Realtime Boost

(goto/realtime-boost)
Status:Final
Authors: felipeg@google.com
Last Updated: 2016-05-12

Demo and Docs

- http://go/realtime-boost-demo
- Realtime Boost presentation: go/realtime-boost-preso
- To learn about how to use from SR: go/realtime-boost-howto
- How to use from QRewrite: go/realtime-boost-qrewrite
- Superroot Design Doc: go/realtime-boost-dd
- How to listen to the Pubsub
- The hivemind UI is here: go/realtime-boost-hm
- Brussels Attacks analysis made by Google Now (rubena@)
- What the spike means: go/realtime-boost-spike

Objective

Detect, under a minute, trending topics and real-world events in realtime. Use this signal to improve realtime results in google search (not only on twitter carousel results).

The spike signal and the correlated entities/salient-terms is available in Superroot and in QRewrite for every query.

Background

It took google SERP page 14 minutes to start showing news documents about the San

Bernardino shooting. The twitter carousel, which uses a system similar to this design doc, took 7 minutes.

In Search, ranking is traditionally based on signals that takes time to accumulate. For example, doc-anchors and navboost clicks takes time to happen and the millwheel joiner also takes time to aggregate it in docjoins. With recent improvements from the Freshness team, instant navboost clicks are available in komodo in 10 minutes and in DocJoins in 30m. Our goal is to improve even further that latency and help in cases where the new document didn't have impressions and clicks yet.

When a breaking-news event happen in the world, we need some signal to tell us that we need to promote newly created (fresh) documents to the top of the ranking even if those docs don't have anchors or clicks, so we can quickly cover the news with up-to-date information.

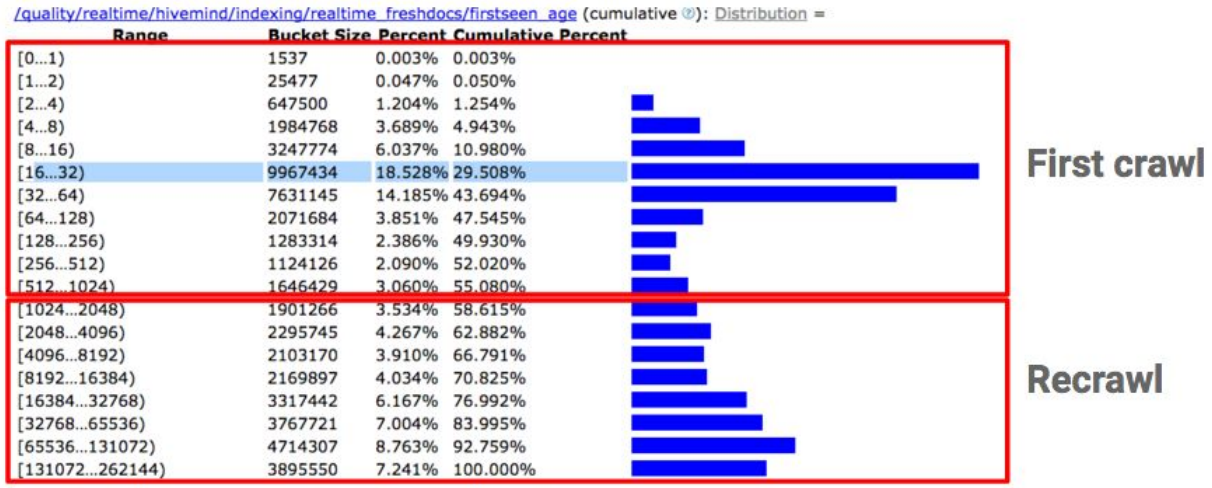
Freshness team in Zurich currently tackles this problem by calculating the number of high-quality news documents that comes up from Muppet in Superroot. They have their own classifier to classify documents as news / high quality. The problem with this approach is that Superroot has access to only so many documents (mostly only the top-1000 docs). If the event is really recent, the new documents most likely won't gather enough signals to be ranked in the top-1000 and for some period of time this approach can miss the opportunity to detect the breaking news.

Overview

We take a different approach, by indexing fresh documents in hivemind and detecting spikes in number of documents published in real time that matches a given query.

We listen to the freshdocs-instant-docs pubsub topic, which gives us the CDoc of all documents crawled from the instant pipeline. Those docs are mostly news websites and the majority is crawled very quickly (under 1 minute) after being published.

Here is a histogram of the document ages that my ingestor receives from the freshdocs-instant-docs pubsub. (ranges are in seconds)



The first bellcurve in the histogram is from the first-crawl of a new document.
The rest is mostly coming from the re-crawls of older documents. These older documents won't impact the speed of detecting trending queries, since they have already been crawled before when they were new. The documents are indexed using the firstseen date, and thus, the spike time is independent from when freshdocs decided to crawl the document. We could miss a new spike because freshdocs didn't crawl fast enough. It can only affect the recall, but not the precision. The speed of crawling cannot generate a fake spike.

The super-fast crawling of these high quality news websites allows us to detect when there is a sudden increase in documents published that mention the query.

The first step after crawling, is indexing these docs into hivemind-instant in realtime.

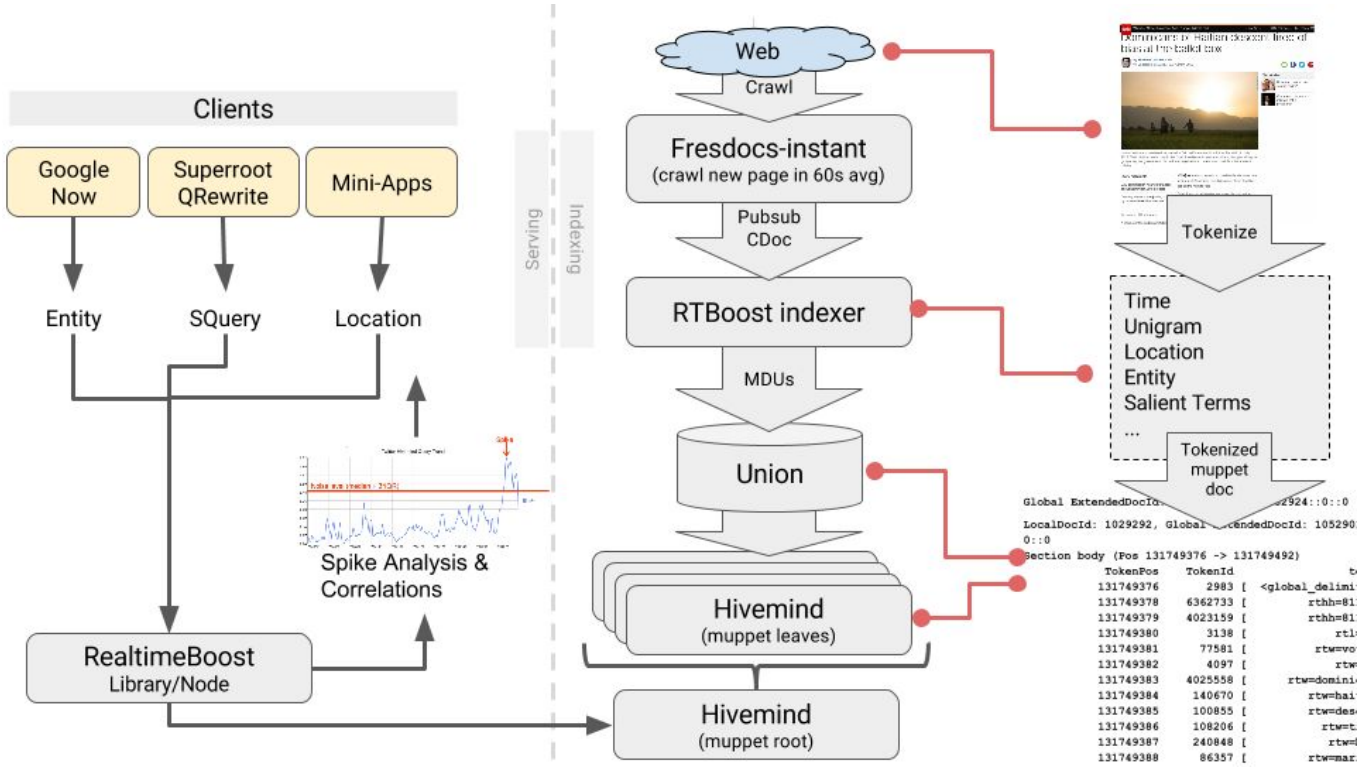
Side note:

The Crawl-to-index SLA for regular web-search-indexing is described in this page

- Instant: 95% of documents we receive have a crawl-to-index latency of ≤ 80 seconds.
- Daily/Instant: 99.9% of documents we receive have a crawl-to-index latency of ≤ 20 minutes

This is different from crawl-to-realtimehivemind as we describe in this document, but serves as a perspective.

System Diagram Overview



Hivemind tokens indexed

Unigrams (RTW)

For each document, we use its title, anchor texts and the first 400 chars from its centerpiece text. These are split into unigrams and added to the hivemind index. The centerpiece usually correctly contains the main content of the article excluding the boilerplate.

Centerpiece text examples

> http://www.eurosport.co.uk/football/matt-taylor-earns-the-plaudits-after-hat-trick-heroics_sto5011668/story.shtml
> Centerpiece: Only a fantastic save from Wycombe goalkeeper Matt Ingram prevented Taylor scoring four times in the space of 12 minutes as he turned a previously close game emphatically in Rovers' favour, sealing a 3-0 win. Manager Clarke said: "I have always had faith in Matty, even when he was taking some unfair stick early in the season. "He is high on confidence at the moment, but he is also a terrific
>
> http://www.usatoday.com/story/travel/flights/todayinthesky/2015/12/01/jetblue-considers-new-approach-training-pilots/76600018/

Centerpiece: 9 Share This Story! Let friends in your social network know what you are reading about FacebookEmailTwitterGoogle+LinkedInPinterest JetBlue considers a new approach to training pilots JetBlue wants to recruit a few flight students with no flying experience and train them to become pilots. Loading… Post to Facebook JetBlue considers a new approach to training pilots JetBlue wants to recruit a

http://www.edmontonsun.com/videos/4640483518001
Centerpiece: RAW: Suspect shot after three found dead in their home

http://www.thetubdoc.com/headphones-that-make-silence-beats-discount
Centerpiece: Headphones That Make Silence Headphones That Make Silence Headphones That Make Silence Headphones That Make Silence Headphones That Make Silence Headphones That Make Silence Headphones That Make Silence Headphones That Make Silence Headphones That Make Silence Headphones That Make Silence Headphones That Make Silence Headphones That Make Silence Headphones That Make Silence Headphones That Make Silence Headphones That Make Silence Headphones That Make

http://www.thehindu.com/news/cities/chennai/thoongavanam-to-release-online-legally/article7931531.ece
Centerpiece: The movie will be available on herotalkies.com on December 4. Actor Kamal Haasan's Deepavali release, Thoongavanam , will be made available for fans living outside India through the movie streaming websiteherotalkies.comon December 4. Despite getting good reviews, the film's box office performance was partially affected by the unprecedented rains that lashed Tamil Nadu in the second week of

Correlated unigrams that are returned for the query can be used to retrieve documents relevant to current news event. E.g. for query spike for query rtw=california returns following correlated unigrams: (san, bernardino, shooting, farook, sayd). There tokens can be added as context terms or as optional terms to Muppet squery.

Half Hours since epoch (TEHH)

The timestamp of the document publishing (or creation) time in number of Half Hours since epoch. We estimate that by assuming it is equal to firstseen-age (the time the crawler first seen the URL for this doc). This token-class is the centerpiece of the spike detection by counting how many documents were created in each half-hour bucket. The Hivemind Lift is used to detect spikes in the current unfinished half-hour bucket, so the resolution is not limited to half hour, we can detect spikes in less than 5 minutes resolution.

KG Entities (RTKG)

The entities with high topicality and high confidence, we index its kg mid into the RTKG token-class and also its human-readable name into the unigrams / rtw token-class.

S2 Cells (S2)

If the document contains a highly topical entity that contains a location related collection with high score, we index the S2Cell associated with the location. This way if the news article is talking about a shooting that happened in San Bernardino, we index the S2Cell for San Bernardino city.

Freshbox Article Score (RTF)

This is the score that the freshness-zrh-team classifier gave to this document. This signal tells

us how likely this document is a high quality news article.
We use this signal to detect how many high quality news article is present in the spike, so we can avoid spurious or spammy spikes, such as spikes caused by low quality sites publishing the same ngrams at the same time, trying to game the system.

Document NSR (RTN)

This is the NSR score of the document. Like Freshbox Article Score, we use this signal to calculate how many documents coming from high quality websites are present in the current spike and avoid spikes caused by spammy websites.

Geographical Dimensions

Allows us to detect spikes that are only relevant to one region, e.g. political events.

- Country (RTC)
- Language (RTL)
- Domain (RTD)

Other ideas

- Produce correlated queries by extracting n-grams, salient terms etc and indexing in separate token class.
- Add correlated terms as optional terms into SQuery in QRewriter. Will allow to bring more documents relevant to the current spike, increasing topicality.

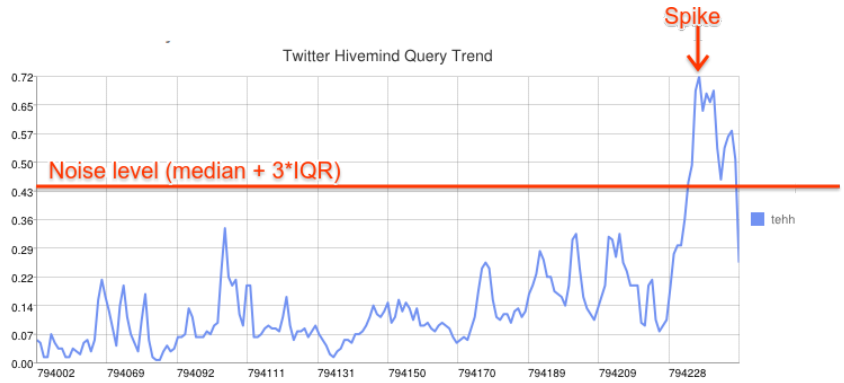
Detecting Spikes

Similarly to realtime-team's twitter design, we detect spikes on the time series that hivemind returns.

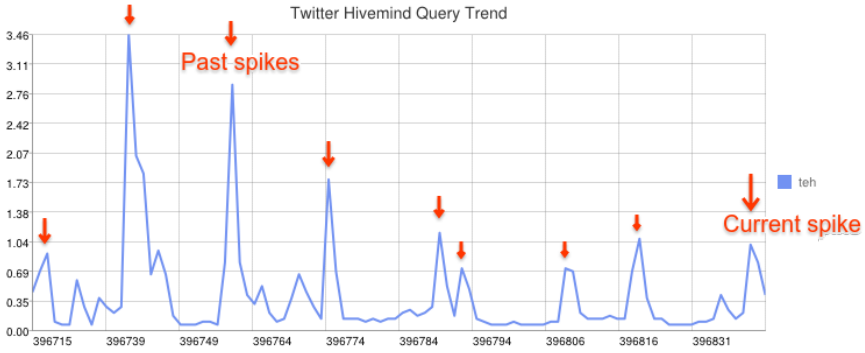
Given a rewritten squery from the query-rewriter in Superroot, we convert it to an mquery to be used in the hivemind RPC.
Hivemind will respond with the counts of all documents that matches the query.
Hivemind response contains the correlated (co-occuring) tokens of each token-class we ask for.
In our case, we get back the counts of documents in the TEHH buckets that we use as a time series to detect spikes. Here is the basic idea:

Using the doc count and lift as a time series we analyse if the past 4 hours contains a spike (on both count and lift) above the noise level (taken from the whole 3-day history):

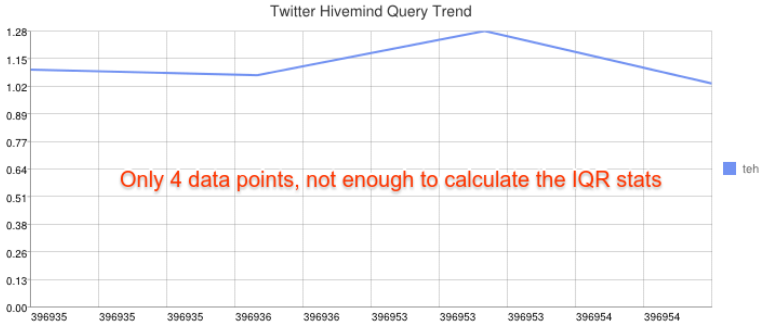
```
If spike > median + 3*IQR
```



The IQR threshold is similar to the Standard Deviation and it protect us from noisy or too frequent spikes.



For queries that have no past history, i.e. a brand new event with no documents about it. In that case, we can't calculate the IQR, so instead we require a minimum Lift and a minimum document count in the recent history.

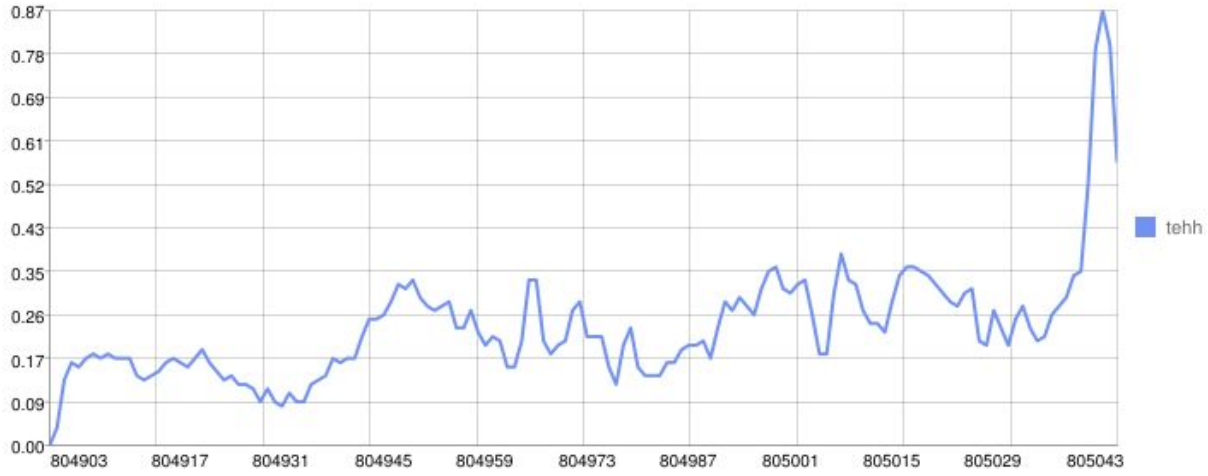


Use Case Analysis

Query: California

Right after the San Bernardino shooting event in 2015-12-02

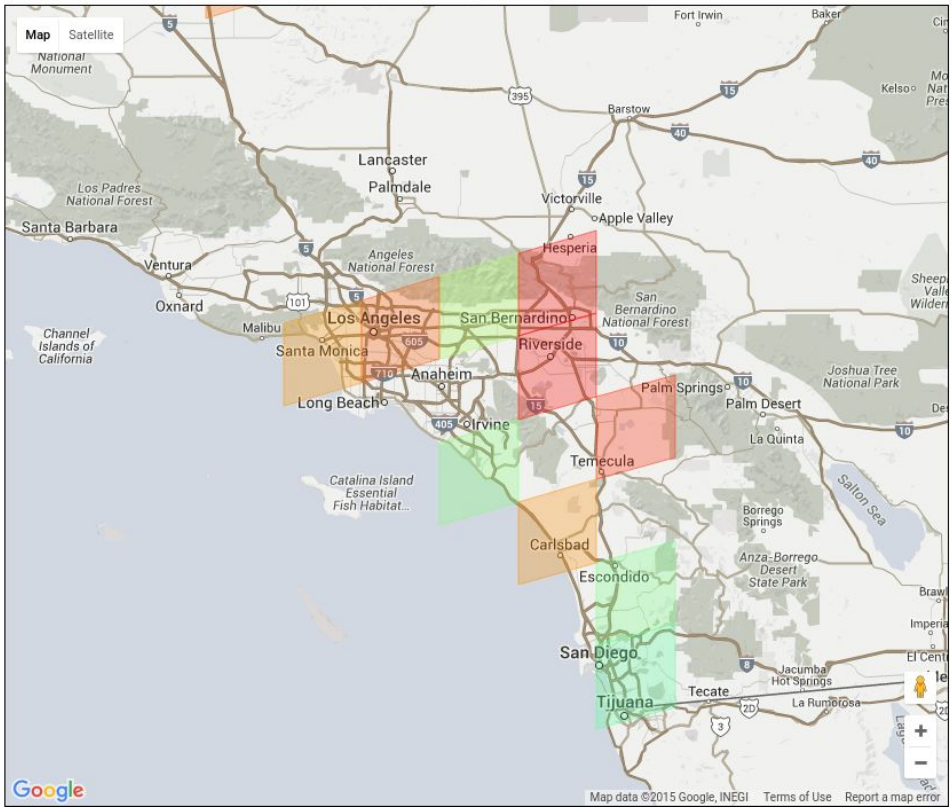
The spike in TEHH (time) is sharp and was created quickly after the event took place.

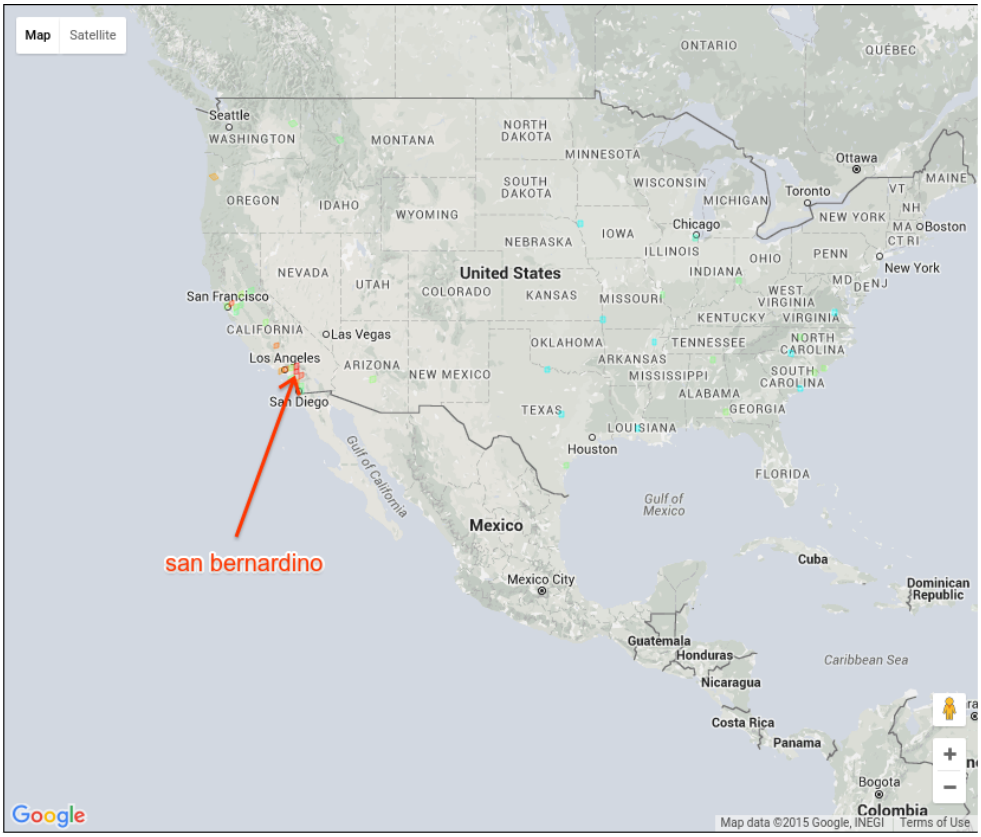


The S2Cell (location) spiking was exactly the San Bernardino cell:

Map

Location (s2) Heatmap | Zoom to fit | Min score: 0 | Refresh





The NSR (site rank score) and Freshdocs Article Score both shows that most documents were high quality news:

Realtime Doc NSR (rtn)

Matched sessions: 10000 of 151826413

Search:

Show 25 **entries**

[First] [Previous] 1 [Next] [Last]

#	score	matched	count	bits	Go	HM	Label↓
15	0.2452	16	57940	10			0.80
1	2.0477	620	935033	28			0.75
2	1.8982	851	1662705	32			0.70
4	1.7100	751	1605521	32			0.65
3	1.7816	778	1595124	32			0.60
5	1.6074	795	1907531	32			0.55
7	1.2959	738	2325082	32			0.50
10	0.7181	938	7693248	32			0.45
11	0.7145	512	3168400	32			0.40
6	1.5492	848	2211247	32			0.35
8	1.0539	673	2726402	32			0.30
9	0.9239	668	3254481	32			0.25
12	0.5857	373	2626250	32			0.20
13	0.4935	282	2205474	31			0.15
14	0.2995	87	741319	24			0.10
16	0.0784	11	145796	5			0.05
17	0.0001	1	27600	1			0.00

Showing 1 to 17 of 17 entries

Realtime Doc Freshbox Article Score (rtf)

Matched sessions: 10000 of 151826413

Search:

Show 25 **entries**

[First] [Previous] 1 [Next] [Last]

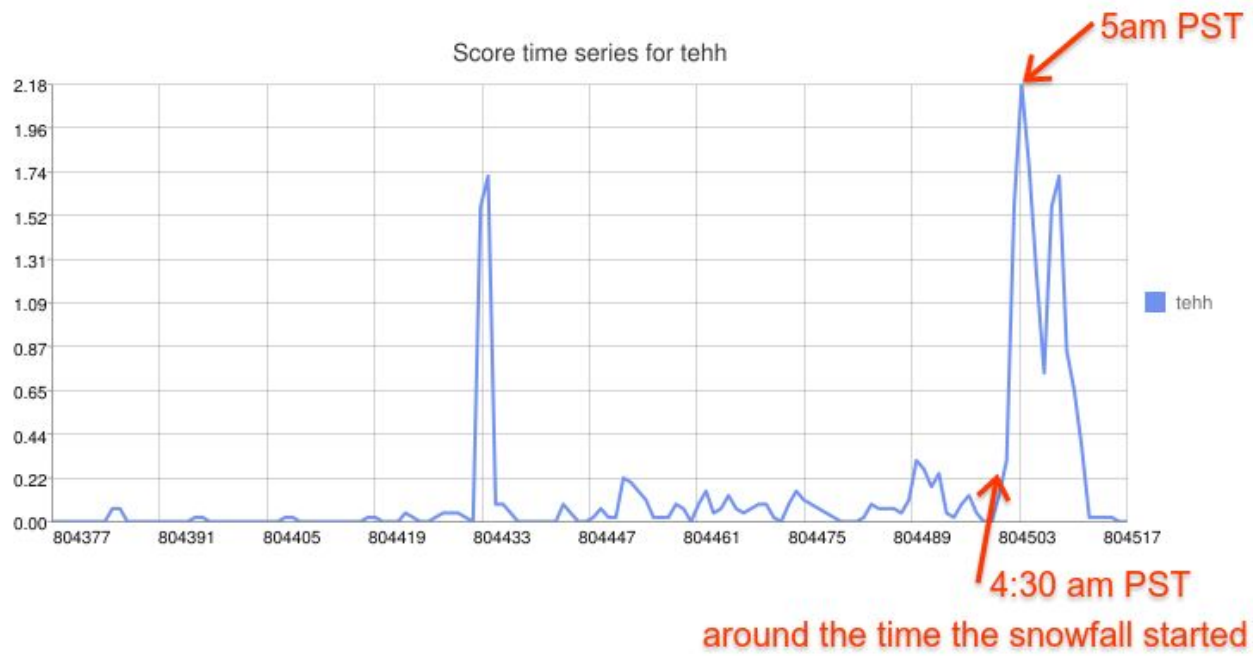
#	score	matched	count	bits	Go	HM	Label↓
1	2.0074	911	1696917	32			0.95
3	1.4179	393	802976	32			0.90
6	0.9744	216	558002	32			0.85
5	1.0135	193	444595	30			0.80
9	0.7988	142	390755	30			0.75
12	0.6849	118	365809	32			0.70
19	0.5434	95	363228	29			0.65
18	0.5474	96	365401	29			0.60
13	0.6587	116	376595	31			0.55
11	0.7296	132	397456	32			0.50
20	0.5343	105	435009	29			0.45
16	0.5848	121	476528	31			0.40
17	0.5578	125	536226	30			0.35
14	0.6248	155	634949	32			0.30
15	0.6198	175	773135	32			0.25
10	0.7404	241	974954	32			0.20
8	0.8059	322	1335805	32			0.15
7	0.9585	508	2058516	32			0.10
4	1.1175	907	3892742	32			0.05
2	1.7191	4927	24858946	32			0.00

Showing 1 to 20 of 20 entries

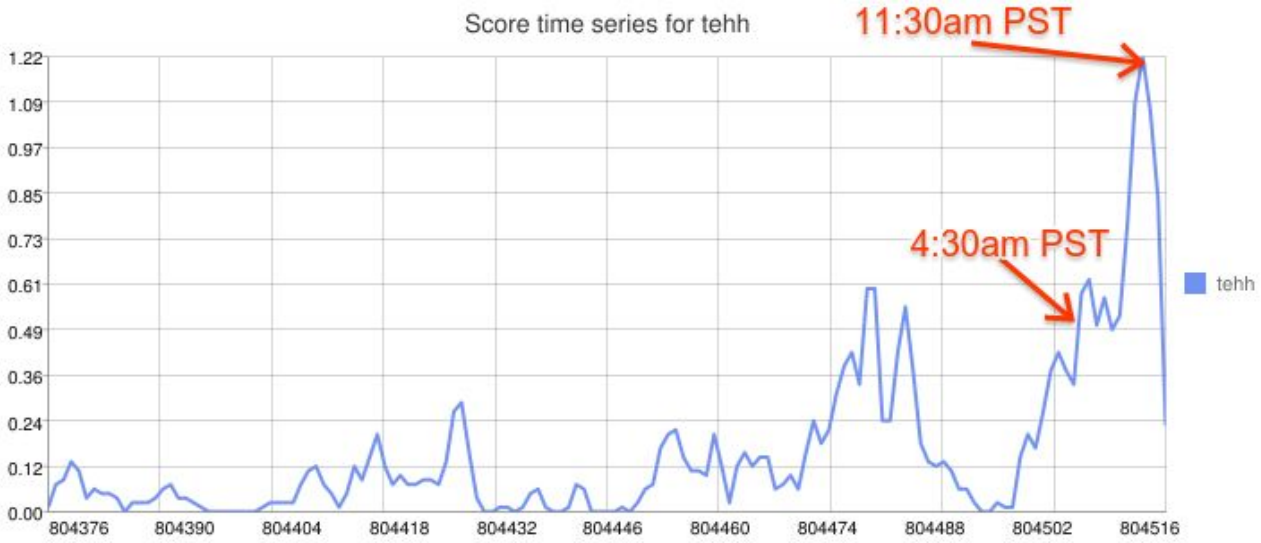
Query [Weather Chicago]

Realtime Boost can be faster than Twitter
Huge snowstorm arrived in the middle of the night

Realtime Boost Spike:

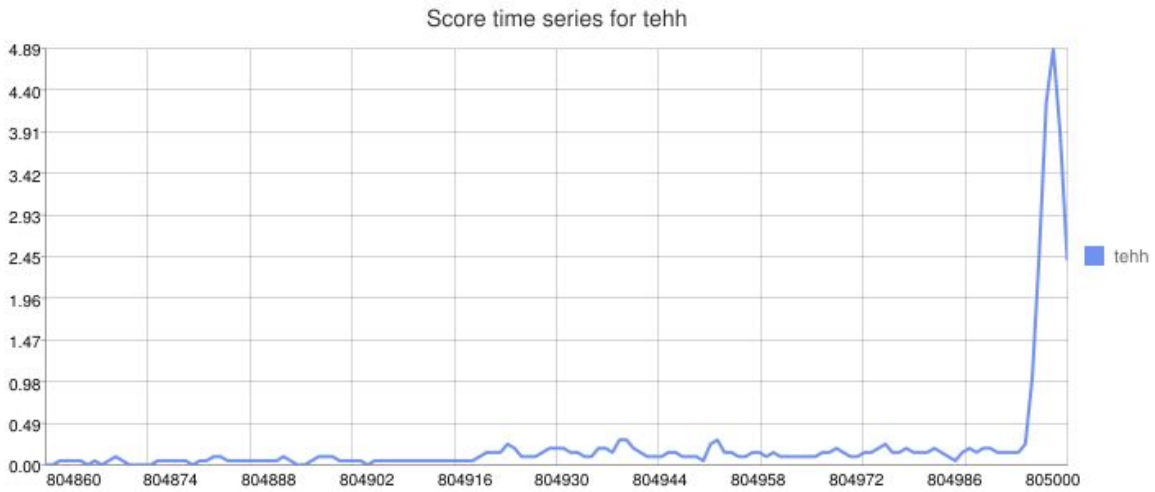


Twitter Hivemind Spike (for comparison):



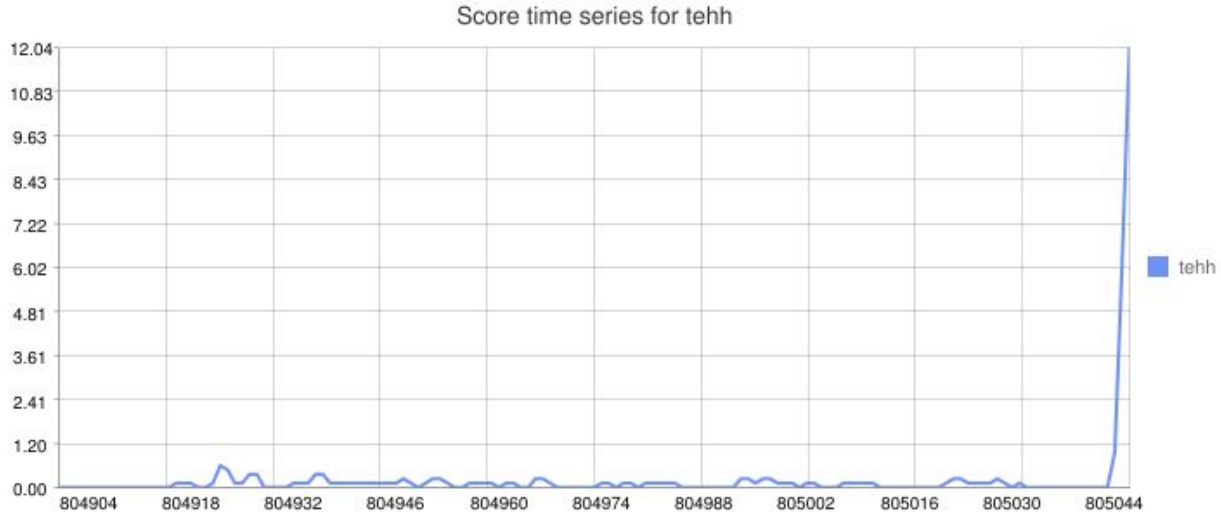
Query: Facebook Stocks

Right after Mark Zuckeberg announced he is giving away 99% if his stocks to charity.



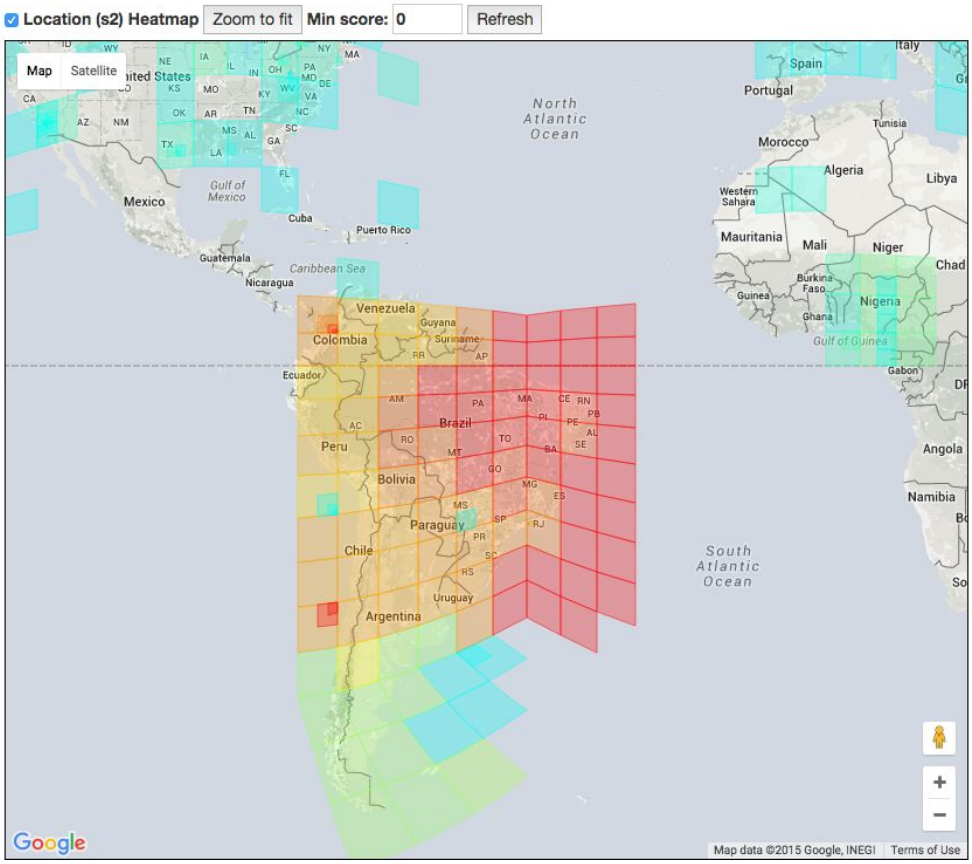
Query: Dilma

Right after Brazil congress announced they will start the impeachment procedure against brazil's president Dilma.
The spike happened only minutes after the announcement was made.



S2Cells for that query, of course, covers the whole Brazil:

Map

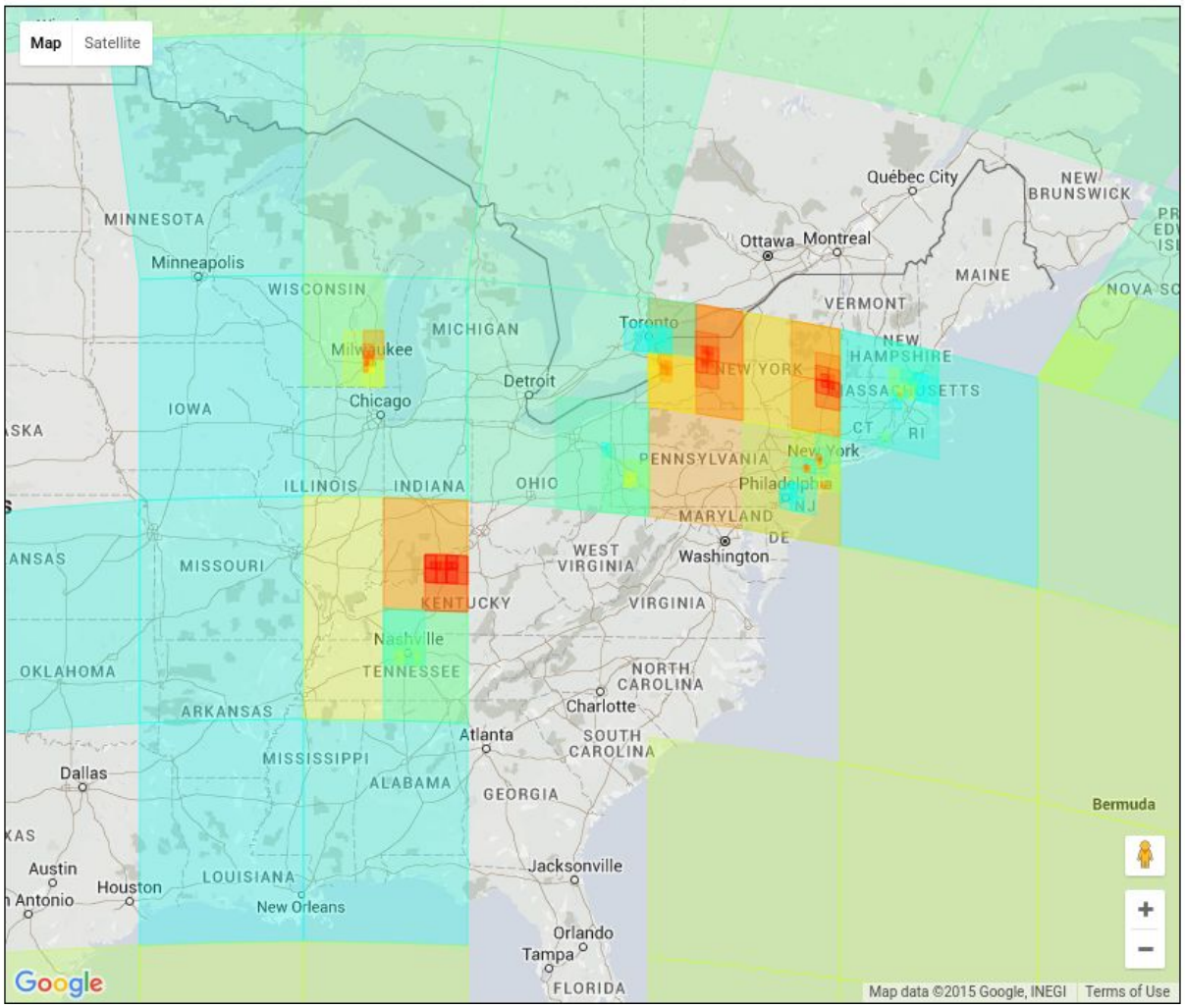


Query: Bruce Springsteen

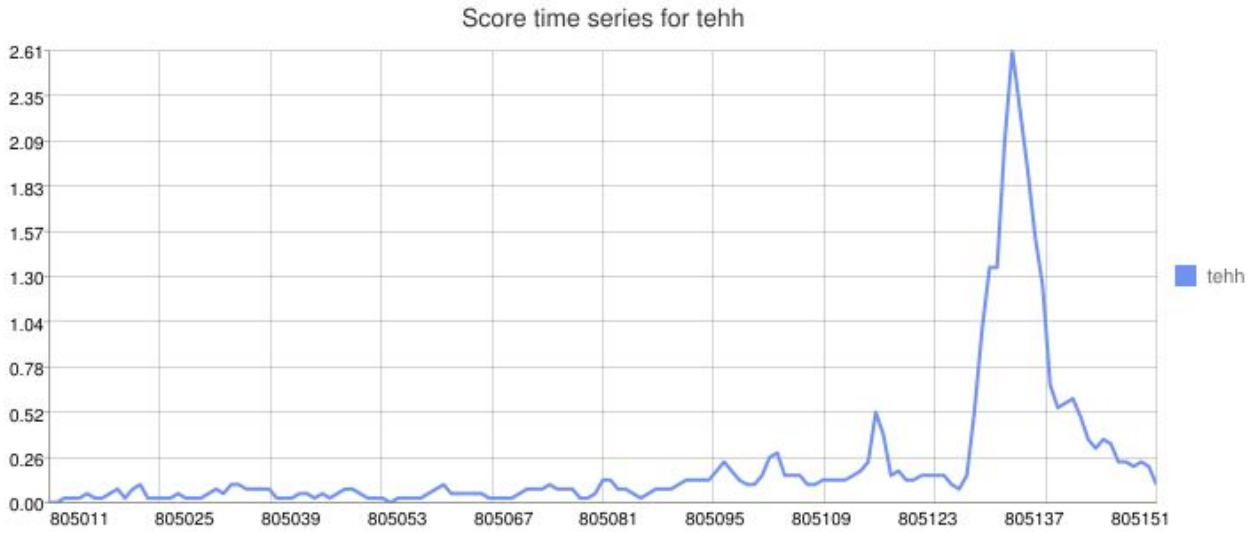
He announced his new tour, map shows most of the places where the tour is going to pass.

Map

Location (s2) Heatmap | Zoom to fit | Min score: 0 | Refresh



Spike:



He was announcing the tour for the Album called The River

Realtime KG entities (rtkg)

Matched sessions: 4260 of 147530340

Search:

Show 25 **entries**

[First] [Previous] 1 2 3 4 5 [Next] [Last]

#↑	score	matched	count	bits	Go	HM	Label
1	1277.1569	3191	3489	32			/m/0gcs9
2	733.3225	790	992	32			/m/09j_c4
3	346.4903	218	423	31			/m/01hmkd9
4	315.8387	299	778	32			/m/0c83_
5	275.6321	130	266	27			/m/07y8z
6	202.1766	133	429	31			/m/0gpvt
7	189.8295	273	1404	32			/m/0f6c3
8	169.7850	1053	12251	32			/m/02py3
9	106.4968	35	127	19			/m/0lhr6r
10	100.9163	32	118	18			/m/04051
11	73.0128	30	168	19			/m/0lhr7v
12	67.5626	26	53	8			/m/0bzj8
13	54.9956	21	64	12			/m/0w9y5
14	49.9862	20	57	14			/m/04zz_
15	49.0301	20	53	2			/m/05f7wv6

NO IMAGE

The River

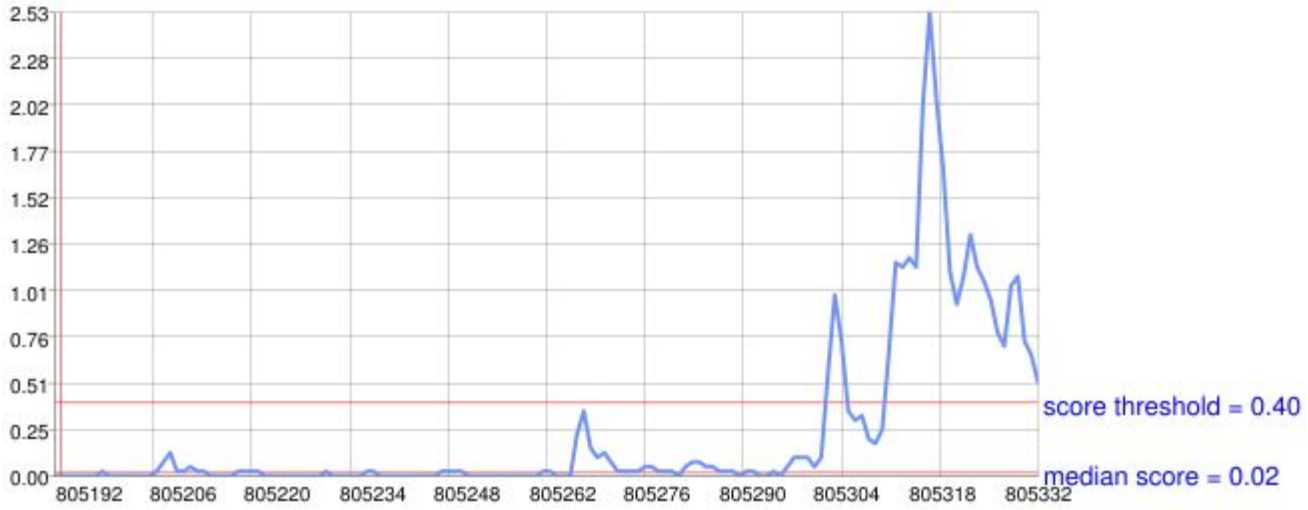
/en/the_river

The River is the fifth studio album, and the first double album by Bruce Springsteen. The album was released on October 17, 1980. Rolling Stone ranked it at number 253 on their list of the greatest albums of all time.

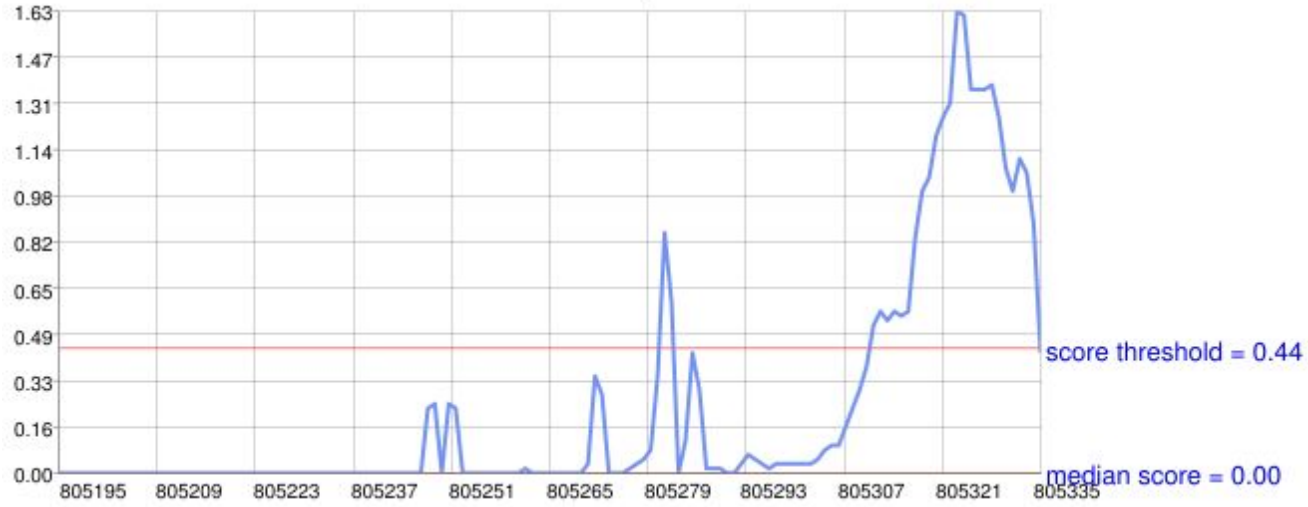
Query [John Lennon]

His death anniversary
Political figures took the opportunity to talk about gun law

Realtime Boost spike:



Twitter Hivemind spike (for comparison):



Correlated Salient Terms

From each article's salient term annotation we extract 20 salient terms to index as token class RTST. Aggregated by query, the list of salient terms returned in RealtimeBoostResult represents the news that is trending.

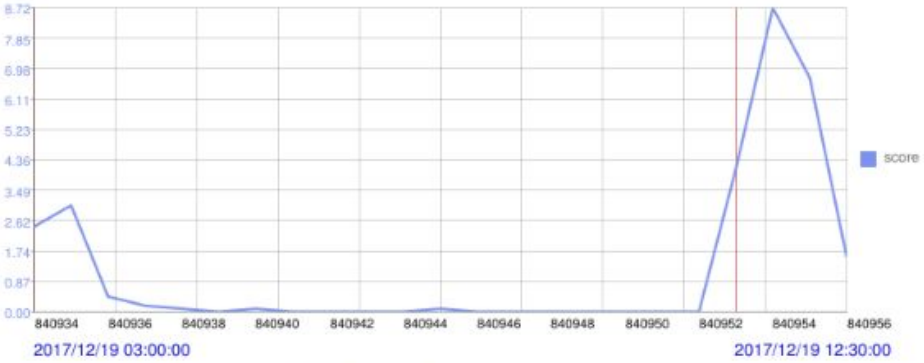
Query

shelly simonds

The Latest: Democrat leads recount in Virginia House race

The Latest on the recount in the Virginia House of Delegates race in Newport News (all times local):

Spikes



Spike 1 hour 34 min ago
Score: 0.65
Confidence: 0.92
Start Time: 12:30:00
End Time: 14:00:00

Spike 11 hours 4 min ago
Score: 4.20
Confidence: 0.70
Start Time: 03:00:00
End Time: 04:00:00

Salient Terms

Name	Score	Numerator	Denominator
virginia house	1017.84	65	173
recount	580.24	74	484
democrat	223.52	68	1631
virginia	102.84	62	4216
race	85.95	57	4717
house	52.51	72	14271
simonds	49.06	5	101
leads	40.91	32	5064
delegates	34.72	7	357
one vote	31.80	4	107
latest	29.79	29	6712
seat	22.11	10	1430
single	17.91	7	919
newport news	14.68	3	143
vote	13.44	23	13908

Correlated NGrams

From the article centerpiece (text) I scan through every 2-to-5 gram window of ngrams and add it to the token class called RTNG.
This way we get the correlated ngrams (up to 5-gram) of all the documents related to the query.
Here are some examples.

Query [Flood]

Los angeles had a weather alert for storms .

Realtime ngrams (rtng)

Matched sessions: 27134 of 140393739

Search:

Show 25 **entries**

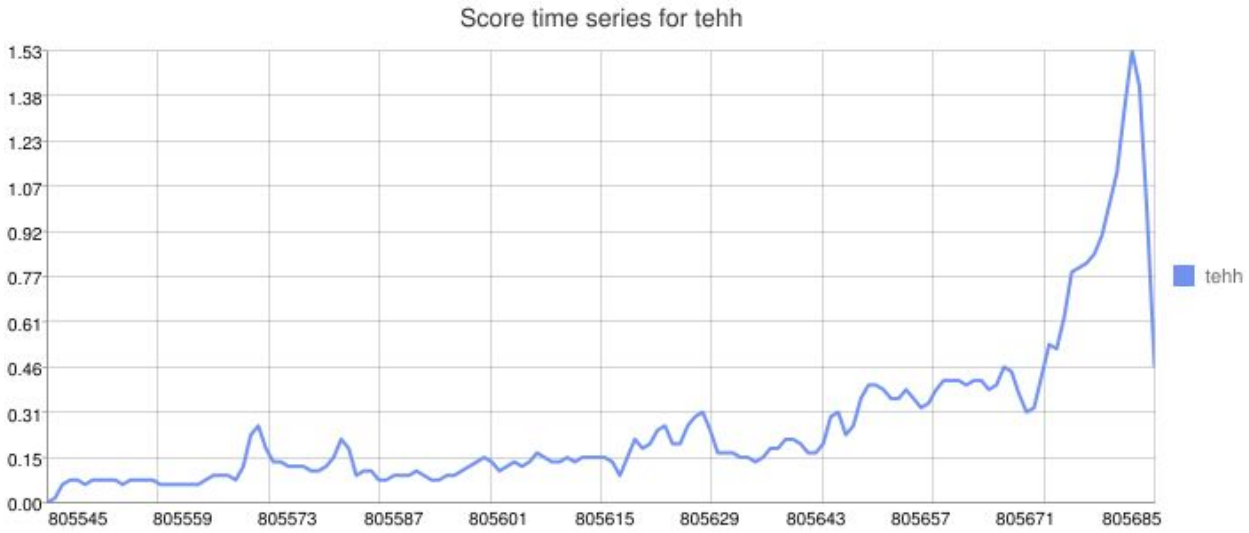
[First] [Previous] 1 2 3 4 5 [Next] [Last]

#↑	score	matched	count	bits	Go	HM	Label
1	154.8009	1353	1353	32			flood
2	90.5414	288	295	10			flood warning
3	89.0965	454	596	32			flooding
4	86.6952	243	243	1			los angeles weather alerts alert detail
5	86.6952	243	243	1			los angeles weather alert detail
6	86.6952	243	243	1			los angeles california alert detail
7	86.6952	243	243	1			los angeles ca alert detail
8	86.6952	243	243	1			los angeles alert detail
9	86.6952	243	243	1			flash flood warning los angeles

Query: Debate

During the republican debate

#↑	score	matched	count	bits	Go	HM	Label
1	93.1586	11306	11465	32			gop debate
2	80.9924	19468	30612	32			republican party
3	80.3338	7029	7139	32			republican debate
4	75.4633	5797	5910	32			c:gop debate
5	69.8955	8189	10857	32			republican presidential
6	69.7598	4663	4827	32			presidential debate
7	69.2916	4233	4237	32			d:gop debate
8	68.6793	4115	4119	32			debate
9	68.0664	5415	6206	32			c:republican presidential
10	67.8712	4146	4235	32			c:republican debate
11	66.8714	12066	20219	32			republican party united
12	65.9270	5471	6593	32			republican party preside
13	65.3374	5519	6763	32			party preside
14	64.4983	12132	21458	32			party united
15	61.3833	2958	2997	32			debate stage
16	60.1336	2967	3100	32			c:presidential debate
17	59.9441	4372	5461	32			presidential candidates
18	59.3864	3224	3560	32			c:presidential candidates
19	59.0497	2741	2836	32			republican presidential debate
20	59.0293	2598	2625	32			republicans take
21	58.7372	2584	2623	32			debate las vegas
22	58.6256	3900	4778	32			gop presidential
23	58.5777	2520	2539	32			take debate
24	58.5477	2516	2535	32			take debate stage
25	58.5307	2598	2657	32			debate las
26	58.3860	2475	2485	32			republicans take debate
27	57.5599	2671	2833	32			gop presidential candidates
28	56.2458	2252	2285	32			c:debate las vegas
29	55.9200	2256	2310	32			c:debate las
30	55.5905	2111	2115	32			d:republican debate
31	54.3992	2041	2077	32			fifth republican
32	53.3512	9712	20422	32			ted cruz
33	53.1176	1823	1823	32			d:debate stage



Query Southwest Airlines

Forced landing of airplane rolls off taxiway.

Realtime ngrams (rtng)

Matched sessions: 2097 of 139192756

Search:

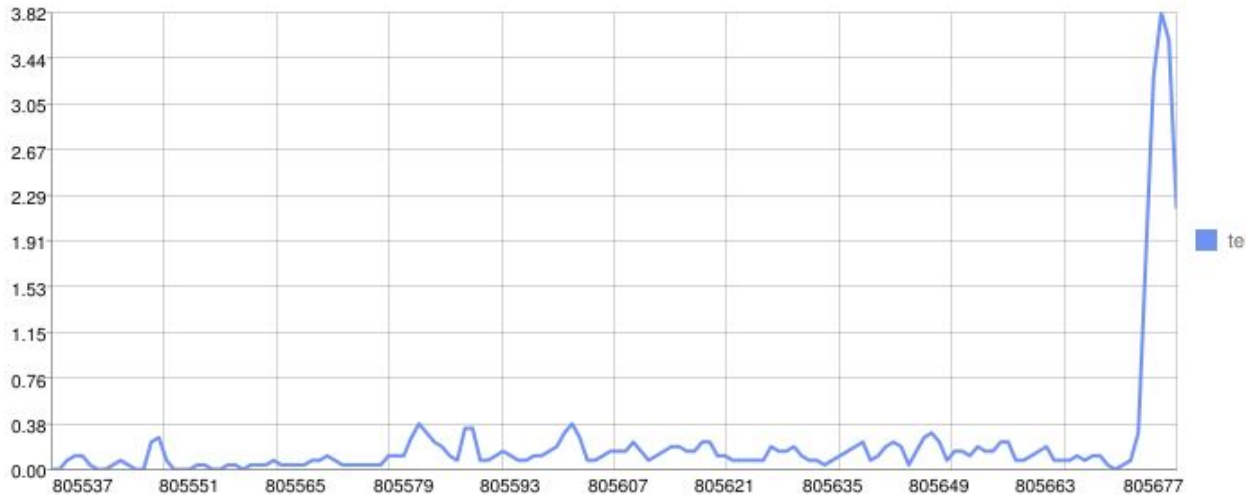
Show 25 **entries**

[First] [Previous] 1 2 3 4 5 [Next] [Last]

#↑	score	matched	count	bits	Go	HM	Label
1	1621.4200	1183	1522	32			southwest airlines
2	1368.3203	449	460	32			c:southwest airlines
3	1184.0359	289	292	32			southwest airlines plane
4	1064.6913	269	305	30			d:southwest airlines
5	1026.9805	202	208	31			c:southwest airlines plane
6	967.2355	230	276	32			off taxiway
7	965.6673	228	273	32			c:off taxiway
8	954.0674	280	379	32			c:airlines plane
9	937.3239	165	174	32			southwest airlines plane rolls
10	937.3239	165	174	32			airlines plane rolls off
11	937.3239	165	174	32			airlines plane rolls
12	903.0739	351	576	32			skids off
13	891.6963	293	447	32			airlines plane
14	871.7590	184	228	32			plane rolls off
15	871.7590	184	228	32			plane rolls
16	828.9980	206	291	30			plane skids
17	822.6434	198	277	32			c:rolls off
18	819.7647	202	287	30			plane skids off
19	815.3528	215	318	32			rolls off
20	789.9816	196	289	30			c:off runway nashville
21	789.6116	152	196	31			southwest plane
22	787.6816	199	297	30			c:runway nashville
23	779.5639	123	144	26			c:after landing
24	769.6324	199	307	30			runway nashville
25	768.4096	197	303	30			off runway nashville

Showing 1 to 25 of 500 entries

Spike:



Happened in Nashvile Airport:

Realtime KG entities (rtkg)

Matched sessions: 2097 of 139192756

Search:

Show 25 ▾ **entries**

[First] [Previous] 1 2 3 4 5 [Next] [Last]

#↑	score	matched	count	bits	Go	HM	Label
1	1973.4531	1617	1828	32			/m/0gztl
2	728.3311	160	238	32			/m/02ckzq
3	395.7611	135	438	30			/m/01ql_l
4	327.3161	58	152	1			/m/0
5	281.9748	263	1967	32			/m/0
6	232.7881	91	506	3			/m/0
7	223.5548	36	80	20			/m/0
8	153.5115	119	1394	17			/m/0
9	139.9204	209	3778	32			/m/0
10	120.2489	309	8475	32			/m/0
11	118.2290	92	1355	28			/m/0
12	93.9277	33	358	18			/m/0
13	72.5776	16	54	7			/m/0
14	63.4859	15	157	1			/m/0
15	62.2516	17	203	7			/m/0
16	57.0977	114	5357	1			/m/0
17	50.7928	11	120	2			/m/0
18	48.2759	10	93	8			/m/0
19	47.7467	60	2511	1			/m/0
20	43.3727	27	768	17			/m/0cb0xm
21	43.0645	18	381	2			/m/01kyln
22	42.5846	10	72	1			/m/01qgw0
23	41.2407	10	66	3			/m/01kyd9
24	40.3274	9	88	1			/m/01my_s
25	39.8602	49	2342	7			/m/01mz2g

Showing 1 to 25 of 500 entries

Nashville International Airport

/en/nashville_international_airport

Nashville International Airport is a public and military use airport in the southeastern section of Nashville in the U.S. state of Tennessee. It is included in the National Plan of Integrated Airport Systems for 2011–2015, which categorized it as a primary commercial service airport. Established in 1937, its original name was Berry Field, from which it its ICAO and IATA identifiers are

Nashville

Unigrams:

Realtime unigrams (rtw)

Matched sessions: 2097 of 139192756

Search:

Show 25 **entries**

[First] [Previous] 1 2 3 4 5 [Next] [Last]

#↑	score	matched	count	bits	Go	HM	Label
1	1034.9033	345	462	32			taxiway
2	616.8649	386	1145	32			skids
3	457.7344	2088	21768	32			southwest
4	411.0762	2071	25030	32			airlines
5	270.9335	130	710	31			a:southwest
6	199.4188	72	436	28			cp:harlingen
7	190.2547	51	268	16			skidded
8	164.4831	57	395	1			thurgood
9	145.6877	274	5375	31			737
10	141.9999	326	7244	32			cp:southwest
11	136.8413	737	25803	32			nashville
12	136.2762	89	1050	30			harlingen
13	131.1796	93	1188	7			b737
14	121.2045	371	11029	32			runway
15	110.3190	287	8576	32			cp:airlines
16	105.4562	67	966	28			cp:987
17	102.8704	92	1653	29			diverted
18	100.8224	77	1285	27			cp:irregular
19	100.0948	17	97	1			airportfaa
20	96.2436	643	34835	32			plane
21	93.6512	356	14985	32			airplane
22	87.0194	66	1241	2			bwi
23	86.0085	80	1713	30			987
24	79.8516	372	20100	32			landing
25	79.5037	278	13051	31			boeing

Showing 1 to 25 of 500 entries

Query: Mia Airport

Bomb threat in airplane in MIA Airport. Bombsquad is called to search the airplane.

Realtime ngrams (rtng)

Matched sessions: 212 of 139212186

Search:

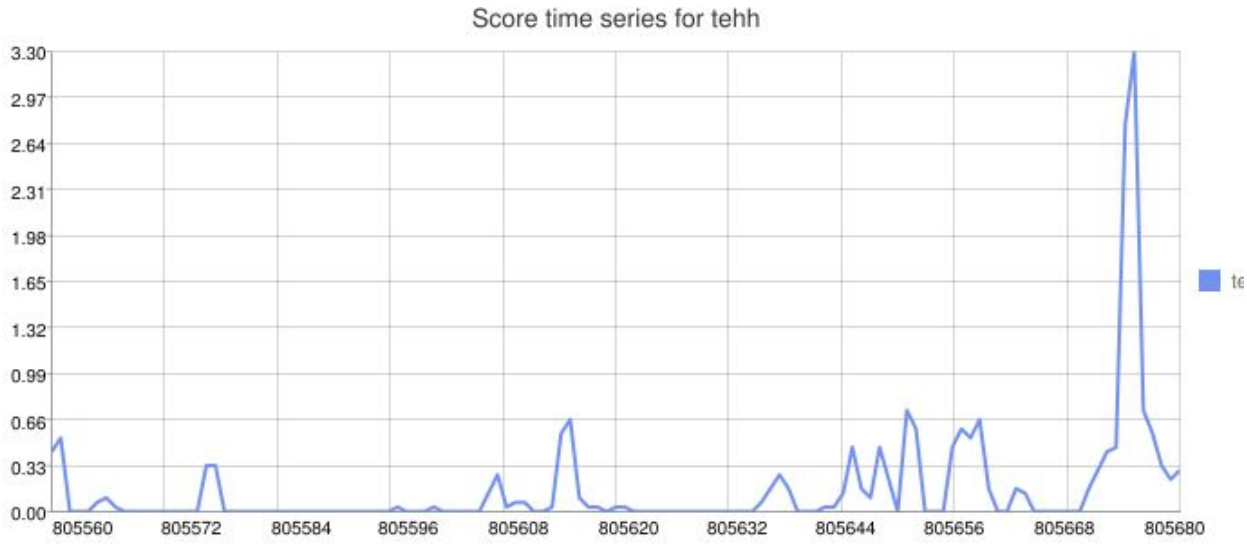
Show 25 ▾ **entries**

[First] [Previous] 1 2 3 4 5 [Next] [Last]

#↑	score	matched	count	bits	Go	HM	Label
1	4377.2720	87	197	29			miami international airport
2	2454.0334	86	442	29			miami international
3	2184.8279	36	125	22			c:searches plane miami
4	2184.8279	36	125	22			c:searches plane
5	1853.6160	30	116	21			c:squad searches plane miami
6	1794.2197	43	223	22			c:bomb squad
7	1791.3423	36	166	22			squad searches plane miami
8	1791.3423	36	166	22			squad searches plane
9	1791.3423	36	166	22			squad searches
10	1791.3423	36	166	22			searches plane miami
11	1791.3423	36	166	22			bomb squad searches plane
12	1791.3423	36	166	22			bomb squad searches
13	1776.3876	36	168	22			searches plane
14	1719.4730	36	176	22			c:plane miami
15	1683.9688	27	111	20			c:bomb squad searches plane
16	1683.9688	27	111	20			c:bomb squad searches
17	1663.0496	27	113	20			c:squad searches plane
18	1663.0496	27	113	20			c:squad searches
19	1587.7883	46	297	23			bomb squad
20	1572.2921	36	200	22			plane miami
21	1330.4365	25	64	17			d:miami international airport
22	1279.1768	26	154	18			d:miami international
23	1082.1969	18	98	14			near miami
24	1067.5865	18	104	14			unit apartment
25	933.3984	18	126	14			building near

Showing 1 to 25 of 500 entries

Spike:



How to use the signal

We call hivemind from QRewrite servlets, to make this signal available in Rewritten query. Anyone in Superroot could consume the RealtimeBoostResponse that is based on the web search query.

The signal is also available in Superroot. Clients can call with a query, and we will reply if that query is trending, and what else is related (entities, places, etc).

Another option is to use our library to analyse Hivemind's response, so you will only need to call hivemind and use the library to receive back a signal with a score of how likely this query is spiking / trending right now (or in the recent past), and the quality of the spike (based on NSR and Freshbox Article Score).

Live Experiment and SxS with Freshbox

Still under development.

We push freshbox up to top3 positions for the queries that are trending according to this Realtime Boost signal.
Preliminary SxS

Other ideas: Add unigrams, ngrams and correlated kg entities to the SQuery as optional terms to help bringing recent documents related to the event that is spiking.