The Distributed Crawler

v2.0-chaika

Service architecture overview

The Hierarchical Cluster Engine project

IOIX Ukraine, 2014-2017

Introduction

The HCE-DC is a multipurpose high productivity scalable and extensible engine of web data mining.

Built on several HCE project's sub-products and technologies:

- The **hce-node** a network transport cluster application.
- The Distributed Crawler (**DC**) service.
- The Distributed Tasks Manager (**DTM**) service.
- Web administration management console.
- The tools and libraries for crawling and scraping algorithms with REST API and bindings for a Python and PHP development environments.

provides a flexible configuration and a deployment automation to get installation more closed with a target project and easy integration.

Main functional purposes

- **Crawling - scan** web sites, **analyze** and **parse** web pages, **detect** and **collect** URLs links and web resources. **Download** resources from web-servers using automatically collected or provided URLs including dynamic JS rendered web-pages and **store** them in a shard local raw file storage.

- **Processing** of a web page content with several customizable applied algorithms like a unstructured textual content ***scraping***, ***statistical*** data mining, ***NLP*** data mining and so on, and **store** results in local SQL DB storage with distributed multi-host and multi-process architecture model.

- **Tasks management** of crawling, processing and data archiving as well as internal distributed data architecture tasks like aging, purging, statistical and so on. Tasks scheduling and balancing using tasks management service of multi-host architecture or real-time multi-threaded load-balancing client-server architecture.

Extensible developer’s architecture

- **Developer's API** – full access for configuration, deployment, monitoring and management processes and data.
- **Applied API** – full featured multi-thread multi-host REST http-based protocol to perform crawling and scraping batch requests.
- **Web administration management console** – for DC and DTM services with support of user's accounts, roles, permissions, crawling&scraping, results collect, aggregation, archiving and configurable custom post-processing, statistical reports, e-mail notifications, triggering and another utility tools.
- **Helper tools and libraries** – several support applied utilities to convert, prepare, parse, format, data used in sequential tasks chains as input and output objects.

Distributed asynchronous nature

The HCE-DC service engine itself is a fully distributed and parallel. It can be deployed and configured as single- and multi-host installation. Key features and properties of a distributed parallel architecture:

- No central database or data storage for crawling and processing. Each physical host unit with the same storage shards portion of data but represented as atomic service.
- Crawling and processing goes in parallel multi-process way on each physical host including JS execution in a browser environment (in case of dynamical fetcher usage) downloading, DOM and raw text parsing, URLs collecting, fields extracting, post-processing and so on tasks.
- Customizable strategies of data sharding and requests execution balancing with minimization of data redundancy and optimization of system resources usage (CPU, RAM, DISK).
- Resulted data merging avoiding of resources duplicates.

Flexible balancing and scalability

The HCE-DC as service can be deployed at set of physical hosts. A number of hosts depends on their hardware productivity rate (CPU cores number, RAM size, disk space and speed, network interface speed and so on) and can to be scaled from one up to IPv4 C class network hosts number (254). Key scalability principles are:

- A computational node is a physical or logical host (any kind of virtualization and containers supported).
- Nodes can be added in to the system and gradually filled with data at run-time. No dedicated data migration.
- Computational tasks can be configured as round-robin (RR) or resource usage (RU) balanced. In case of RU-balancing tasks scheduler selects node with maximum free resources using customizable estimation formula. Different system indicators available: CPU, RAM, DISK, IO wait, processes number, threads number and so on.

Scalable software and algorithms

The HCE-DC service for the Linux OS platform has three main parts:

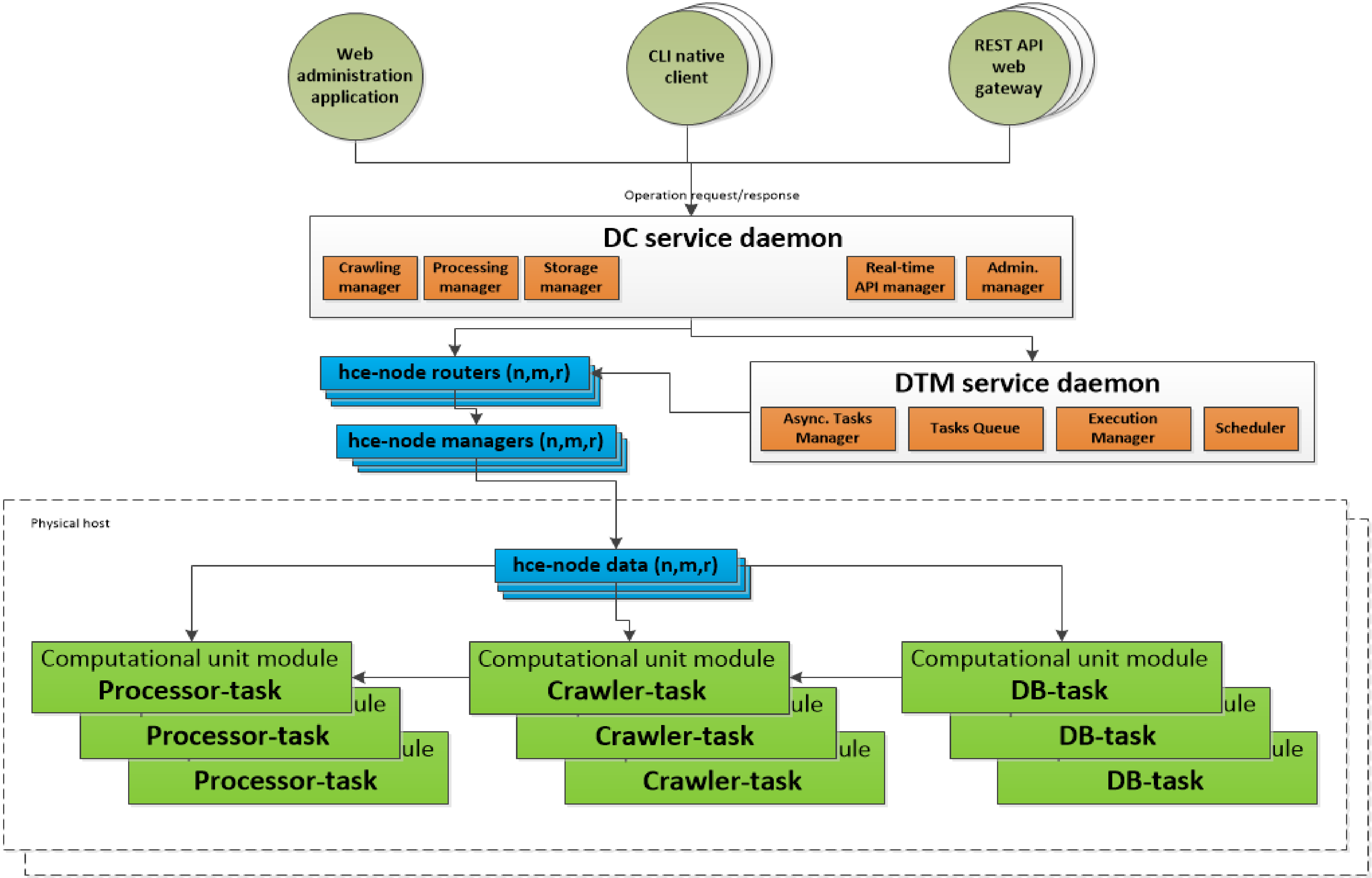
- A core daemon module with functionality of: scheduler of crawling and processing tasks, managers of: tasks queues, periodical processes, computational nodes storage, real-time API requests tasks and so on. Typically the core daemon process runs on a dedicated physical host and represents service itself.
- The computational unit modules started on a computational node in a session-based manner to perform a batch processing including crawling, scraping, statistical post-processing, storage SQL DB and raw file algorithms as well as some helper utilities.
- Administration management web application provides a standard web UI with support of projects, data collectors, notifications, stat and another reports. It can be configured to work wits several core daemons and switched on demand.

Open processing architecture

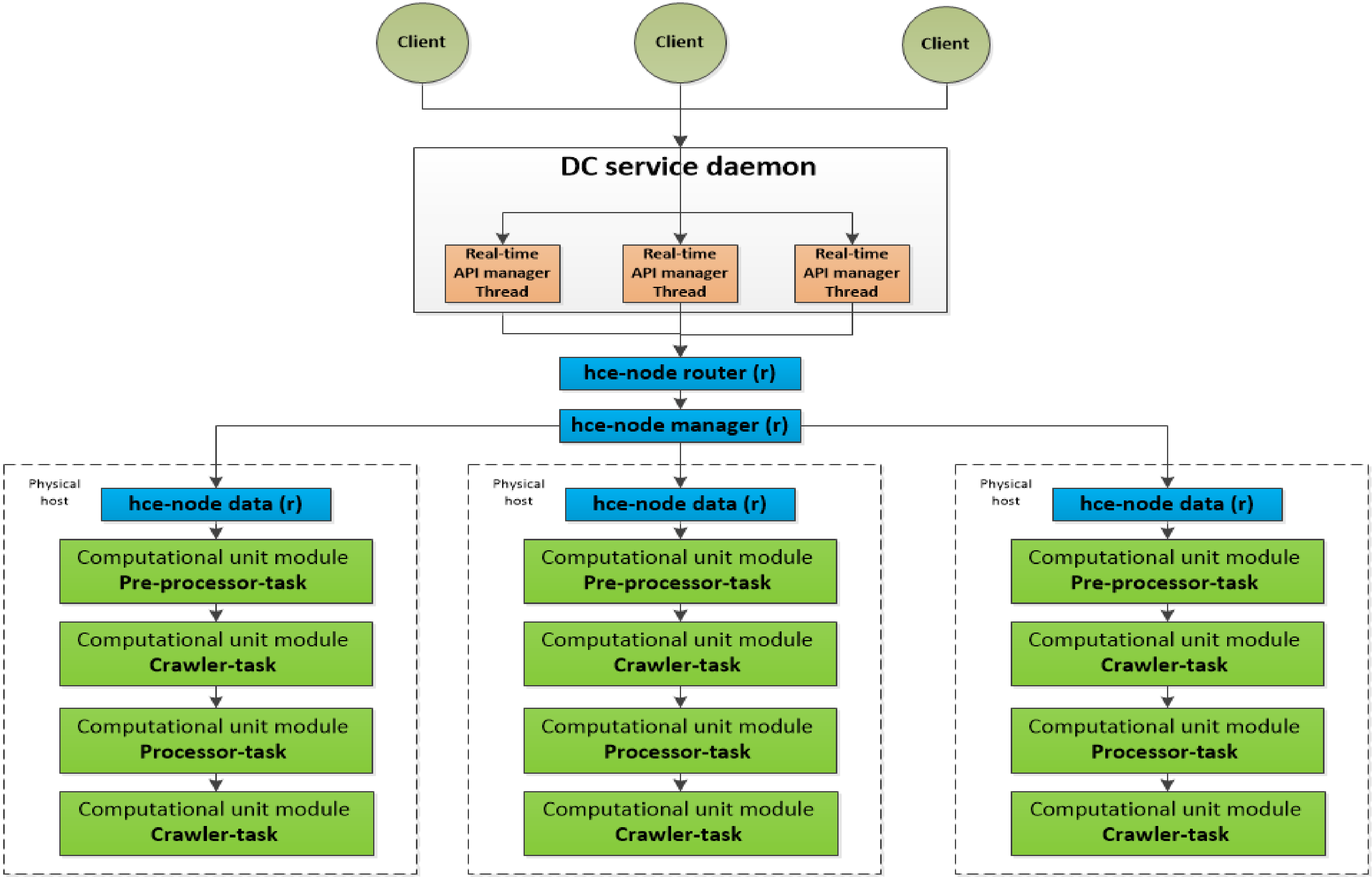
The computational nodes modules set can be extended with any kind of a custom algorithms, libraries and frameworks for any development frameworks and programming languages. The limitation is only an API interaction translation that typically needs some adapters or converters. Key principles are:

- Data processing modules involved as native OS processes or via API including REST and CLI.
- Process instances are isolated by Linux OS.
- A CLI API is default for processes chains and data exchange or simulated by converter utilities.
- Open input/output protocol used to process interaction objects like a batches sequential way step by step by each processing chain.
- Closed data formats can be easily serialized – json, xml and so on.

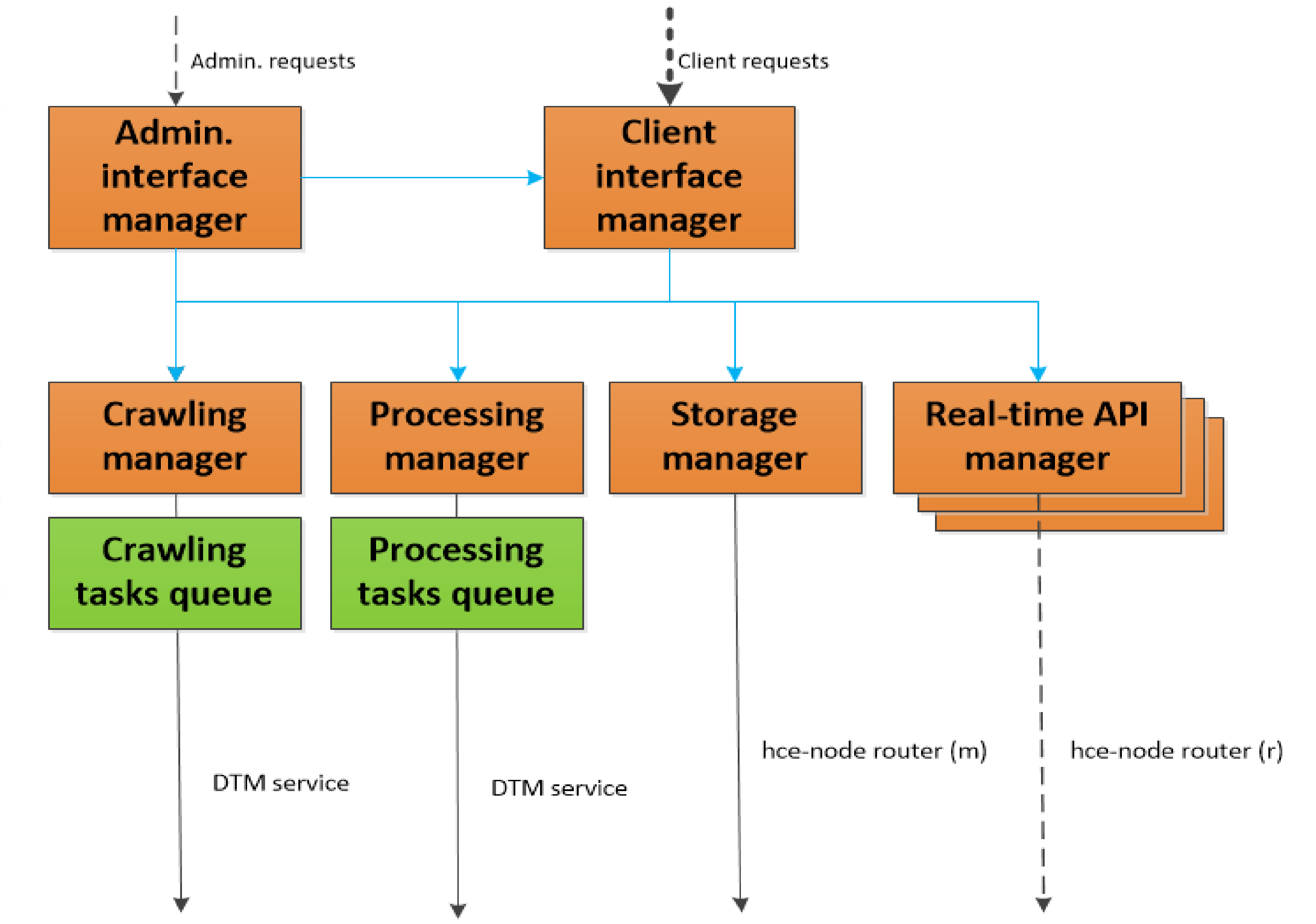
General DC service architecture



Real-time client-server architecture


Client
Client
Client
DC service daemon
Real-time API manager Thread
Real-time API manager Thread
Real-time API manager Thread
hce-node router (r)
hce-node manager (r)
Physical host
hce-node data (r)
Computational unit module Pre-processor-task
Computational unit module Crawler-task
Computational unit module Processor-task
Computational unit module Crawler-task
Physical host
hce-node data (r)
Computational unit module Pre-processor-task
Computational unit module Crawler-task
Computational unit module Processor-task
Computational unit module Crawler-task
Physical host
hce-node data (r)
Computational unit module Pre-processor-task
Computational unit module Crawler-task
Computational unit module Processor-task
Computational unit module Crawler-task

Internal DC service architecture



Brief list of main DC service features

A fully automated distributed web crawling with: projects with more than 100 configuration options including sets of root URLs, periodical re-crawling, HTTP and HTML redirects, http timeouts, dynamic JS rendering, priorities, limits (size, pages, contents, errors, URLs, redirects, content types), requests delaying, robots.txt, rotating proxies, RSS (1, 2, RDF, Atom), scanning depth, complex filters, splitted html pages or chains, batching, HTTP header optimizations.

A fully automated distributed web-pages processing with: News™ (pre-defined sequential scrapers and extractors based on Goose, Newspaper and Scrapy) and Template™ (universal rules definitions to extract data from pages based on xpath and csspath, content parts joining, merging, best result selection with metrics, regular expressions post processing, multi-item pages split and join (product, search results, articles split on parts, etc...), multi-rule, multi-template compositions and so on) scraping engines, WYSIWYG templates editor, processed contents merging and so on.

A fully automated resources management: periodic operations, aging, purging, update, re-crawling and re-processing.

A web administration console: full CRUD of projects for data collect and process with set of parameters per project, users with roles and permissions ACL, DC and DTM service's statistics, crawling and processing project's statistics.

A real-time API: native CLI client, asynchronous and synchronous REST requests support.

Applied texts mining solutions

On a basis of a DC core engine architecture several applied textual data mining solutions with implementation of several known and new algorithms available for a target projects.

- ***General Mining*** – basic statistics computation and basic text analysis of corpus of texts like Sentences, words, characters average and frequencies, parsers, stemming, stochastic filters, languages detection, and so on.

- ***Classification and Entities Detection*** (HCE CED^{TM}) – computation, Bayesian vocabulary-based algorithms with NLP elements. Provides a possibility to classify a textual data content (an article, or example) on belonging to a category with basic terms vocabulary definition and statistical threshold.

- ***Sentiment Analysis*** (HCE SA^{TM}) – computation, Bayesian vocabulary-based algorithms with NLP elements. Provides a possibility to detect a sentiment (positive, negative or neutral) and amount of it for a textual content.

- ***Similarity Rate*** (HCE SR^{TM}) – computation and grouping, Bayesian vocabulary-based algorithms with NLP elements. Provides a possibility to calculate a rate of a similarity between textual contents and to group similar contents.

Applied texts mining solutions

- ***Popular Words*** (HCE PW^{TM}) detection most popular key phrases, terms, entities especially for news articles based on HCE SR. Provides a possibility to detect short (single or several words) phrases that are key sense or main point of an article. Also, a possibility to detect lists of related by Popular Words or Alternate articles in mass set of collected data for a period of time like a six hours or daily.
- ***Social Networks Data Mining*** (NewsHub SNDM^{TM}) (Tw, Fb, G+, Li, and so on), messages bodies and indicators (messages numbers, likes, shares, etc…) scraping and analysis including all described statistical algorithms techniques.
- ***Multi-lingual engine*** including parsers, tokenizers, stemmers, with support of English, Japan, Polish, Russian, Ukrainian languages with NLP elements and synonyms support, potentially expendable for any language in the world.

Applied technologies and tools

On a basis of a text mining solutions several applied technologies and tools created and used for target projects.

- ***Popular Words Time-line tool*** – a visual tool with possibility to get a report with chart (area, line, bar, pie, gantt, and so on) to visualize a time-line dynamics of changes of a pop-word frequency, number of articles, social indicators (posts, re-posts, likes, shares, etc), pop-words trends detection by direction of a process (rise, fall, both), filtration of noise to show new, sparse, not periodical, periodical, single tracks of pop-words. Reports can be used for mechanisms seek and research in area of dynamics of popularity of descriptive entities, acronyms, goods, trade marks, brands, persons, political parties, news media sources analysis and so on and visual data representation.

Applied technologies and tools

- ***Hot topics visualization tool*** – a form of a representation of a pop-words detected and collected for a period of time with filtration by multi-level classification with HCE CED.
- ***Modeling*** - Investigations of a topics probability expectation, predictions and probability hypothesis checks related on social networks textual data or based on them. Modeling and an automations of statistical researches of a public echo on topic events and relations with public media information.
- ***NewsHub*** – a news and information web-site engine with integration of any HCE's tools and solutions. WordPress-based front-end with plug-ins and widgets oriented statistical data representation provides a lists of news articles ranged with HCE custom ranks like Similarity Rank, Social Networks Data Rank and Synthetic Composite Rank; list of Similar Articles for each article, list of Alternate Articles for each article and list of Related Articles; Sentiment Rate; categories of articles filled with usage of HCE Classifiers.

Possible solutions

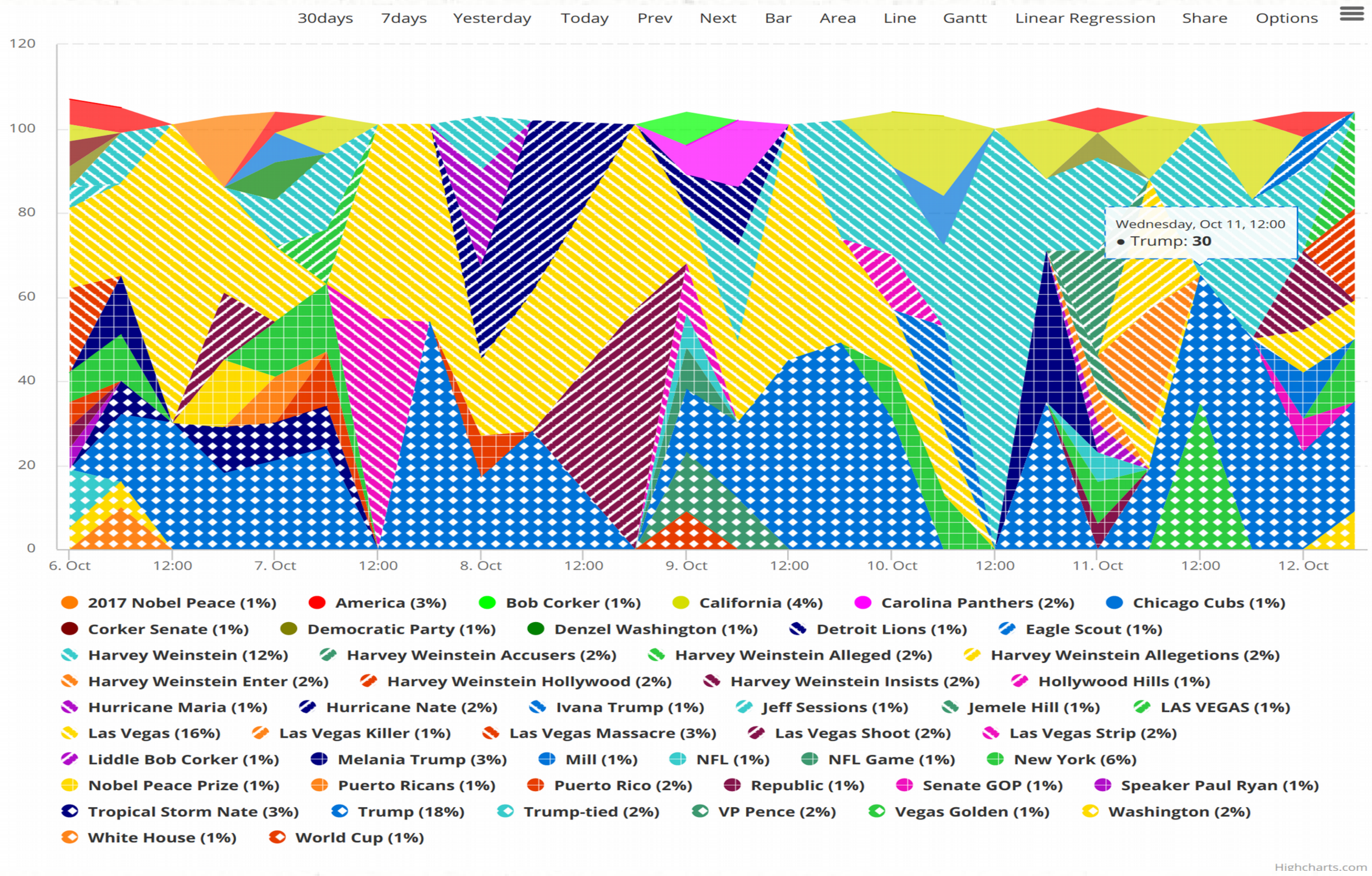
- Deep mutual documents **Citation and Intersection** detection.
- **POP-words Correlation** detection and probability estimations.
- **POP-words Trends** detection including time series analysis harmonic decomposition, Fourier transformation and synthesis, prediction and pair and multiple linear correlation.
- **Custom Classifications** with dynamically generated or external classifier vocabularies including informational events detection and watchdogs tools, filtration and screening tools, general subjects verification and detection tools and related solutions.
- **Social Networks Messages Mining** including a natural (number of messages like posts, twits and so on, re-posts, shares, likes and so on) and synthetic (sentiment rate, POP-words rate, etc) indicators computation, periodical POP-words tracking, topics and themes relations detection and customizable visualization tools in combination with Custom Classification – of a specially tuned tools for social networks messages.

Possible engines and frameworks

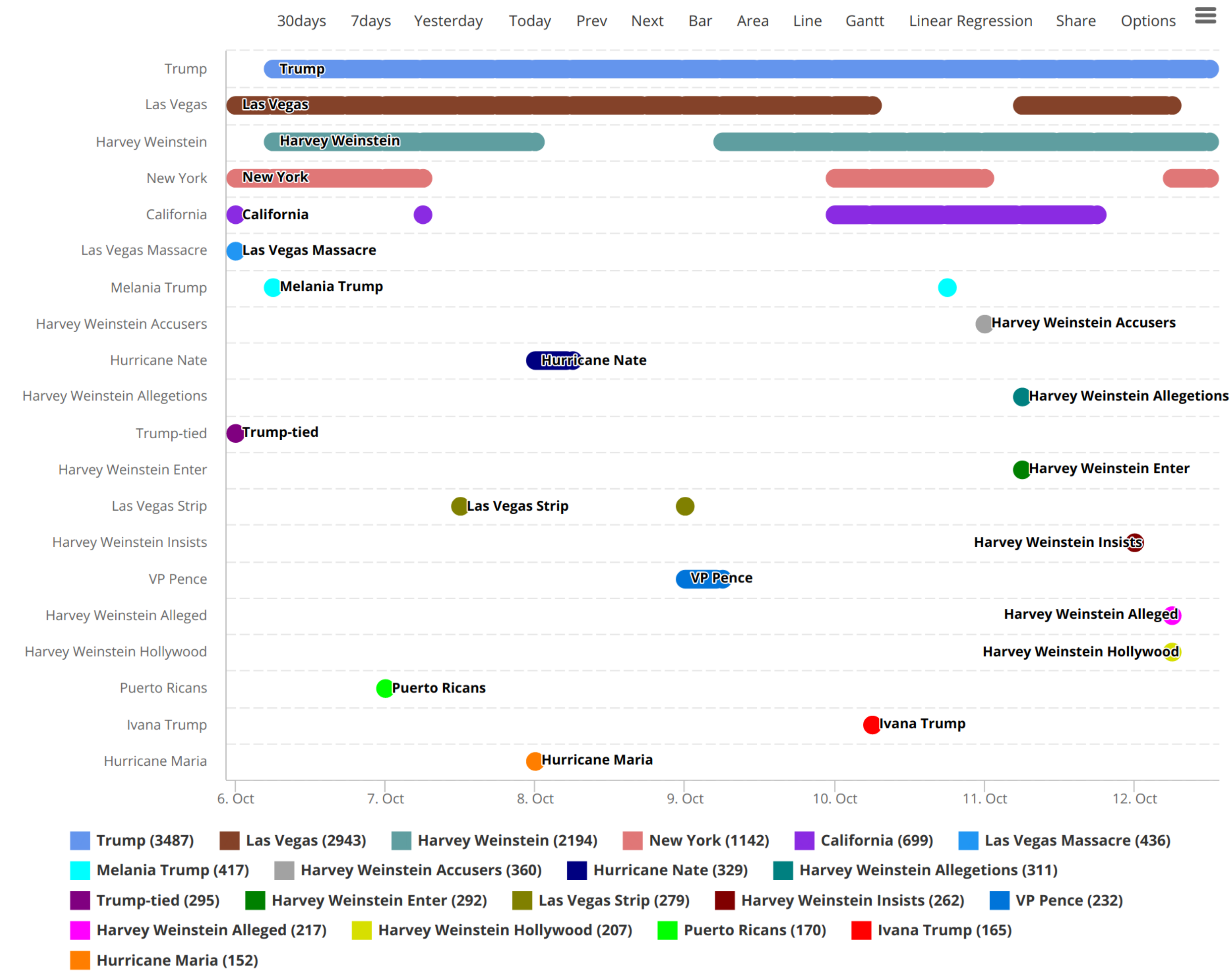
- News media aggregation and analysis informational **web-site engine** with full cycle of data mining from articles harvesting to many different ways of visualization including an articles lists - sorted and filtered with many criteria, POP-words, personalized outputs including reports, mail subscriptions and PDF digests.
- Textual data analysis web-service tools with UI and API in SaaS manner.
- Custom configured and supported textual data collect and/or processing service back-end engine with distributed multi-host, multi-task, multi-process, multi-threaded, multi-node, extensible computations and storage architecture.
- Web Events Watchdog services and framework engine of common (completely freely customizable schedule, change detection and reaction rules and actions) and special/custom edition (e-commerce e-shops prices watchdogs, stock or currency exchange indexes watchdogs and so on web-sites).

Examples of graphical visualizations and data representation views

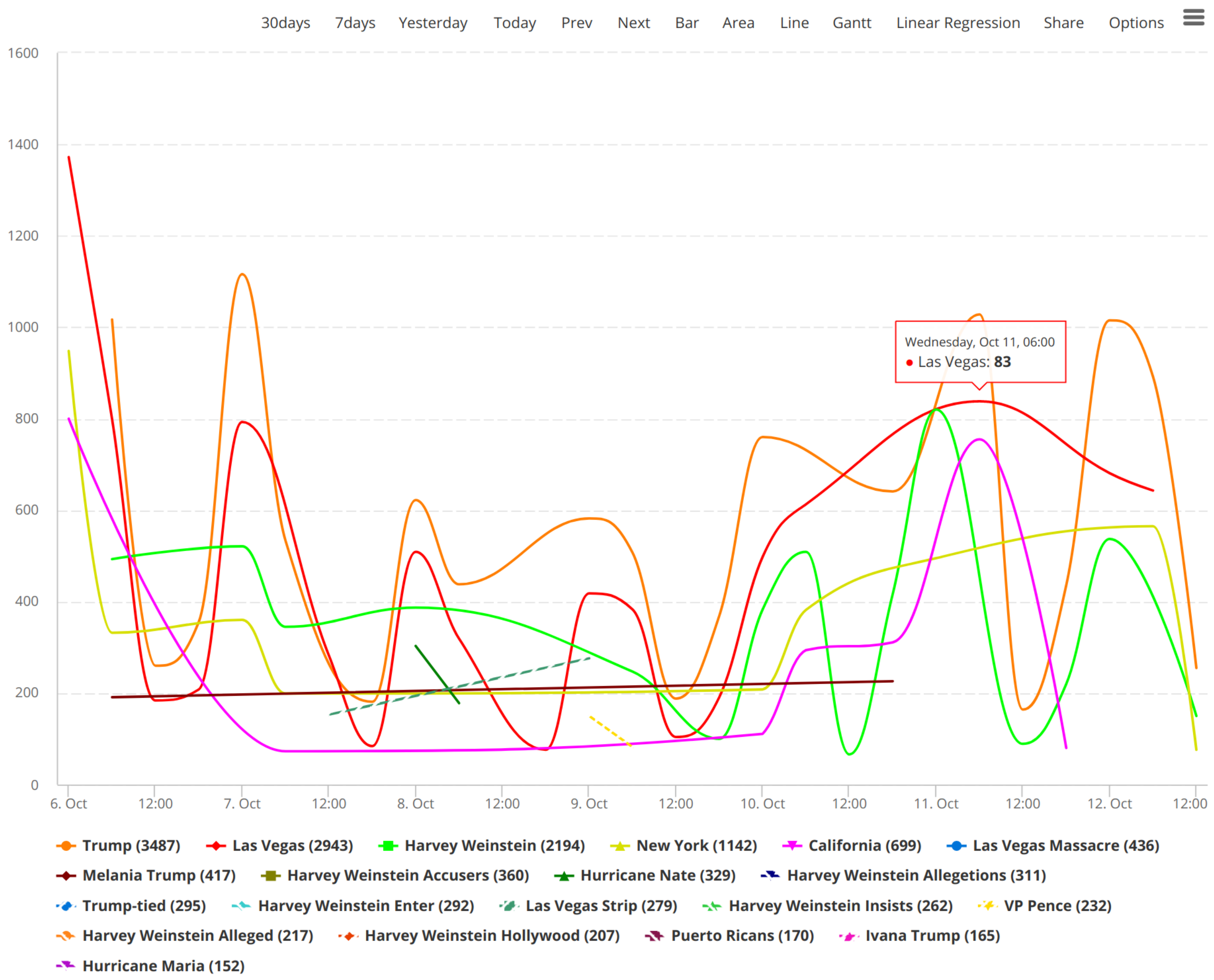
A Popular Words time-line tool: Area chart for 7 days top 50 words dynamic



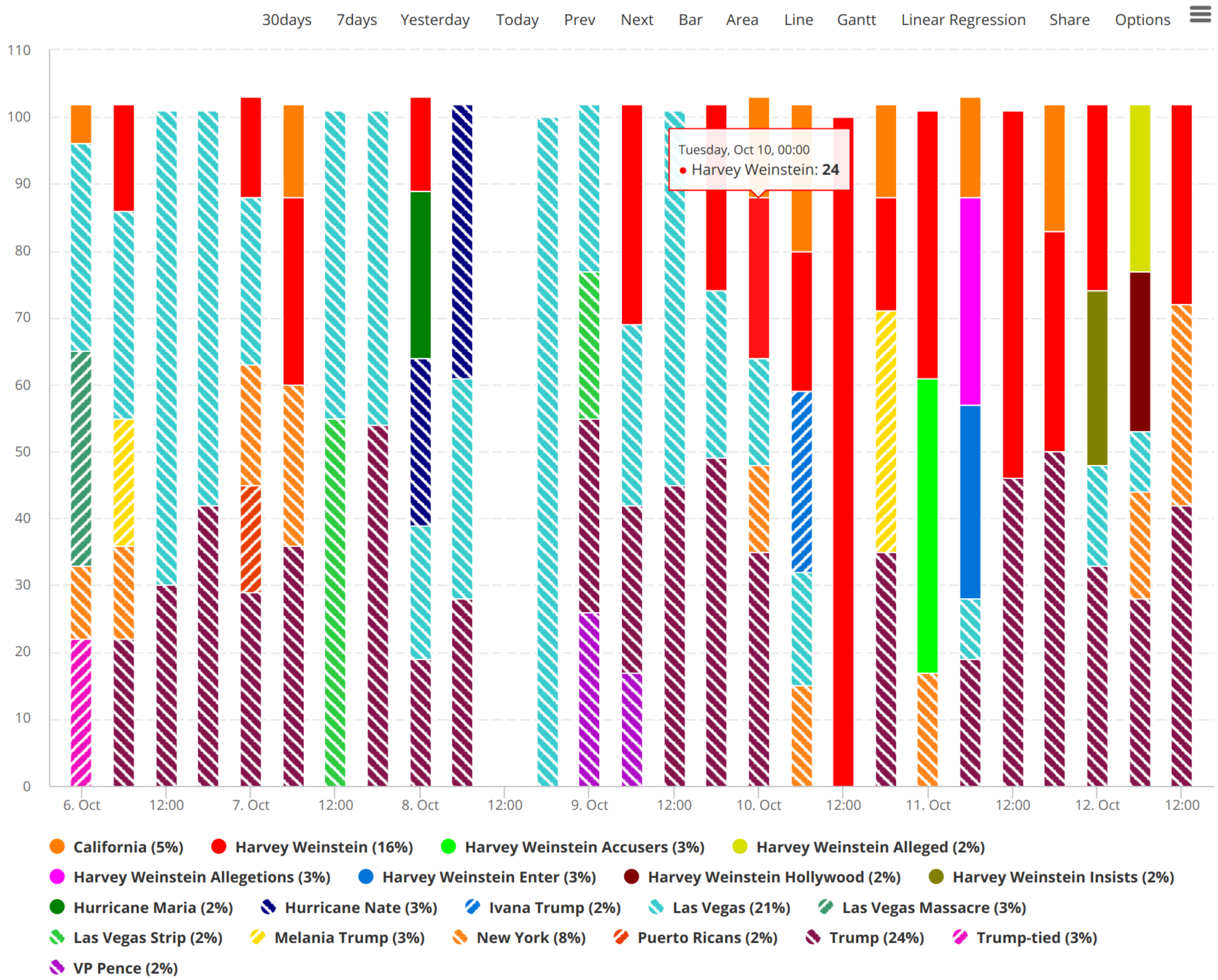
A Popular Words time-line tool: Gantt chart for 7 days top 50 words dynamic



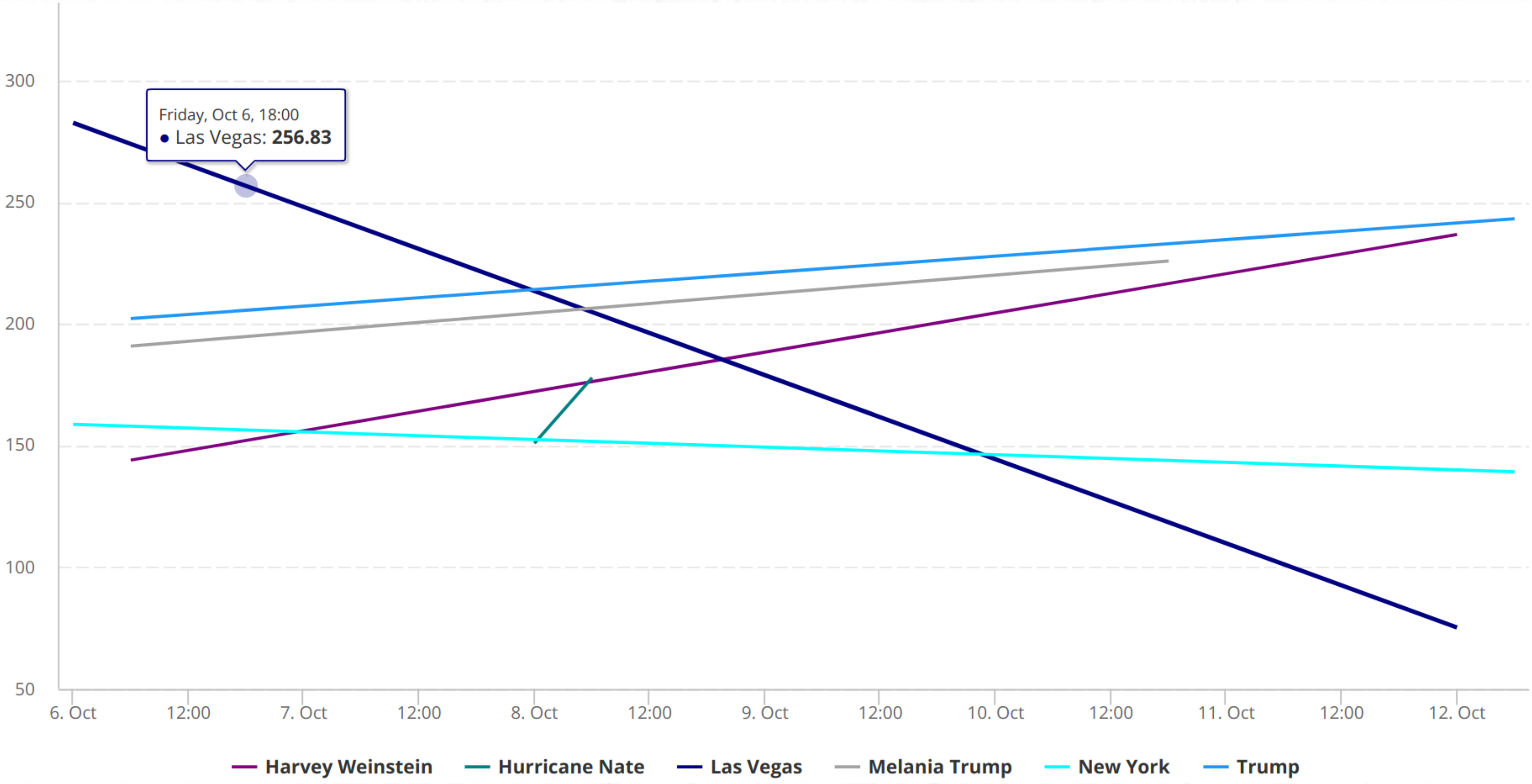
A Popular Words time-line tool: Line chart for 7 days top 50 words dynamic



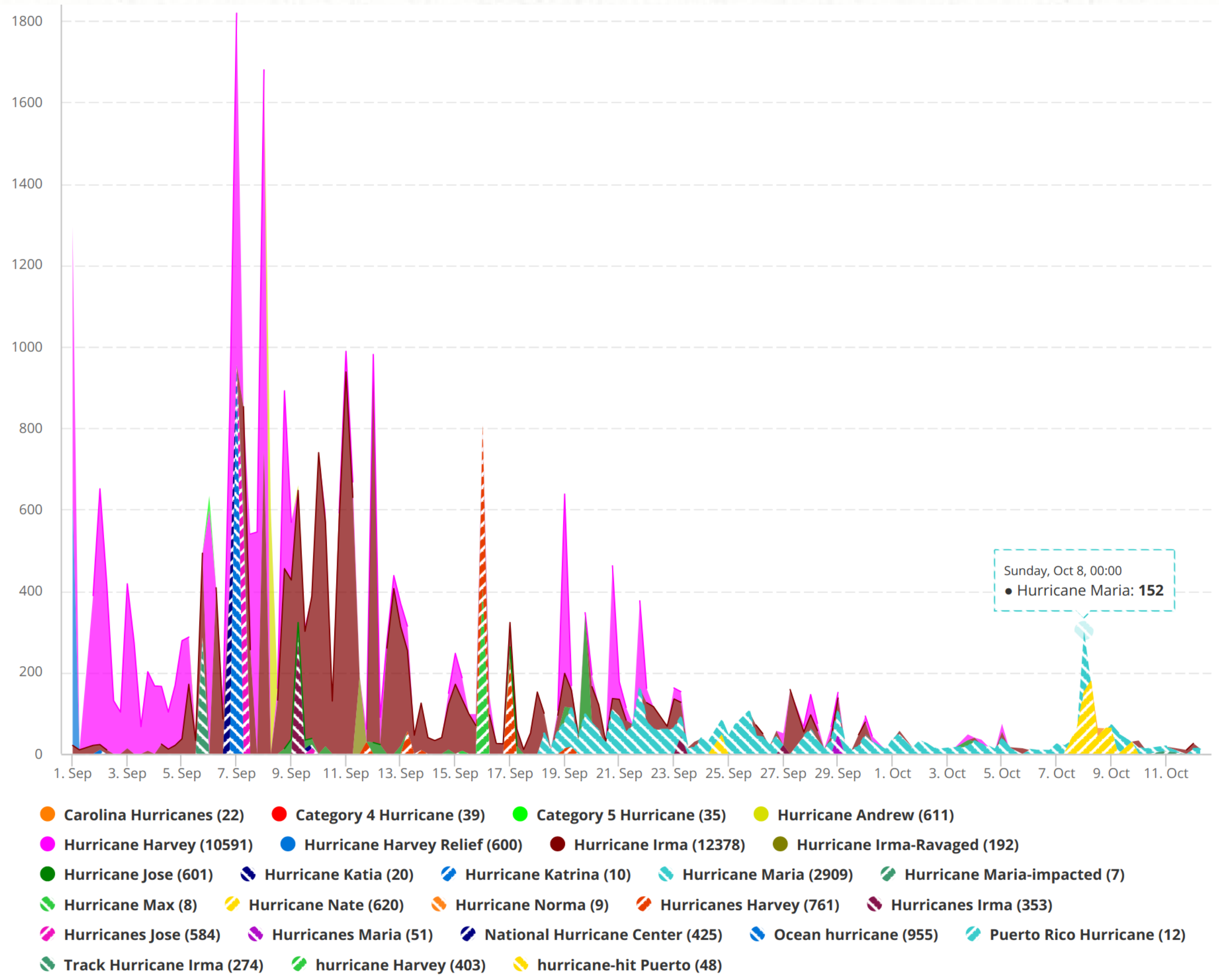
A Popular Words time-line tool: Bar chart for 7 days top 50 words dynamic



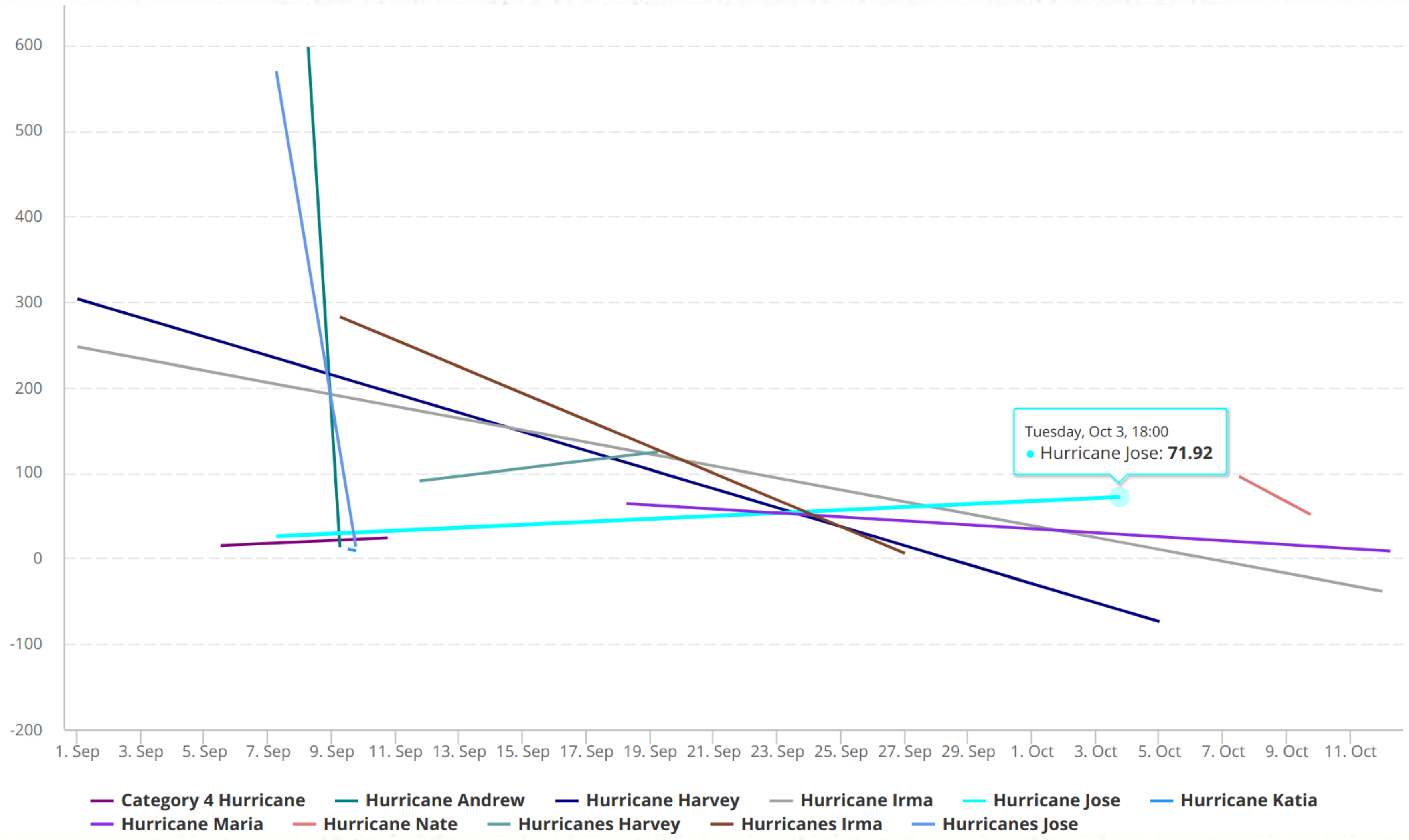
A Popular Words time-line tool: Trends chart grow & fade for 7 days top 50 words



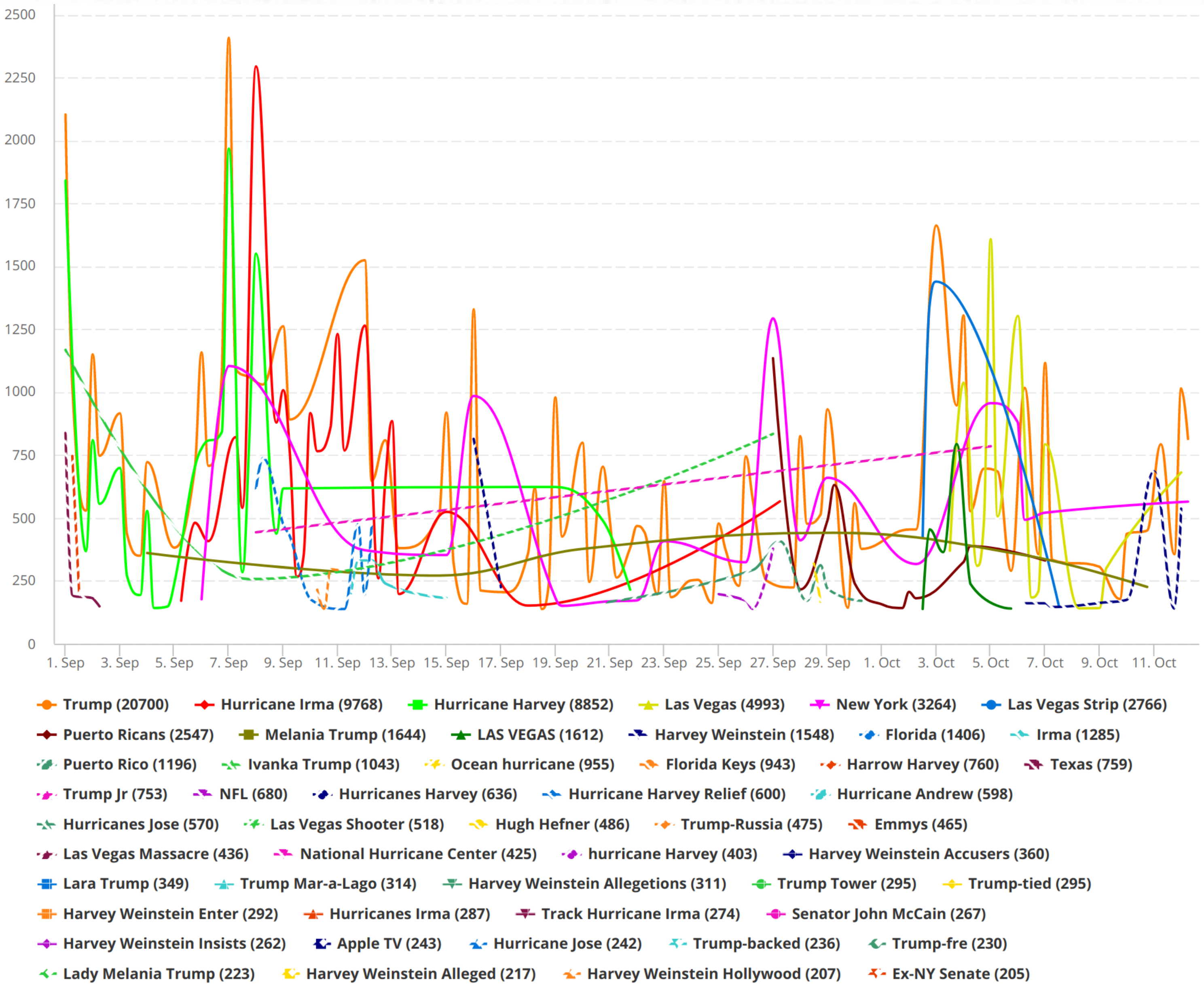
A Popular Words time-line tool: Area chart for 1 month a “Hurricane” pop-words



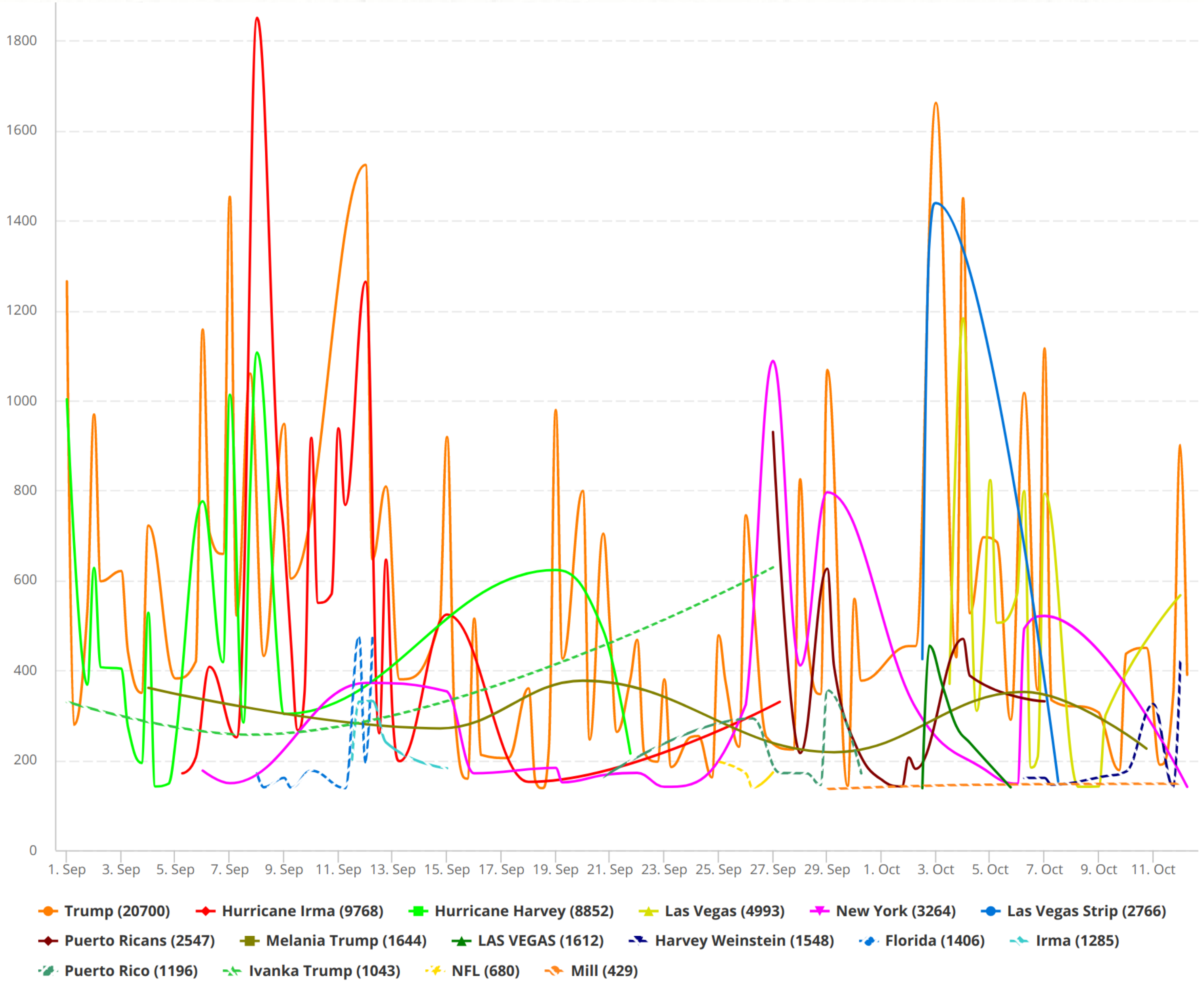
A Popular Words time-line tool: Trends chart grow&fade for 1 month a “Hurricane” pop-words



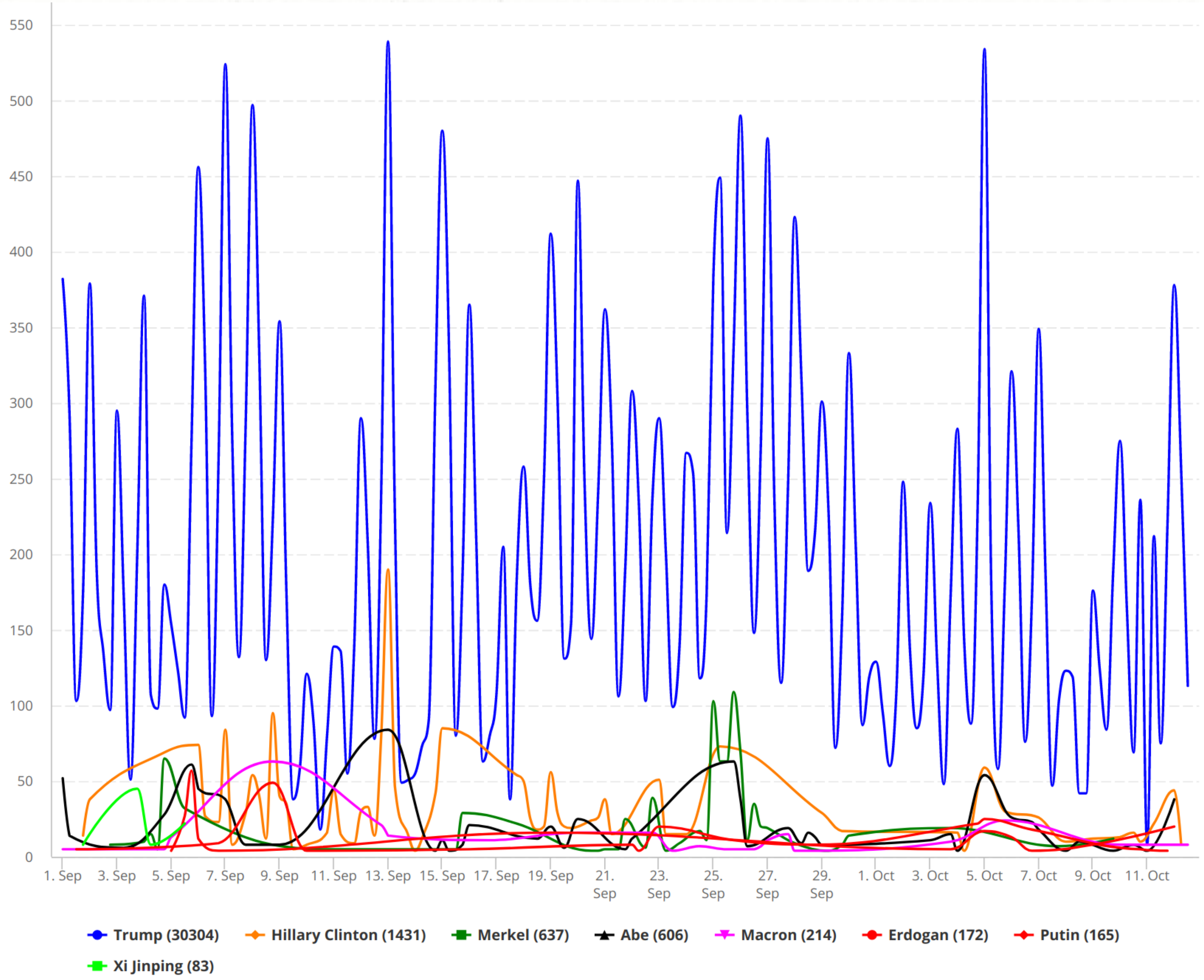
A Popular Words time-line tool: Line chart for 1 month regular min 1 day dynamic



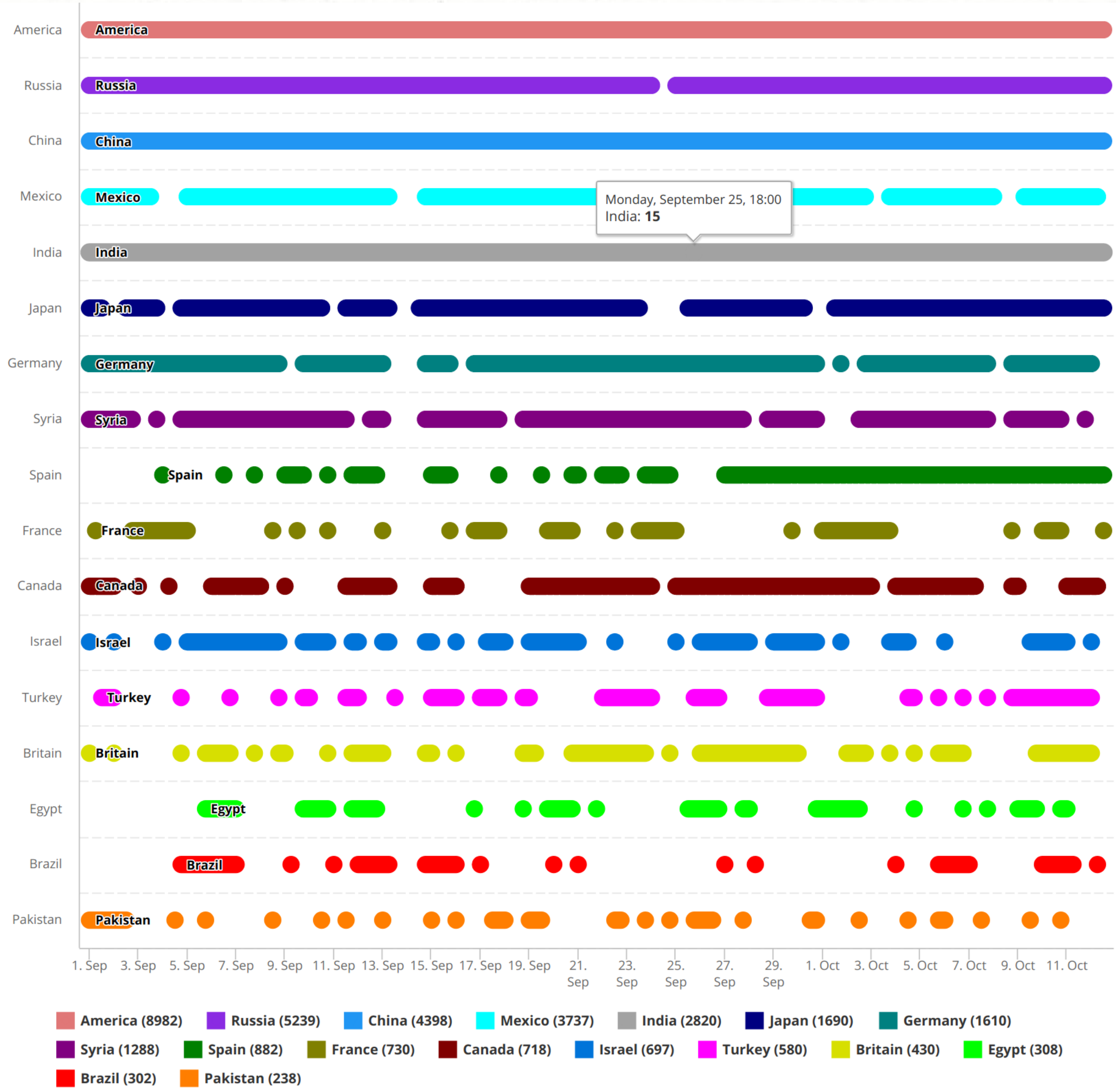
A Popular Words time-line tool: Line chart for 1 month regular min 3 days dynamic



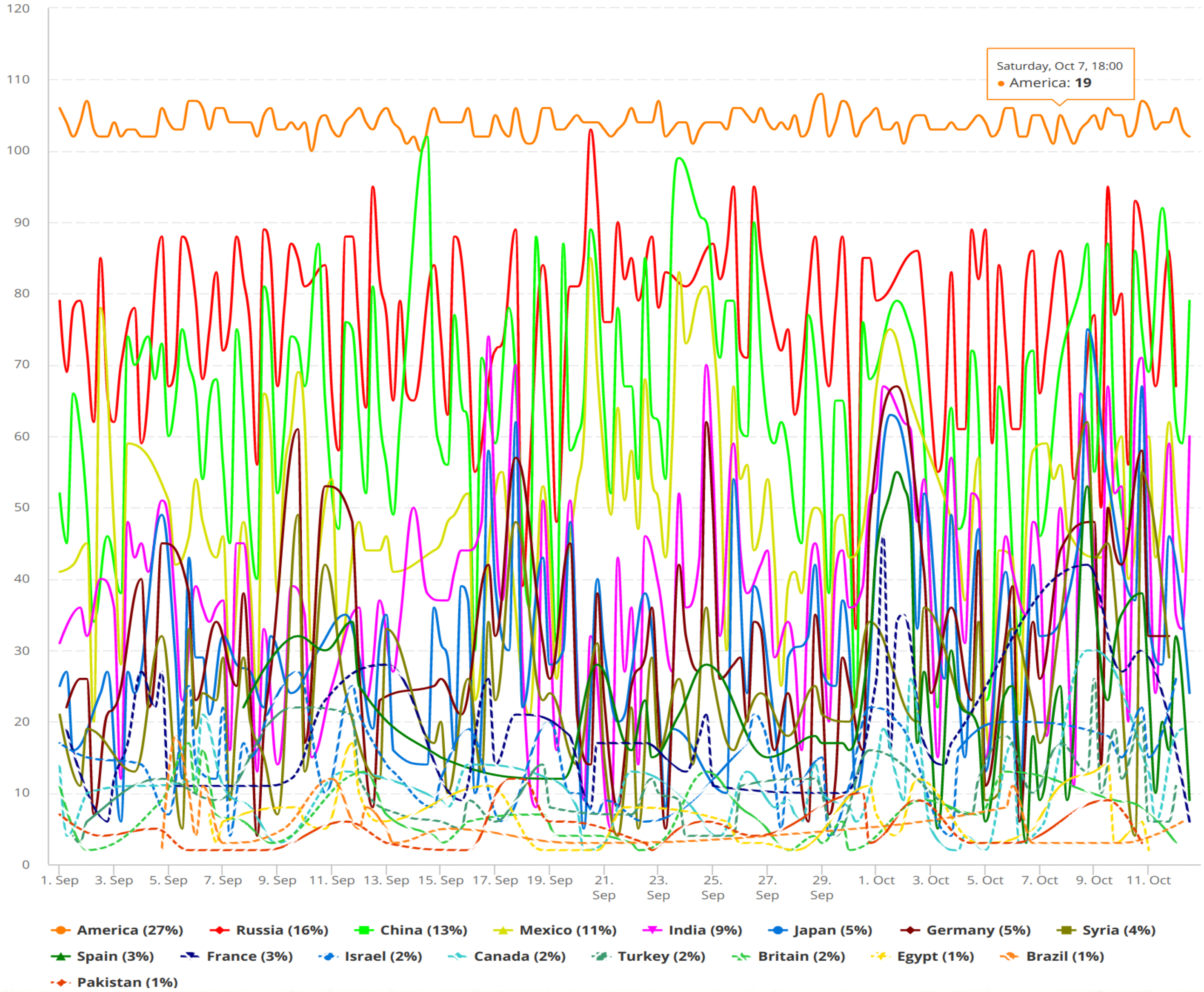
A Popular Words time-line tool: Line chart for one month list of persons dynamic



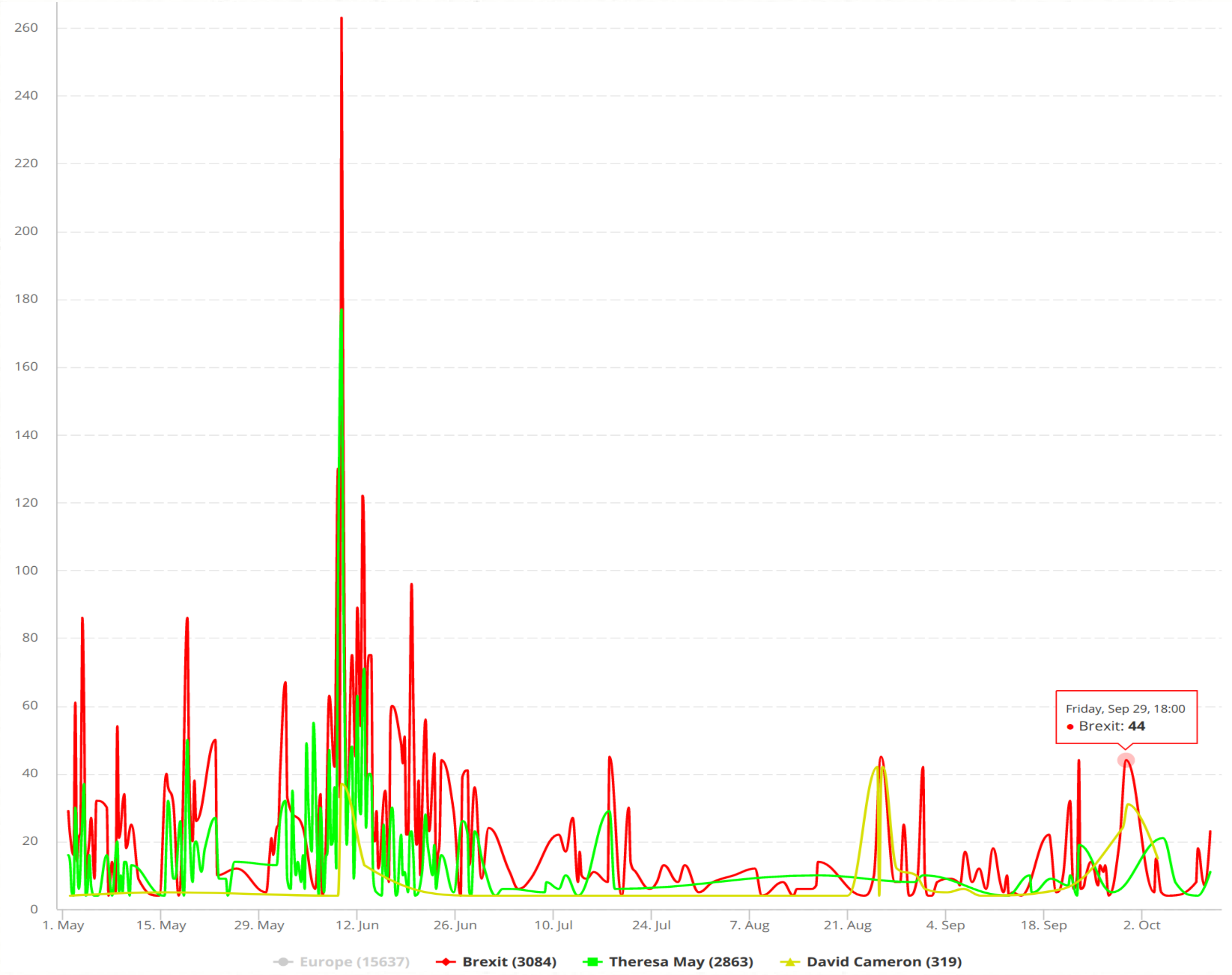
A Popular Words time-line tool: Gantt chart for one month some countries pop-words



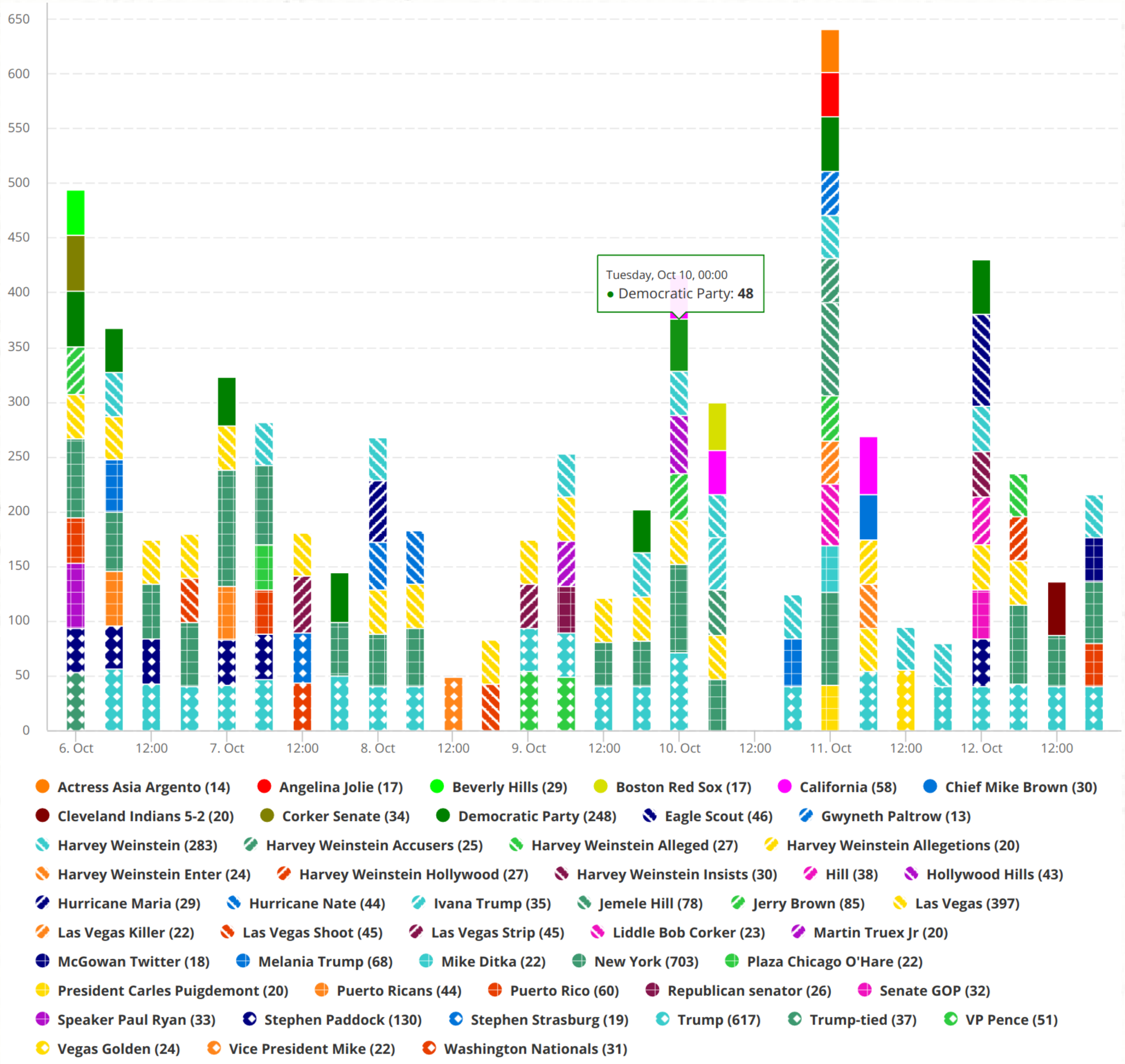
A Popular Words time-line tool: Line chart for one month some countries % pop-words



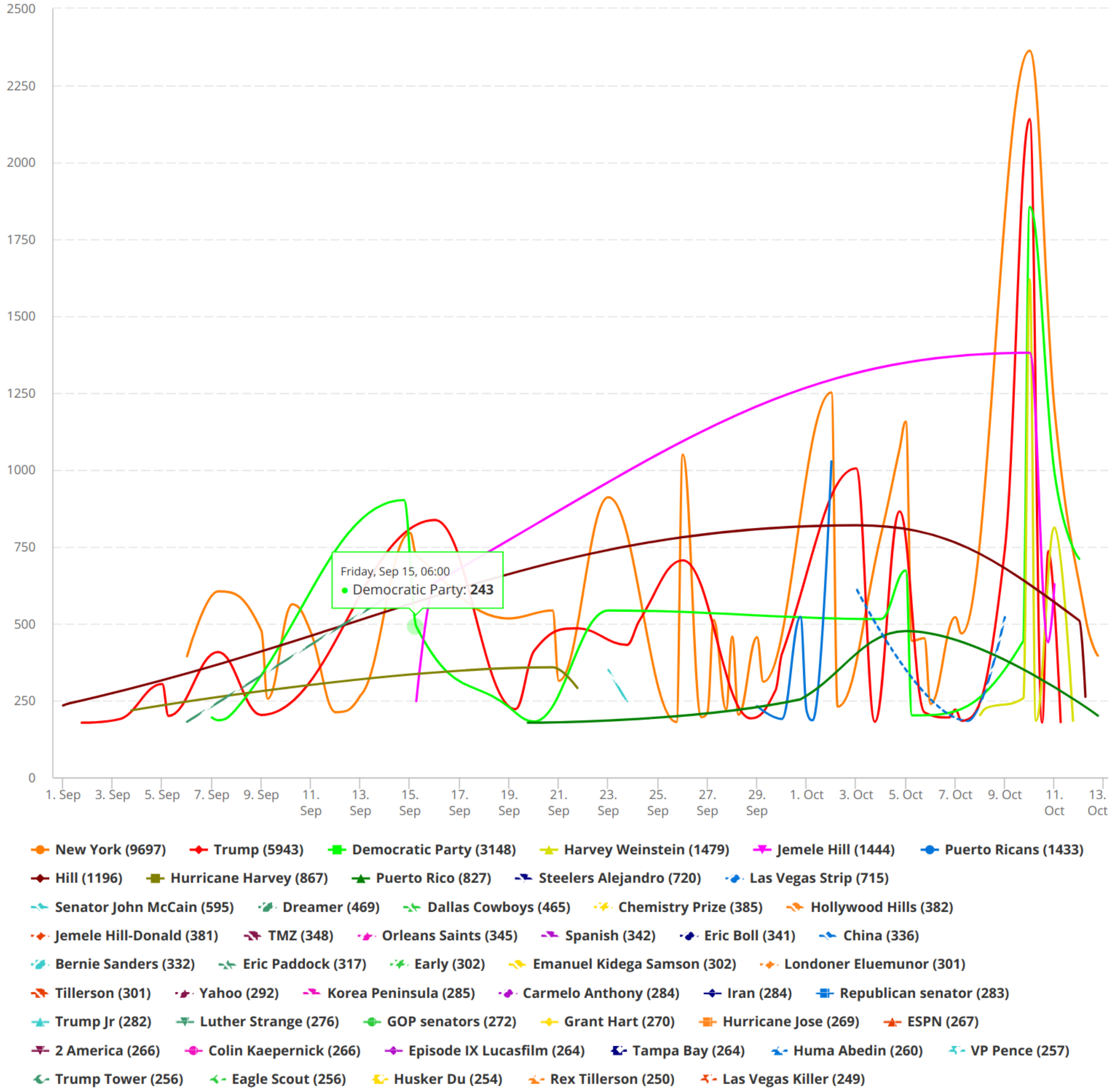
A Popular Words time-line tool: Line chart for half of a year some persons and events



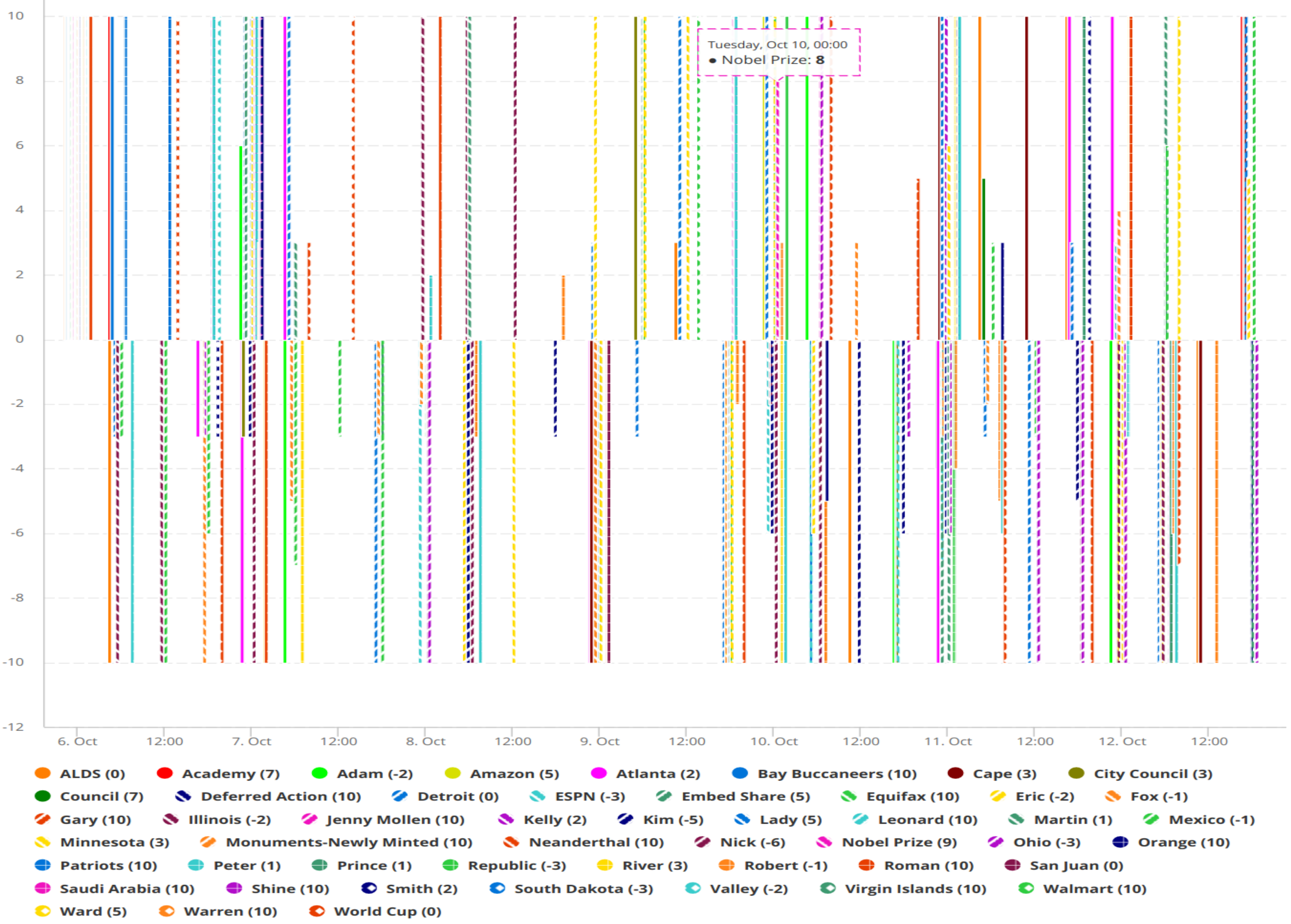
A Popular Words time-line tool: Bar chart 7 days number of articles dynamic



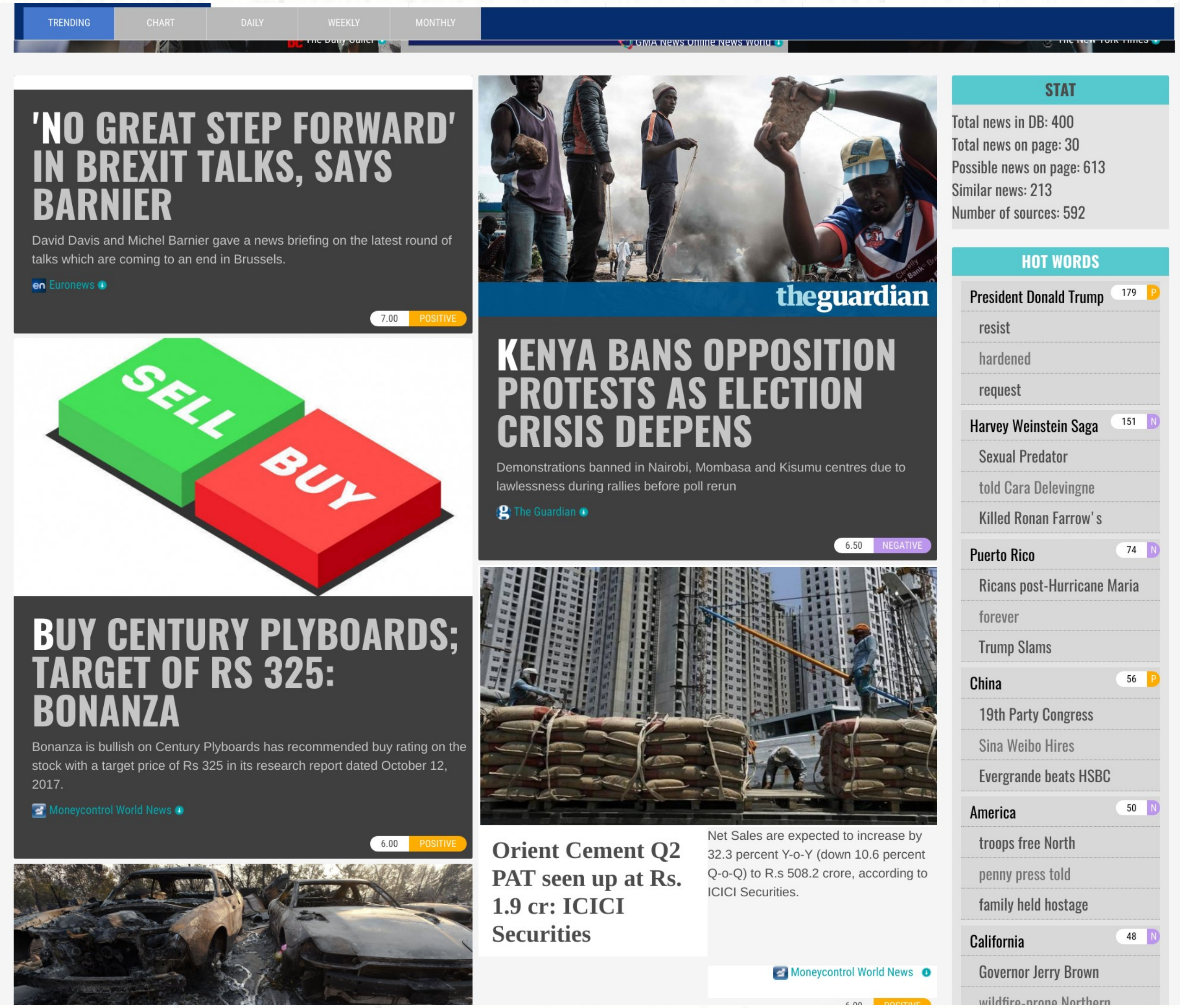
A Popular Words time-line tool: Line chart for 7 days top 50 Twitter posts dynamic



A Popular Words time-line tool: Bar chart for 7 days top 50 Sentiment Rate dynamic



Web-site engine front-end: articles, pop-words, Similarity and Sentiment ranks



'NO GREAT STEP FORWARD' IN BREXIT TALKS, SAYS BARNIER

David Davis and Michel Barnier gave a news briefing on the latest round of talks which are coming to an end in Brussels.

Euronews

7.00 POSITIVE

BUY CENTURY PLYBOARDS; TARGET OF RS 325: BONANZA

Bonanza is bullish on Century Plyboards has recommended buy rating on the stock with a target price of Rs 325 in its research report dated October 12, 2017.

Moneycontrol World News

6.00 POSITIVE

KENYA BANS OPPOSITION PROTESTS AS ELECTION CRISIS DEEPENS

Demonstrations banned in Nairobi, Mombasa and Kisumu centres due to lawlessness during rallies before poll rerun

The Guardian

6.50 NEGATIVE

Orient Cement Q2 PAT seen up at Rs. 1.9 cr: ICICI Securities

Net Sales are expected to increase by 32.3 percent Y-o-Y (down 10.6 percent Q-o-Q) to R.s 508.2 crore, according to ICICI Securities.

Moneycontrol World News

6.00 POSITIVE

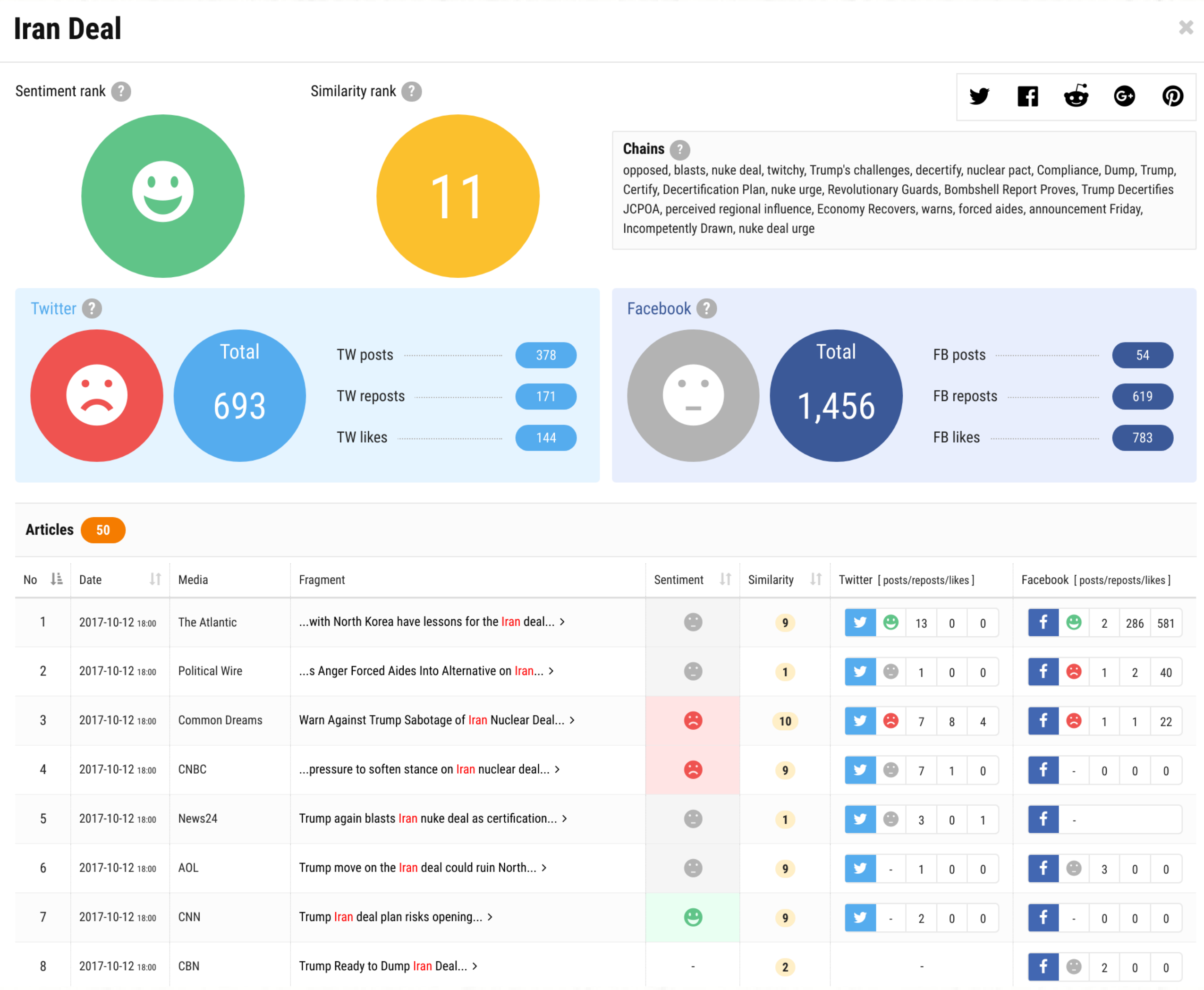
STAT

Total news in DB: 400
Total news on page: 30
Possible news on page: 613
Similar news: 213
Number of sources: 592

HOT WORDS

President Donald Trump	179	P
resist		
hardened		
request		
Harvey Weinstein Saga	151	N
Sexual Predator		
told Cara Delevingne		
Killed Ronan Farrow's		
Puerto Rico	74	N
Ricans post-Hurricane Maria		
forever		
Trump Slams		
China	56	P
19th Party Congress		
Sina Weibo Hires		
Evergrande beats HSBC		
America	50	N
troops free North		
penny press told		
family held hostage		
California	48	N
Governor Jerry Brown		
wildfire-prone Northern		

Web-site engine front-end: pop-word's social networks stats, article's fragments



Iran Deal

Sentiment rank ?

🙂

Similarity rank ?

11

Chains ?

opposed, blasts, nuke deal, twitchy, Trump's challenges, decertify, nuclear pact, Compliance, Dump, Trump, Certify, Decertification Plan, nuke urge, Revolutionary Guards, Bombshell Report Proves, Trump Decertifies JCPOA, perceived regional influence, Economy Recovers, warns, forced aides, announcement Friday, Incompetently Drawn, nuke deal urge

Twitter ?

☹

Total	
693	
TW posts	378
TW reposts	171
TW likes	144

Facebook ?

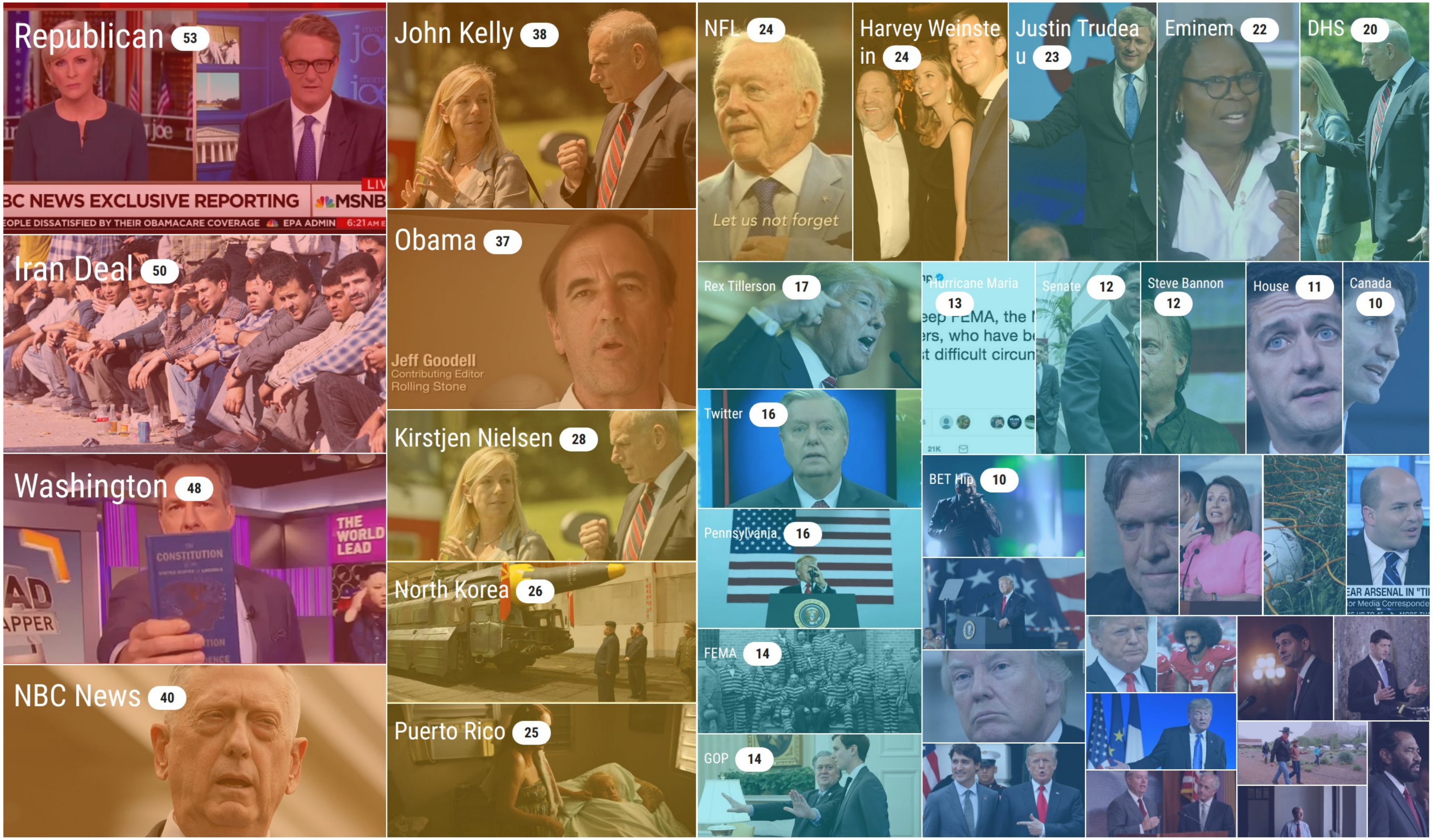
😐

Total	
1,456	
FB posts	54
FB reposts	619
FB likes	783

Articles 50

No	Date	Media	Fragment	Sentiment	Similarity	Twitter [posts/reposts/likes]	Facebook [posts/reposts/likes]
1	2017-10-12 18:00	The Atlantic	...with North Korea have lessons for the Iran deal... >	😐	9	🙂 13 0 0	🙂 2 286 581
2	2017-10-12 18:00	Political Wire	...s Anger Forced Aides Into Alternative on Iran... >	😐	1	😐 1 0 0	☹ 1 2 40
3	2017-10-12 18:00	Common Dreams	Warn Against Trump Sabotage of Iran Nuclear Deal... >	☹	10	☹ 7 8 4	☹ 1 1 22
4	2017-10-12 18:00	CNBC	...pressure to soften stance on Iran nuclear deal... >	☹	9	😐 7 1 0	- 0 0 0
5	2017-10-12 18:00	News24	Trump again blasts Iran nuke deal as certification... >	😐	1	😐 3 0 1	-
6	2017-10-12 18:00	AOL	Trump move on the Iran deal could ruin North... >	😐	9	- 1 0 0	😐 3 0 0
7	2017-10-12 18:00	CNN	Trump Iran deal plan risks opening... >	🙂	9	- 2 0 0	- 0 0 0
8	2017-10-12 18:00	CBN	Trump Ready to Dump Iran Deal... >	-	2	-	😐 2 0 0

Web-site engine front-end: pop-words and Similarity Ranks with images



Statistics of typical three physical hosts installation for one month

- Projects: 8
- Pages crawled: 6.2M
- Crawling batches: 60K
- Processing batches: 90K
- Purging batches: 16K
- Aging batches: 16K
- Projects re-crawling: 30K
- CPU Load Average: 0.45 avg / 3.5 max
- CPU utilization: 3% avg / 30% max
- I/O wait time: 0.31 avg / 6.4 max
- Network connections: 250 avg / 747 max
- Network traffic: 152Kbps avg / 5.5Mbps max
- Hosts data: 2, manage: 1

- Load-balancing of system OS resources linear managed CPU load average, I/O wait and RAM usage without excesses and overloads.
- Linear scalability of real-time requests per physical host.
- Linear scalability of automated crawling, processing and aging per physical host.

Statistics of typical five physical hosts installation for one month

- **Projects total: 52, active 31**
- **Pages crawled: 8M**
- **Crawling batches: 140K**
- **Processing batches: 120K**
- **Purging batches: 20K**
- **Aging batches: 20K**
- **Projects re-crawling: 30K**
- **CPU Load Average: 1.7 avg / 4 max**
- **CPU utilization: 8% avg / 30% max**
- **I/O wait time: 1.0 avg / 12 max**
- **Network connections: 500 avg / 2000 max**
- **Network traffic: 4Mbps avg / 45Mbps max**
- **Hosts data: 4, manage: 1**

- **Load-balancing of system OS resources linear managed CPU load average, I/O wait and RAM usage without excesses and overloads.**
- **Linear scalability of real-time requests per physical host.**
- **Linear scalability of automated crawling, processing and aging per physical host.**