newsLens

Philippe Laban, Marti Hearst

PhD Candidate, Professor

{phillab,hearst}@berkeley.edu

EventStory: Events and Stories in the News, ACL 2017

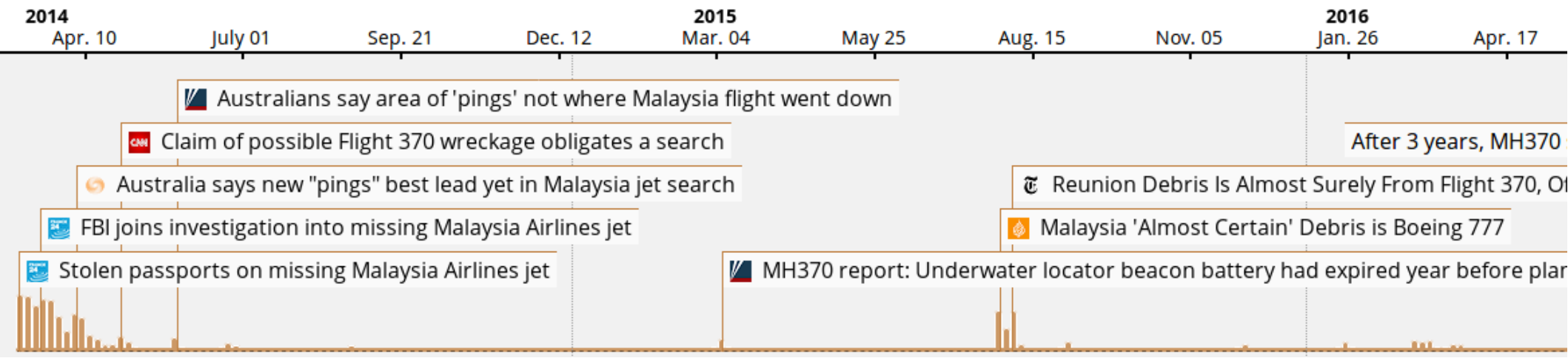
August 04th, 2017

Not so long ago...

- Do you remember the Malaysia Airline plane disappearance?
- When was it?
- They found a flaperon of the plane at some point. When? Where?

Not so long ago...

- Do you remember the Malaysia Airline plane disappearance?
- When was it?
- They found a flaperon of the plane at some point. When? Where?



Motivation

Objective: Build an interface to present **long, complex** news stories.

Emphasis on:

- Using multiple sources, transparency and easy access to the sources
- Methods can run as news comes in, **online**
- Scale to several million news articles

Contributions

- Use Internet Archive to build a decade-long news article dataset
- Building stories that can include long periods of inactivity
- Naming stories with noun phrases
- Highlighting important quotes in stories

Related work

Topic detection and Tracking

- Topic detection and Tracking (TDT, TDT2) Swan and Allan (2000)
- Europe Media Monitor Pouliquen et al. (2008)
- Unified analysis of streaming news. Ahmed et al. 2011

Visualizing news stories

- “Metro maps” by Shafaf et al. 2012
- Creation, Visualization and Edition of Timelines for Journalistic Use. Tannier and Vernier.

Source Acquisition Strategy

- Data needed for each news article:

  title, content, publish date, URL
- *Using patterns in URLs on Internet Archive:*
  - *http://cnn.com/yyyy/mm/dd/**
  - *http://france24.com/en/yyyymmdd**
  - *…*

- *Using this pattern on 20 news sources*

Source Acquisition Strategy

INTERNET ARCHIVE
WayBackMachine

http://www.france24.com/en/20140323

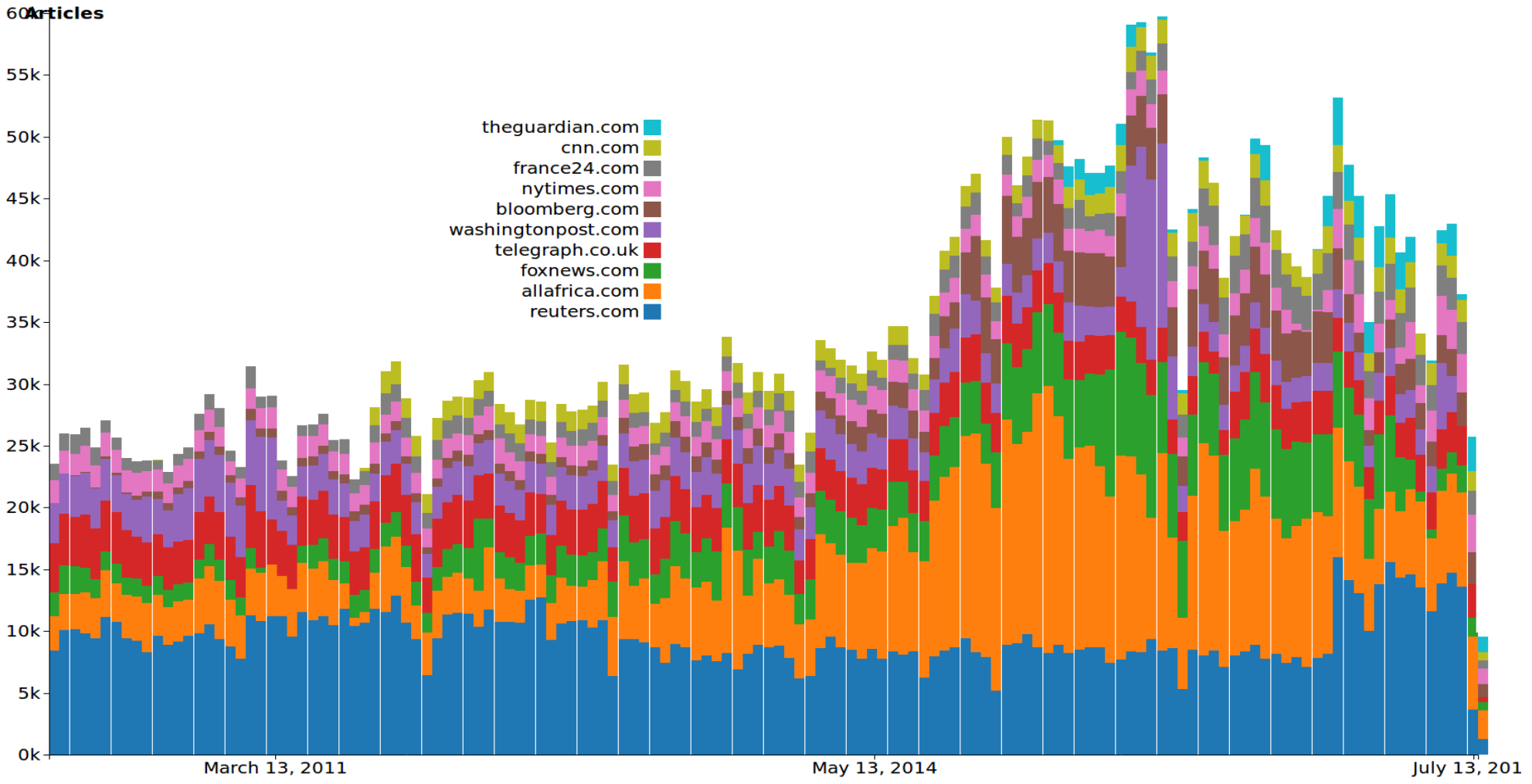
Go Wayback!

78 URLs have been captured for this domain.

Filter results (i.e. '.txt'): URL or MIME Type

URL	MIME TYPE	FROM	TO	CAPTURES	DUPLICATES	UNIQUES
http://www.france24.com/en/20140323-anti-austerity-march-violence-spain-madrid-finance-economy/	text/html	Mar 23, 2014	Jun 26, 2017	50	1	49
http://www.france24.com/en/20140323-anti-austerity-march-violence-spain-madrid-finance-economy/emission.focus	text/html	Mar 26, 2014	Jun 3, 2014	34	30	4
http://www.france24.com/en/20140323-arab-ministers-approve-summit-resolutions-avoid-rifts/	text/html	Mar 23, 2014	Mar 23, 2014	1	0	1
http://www.france24.com/en/20140323-atletico-move-top-ahead-el-clasico/	text/html	Mar 23, 2014	Mar 23, 2014	1	0	1

Dataset Statistics



Number of articles in dataset over time. Bins are 20 days in size. Top 10 sources shown.

Building stories overview

- Step 1: extract features (keywords, entities) for articles
- Step 2: Build topics: local clusters in time using extracted features
- Step 3: Build final stories by combining the local topics

Note: we want a method that is **online**. As new articles come in, topics and stories can be updated accordingly without reprocessing all articles.

Keywords and entities

A story is defined both by its **keyword** and **entities**.

Keywords

- Build bag of words vector for each article
- Normalize vectors with tf-idf
- Extract words for each article with high tf-idf score

Note: this is run on random batches of articles, not the entire collection

Entities

- Use **NER** system to extract people, places and companies from title and content
- Matched strings found with **Wikidata** entries.
- Fetch additional information about entities from Wikidata

Note: the **NER** system used is provided by spaCy

Local graph clustering

It is easier to deal with small frames of time, build local topics, and piece the topics together into large stories.

For all articles in a small time window (e.g. 7 days):

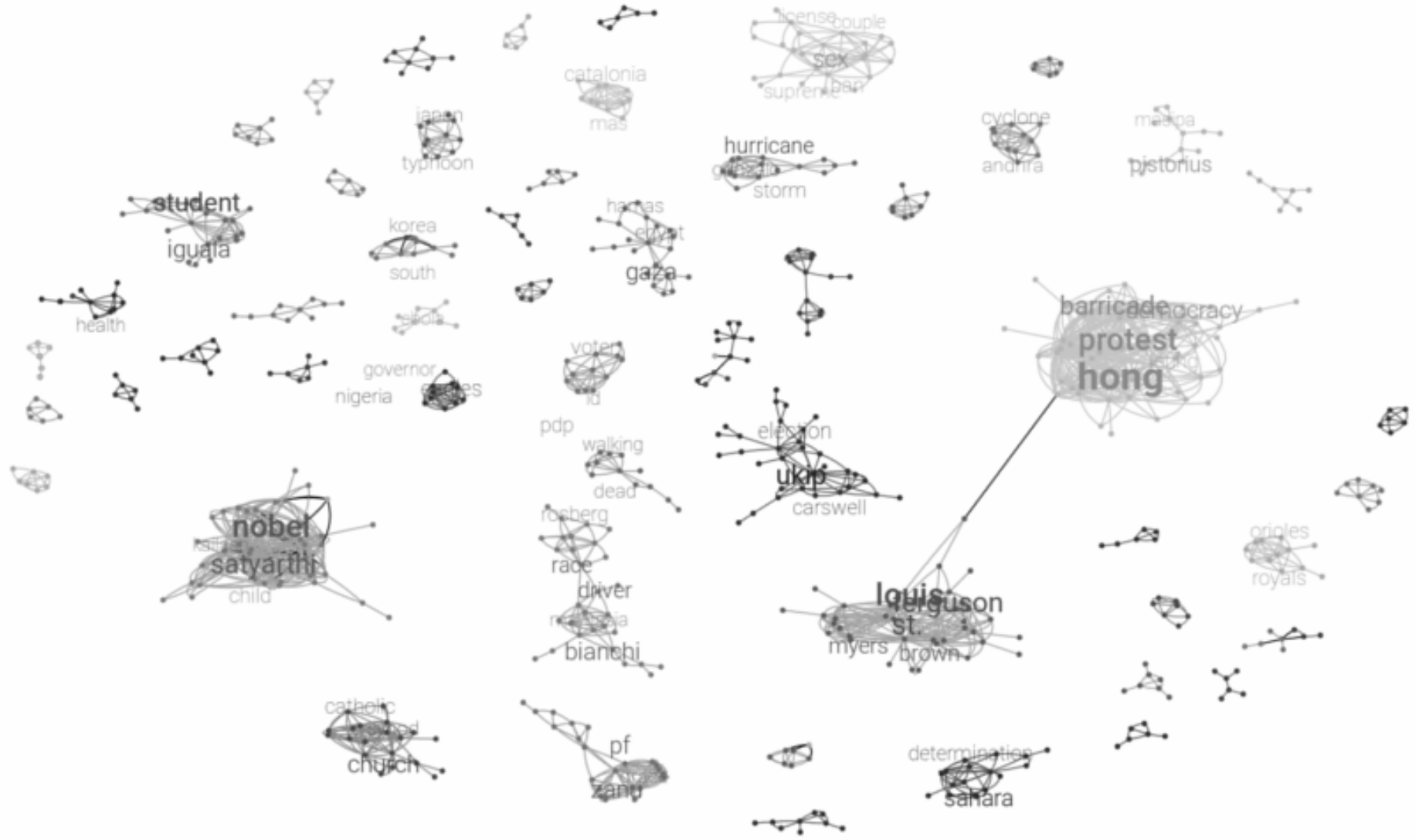
Given 2 articles, their keywords and entities are: kw_i, kw_j, e_i, e_j

We declare articles related if they overlap above a threshold:

$$||kw_i \cap kw_j|| + ||e_i \cap e_j|| \geq T_1$$

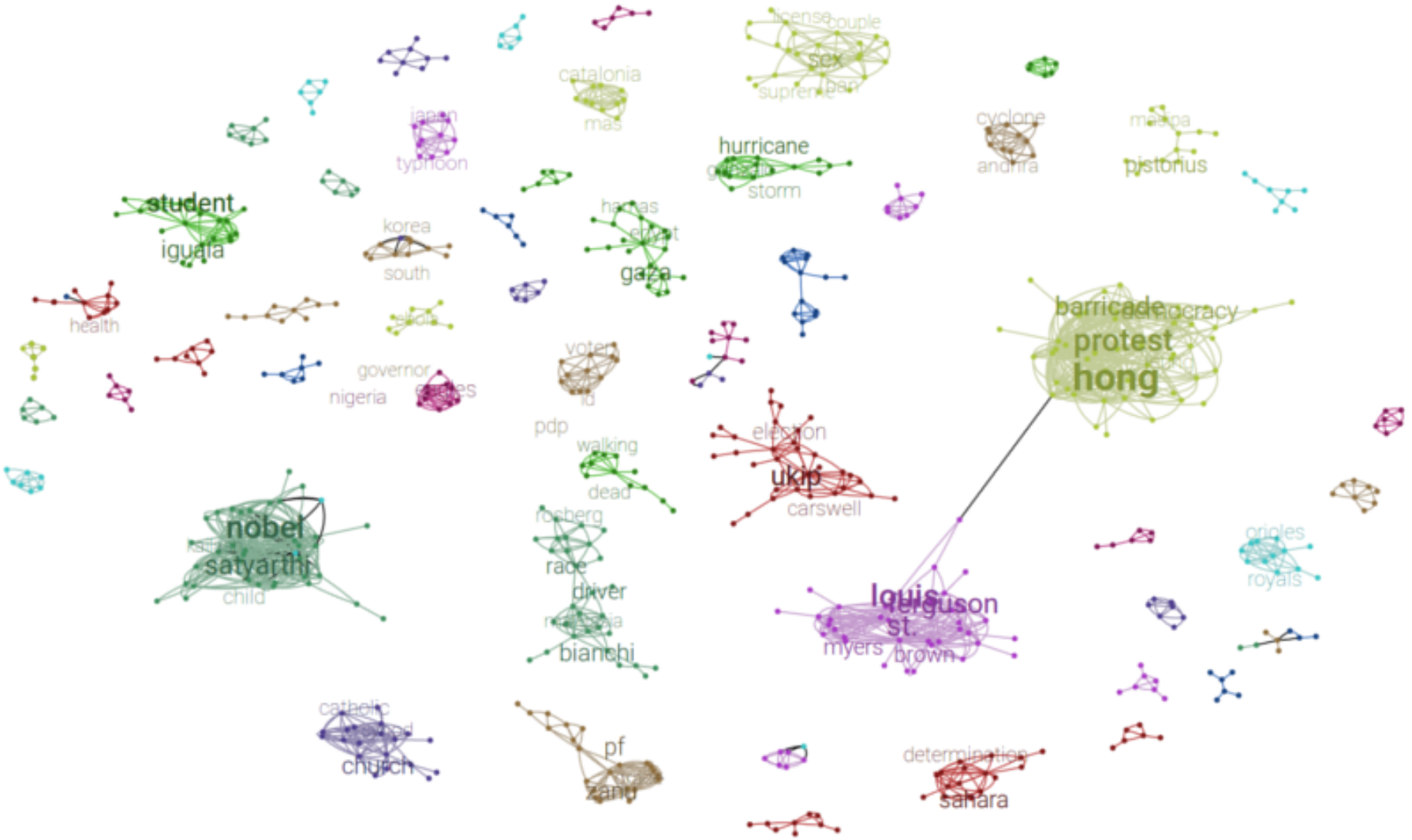
We build a graph: each article is a node. An edge is placed when the condition is satisfied.

Local graph clustering



Graph obtained from June 10th to June 16th 2014

Local graph clustering



Using community detection (Louvain method) to find local topics

Local graph clustering

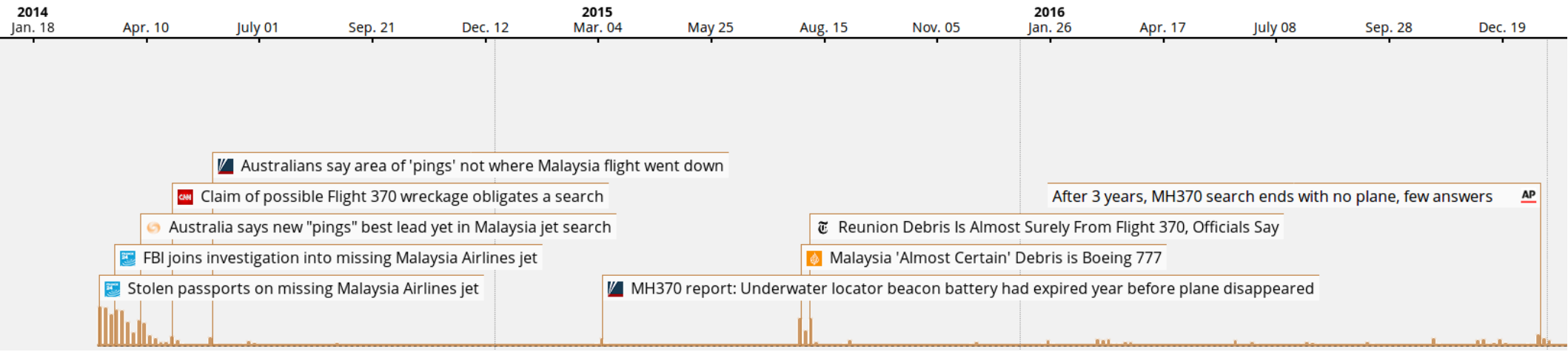
Topics are created by running a **sliding window** of the graph shown above. As time passes:

- Older articles are removed. Incoming ones are added.
- Community detection is run, again. New articles can join an existing topic, or create a new one.
- Note: Clusters can also merge and split over time. See paper for detail.

Story discontinuity limitation

This method works as long as the longest break in a story is **smaller than** the chosen window size.

This is limiting, as many stories have large gaps.



From topics to stories

We create stories from the topics: when aggregating articles, the keyword distribution is less noisy.

Most common keyword in first part of MH370 story:

('plane', 374) ('mh370', 362) ('search', 352) ('malaysia', 296) ('flight', 220)

Most common in second part of story:

('mh370', 140) ('debris', 139) ('plane', 112) ('reunion', 111) ('malaysia', 87)

We use a simple keyword similarity
to merge topics
into the final stories we obtain.

Stories statistics

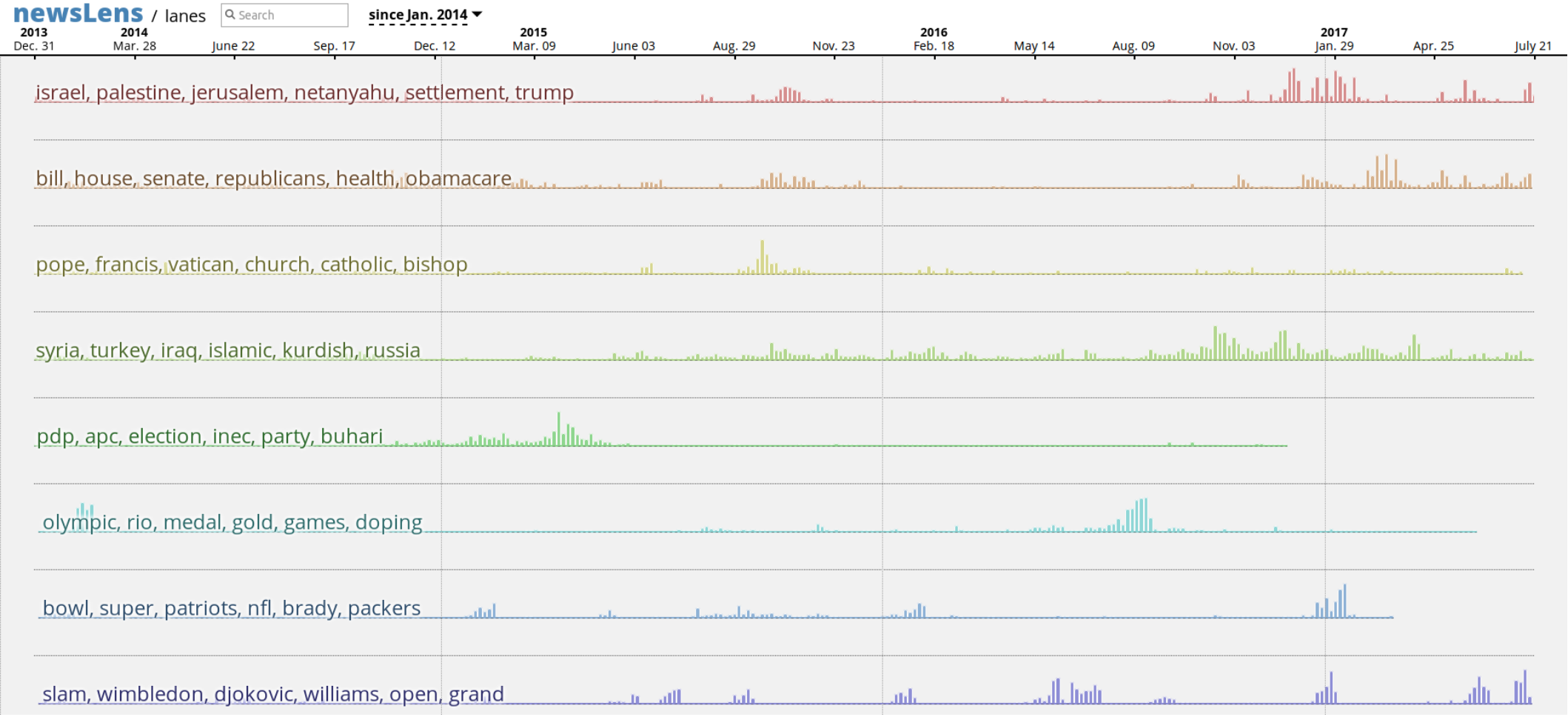
We obtain 100,000 stories from 2010 to 2017.

Size of story (in articles)	Number of stories
1000+	100
100+	800
3+	100,000

About 30% of articles are matched to a story. This varies with threshold T1

Timelines of stories

Typically: stories are named by list of common keywords.



These stories would sure be better with names...

Naming stories: intuition

- What we want story names to be

North Korea nuclear tests

Ukraine crisis

Ebola outbreak

Brexit vote

Paris attacks

Naming stories: intuition

- What we want story names to be:

North Korea nuclear tests Ukraine crisis

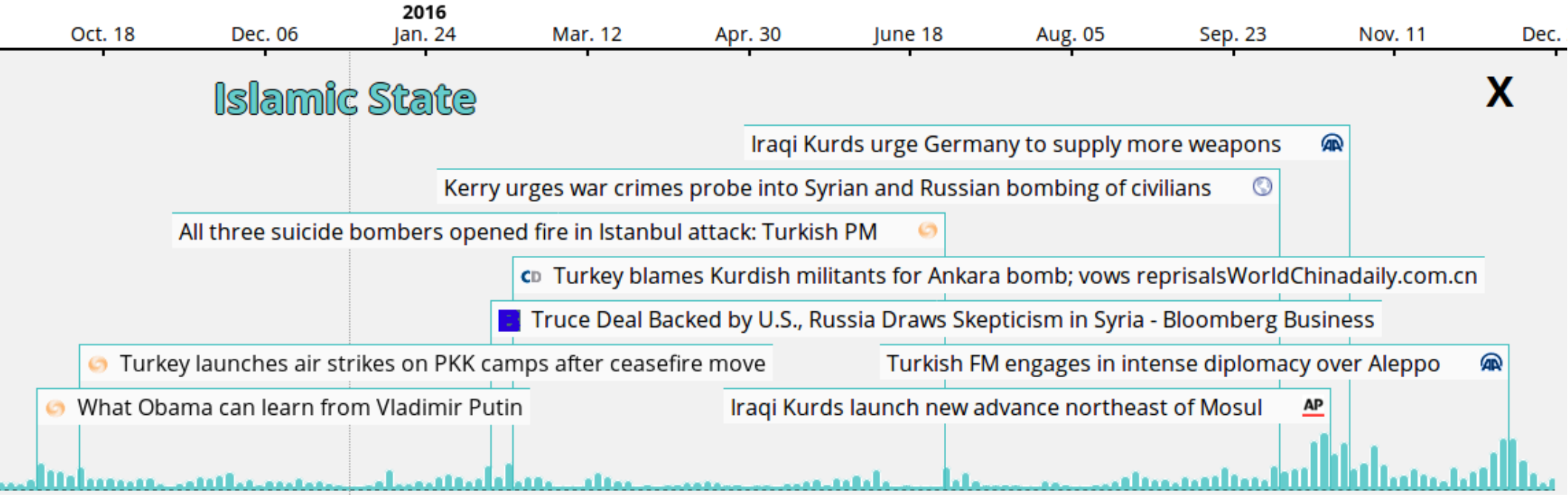
Ebola outbreak Brexit vote Paris attacks

- Made of proper nouns, and common nouns.
- Some words are “abstract”. (Kato et al. 2008)
- The entity is important to the story

Note: phrases are extracted from titles of articles, ranked, and one is chosen.

Same stories, with names

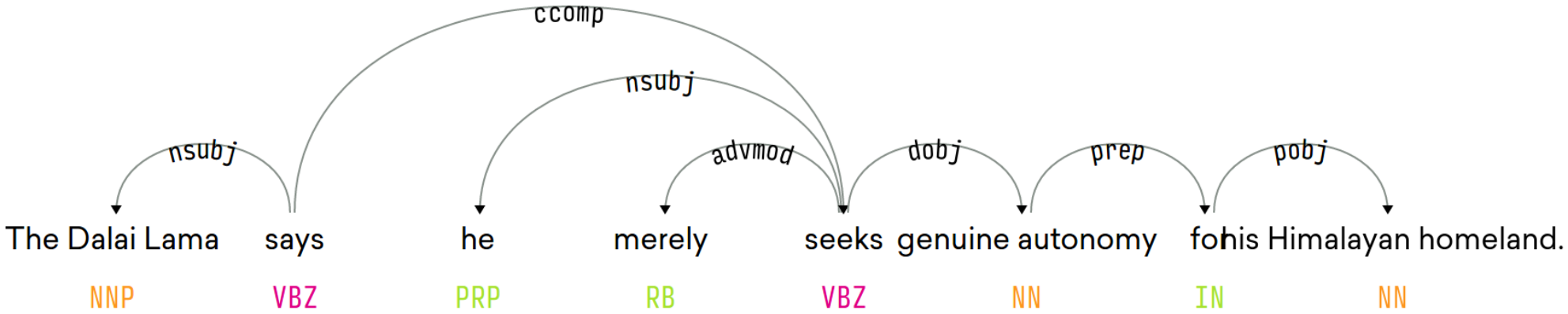
Demo II – opening a story



Showing headlines is a good start. What else can we show? Quotes...

Quote extraction

Very simple method to extract quotes from articles using dependency parse.



The quote extracted here is: (Dalai Lama, “he merely seeks genuine autonomy ...”)

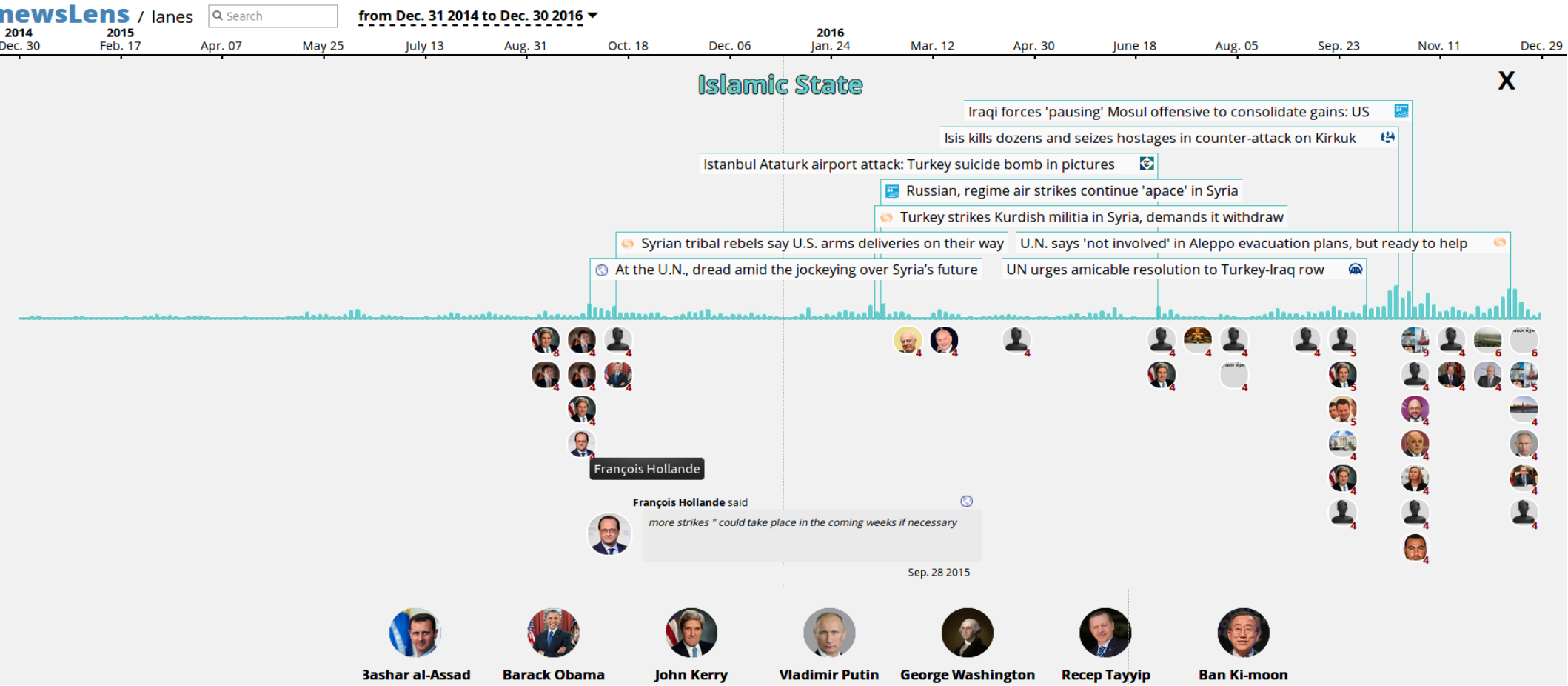
Quotes extracted are attributed to entities, that have been tagged to Wikidata entities.

Quote selection

- On average, 0.75 quotes / article are extracted.
- Thousands of quotes for certain stories.

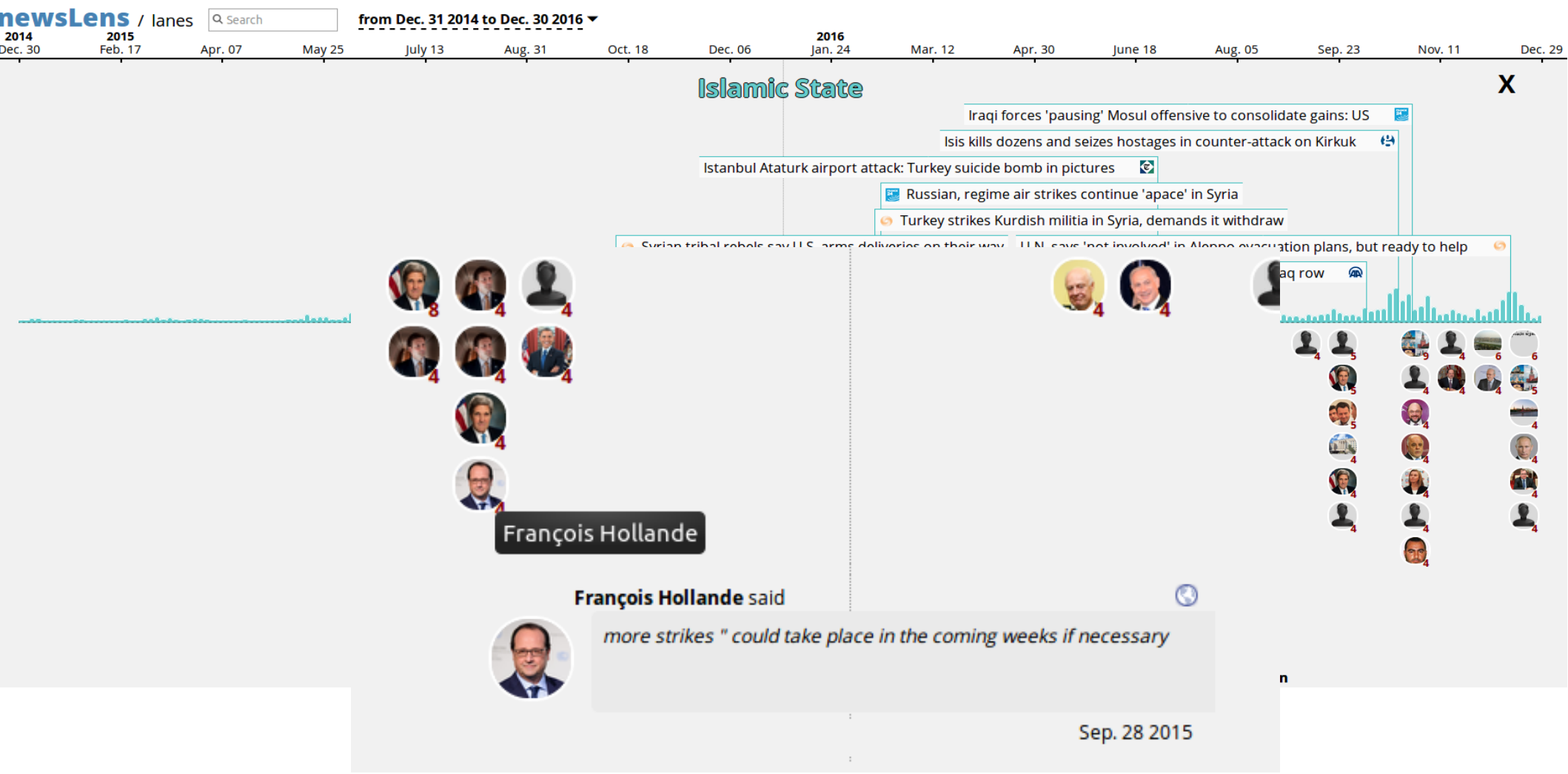
- Quotes are clustered based on:
  - If they are from a same range of time **and** share several words.
- Size of the clusters help determine quote importance.

Quotes are in



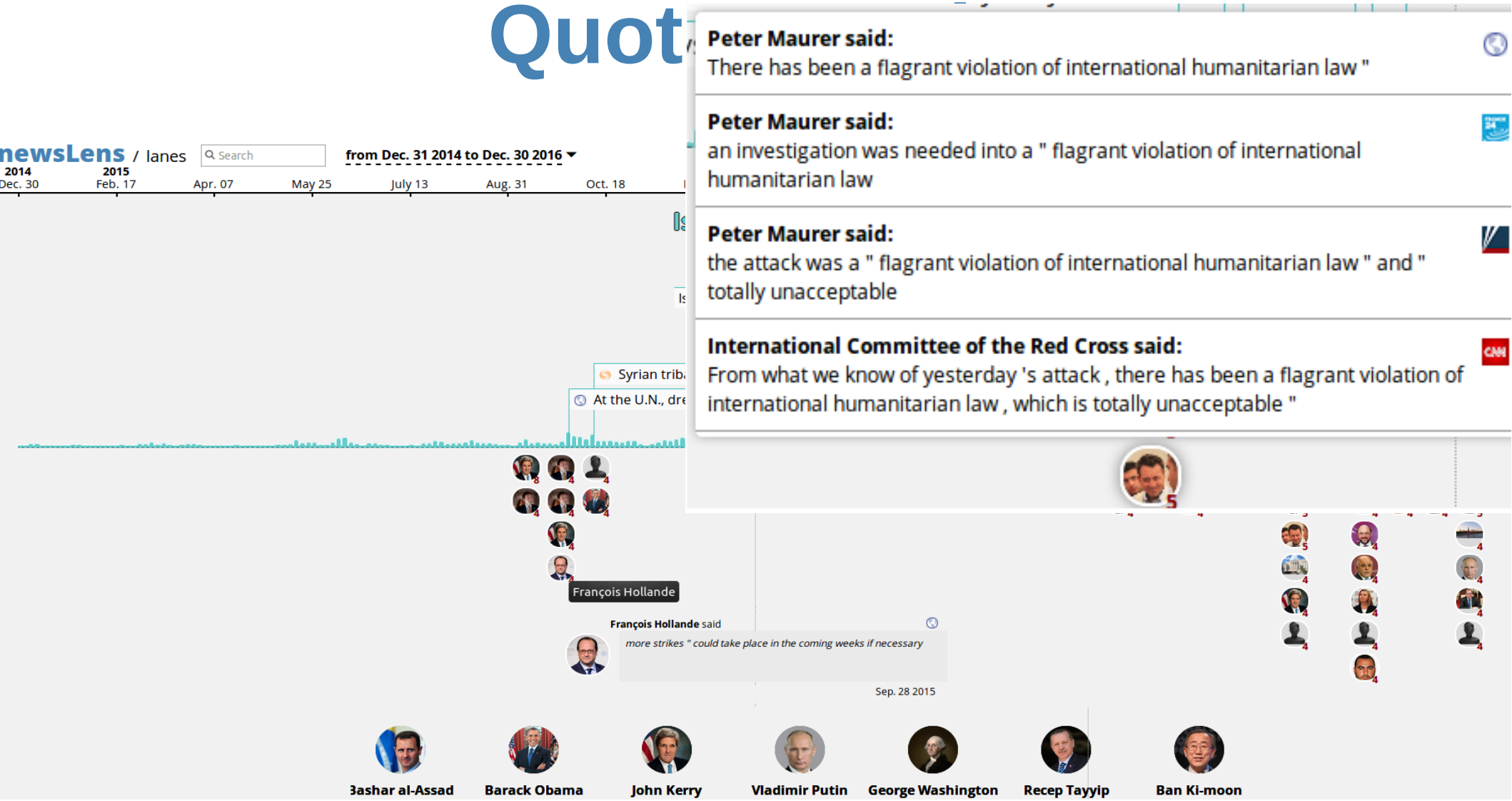
Get a glance at most important quotes of the story.
What if I want all of John Kerry's quotes?

Quotes are in



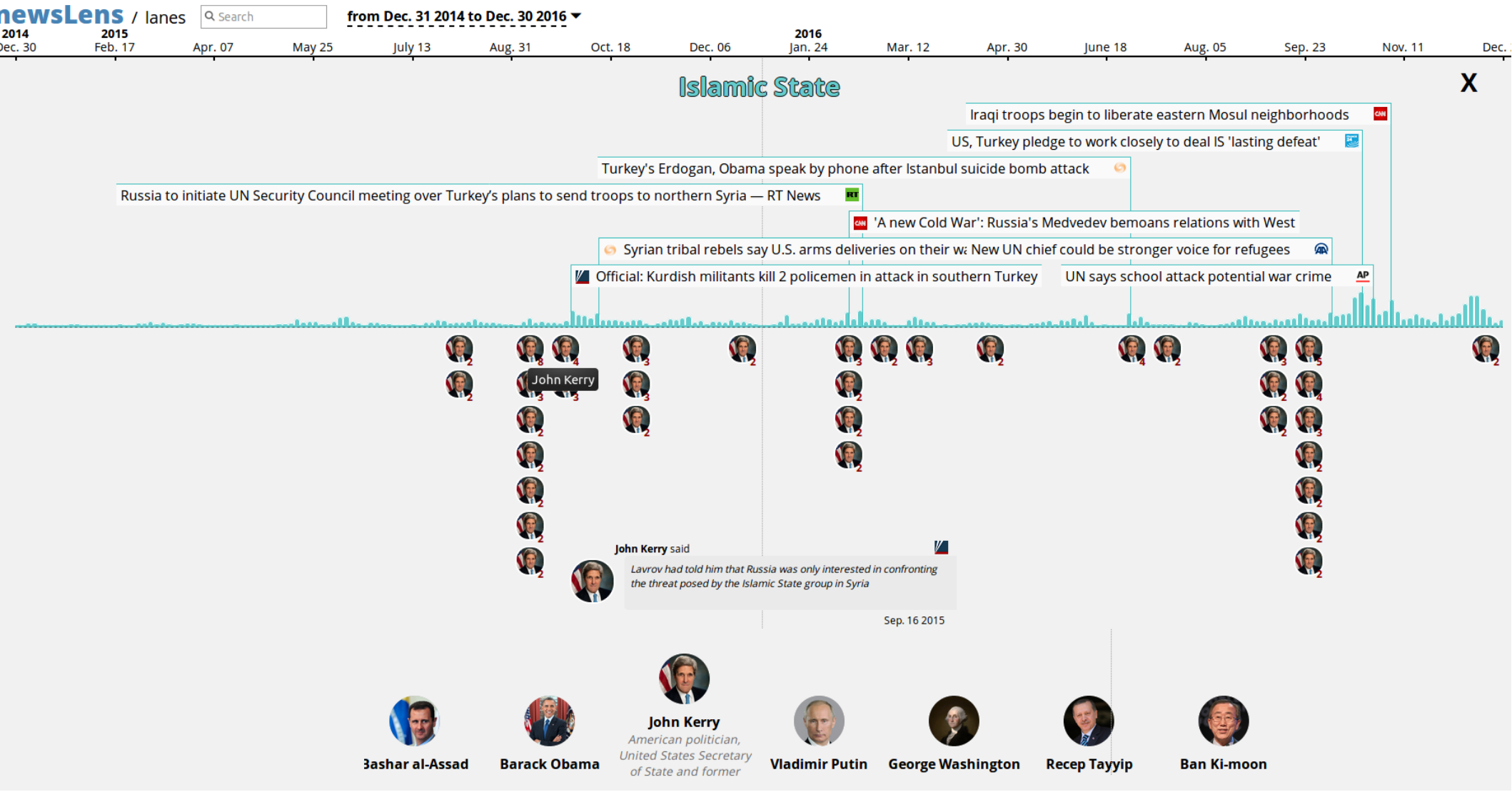
Get a glance at most important quotes of the story.
What if I want all of John Kerry's quotes?

Quot



Get a glance at most important quotes of the story.
What if I want all of John Kerry's quotes?

All John Kerry quotes



A good way to get different geopolitical perspectives

Future directions

This is work in progress. Here is what comes next:

- Evaluation of what we have built:
  - Usage study of our interface
  - Evaluation of our story dataset
- Going beyond headlines:
  - Using the content. Structured events.
- Focus on the breaking/trending news.

Want to help, here are possible ways:

- Test and share our demo and give us feedback. The demo is public.
- Share a dataset with us. News articles, stories etc.
- Tell us what you think and what we can do better.

Questions

- Thanks for listening.

- The demo is publicly available at:

http://newslens.berkeley.edu

References

This project would not have been possible without