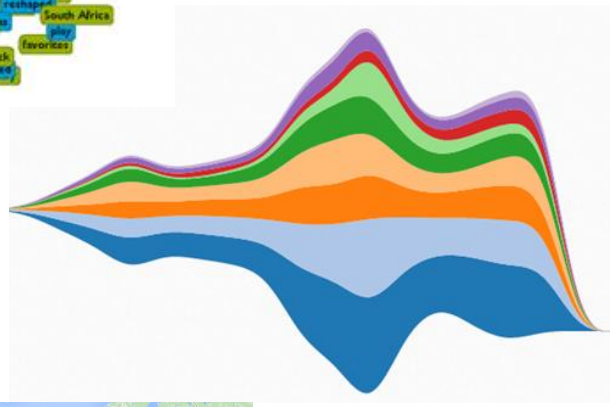
Introduction

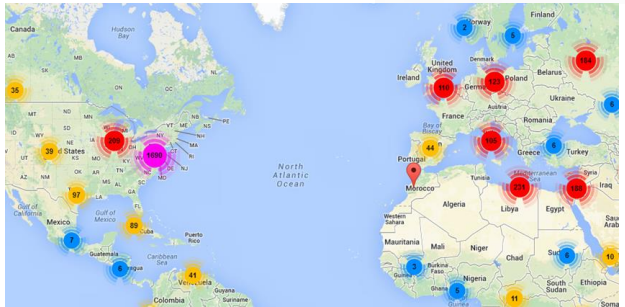
What questions we'll try to answer?

- 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

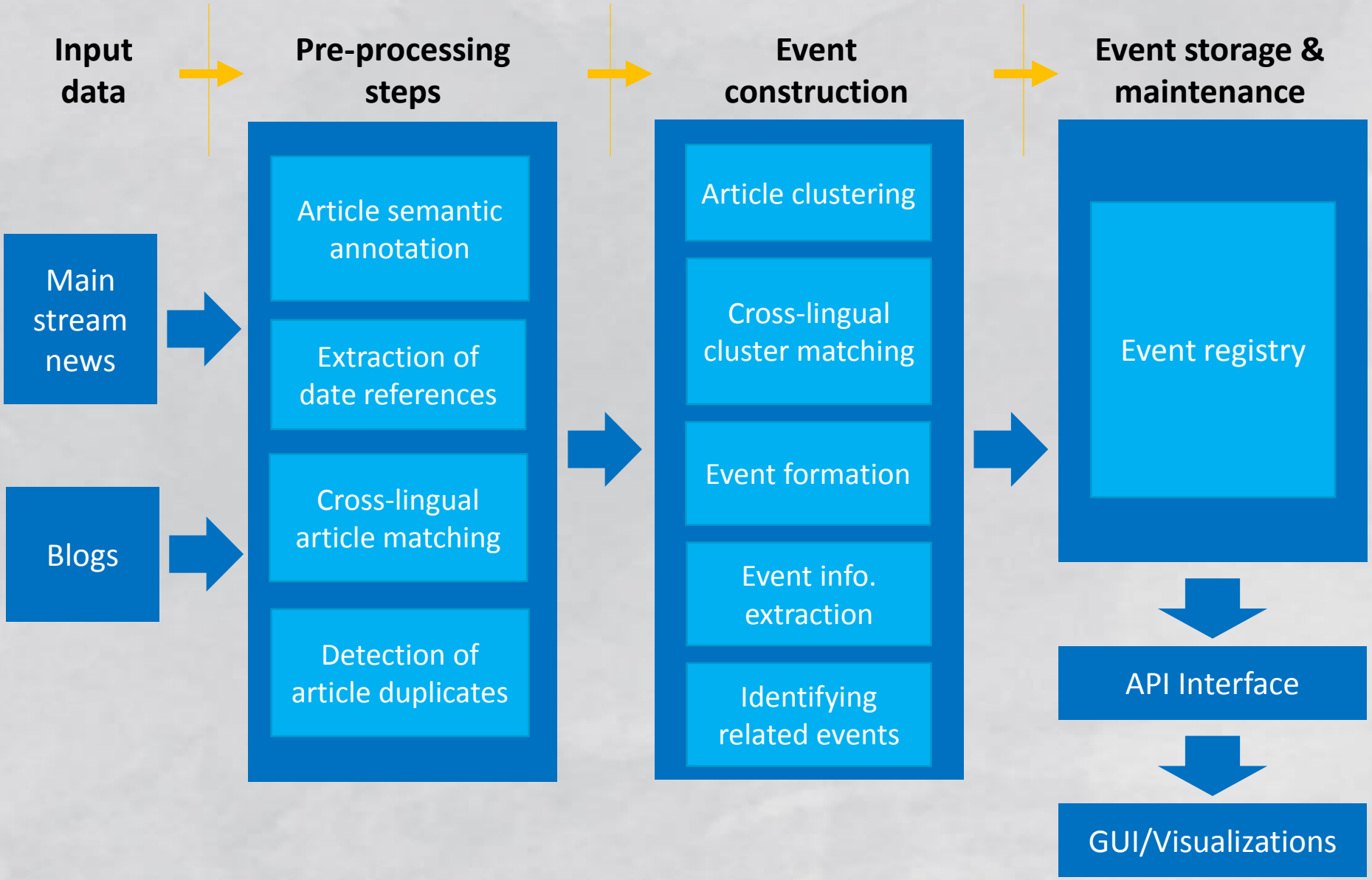




The overall goal

- 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



http://EventRegistry.org

Collecting Media Data

http://newsfeed.ijs.si/

Where to get references to news publishers?

- Good start is Wikipedia list of newspapers:
  - http://en.wikipedia.org/wiki/Lists_of_newspapers

Lists of newspapers

From Wikipedia, the free encyclopedia

Below are **lists of newspapers** organized by continent.

Contents [hide]

1 Africa
2 Asia
3 Europe
4 North America
5 Oceania
6 South America

Europe [edit]

Sovereign states	Albania · Andorra · Armenia · Austria · Azerbaijan · Belarus · Belgium · Bosnia and Herzegovina · Bulgaria · Croatia · Cyprus · Czech Republic · Denmark · Estonia · Finland · France · Georgia · Germany · Greece · Hungary · Iceland · Ireland · Italy · Kazakhstan · Latvia · Liechtenstein · Lithuania · Luxembourg · Macedonia · Malta · Moldova · Monaco · Montenegro · Netherlands · Norway · Poland · Portugal · Romania · Russia · San Marino · Serbia · Slovakia · Slovenia · Spain · Sweden · Switzerland · Turkey · Ukraine · United Kingdom · Vatican City
States with limited recognition	Abkhazia · Kosovo · Nagorno-Karabakh · Northern Cyprus · South Ossetia · Transnistria
Dependencies and other territories	Åland · Faroe Islands · Gibraltar · Guernsey · Jersey · Isle of Man · Svalbard

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

Daily political newspapers [edit]

- *Alithia Online*[2]
- *Adesmeftos Typos* ('Unbound Press')[3]
  - Political affiliation: New Democracy (center-right)
- *Ano Kato* ('Upside Down')
- *I Avgi* ('The Dawn')[4] **(Greek)**
  - Political affiliation: Synaspismos (left)
- *Avriani* ('Tomorrow's [Daily]')[5]
  - Political affiliation: PASOK (center-left)
- *Eleftheri Ora* ('Free Hour')[6]
  - Political affiliation: Nationalist
- *Eleftheros* ('[The] Free [One]')
  - Political affiliation: New Democracy (Right-Wing)
- *Eleftheros Typos* ('Free Press')[7]
  - Political affiliation: New Democracy (center-right)
- *Eleftherotypia* ('Freedom of the Press')[8]
  - High circulation
  - Political orientation: Left-wing
- *Espresso*
  - Political affiliation: none (tabloid)
- *Estia* ('Hestia')[9]
  - Political affiliation: New Democracy (center-right)
- *Ethnos* ('Nation')[10]

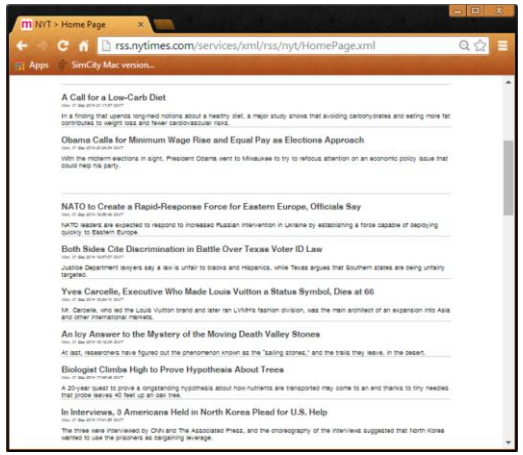
From a newspaper home-page to an article

http://www.nytimes.com/

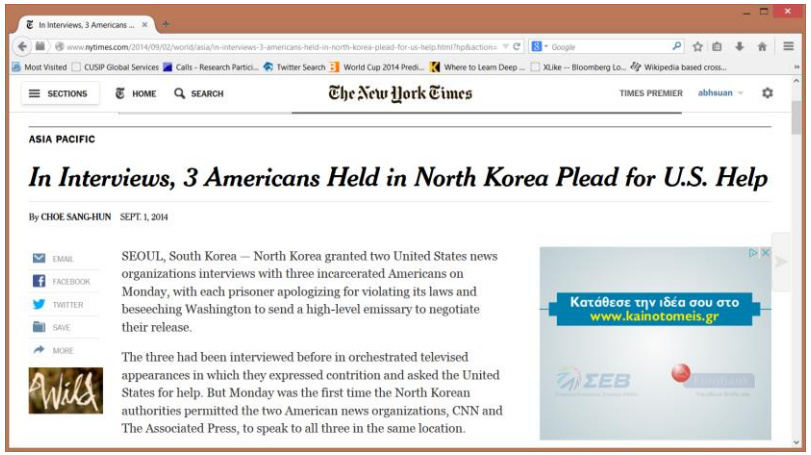
HTML

```
<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:

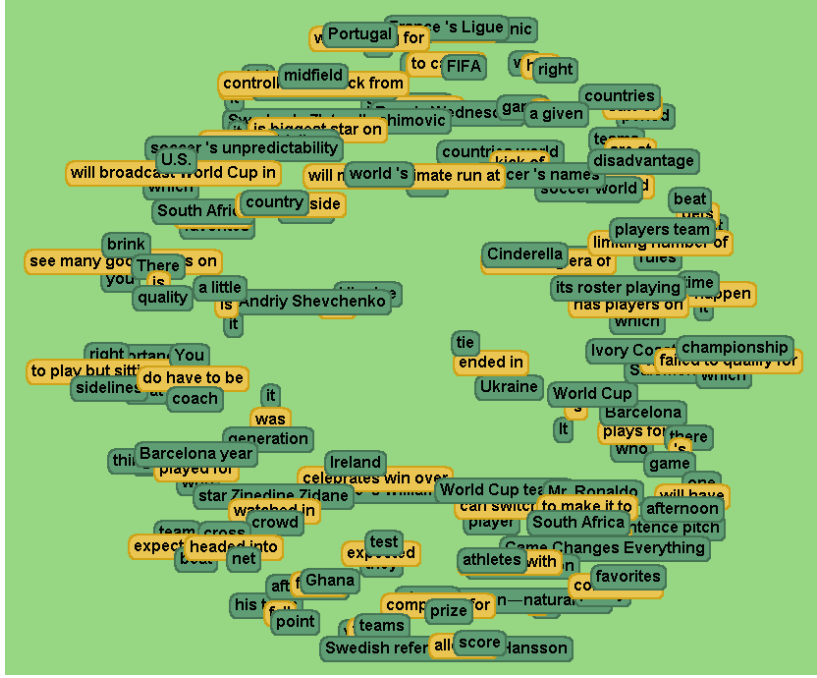
```
timestamp = [when you want to start downloading, e.g. now() minus 1 hour]
while True:
   fetch http://newsfeed.ijs.si/stream/?after=timestamp
   if [404 error]:
      # there is no new data
      pause 1 minute
   else:
      save data
      timestamp = [extract it from the Content-Disposition HTTP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

```
<article id="internal article ID; consistent across streams">
   <source>
      <hostname> Publisher hostname </hostname>
      <title> Name of the publisher; failing that, title of the RSS feed </title>
      <location?>
         <longitude?> publisher longitude in degrees </longitude>
         <latitude?> publisher latitude in degrees </latitude>
         <city?> publisher city </city>
         <country?> publisher country </country>
      </location>
      <tags?>
         <tag+> a tag for the publisher; the vocabulary is not controlled </tag>
      </tags>
   </source>
   <feed+>
      <uri> URL from which the article was discovered; typically the RSS feed </uri>
   </feed>
   <uri> URL from which the article was downloaded </uri>
   <publish-date?> The publication time and date. </publish-date>
   <retrieve-date> The retrieval time and date. </retrieve-date>
   <lang> 3-letter ISO 639-2 language code </lang>
   <story_id?> story cluster this article was grouped into (at download time) </lang>
   <location?+>
      <longitude?> story content longitude in degrees </longitude>
      <latitude?> story content latitude in degrees </latitude>
      <city?> story city </city>
      <country?> story country </country>
   </location>
   <tags?>
      <tag+> a tag for the article; the vocabulary is not controlled </tag>
   </tags>
   <img?> The URL of a related image, usually a thumbnail. </img>
   <title> Title. Can be empty if we fail to identify it. </title>
   <body-cleartext>
      Clear text body of the article, formatted only with <p> tags
   </body-cleartext>
   <body-rych?; only English, Slovene>
      Enriched article body; an XML subtree as returned by Enrycher.
   </body-rych>
   <body-xlike?; only English, Spanish, Catalan>
      Enriched article body; an XML subtree as returned by iSOCO; experimental.
   </body-xlike>
</article>
```

- …more details:
  - Mitja Trampus, Blaz Novak: *The Internals Of An Aggregated Web News Feed*. Proceedings of 15th Multiconference on Information Society 2012 (IS-2012). [PDF]



Document Enrichment

http://enrycher.ijs.si/

What can 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/)

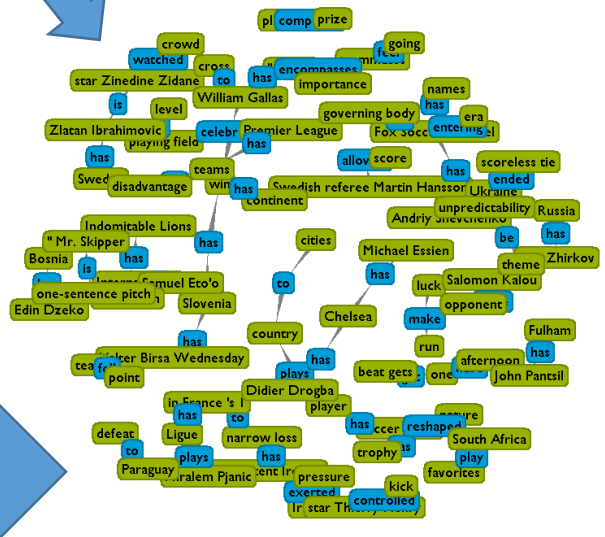
Plain text

Slovenia's dramatic win over Russia Wednesday, and to a lesser extent Ireland's narrow loss to France, capped off a grueling two-year qualifying period that saw some of the smallest countries in the world kick some of soccer's biggest names in the teeth. After a century of near domination from the likes of Brazil, Italy and Germany, international soccer is entering the era of the Cinderella. It may not happen this time around, but given the increasing flow of talent, training and information across borders, it's almost certain that a small upstart nation blessed with good athletes and better luck will make a legitimate run at the wor coveted trophy.

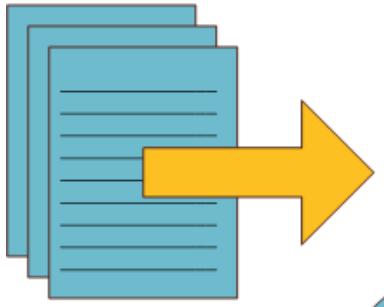
Russia's Yuri Zhirkov, right, fights for the ball with Slovenia's Wednesday.

Text Enrichment

Extracted graph of triples from text

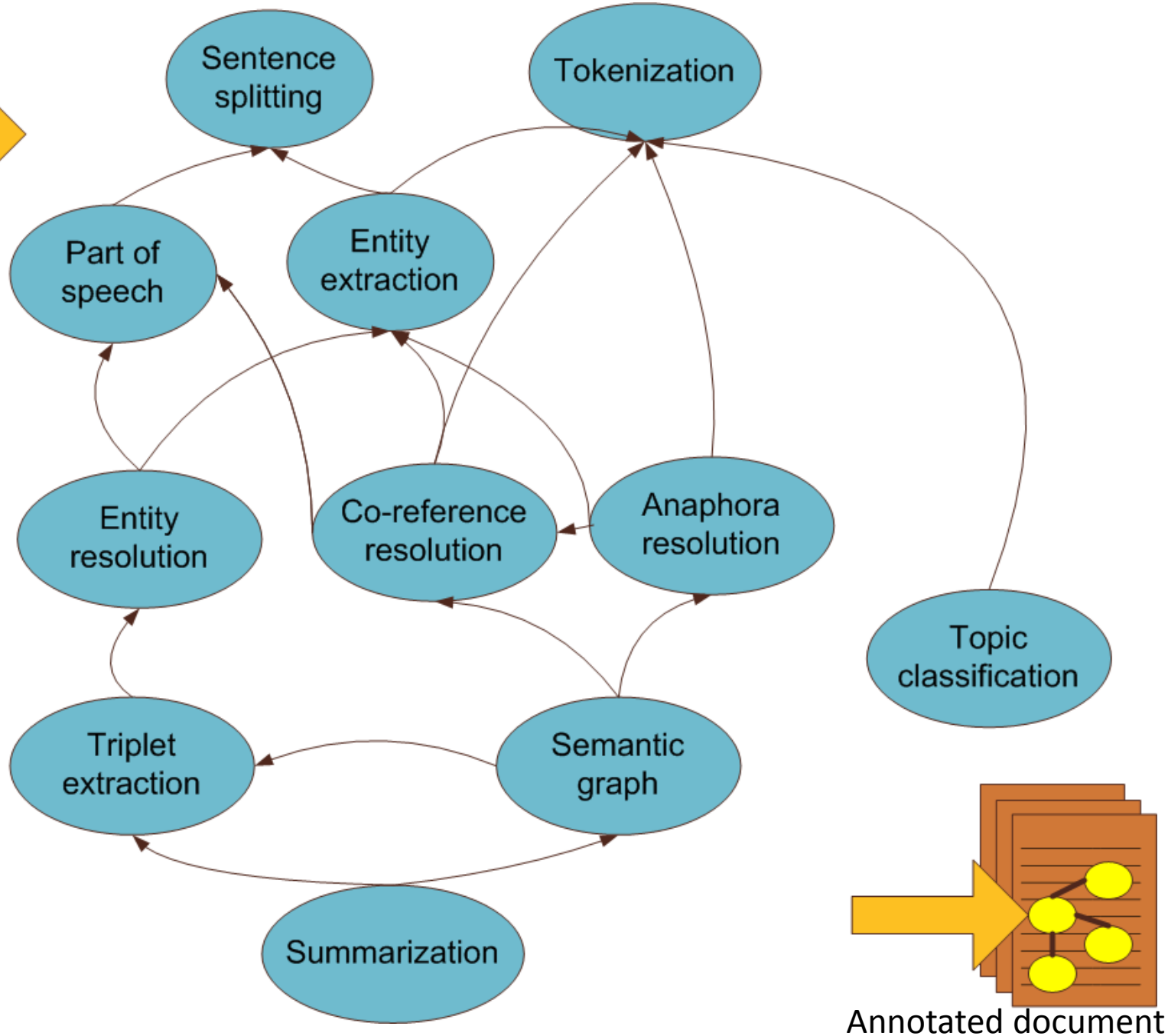

Zlatan Ibrahimovic
William Gallas
Premier League
Swedish referee Martin Hansson
Indomitable Lions
Michael Essien
Samuel Eto'o
Slovenia
Chelsea
Didier Drogba
Salomon Kalou
John Pantsil
South Africa
Fulham
Zhirkov
Russia
Bosnia
Edin Dzeko
Paraguay
Ligue
narrow loss
pressure
controlled
kick
favorites
trophy
Cities
country
Crowd
watched
star Zinedine Zidane
importance
governing body
score
scoreless tie
ended
unpredictability
opponent
make
run
afternoon
beat gets
one-sentence pitch
disadvantage
continent
Swed
level
names
era
Fox soccer
Andriy Shevchenko
Ukraine
Mr. Skipper
Bosnia
Ireland
France
Wednesday
point
defeat
Ibrahimovic
Pjanic
Ibrahimovic
Ibrahimovic
Ibrahimov

Enrycher Architecture



Plain text

- **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`



Annotated document

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

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

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

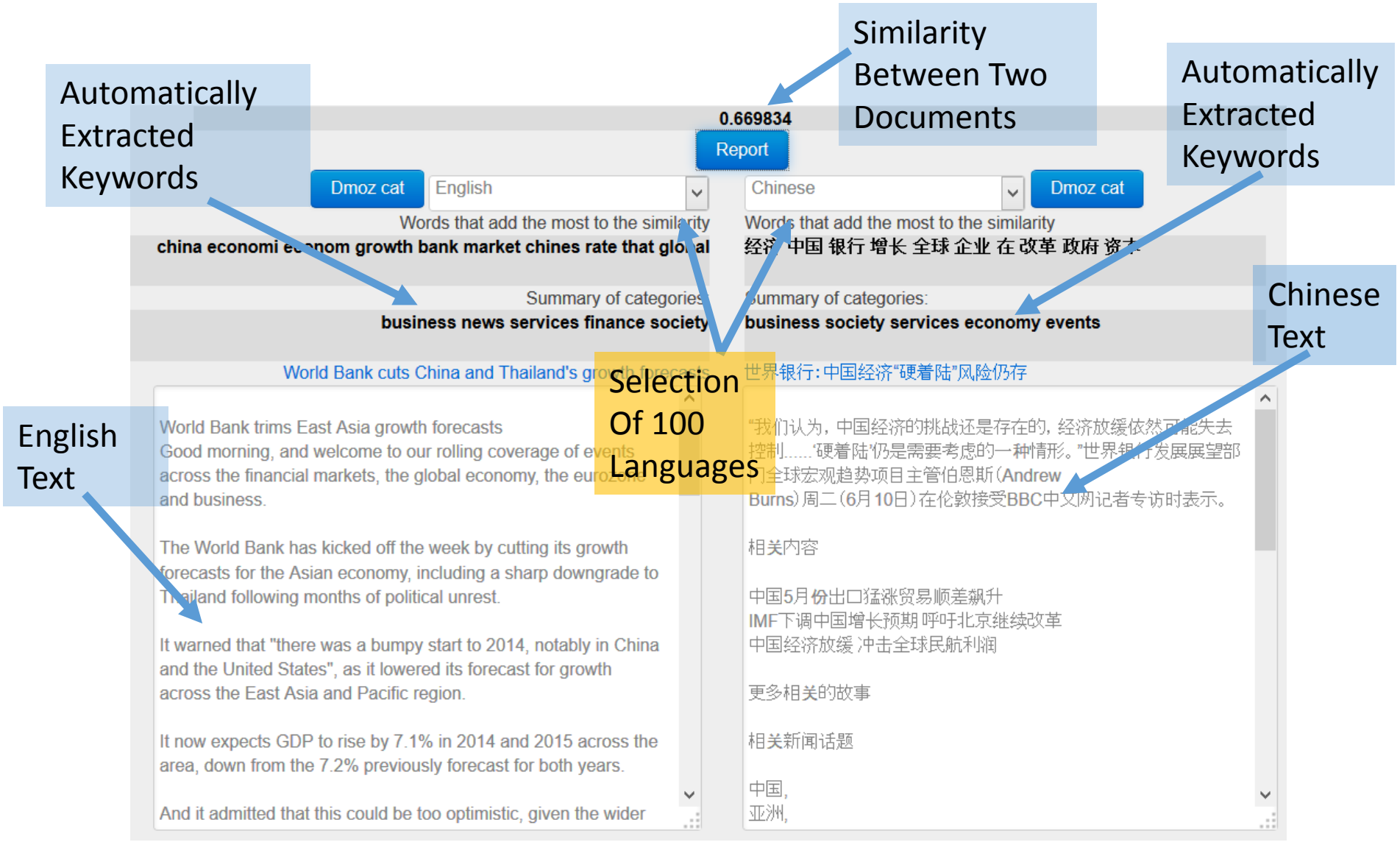
- The key building block is the function for comparing and categorization of documents in different languages
  - http://XLing.ijs.si is an open web service to bridge information across 100 languages

Languages covered by XLing (top 100 Wikipedia languages)

Afrikaans
Albanian
Arabic
Aragonese
Armenian
Asturian
Azerbaijani
Bashkir
Basque
Belarusian
Belarusian (Taraškievica)
Bengali
Bishnupriya Manipuri
Bosnian
Breton
Bulgarian
Burmese
Cantonese
Catalan
Cebuano
Chinese
Chuvash
Croatian
Czech
Danish
Dutch
English
Esperanto
Estonian
Finnish
French
Galician
Georgian
German
Greek
Gujarati
Haitian
Hebrew
Hindi
Hungarian
Icelandic
Ido
Indonesian
Irish
Italian
Japanese
Javanese
Kazakh
Kirghiz
Korean
Kurdish
Latin
Latvian
Lithuanian
Lombard
Low Saxon
Luxembourgish
Macedonian
Malagasy
Malay
Malayalam
Marathi
Nepali
Newar / Nepal Bhasa
Norwegian (Bokm?l)
Norwegian (Nynorsk)
Occitan
Persian
Piedmontese
Polish
Portuguese
Quechua
Romanian
Russian
Serbian
Serbo-Croatian
Sicilian
Simple English
Slovak
Slovenian
Spanish
Sundanese
Swahili
Swedish
Tagalog
Tamil
Tatar
Telugu
Thai
Turkish
Ukrainian
Urdu
Uzbek
Vietnamese
West Frisian
Western Panjabi
Yoruba

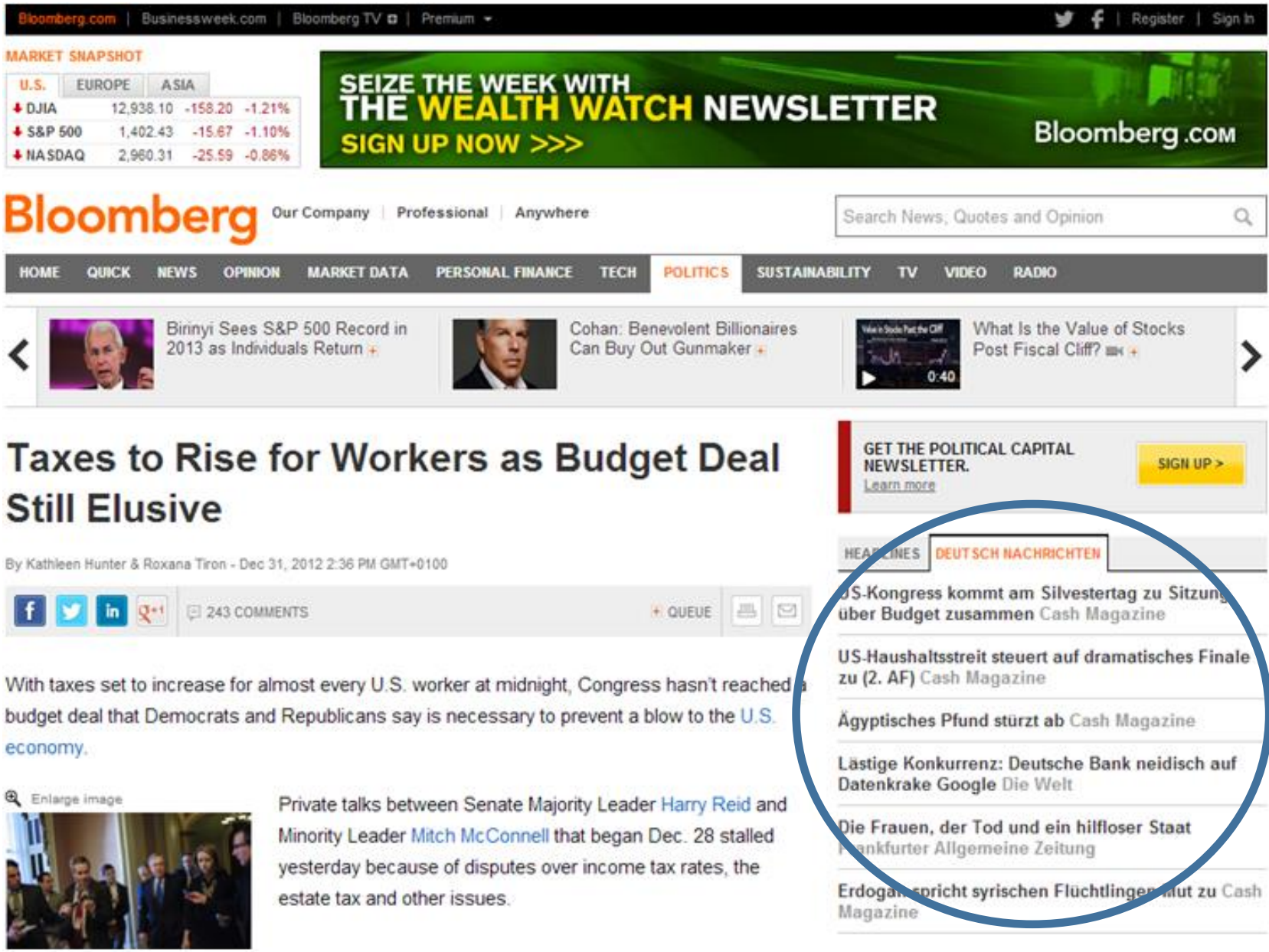
XLing (http://XLing.ijs.si)

service for comparing and categorization of documents across 100 languages



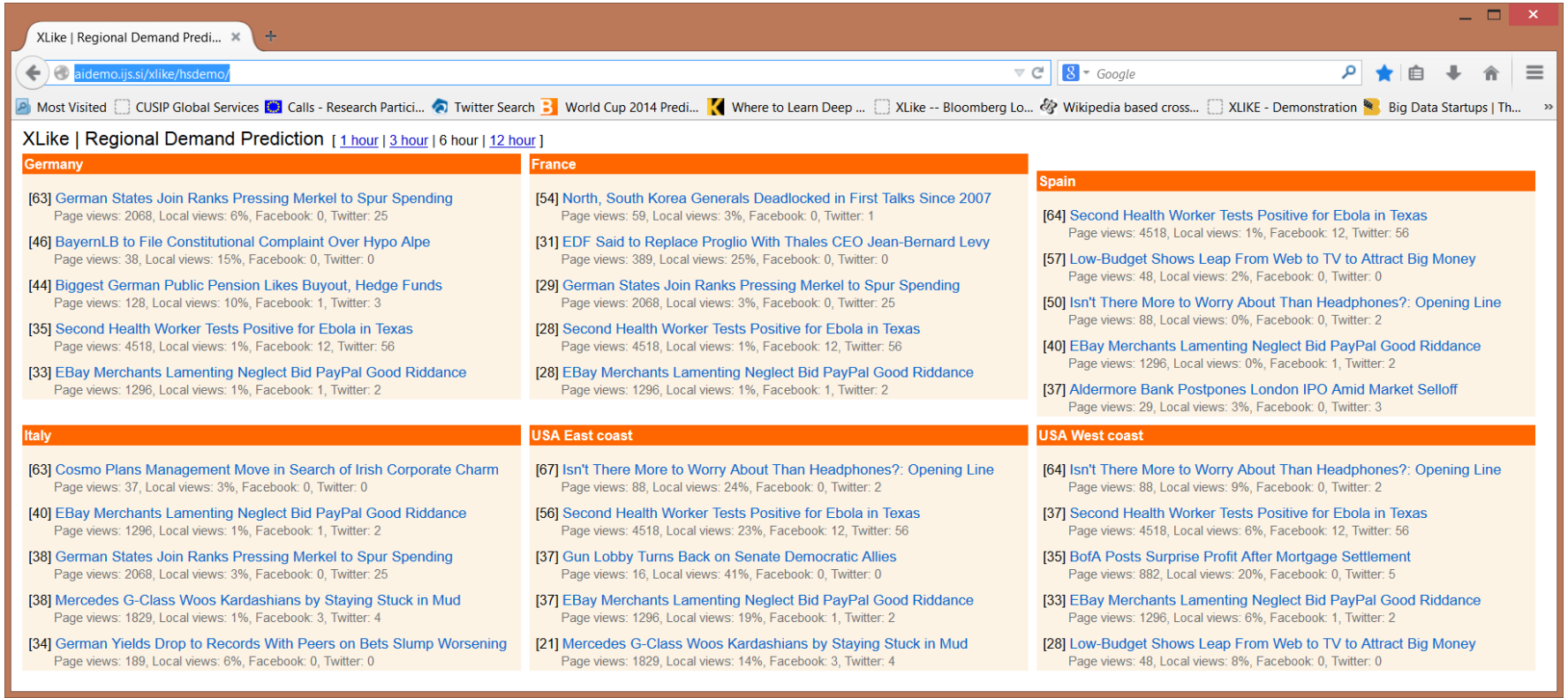
Example: Cross-lingual News Recommendation

- 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/

News Reporting Bias example

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

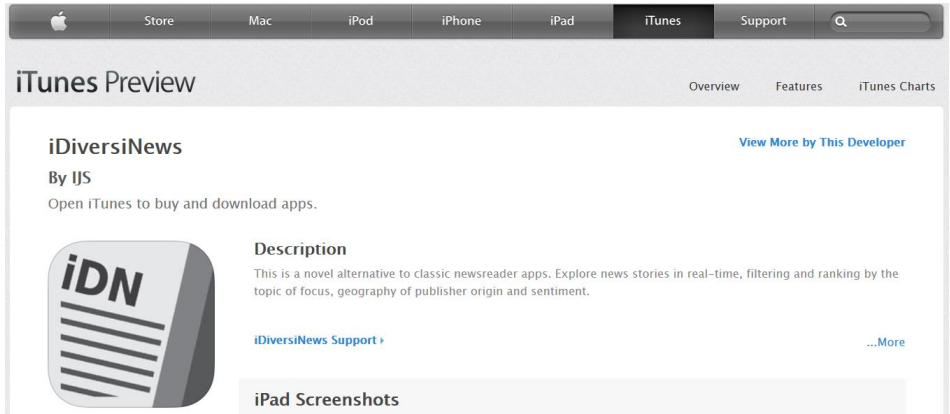
- The task:
  - Given a news story, are we able to say from which news source it came?

- 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

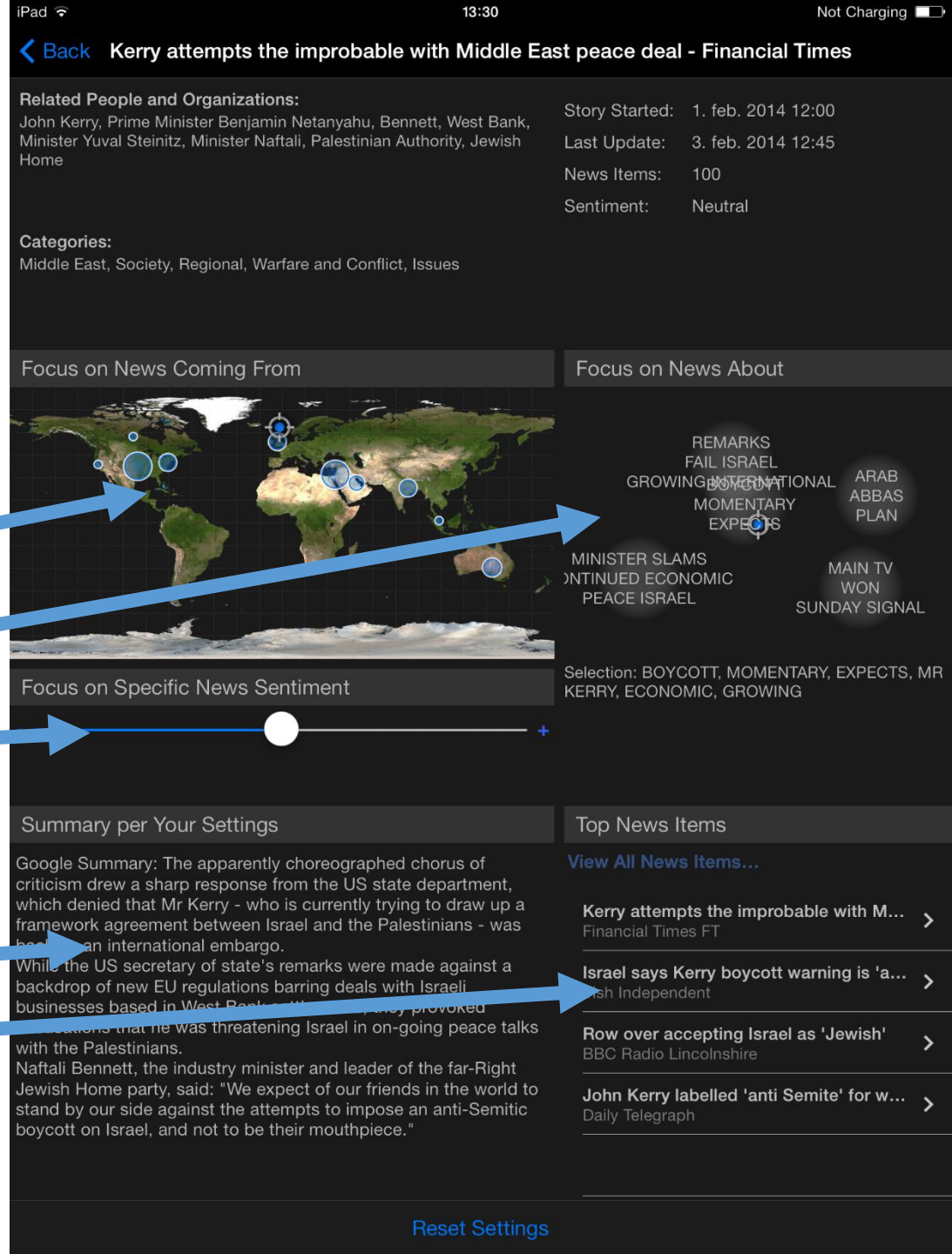


Hot Topics

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**



Geography

Subtopics

Sentiment

Summary

Articles

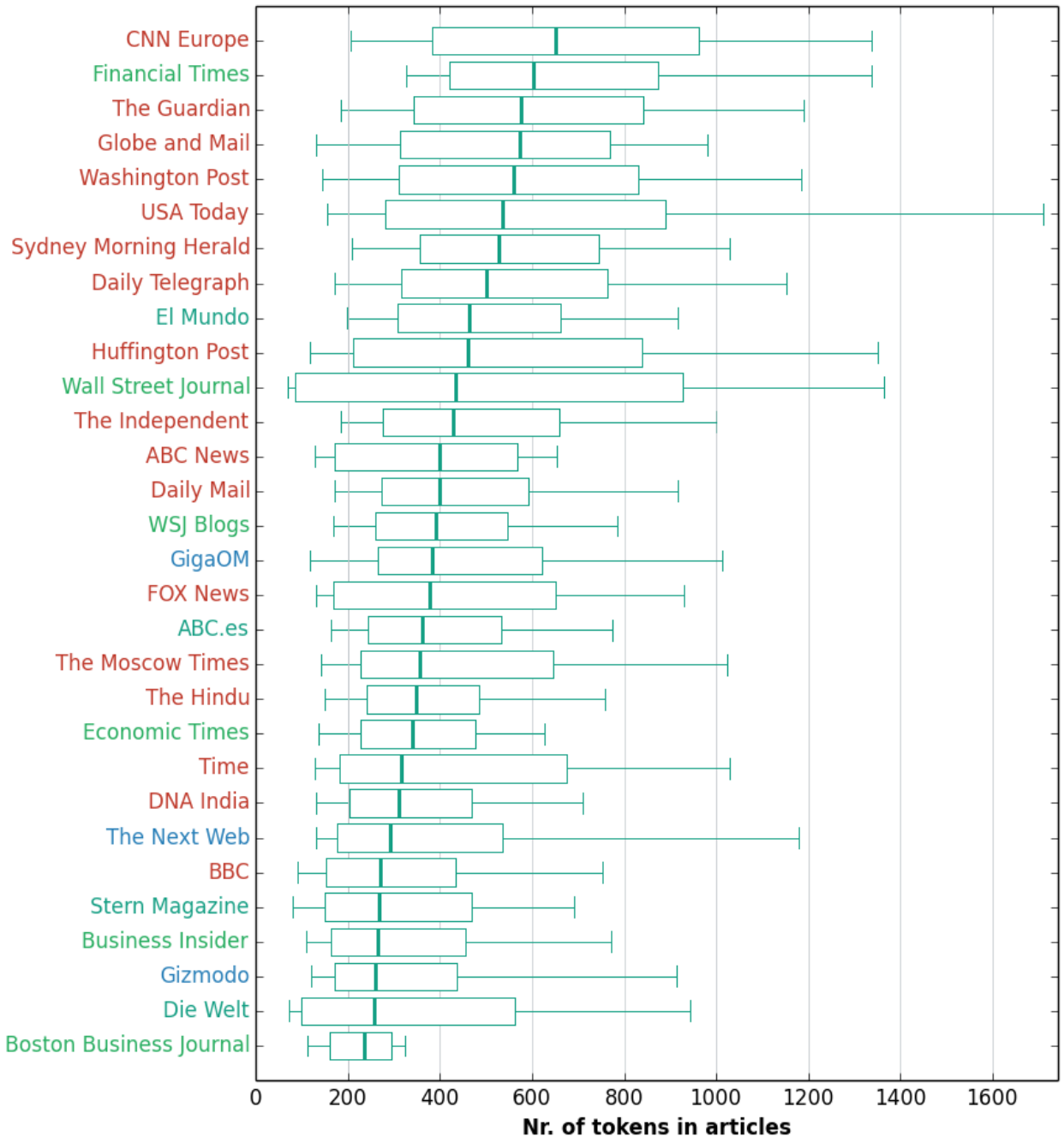
New Bias Experiment

- 30 news publishers selected
- All articles from Dec 15th 2014 – Aug 15 2014
- **General**: BBC, USA Today, Daily Mail, The Guardian, Globe and Mail, Huffington Post, Washington Post, CNN Europe, The Independent, Time, The Hindu, The Moscow Times,
- **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

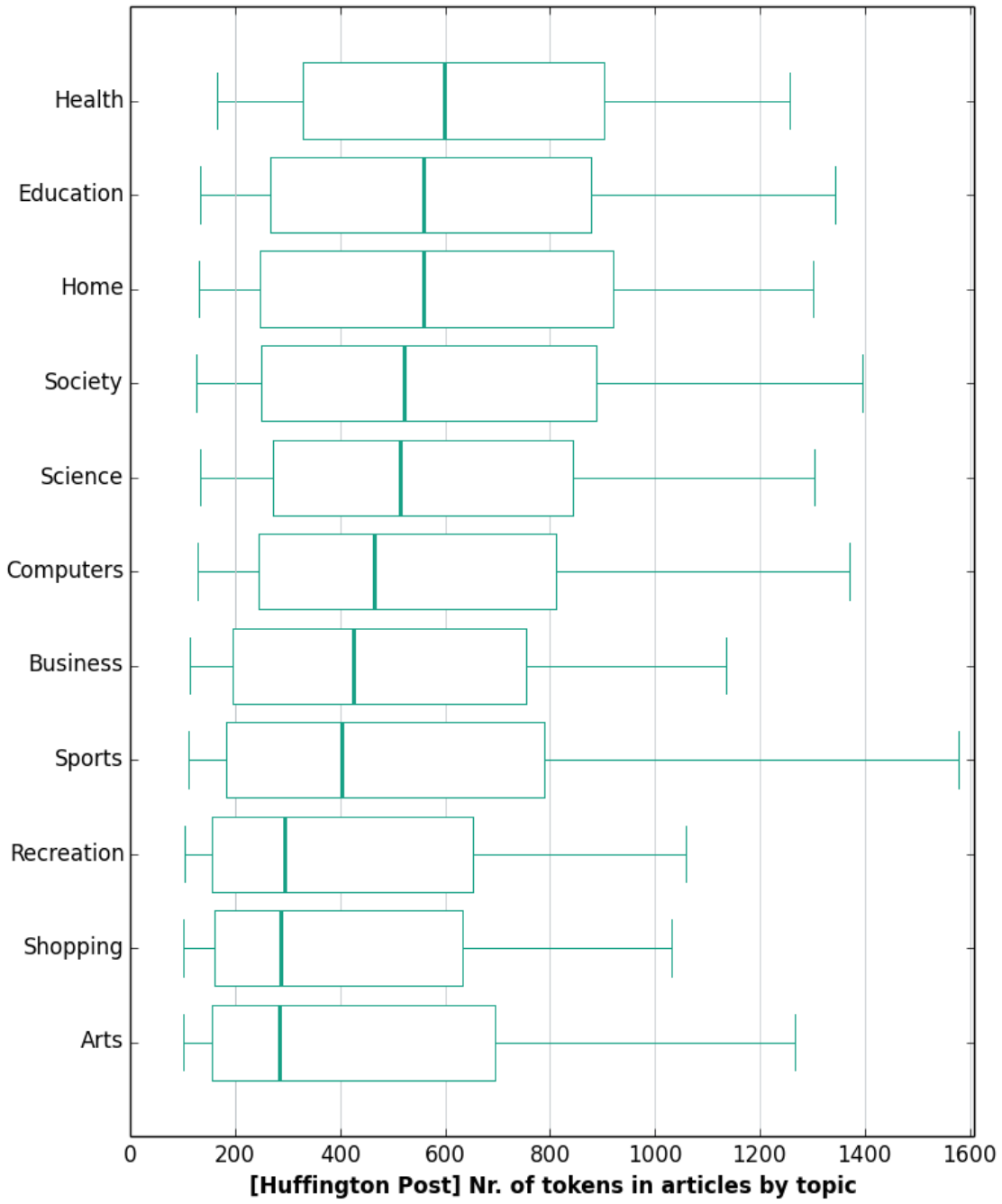
Dimensions of news bias to be observed

- 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

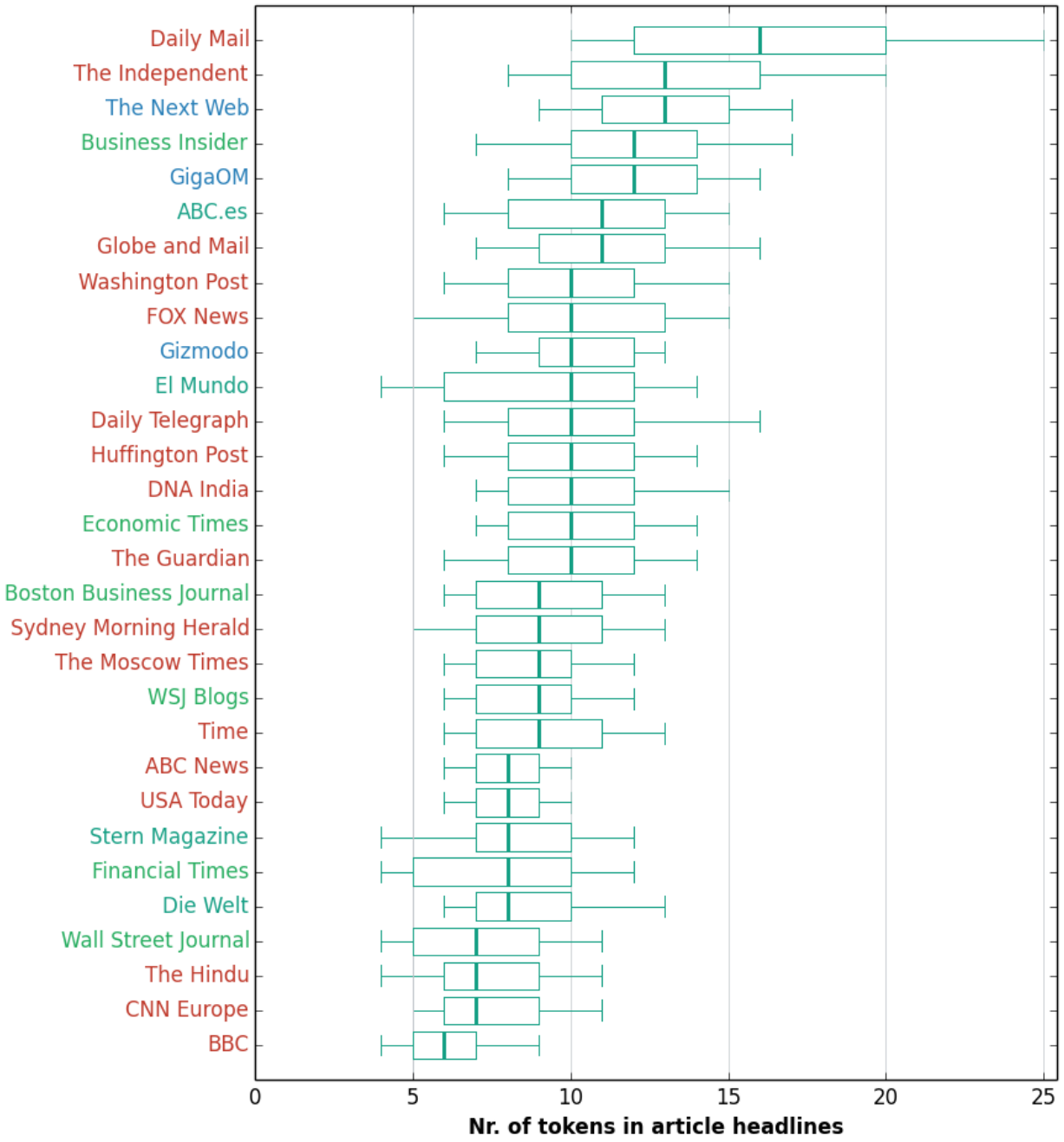
Article length differences



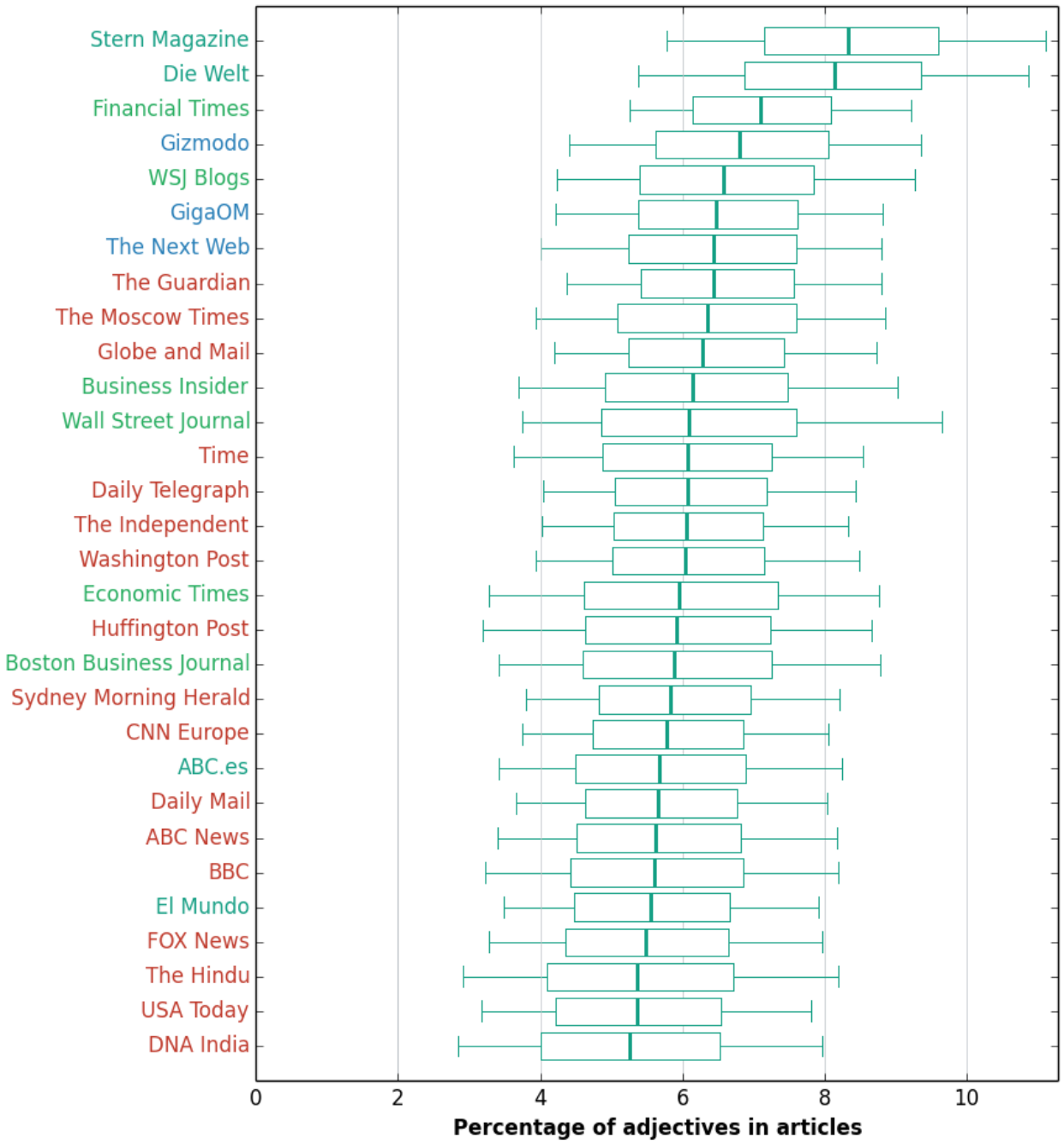
Number of tokens in articles per topic



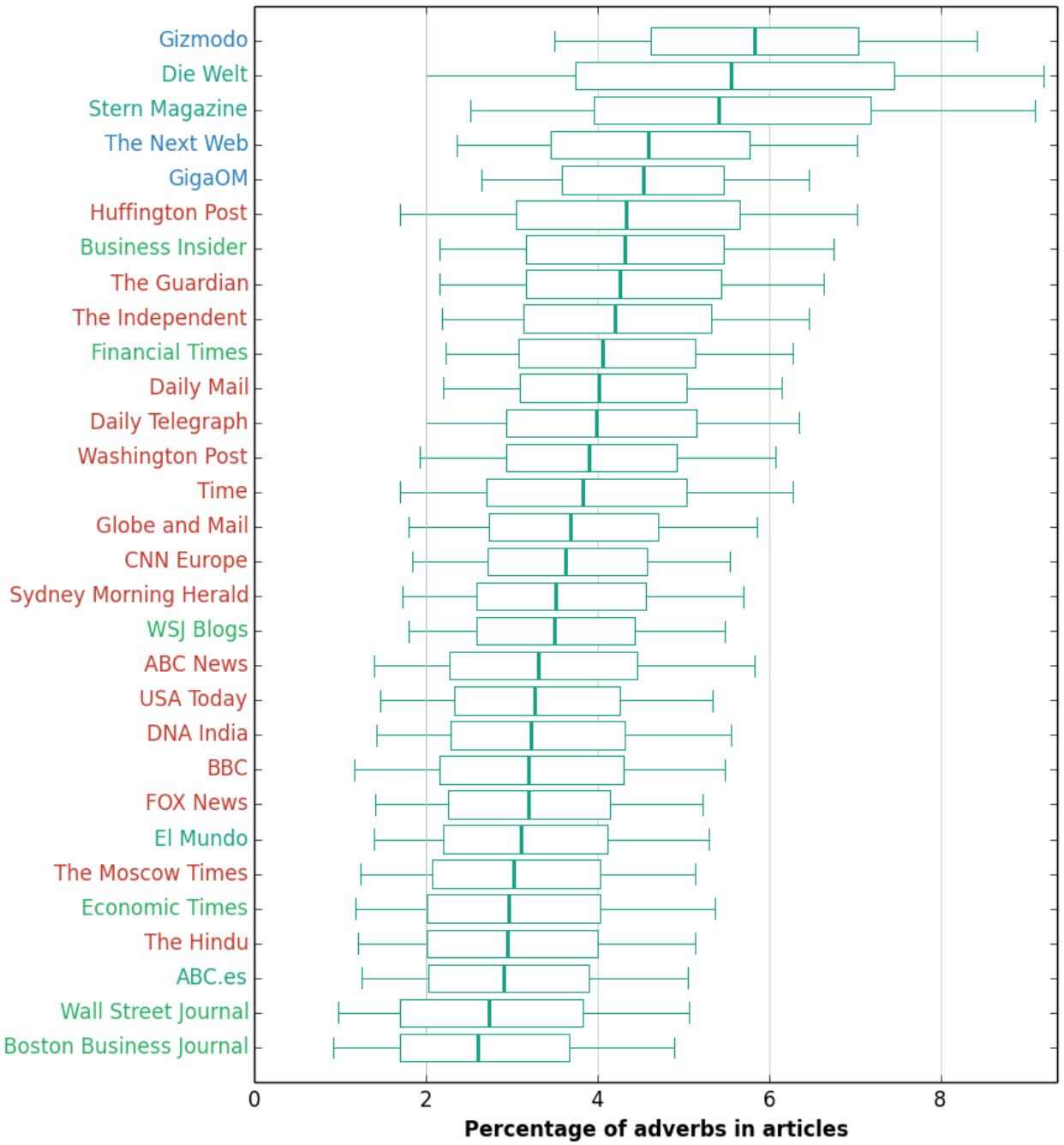
Number of tokens in article headlines



Grammatical differences (% of adjectives in articles)



Grammatical differences (% of adverbs in articles)



Readability differences

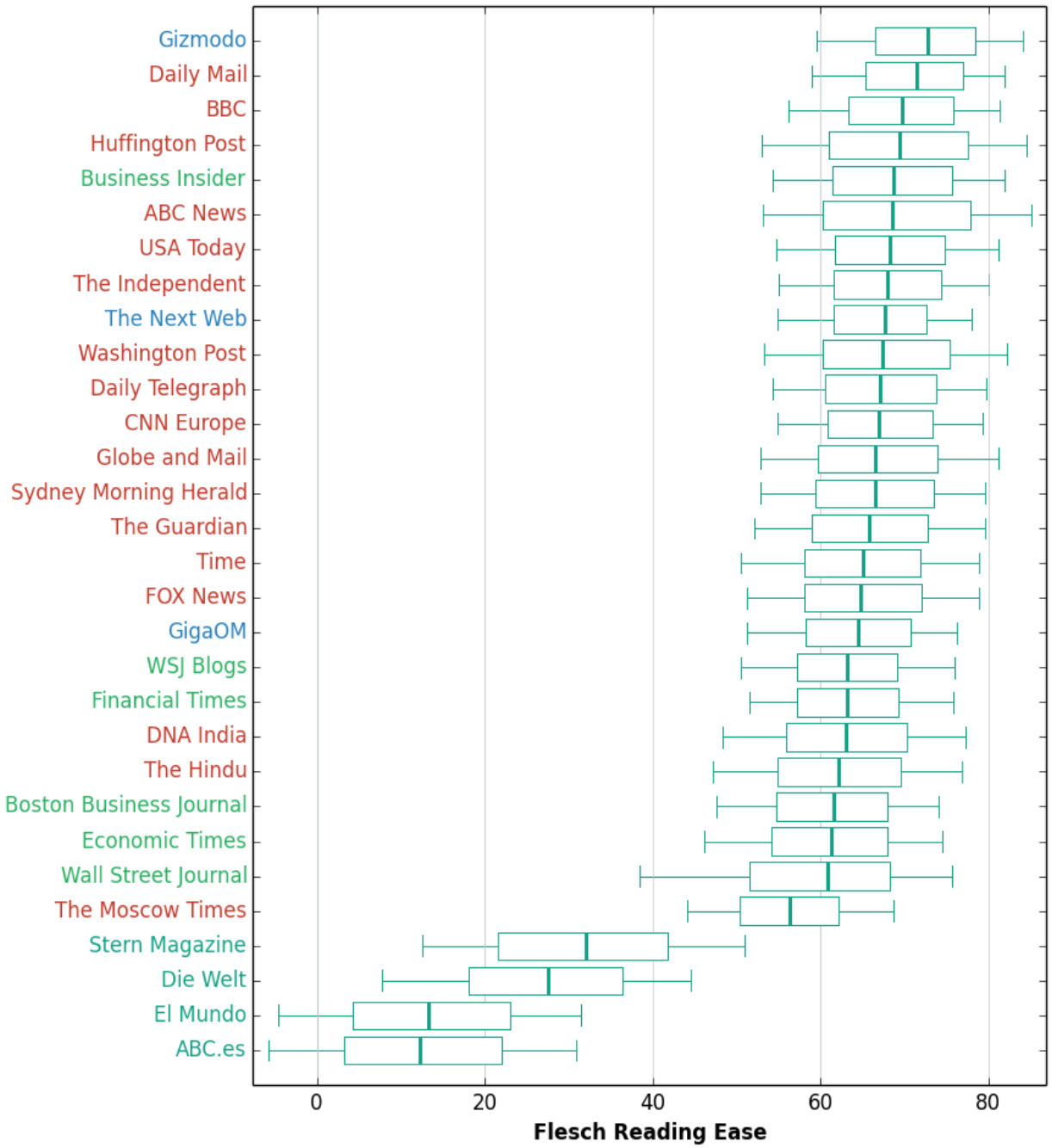
- **Dale-Chall readability formula**.

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

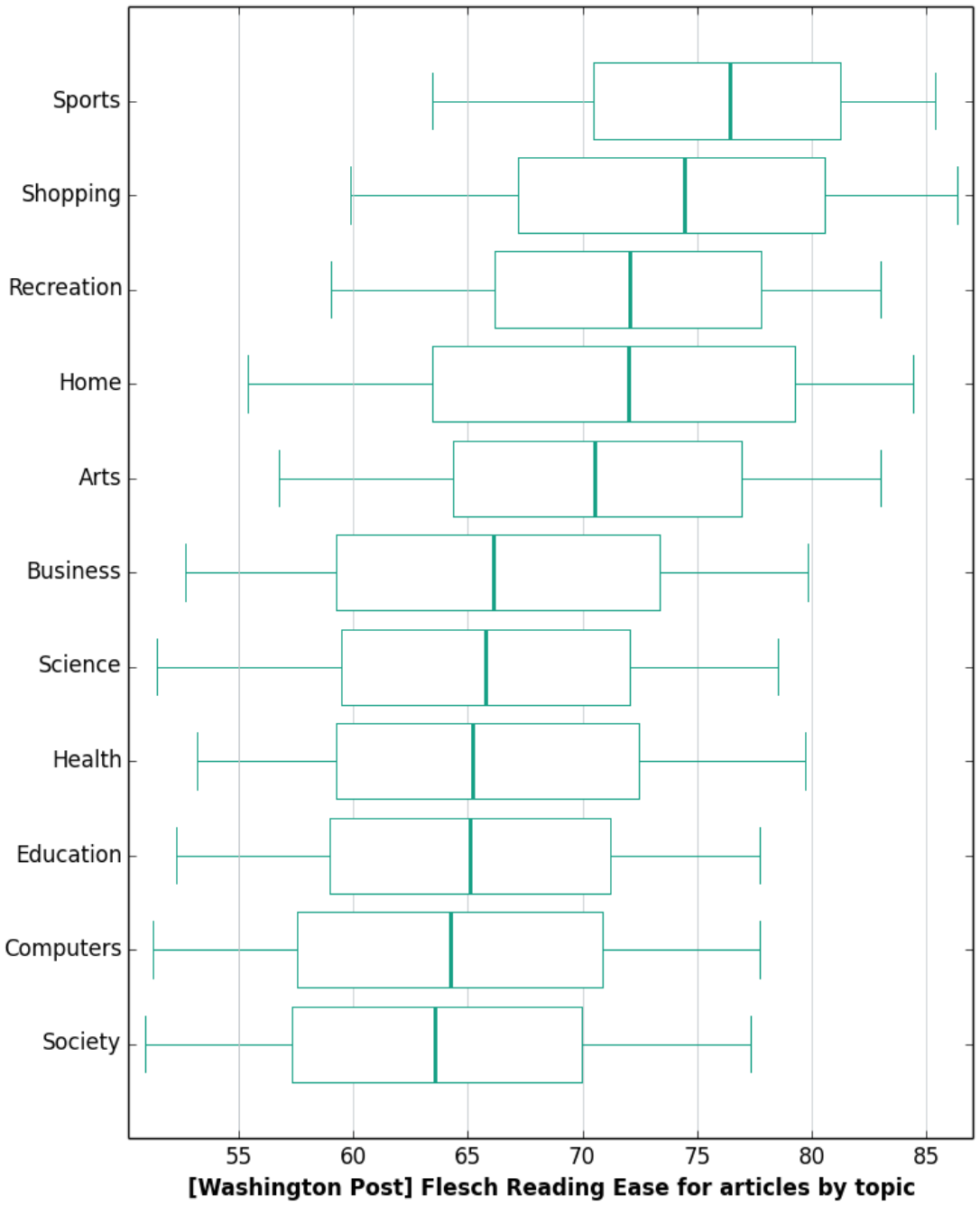
- **Flesch-Kincaid Reading Ease**
  - This measure is a standard way for many US agencies for evaluating technical documents.

$$206{,}835-1,015\ \left(\frac{total\ words}{total\ sentences}\right)-84,6\ \left(\frac{total\ syllables}{total\ words}\right)$$

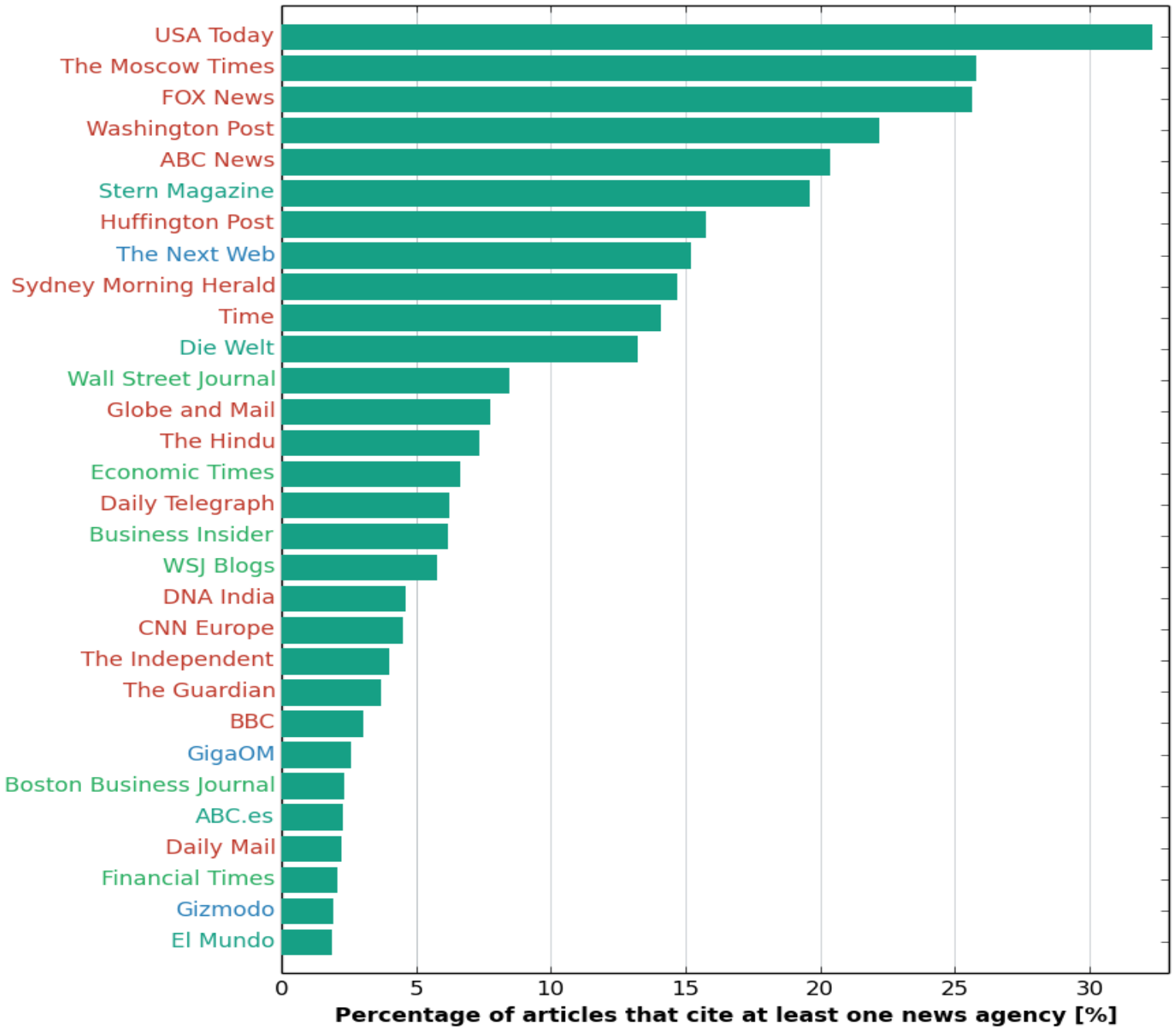
Readability in articles (**Flesch-Kincaid test**)



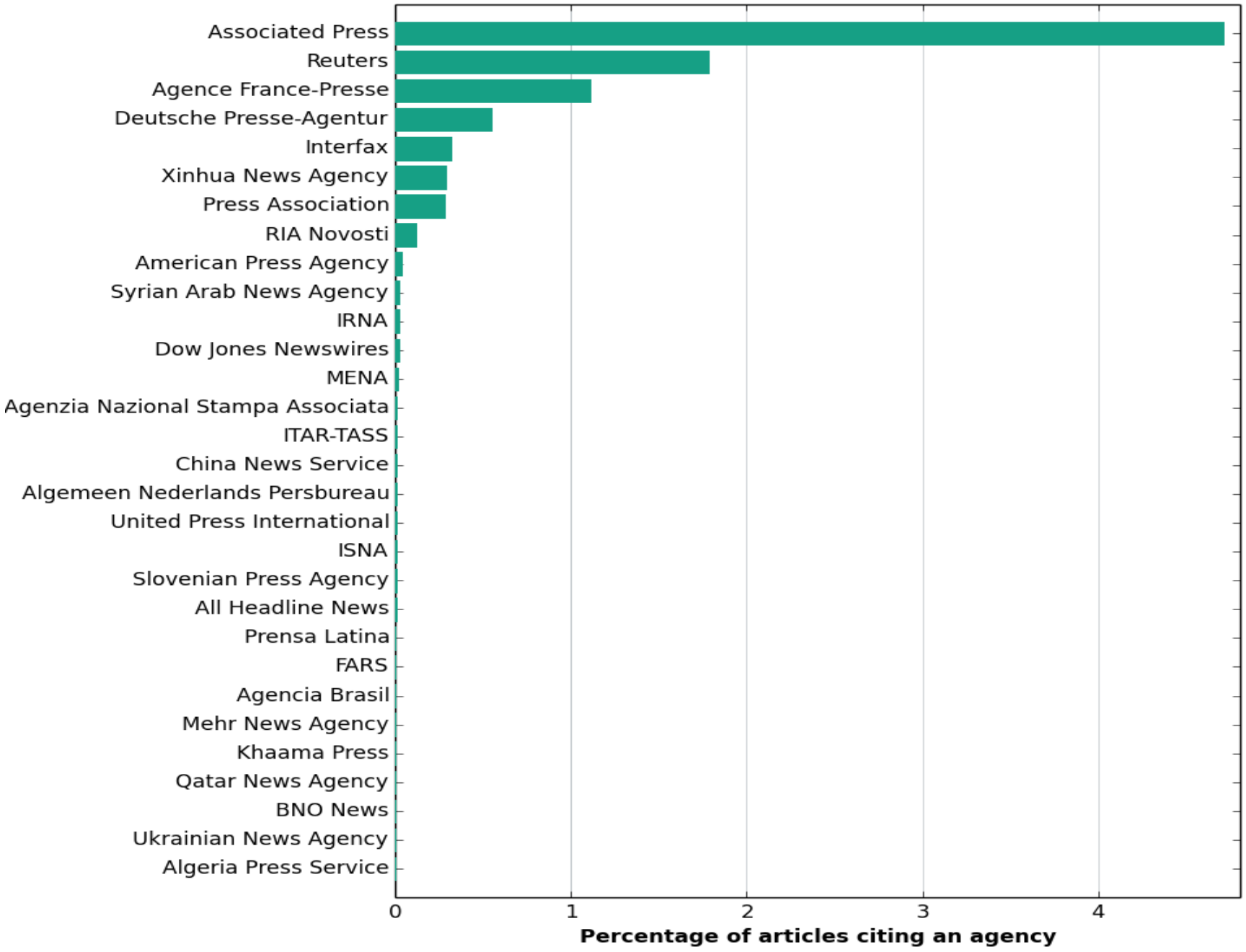
Readability in articles (Flesch-Kincaid test on Washington Post topics)



Publishers citing news agencies



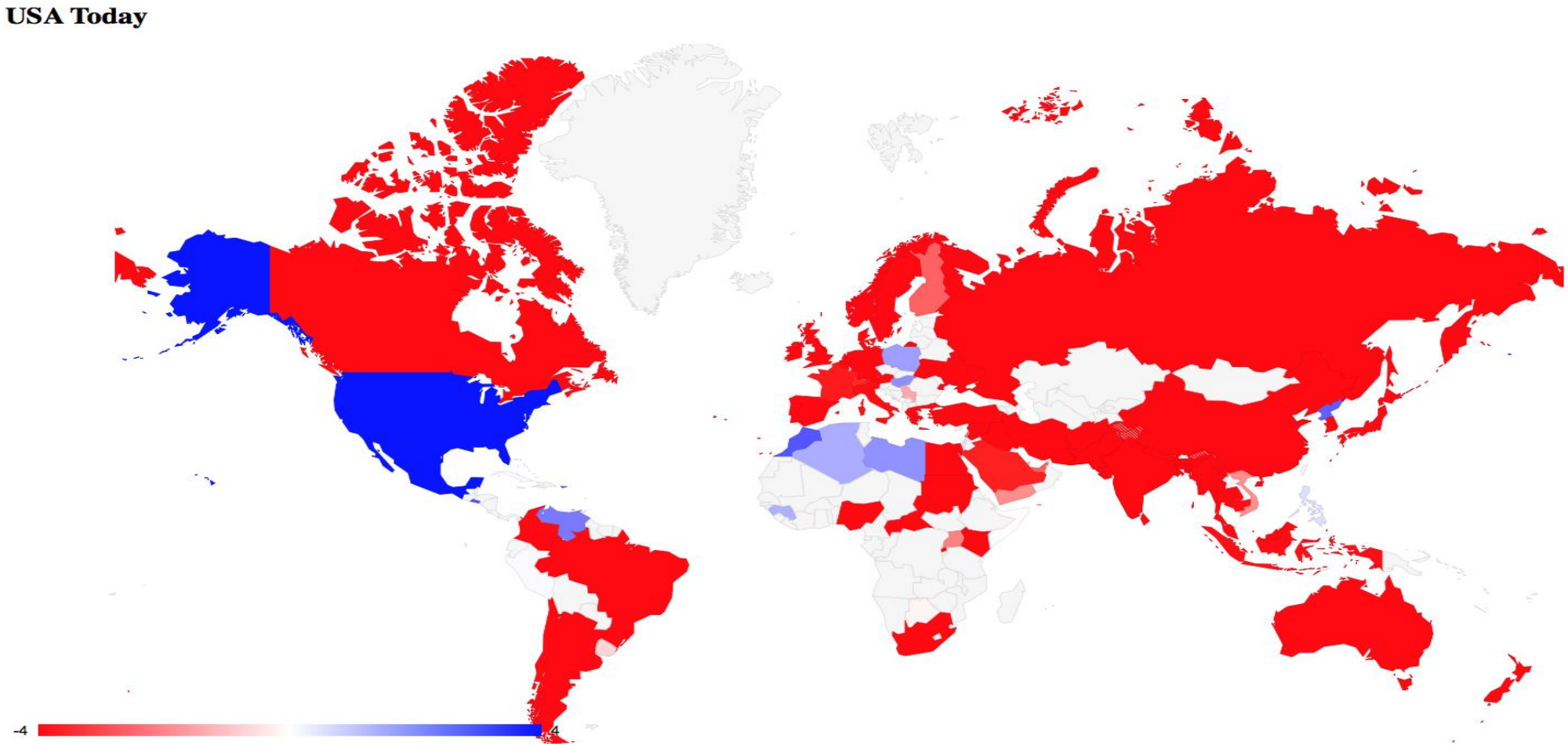
Agencies Citations



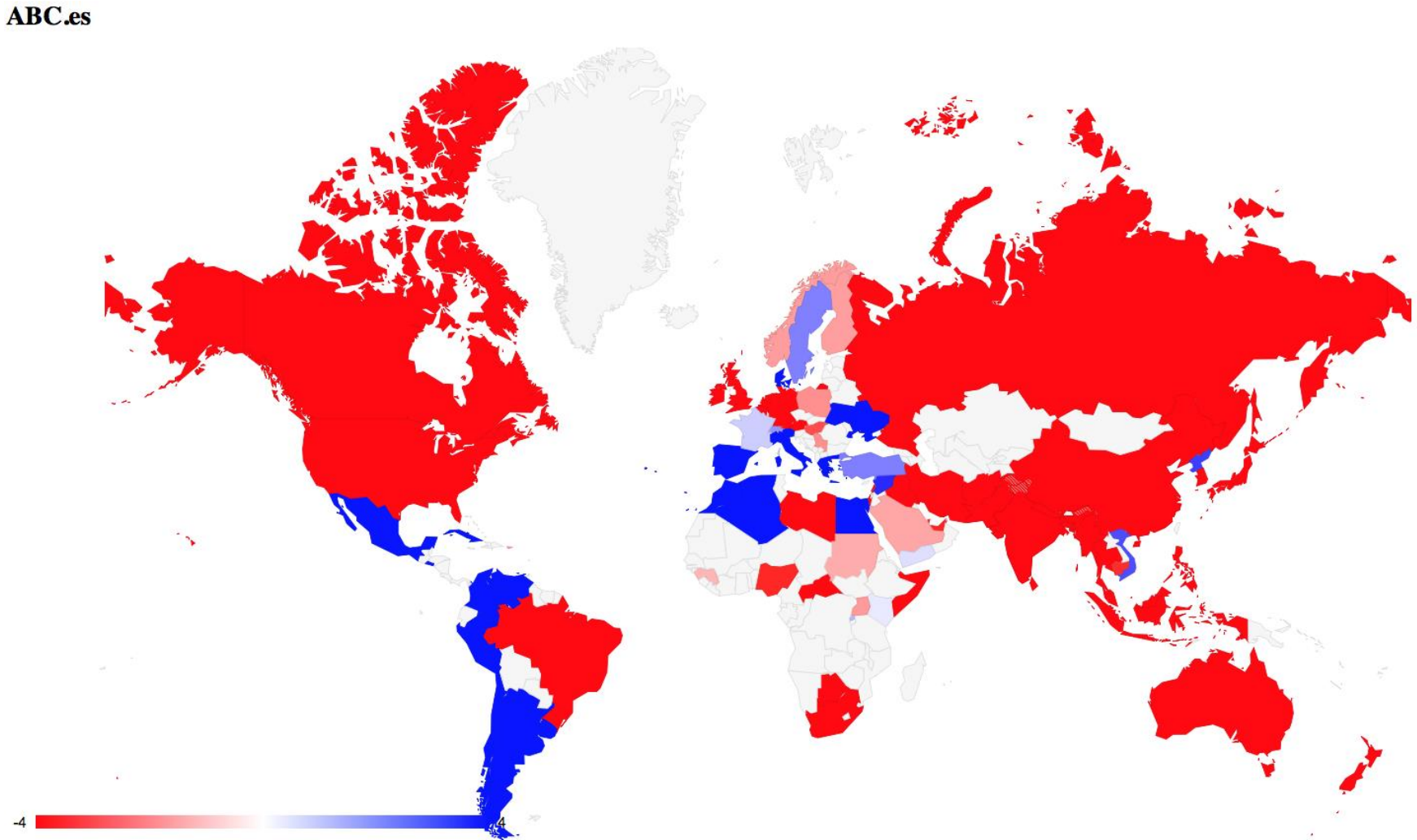
Geographical bias

- What events are being reported by news publishers?

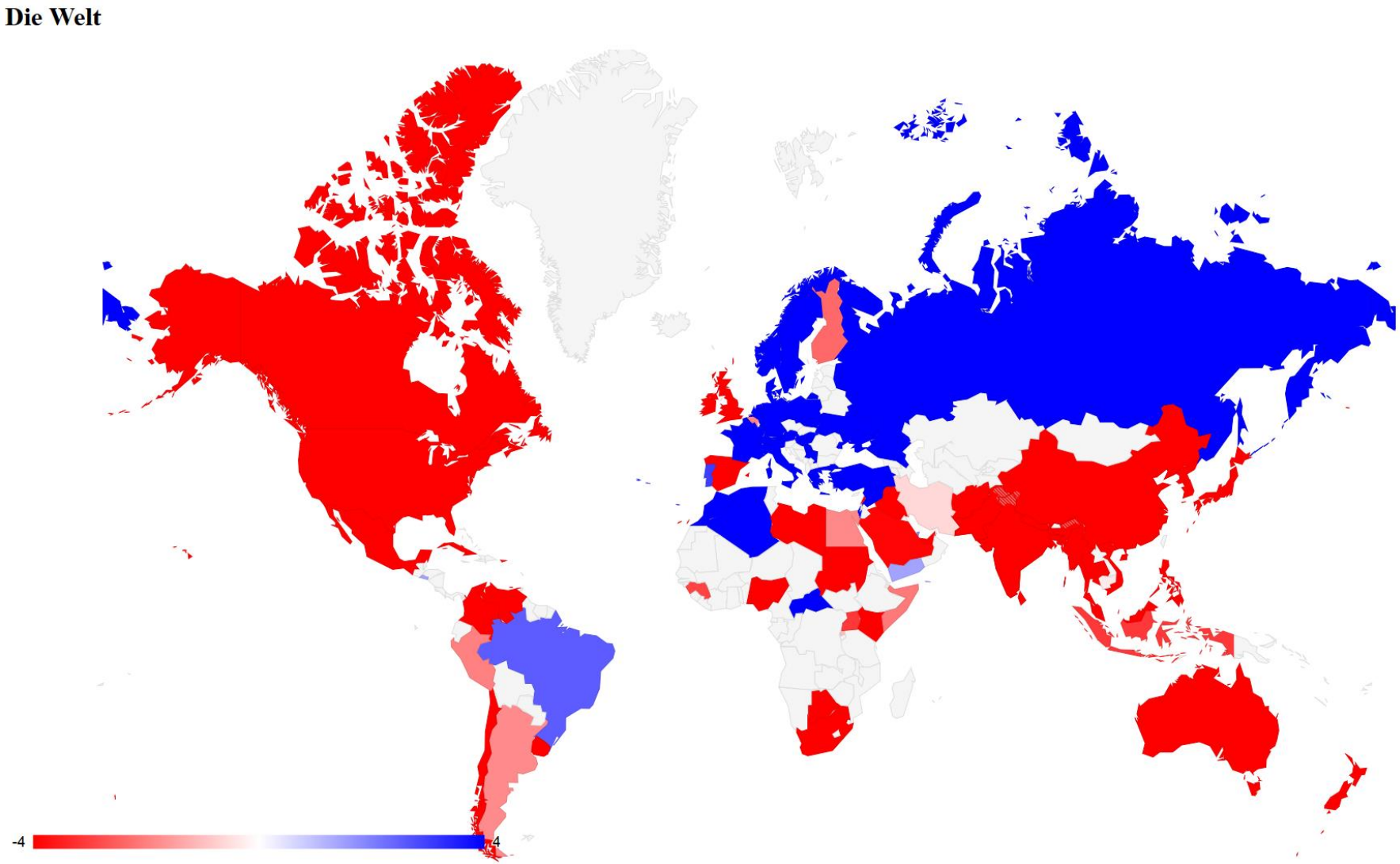
USA Today geographical coverage



ABC.es geographical coverage

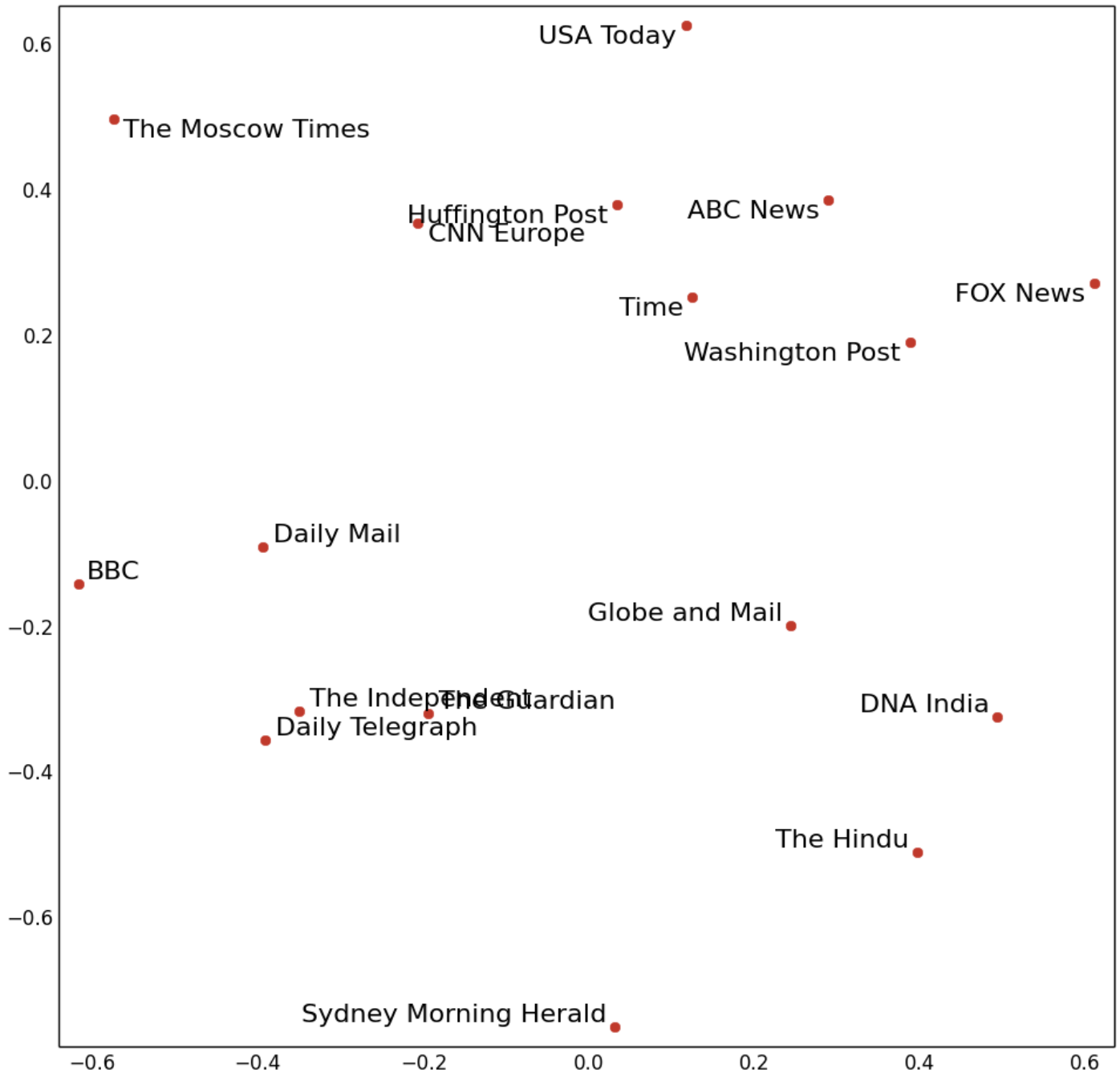


“Die Welt” geographical coverage



Separability between news sources

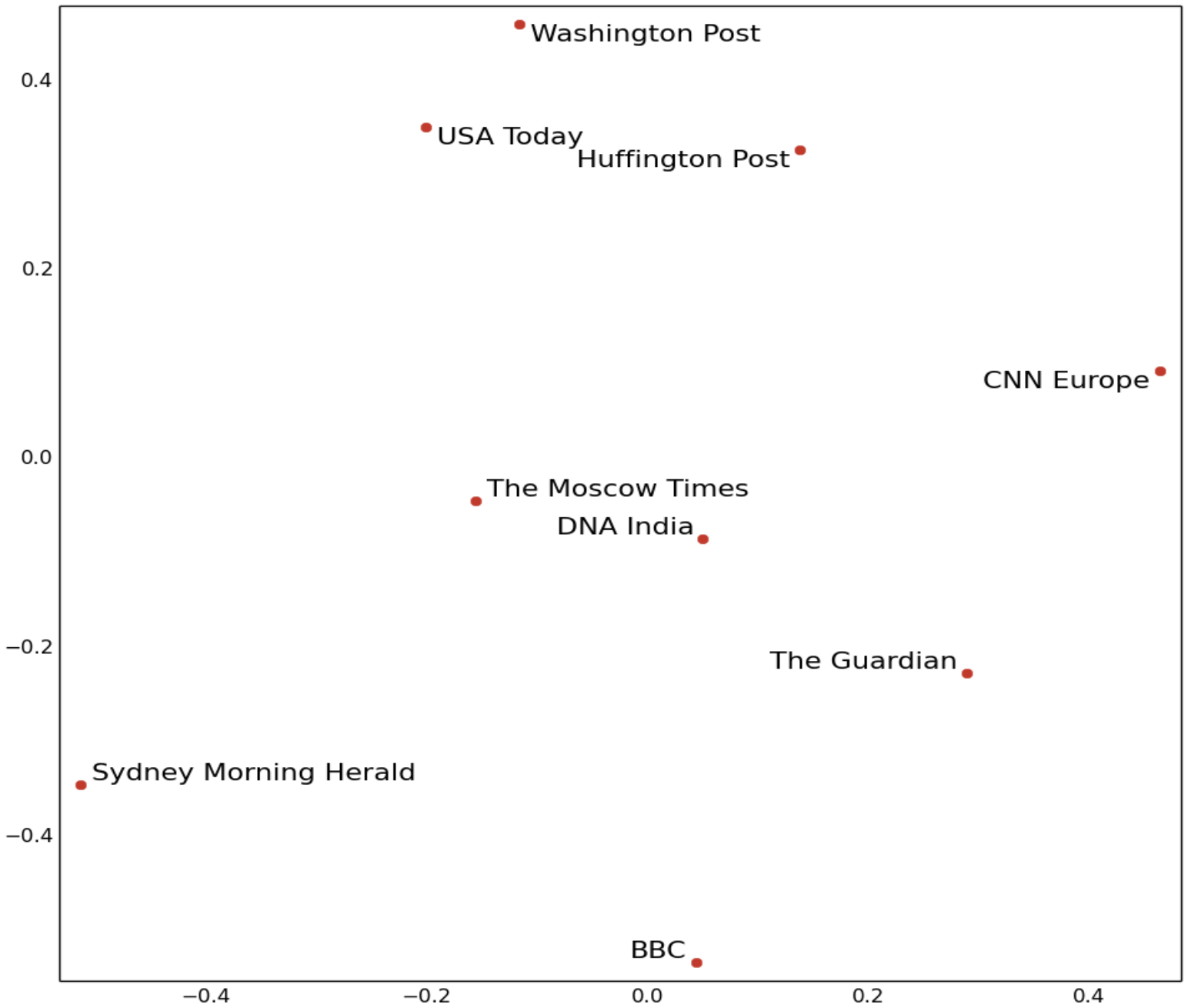
- How well we can determine which news publisher is the author of the article
- Tested on pairs of news publishers
- Used 10.000 articles from each news publisher
- Removed a custom-built set of tokens (like huffpost, mailonline, ...)
- Used keywords as features, used 1.000 top ranking features
- Linear SVM
- Used 70% articles for training, 30% for testing
- Use accuracy as a distance measure → use MDS for plotting


USA Today
The Moscow Times
Huffington Post
CNN Europe
ABC News
FOX News
Time
Washington Post
Daily Mail
BBC
Globe and Mail
The Independent
The Guardian
Daily Telegraph
DNA India
The Hindu
Sydney Morning Herald
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6

News source	Keywords
Time	selfie, uber, rapper, minister, researchers, Hillary, sex, party, relationships
Washingon post	Maryland, redskins, Georgetown, Montgomery, reelection, capital, Arlington, officials, capitol
FOX News	Jerusalem, Christians, minister, dynasty, officials, prosecutors
Daily mail	Mr, uk, centre, behavior, football, Scotland, British, premier, Cameron
BBC	Correspondents, Belfast, council, Scotland, UK, Wales, Ireland
CNN	Editor, commentary, opinions, affiliate, according, London, official, Pakistan
Sydney Morning Herald	Australian, Australia, Melbourne, Fairfax, Canberra, rugby, Queensland, Brisbane,
Moscow times	Russian, Russia, Interfax, soviet, Barack, Petersburg, Siberia, Sergei, Kremlin, Duma, Yury, Oleg
DNA India	Mumbai, indian, singh, Bollywood, Kapoor, cricket, crore, Bangalore, lakh, Pune, Kumar, Tata, Gandhi

Separability on reports about Ukraine

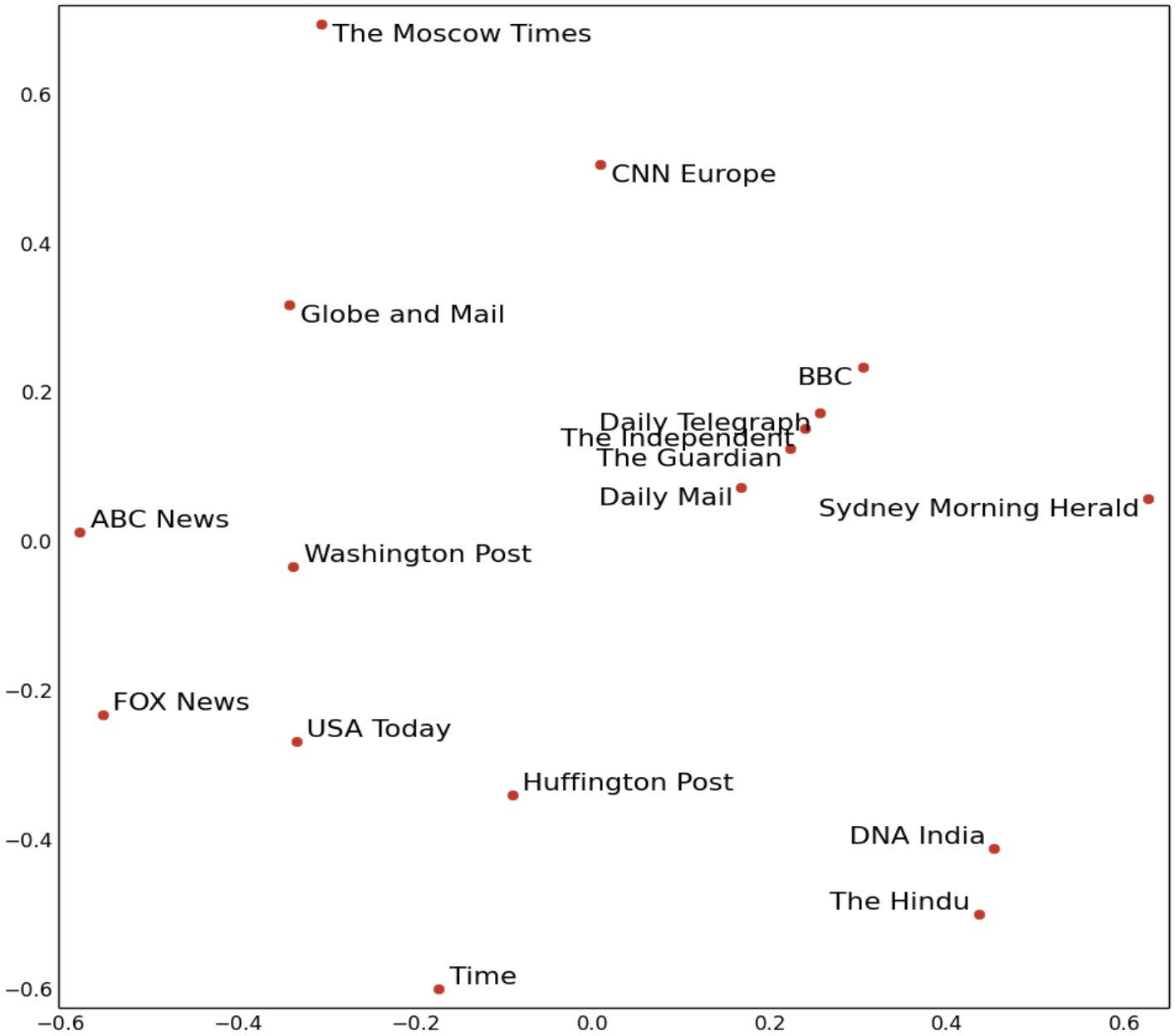
- Analyze articles reporting about Russia-Ukraine incident
- Selected articles
  - Mention Ukraine and Russia
  - between Jan 15th, 2014 and June 1st, 2014
  - Categorized into appropriate Dmoz category
- Selected publishers that reported about 500 articles


Washington Post
USA Today
Huffington Post
CNN Europe
The Moscow Times
DNA India
The Guardian
Sydney Morning Herald
BBC
0.4
0.2
0.0
−0.2
−0.4
−0.4
−0.2
0.0
0.2
0.4

News source	Keywords
Non-Russian sources	troops, allies, invasion, aggression, excellency, insurgents, investigation, defense, ethnic, seized
The Moscow Times	overthrow, authorities, western, election, khodorkovsky, detained, annexation, pressure, seizure, fanning, imposed, domestic

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


The Moscow Times
CNN Europe
Globe and Mail
BBC
Daily Telegraph
The Independent
The Guardian
Daily Mail
Sydney Morning Herald
ABC News
Washington Post
FOX News
USA Today
Huffington Post
DNA India
The Hindu
Time
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6

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 od 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

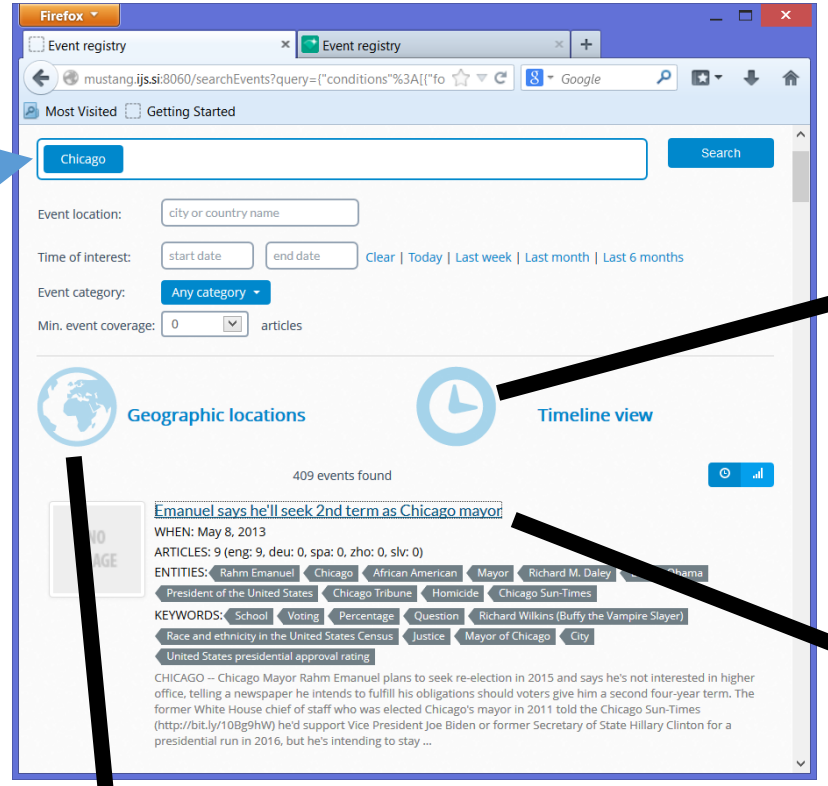
Feature vector event representation

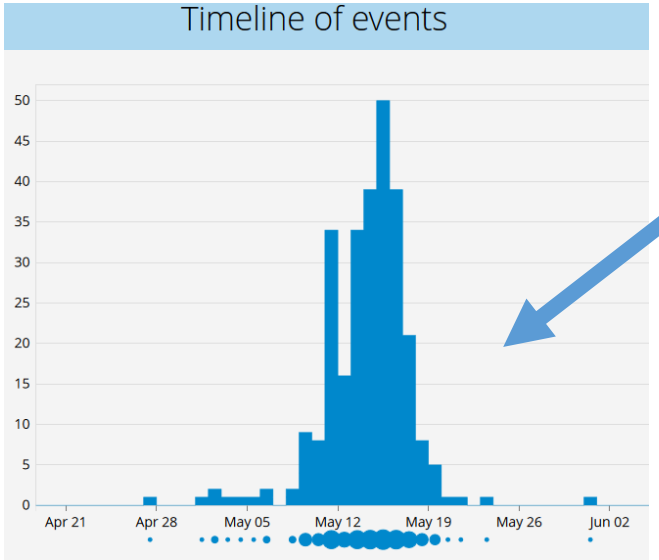
- Feature vectors easily extractable from news documents:
  - **Topical dimension** – what is being talked about? (keywords)
  - **Social dimension** – which entities are mentioned? (named entities)
  - **Temporal aspect** – what is the time of an event? (temporal distribution)
  - **Geographical aspect** – 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

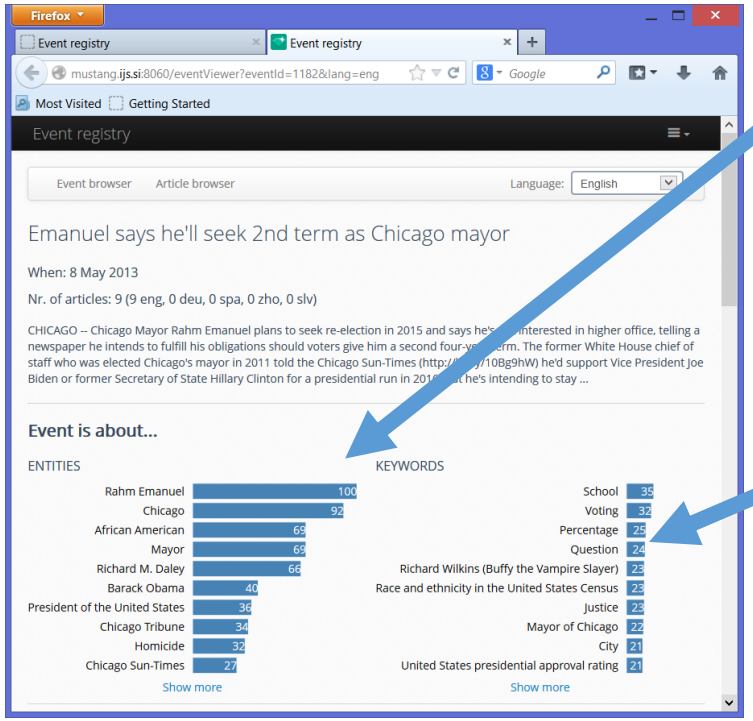
Example of “feature vector” event representation: Event Registry “Chicago” related events

Query:
“Chicago”



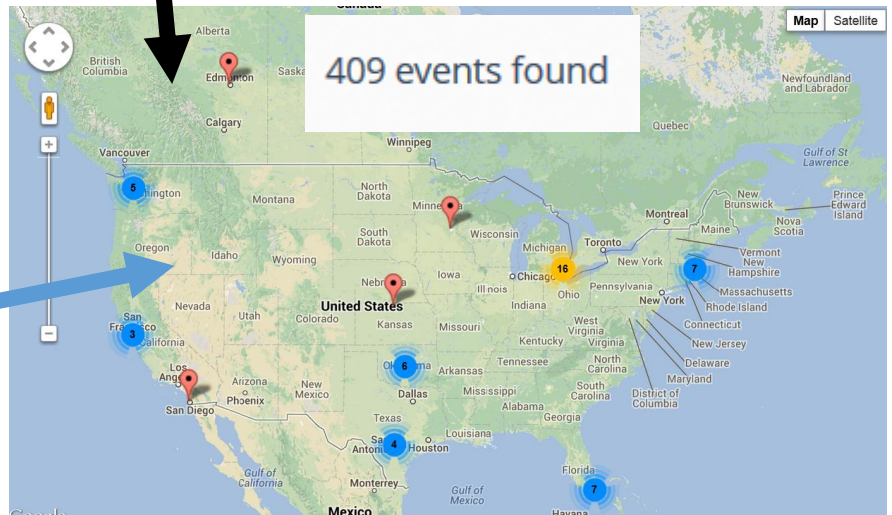


When?
(temporal distribution)



Who?
(named entities)

What?
(keyword/ topics)



Where?
(geography)

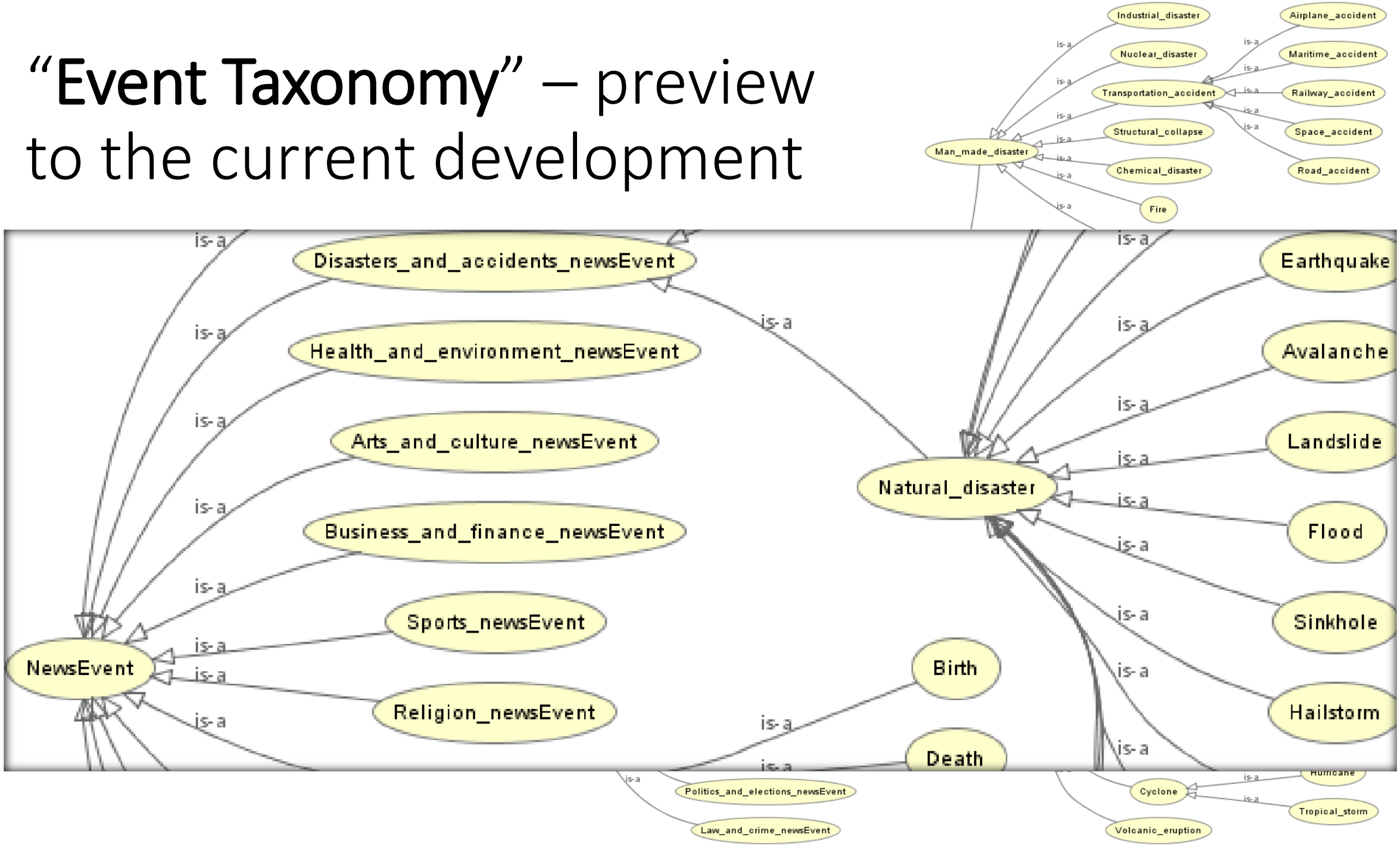
Structured event representation

- 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**

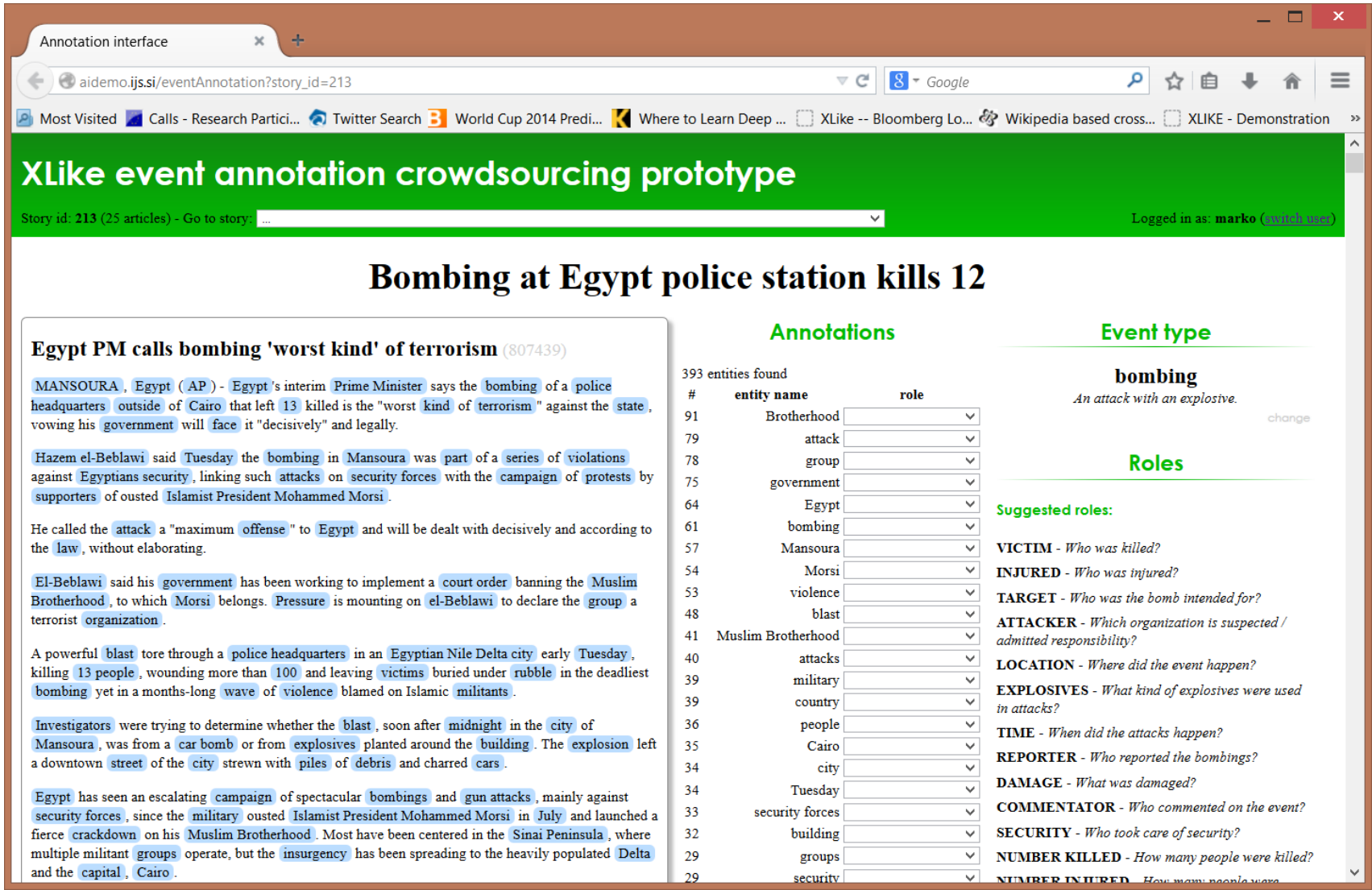
Date	11 March 2011 (3 years, 5 months and 2 weeks)
Origin time	14:46:24 JST (UTC+09:00)
Duration	6 minutes[1]
Magnitude	9.0 (M_w)[2][3]
Depth	30 km (19 mi)
Epicenter	38.322°N 142.369°E
Type	Megathrust earthquake
Areas affected	Japan (primary) Pacific Rim (tsunami, secondary)
Total damage	Tsunami wave, flooding, landslides, fires, building and infrastructure damage, nuclear incidents including radiation releases
Max. intensity	IX - Violent
Peak acceleration	2.99 *g*
Tsunami	Up to 40.5 m (133 ft) in Miyako, Iwate, Tōhoku
Landslides	None
Foreshocks	7
Aftershocks	11,044 (as of 3 August 2014)[4]
Casualties	15,889 deaths,[5] 6,152 injured,[6] 2,609 people missing[7]

“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

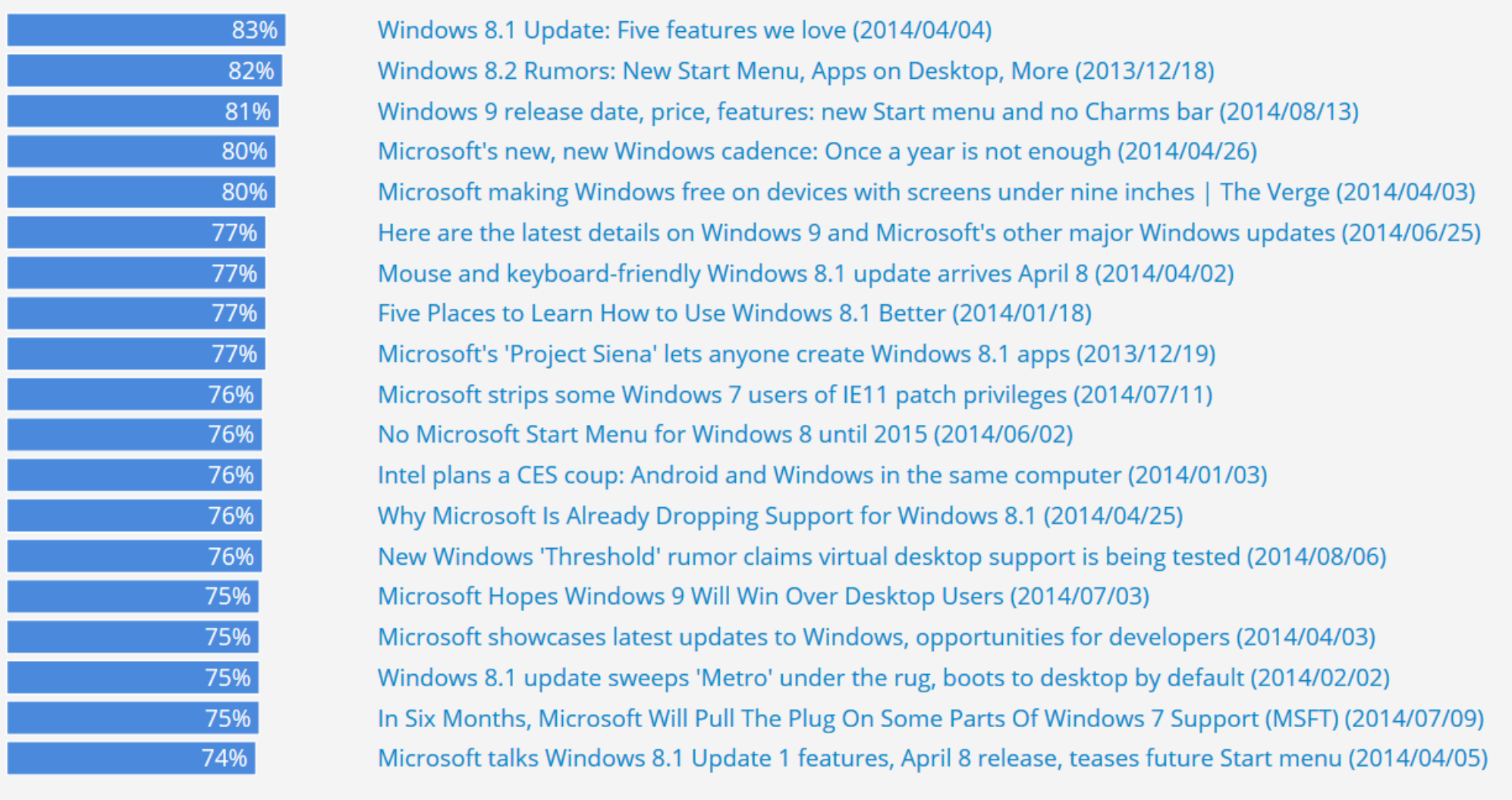
ENTITIES: Microsoft | Twitter | Google | Apple Inc. | Computerworld | Gartner | Facebook | Universal Studios | Nokia | United States

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

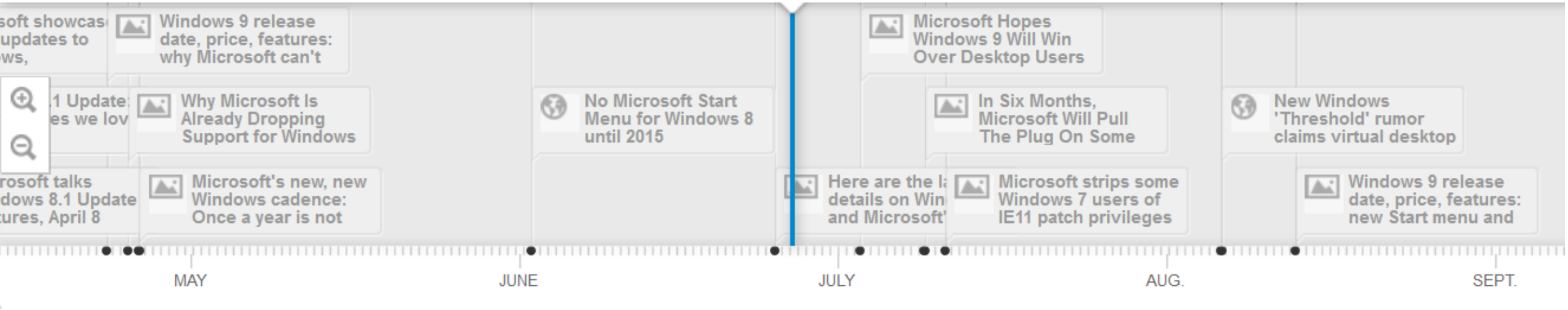
ARTICLES: 239 (239 eng, 0 ger, 0 spa, 0 chi, 0 slo)

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....

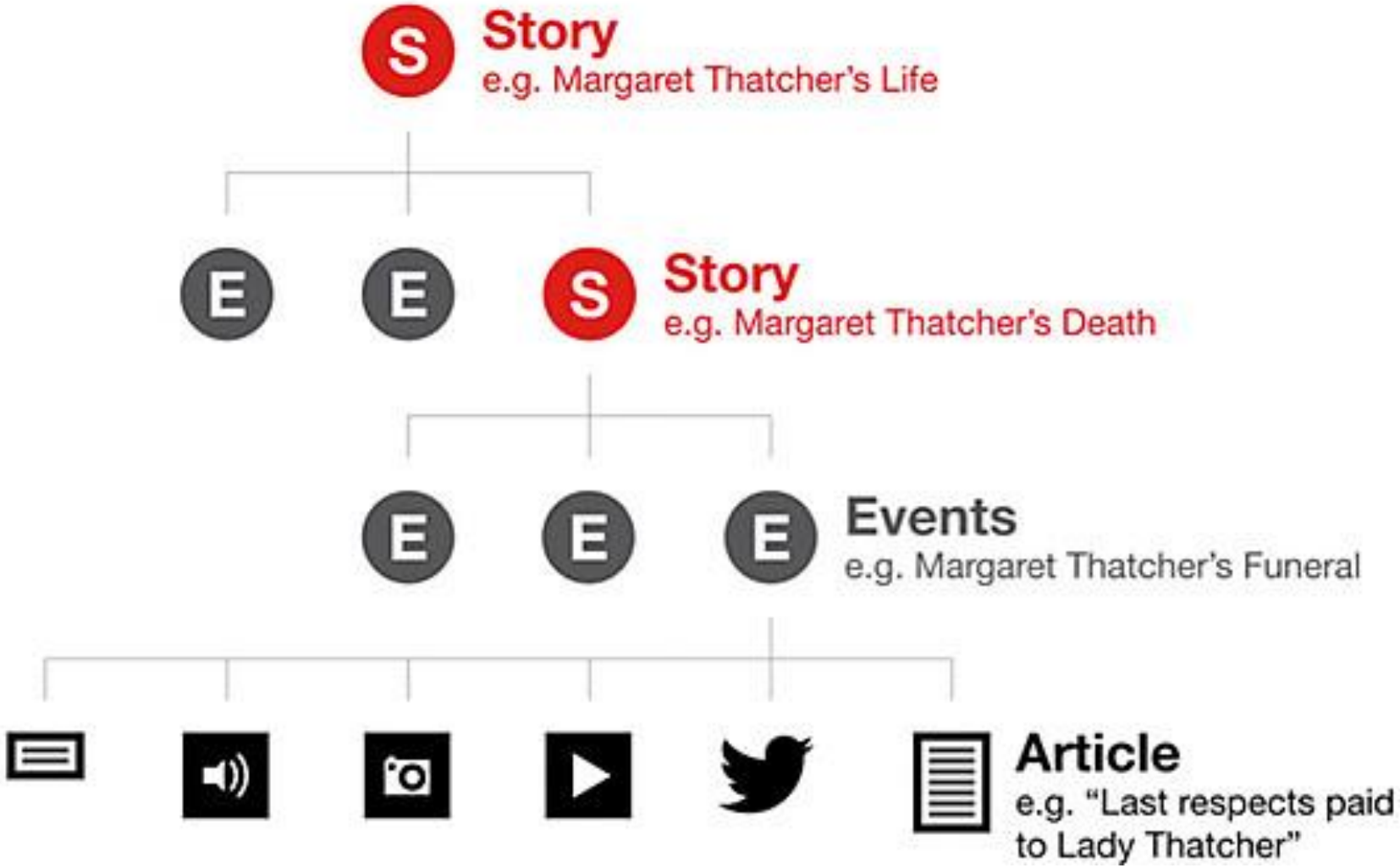
Similar events example: similar events to Microsoft Windows 9 event



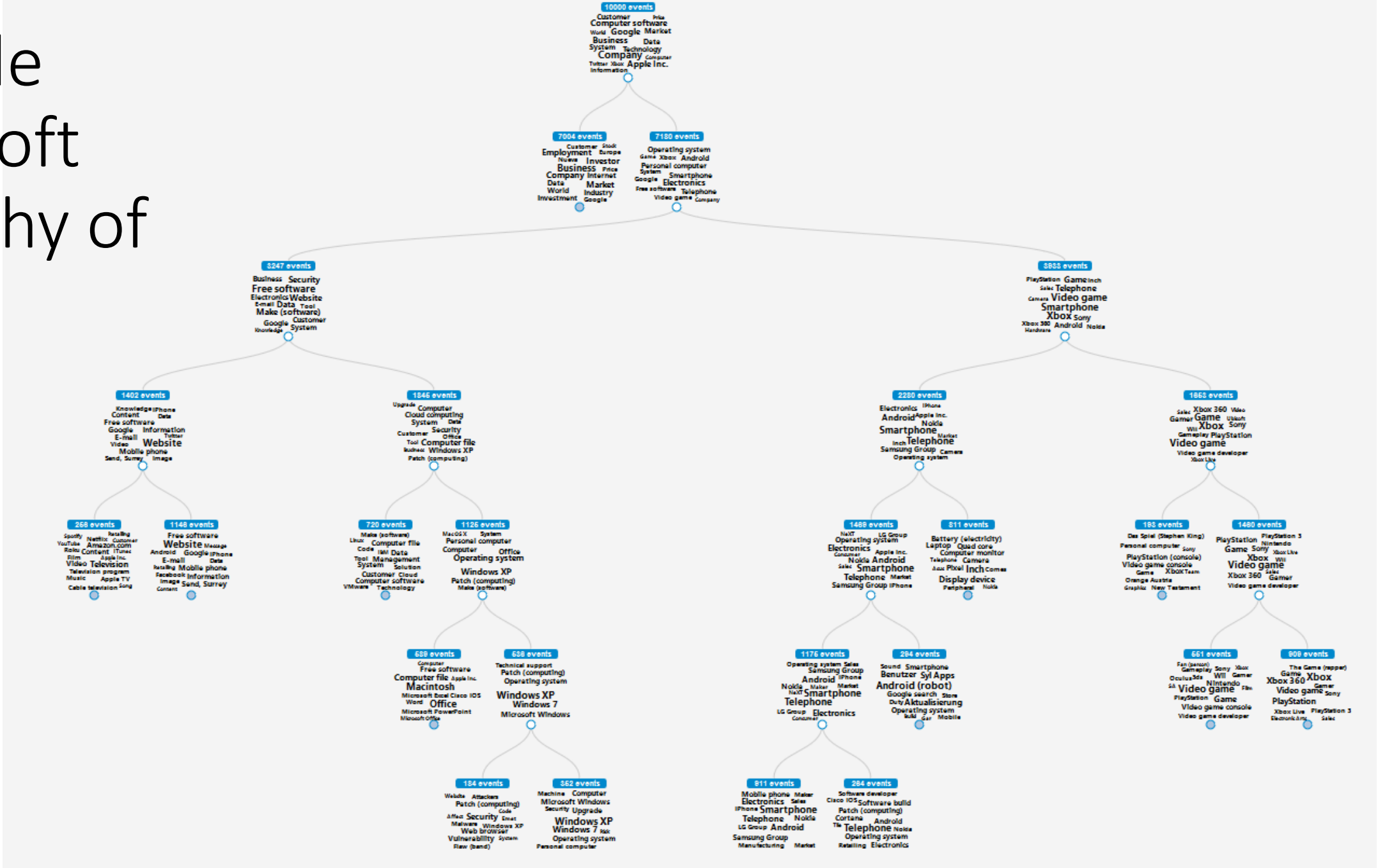
Event sequence identification



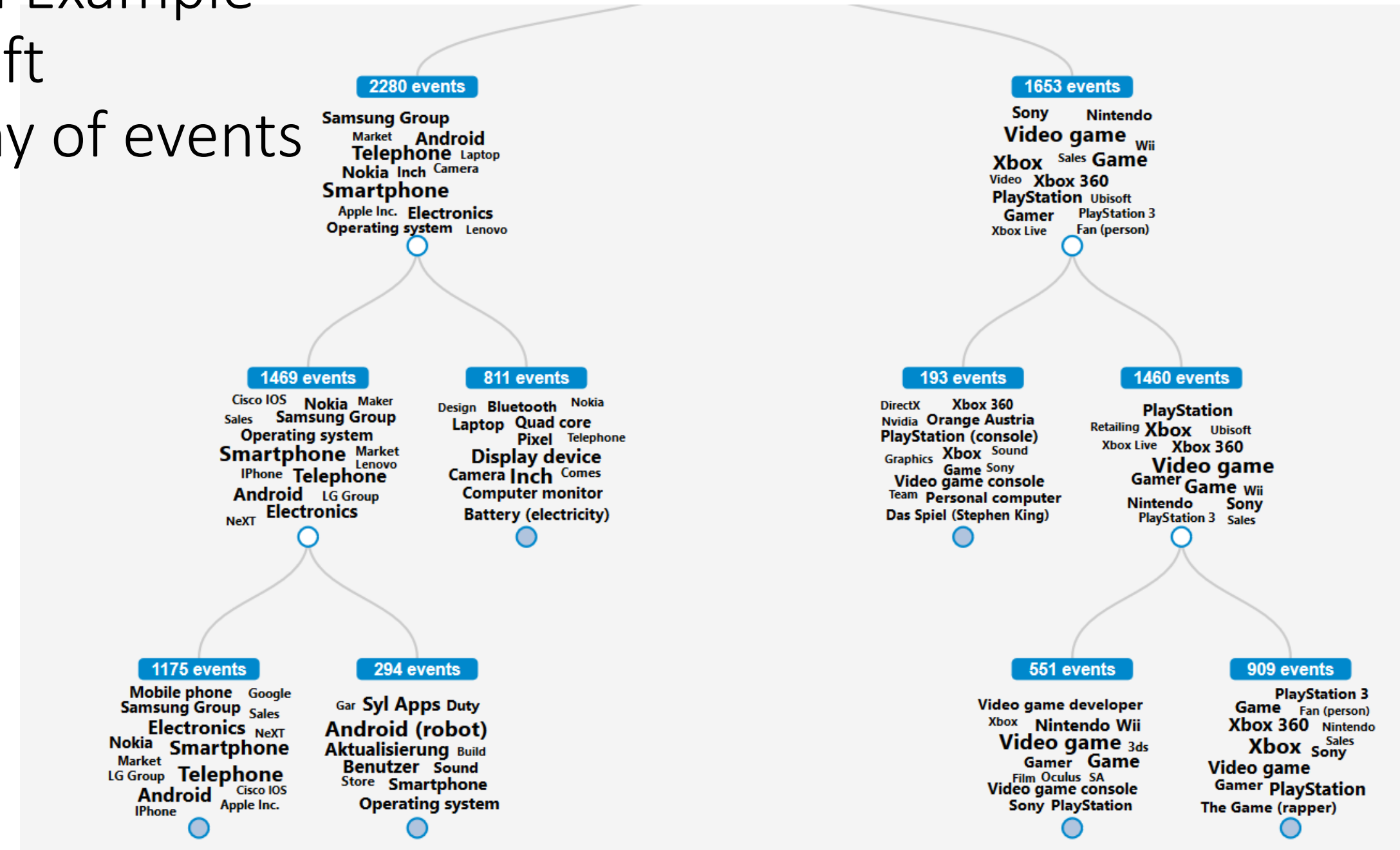
Hierarchy of events



Example Microsoft hierarchy of events



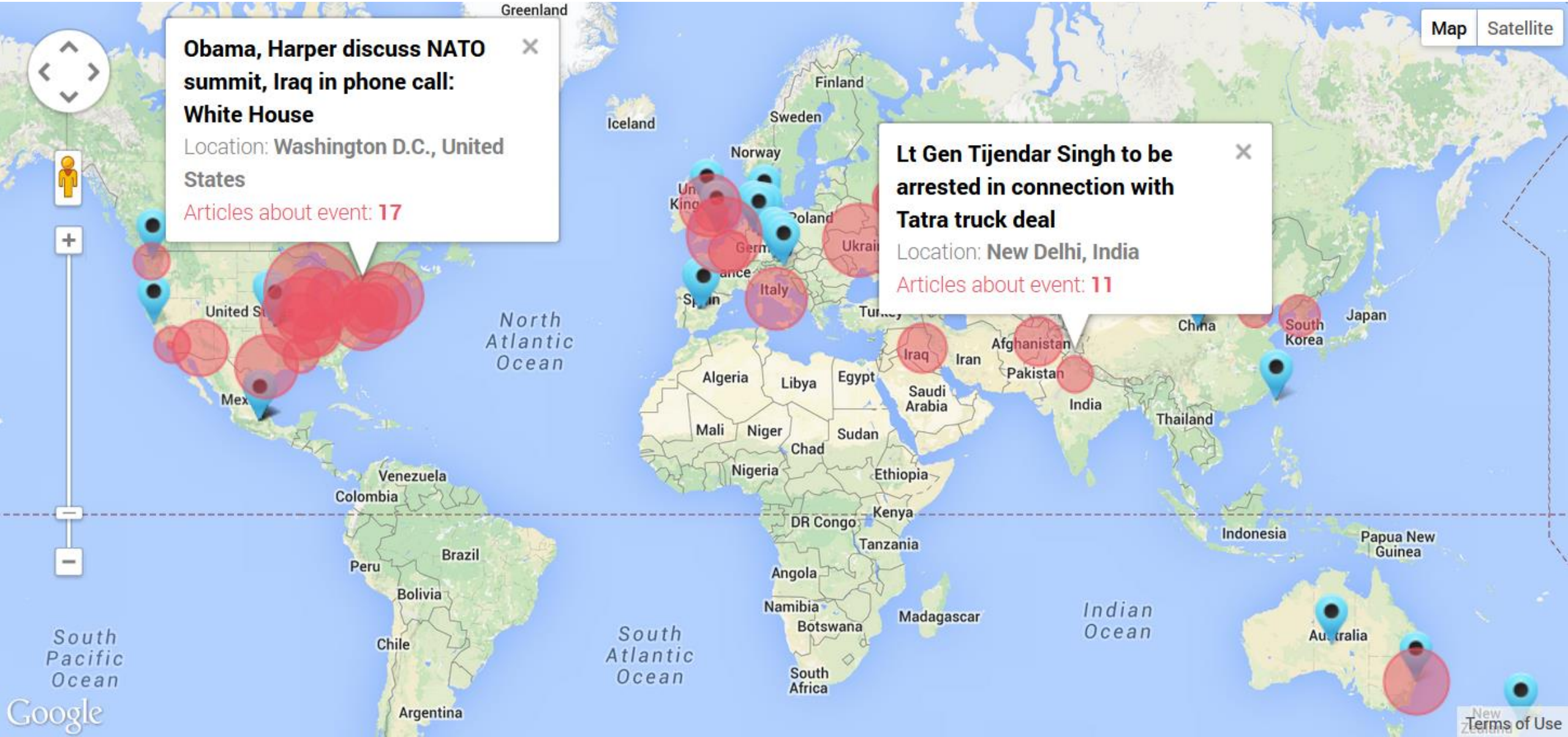
Zoom-in Example Microsoft hierarchy of events



Event Visualization

http://eventregistry.org/

Live Event tracking with http://EventRegistry.org/

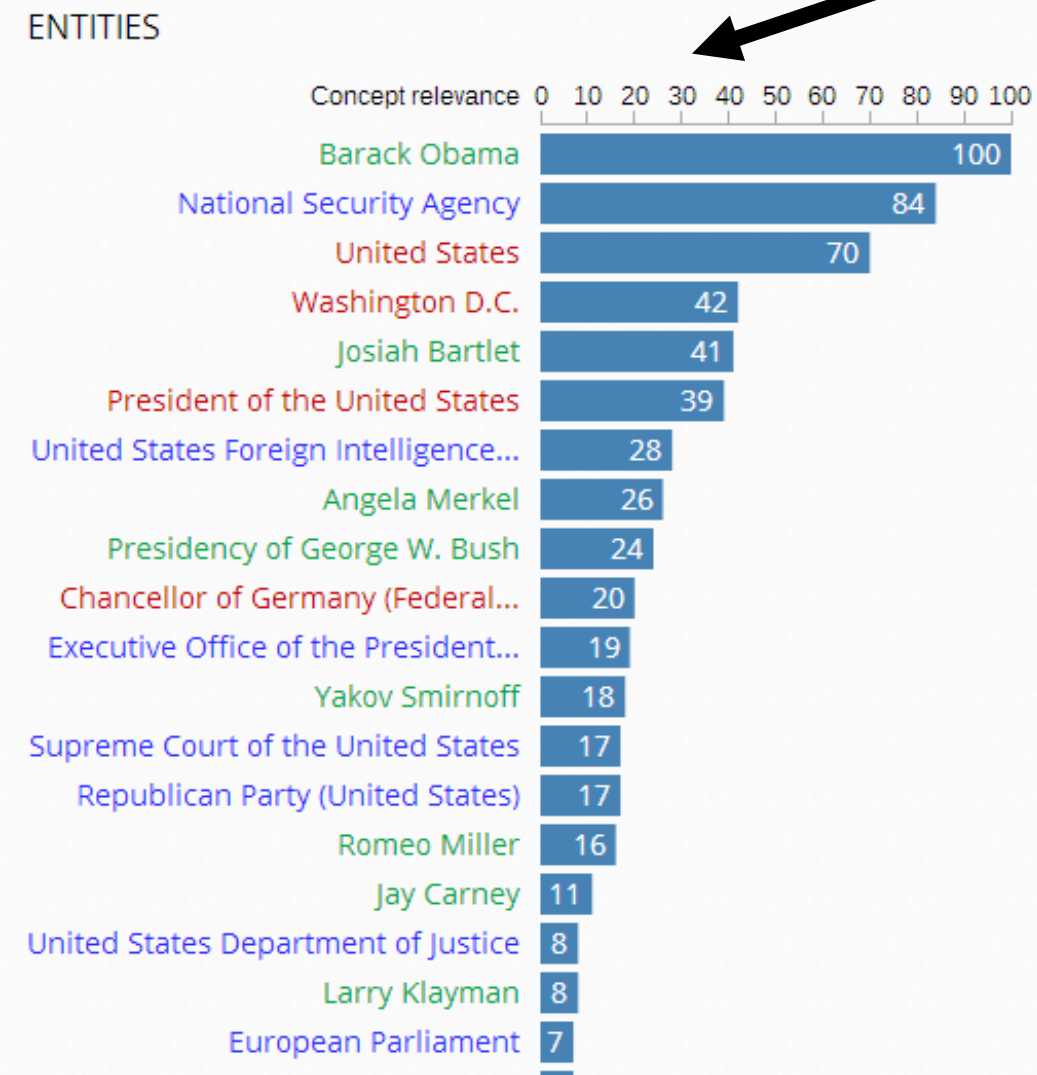


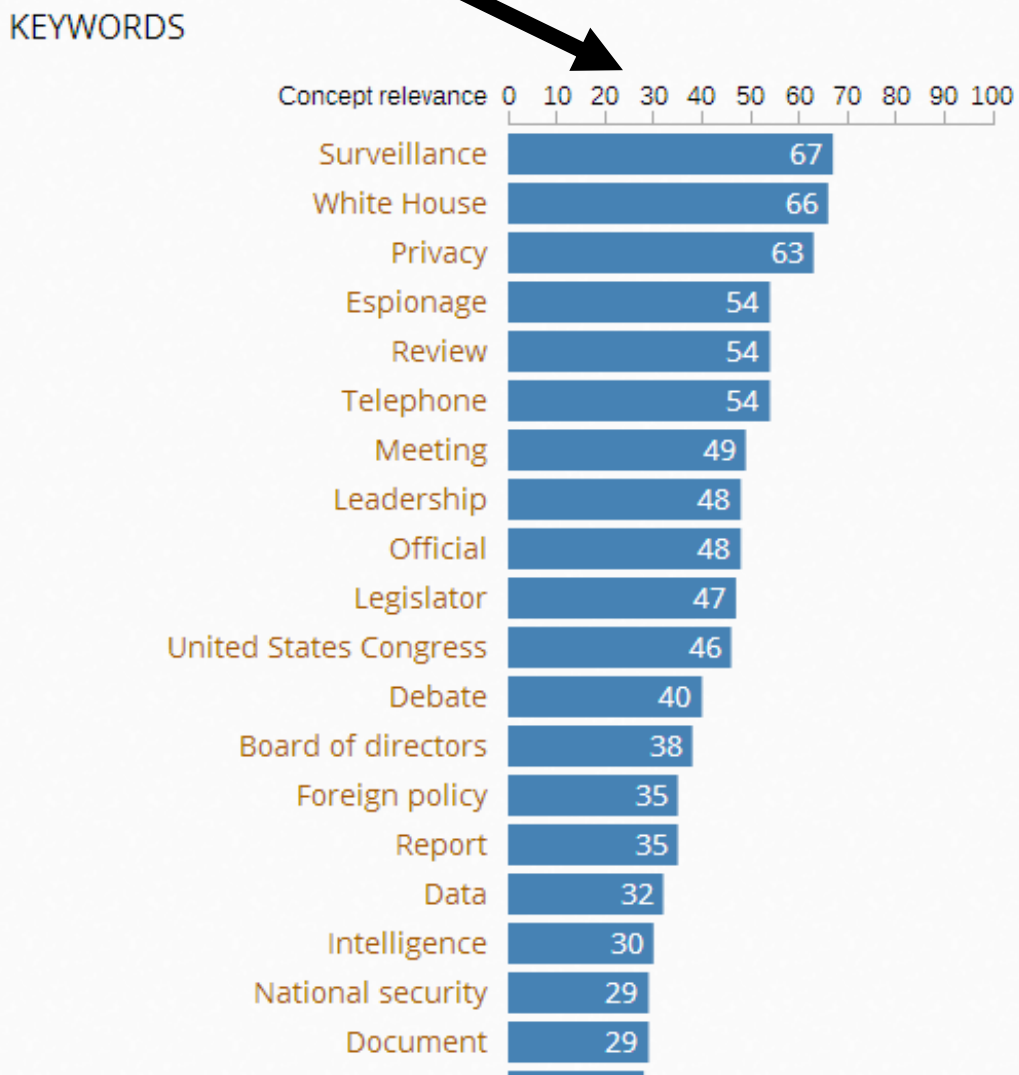
Representatives from tech companies are meeting with White House staff on Friday.

The White House...

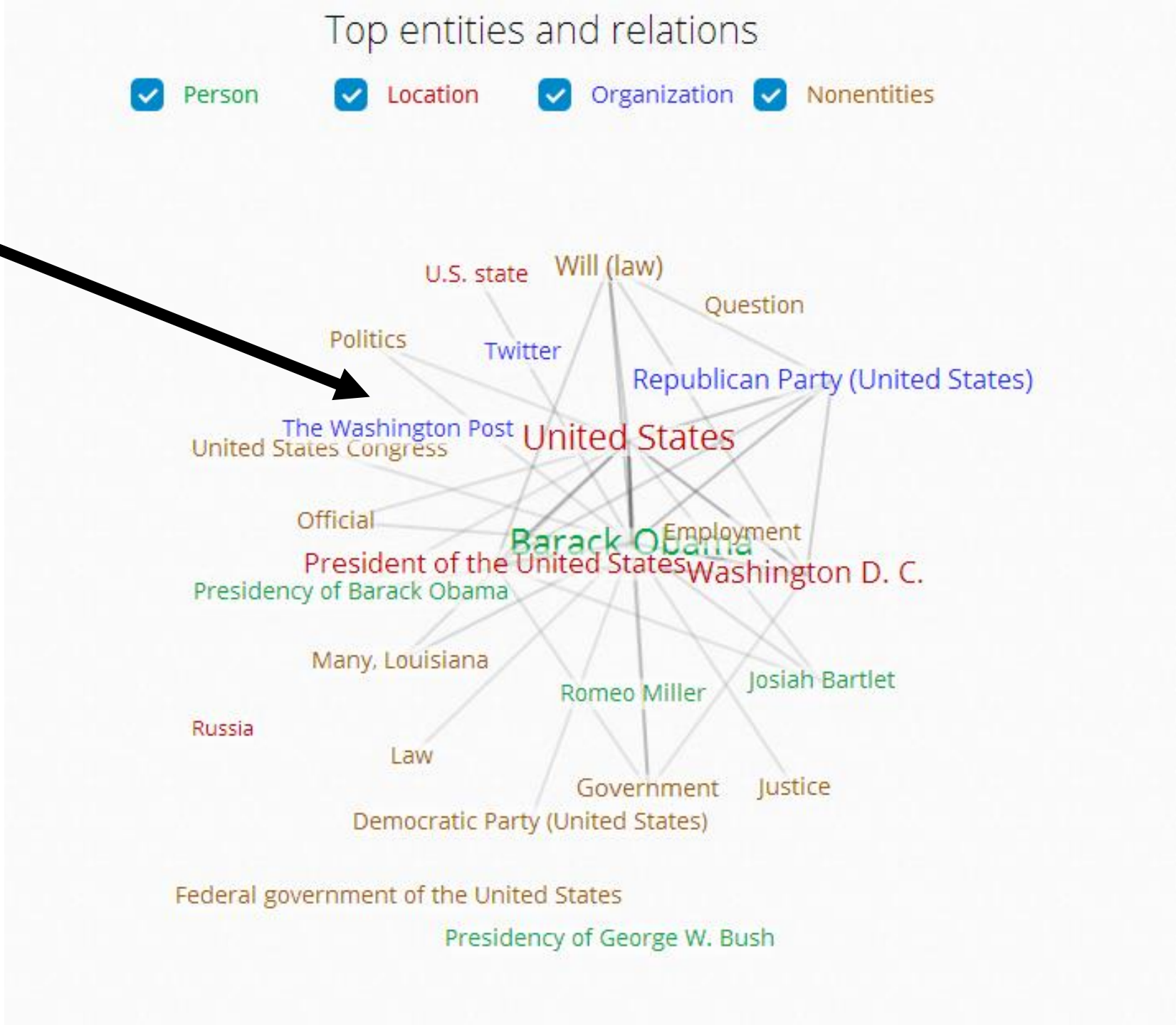
Nr. of articles: 105 (89 eng, 5 ger, 11 spa, 0 chi, 0 slo)

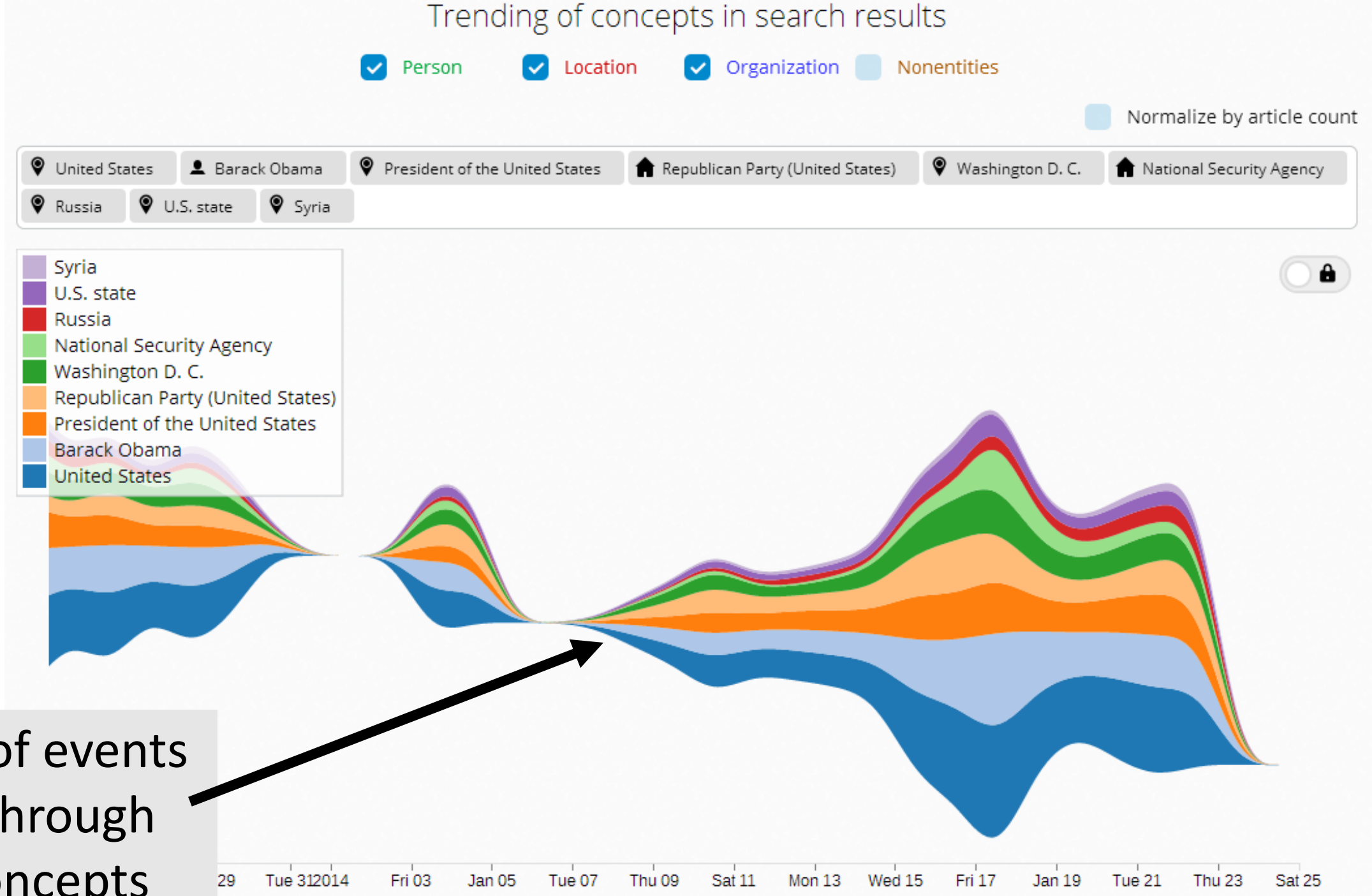
Event description through entities and Semantic keywords



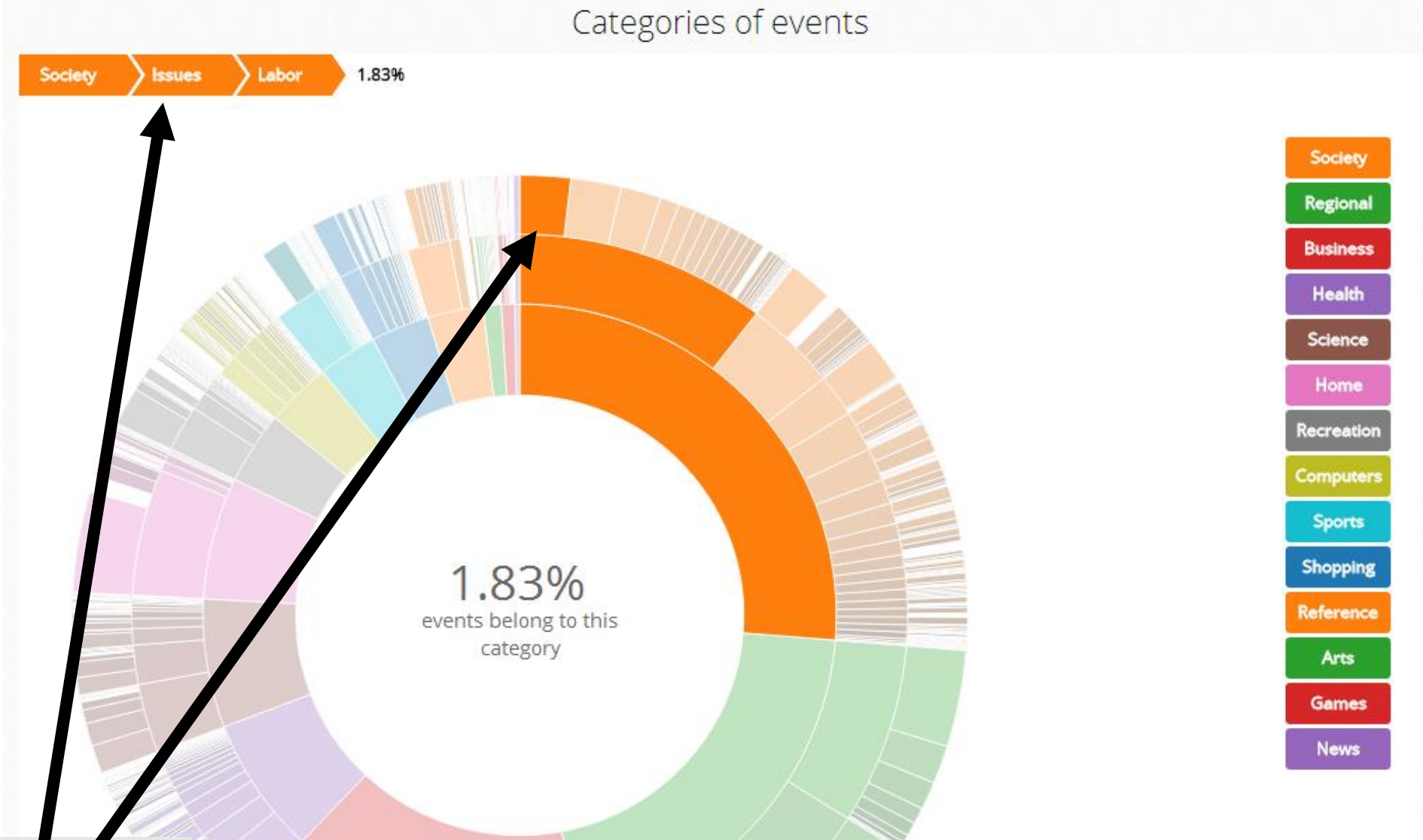


Collection of events described through Entity relatedness





Collection of events described through trending concepts



Collection of events described through three level categorization

Obama ponders limiting NSA access to phone records

9 Jan 2014

Washington D.C., United States

Washington -- President Barack Obama is expected to rein in spying on foreign leaders and is considering restricting National Security Agen
the government's surv

Obama could unveil hi
discuss his review with congressional lawmakers, while his top lawyer plans t meet with privacy groups.
Representatives from tech companies are meeting with White House staff on Frida

The White House...

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

Events identified across languages

Articles

89 eng | 11 spa | 5 ger | 0 chi | 0 slo

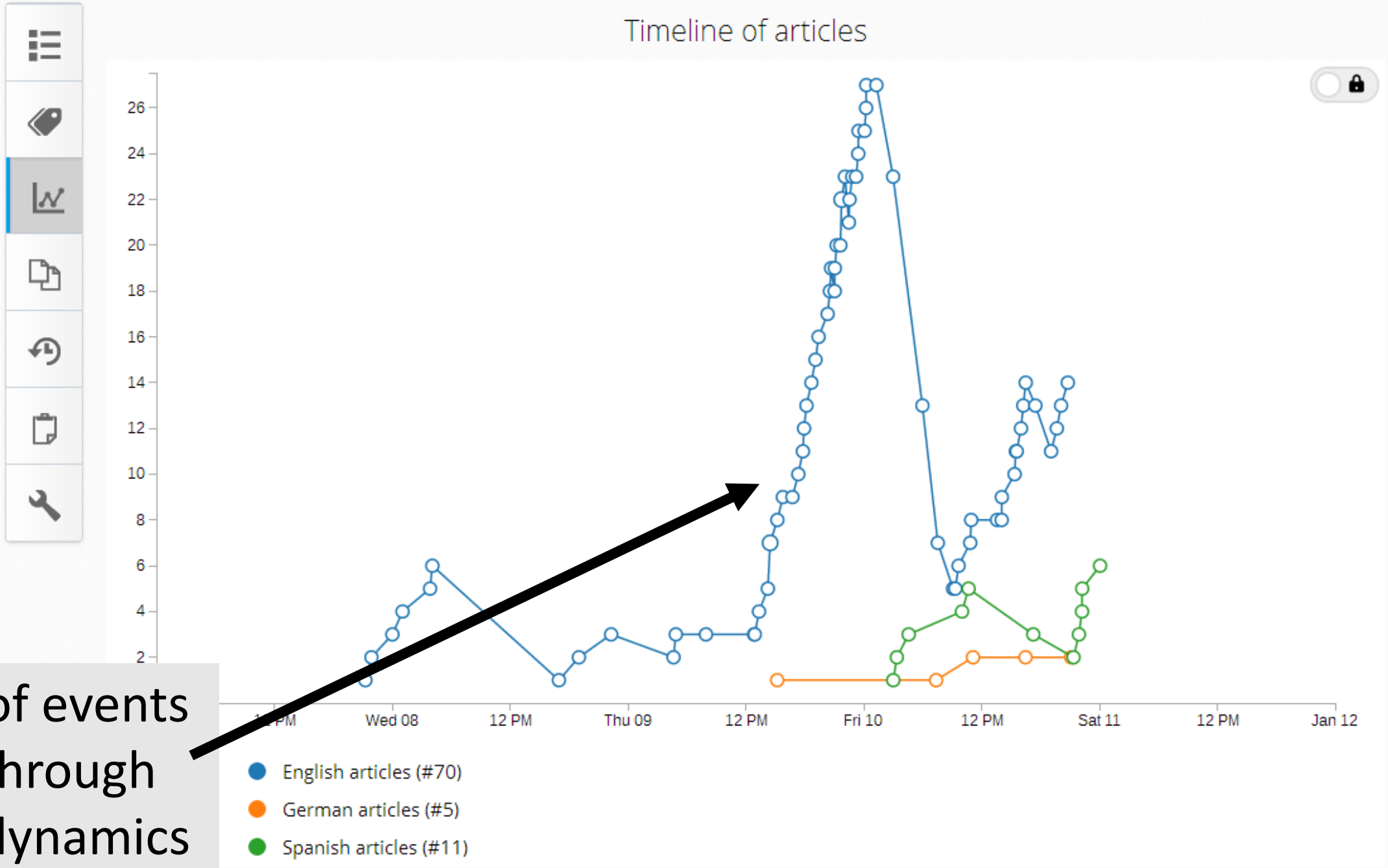
89 eng 11 spa 5 ger

Obama ponders limiting N
Washington -- President Barack Ol
National Security Agency access tc
government's surveillance progra
Obama could unveil his highly ant
congressional lawmakers, while his top lawyer plans to meet with privacy groups. ...
Jan. 9, 12:51
W.DETROITNEWS.COM
view with

Obama likely to adjust US spying on foreign leaders, nearing decision on intelligence changes
Jan. 9, 07:57
CKGL AM 570
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
Jan. 10, 14:04
WWW.IMPACTOUSA.COM
President Barack Obama speaks at a press conference on Dec. 20, 2013. (Pablo Martinez Monsivais/AP)

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



Collection of events described through Reporting dynamics

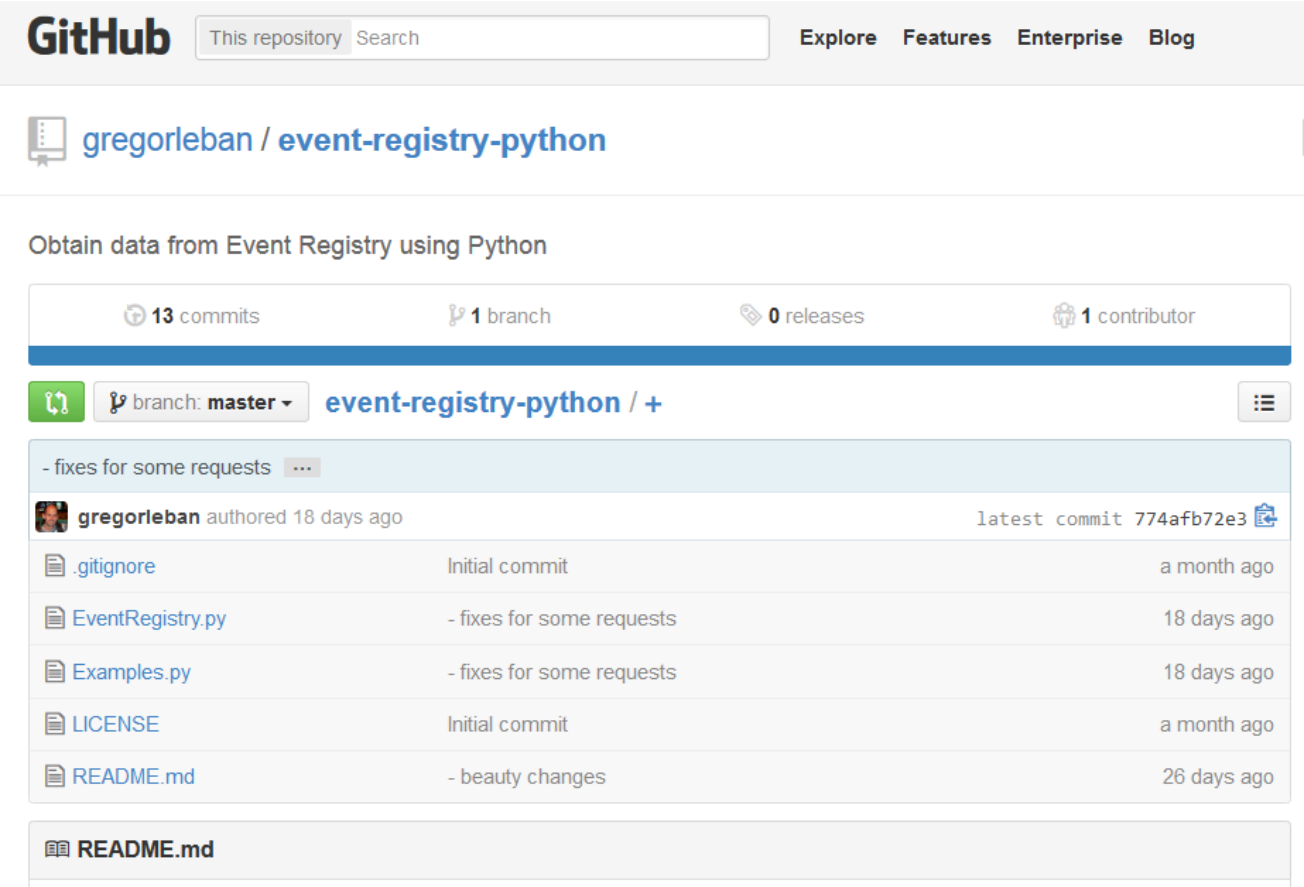
Nr. of articles: 105 (89 eng, 5 ger, 11 spa, 0 chi, 0 slo)

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


```
‘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.execQuery(q);
```

Searching for articles

```
>>> 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
  - http://www.bbc.co.uk/ontologies/storyline/2013-05-01.html

- SPARQL endpoint:
  - http://eventregistry.org/rdf/search

  - http://eventregistry.org/rdf/event/{eventID}
  - http://eventregistry.org/rdf/article/{articleD}
  - http://eventregistry.org/rdf/storyline/{storylineID}
  - Example: http://eventregistry.org/rdf/event/1234