FED4FIRE
Federation for FIRE Plus

Scaling NewSum

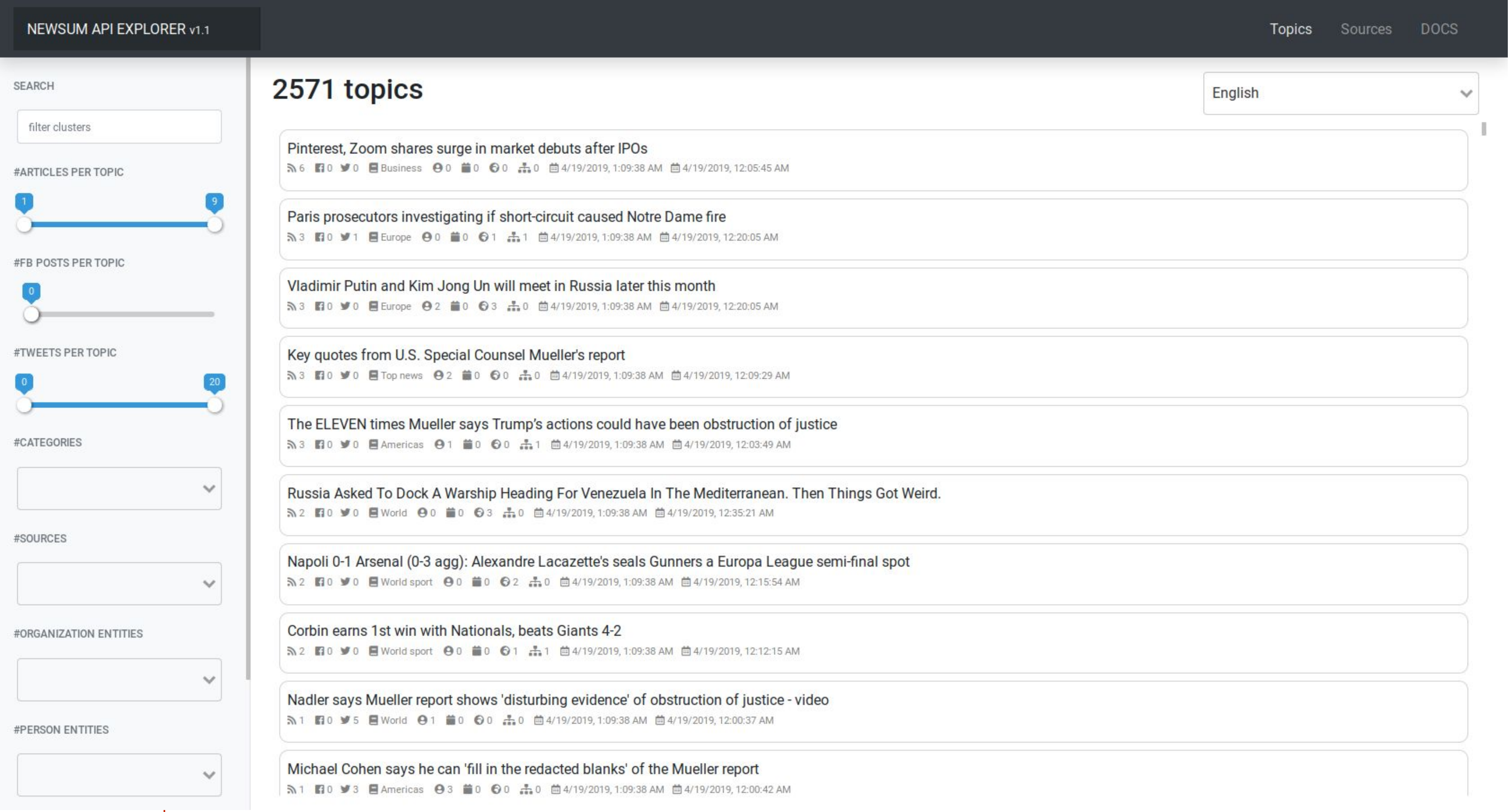
Big data text Clustering and Summarization using N-Gram graphs

SC!FY
SCIENCE FOR YOU
https://www.scify.org

Alexandros Tzoumas | a.tzoumas@scify.org

NewSum

What’s our product about?



NEWSUM API EXPLORER v1.1

SEARCH

filter clusters

#ARTICLES PER TOPIC

1 9

#FB POSTS PER TOPIC

0

#TWEETS PER TOPIC

0 20

#CATEGORIES

#SOURCES

#ORGANIZATION ENTITIES

#PERSON ENTITIES

2571 topics

English

Pinterest, Zoom shares surge in market debuts after IPOs
6 0 0 Business 0 0 0 0 4/19/2019, 1:09:38 AM 4/19/2019, 12:05:45 AM

Paris prosecutors investigating if short-circuit caused Notre Dame fire
3 0 1 Europe 0 0 1 1 4/19/2019, 1:09:38 AM 4/19/2019, 12:20:05 AM

Vladimir Putin and Kim Jong Un will meet in Russia later this month
3 0 0 Europe 2 0 3 0 4/19/2019, 1:09:38 AM 4/19/2019, 12:20:05 AM

Key quotes from U.S. Special Counsel Mueller's report
3 0 0 Top news 2 0 0 0 4/19/2019, 1:09:38 AM 4/19/2019, 12:09:29 AM

The ELEVEN times Mueller says Trump's actions could have been obstruction of justice
3 0 0 Americas 1 0 0 1 4/19/2019, 1:09:38 AM 4/19/2019, 12:03:49 AM

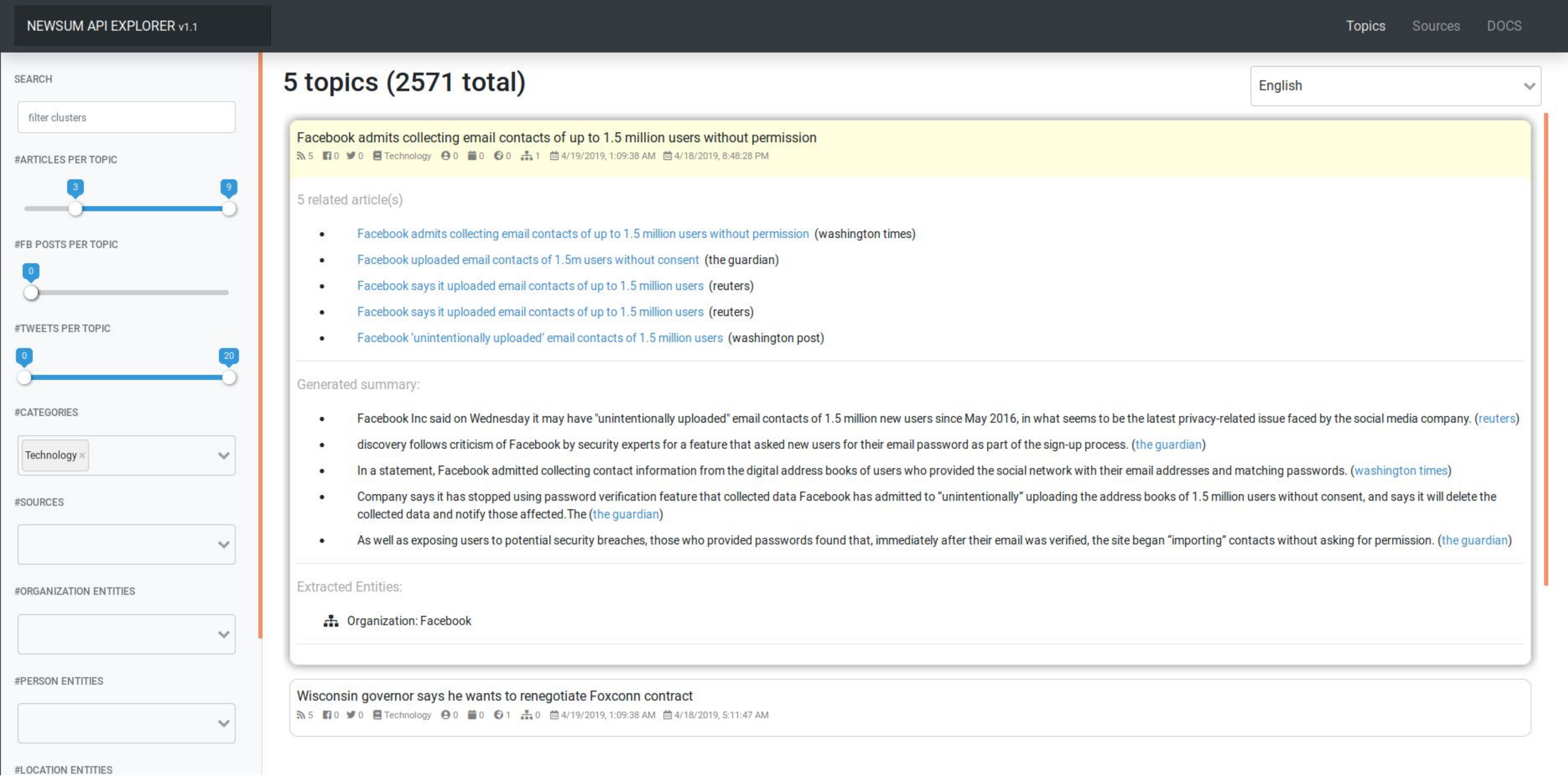
Russia Asked To Dock A Warship Heading For Venezuela In The Mediterranean. Then Things Got Weird.
2 0 0 World 0 0 3 0 4/19/2019, 1:09:38 AM 4/19/2019, 12:35:21 AM

Napoli 0-1 Arsenal (0-3 agg): Alexandre Lacazette's seals Gunners a Europa League semi-final spot
2 0 0 World sport 0 0 2 0 4/19/2019, 1:09:38 AM 4/19/2019, 12:15:54 AM

Corbin earns 1st win with Nationals, beats Giants 4-2
2 0 0 World sport 0 0 1 1 4/19/2019, 1:09:38 AM 4/19/2019, 12:12:15 AM

Nadler says Mueller report shows 'disturbing evidence' of obstruction of justice - video
1 0 5 World 1 0 0 0 4/19/2019, 1:09:38 AM 4/19/2019, 12:00:37 AM

Michael Cohen says he can 'fill in the redacted blanks' of the Mueller report
1 0 3 Americas 3 0 0 0 4/19/2019, 1:09:38 AM 4/19/2019, 12:00:42 AM



NEWSUM API EXPLORER v1.1

Topics Sources DOCS

SEARCH

filter clusters

#ARTICLES PER TOPIC

3 9

#FB POSTS PER TOPIC

0

#TWEETS PER TOPIC

0 20

#CATEGORIES

Technology ×

#SOURCES

#ORGANIZATION ENTITIES

#PERSON ENTITIES

#LOCATION ENTITIES

5 topics (2571 total)

English

Facebook admits collecting email contacts of up to 1.5 million users without permission

5 0 0 Technology 0 0 0 1 4/19/2019, 1:09:38 AM 4/18/2019, 8:48:28 PM

5 related article(s)

- Facebook admits collecting email contacts of up to 1.5 million users without permission (washington times)
- Facebook uploaded email contacts of 1.5m users without consent (the guardian)
- Facebook says it uploaded email contacts of up to 1.5 million users (reuters)
- Facebook says it uploaded email contacts of up to 1.5 million users (reuters)
- Facebook 'unintentionally uploaded' email contacts of 1.5 million users (washington post)

Generated summary:

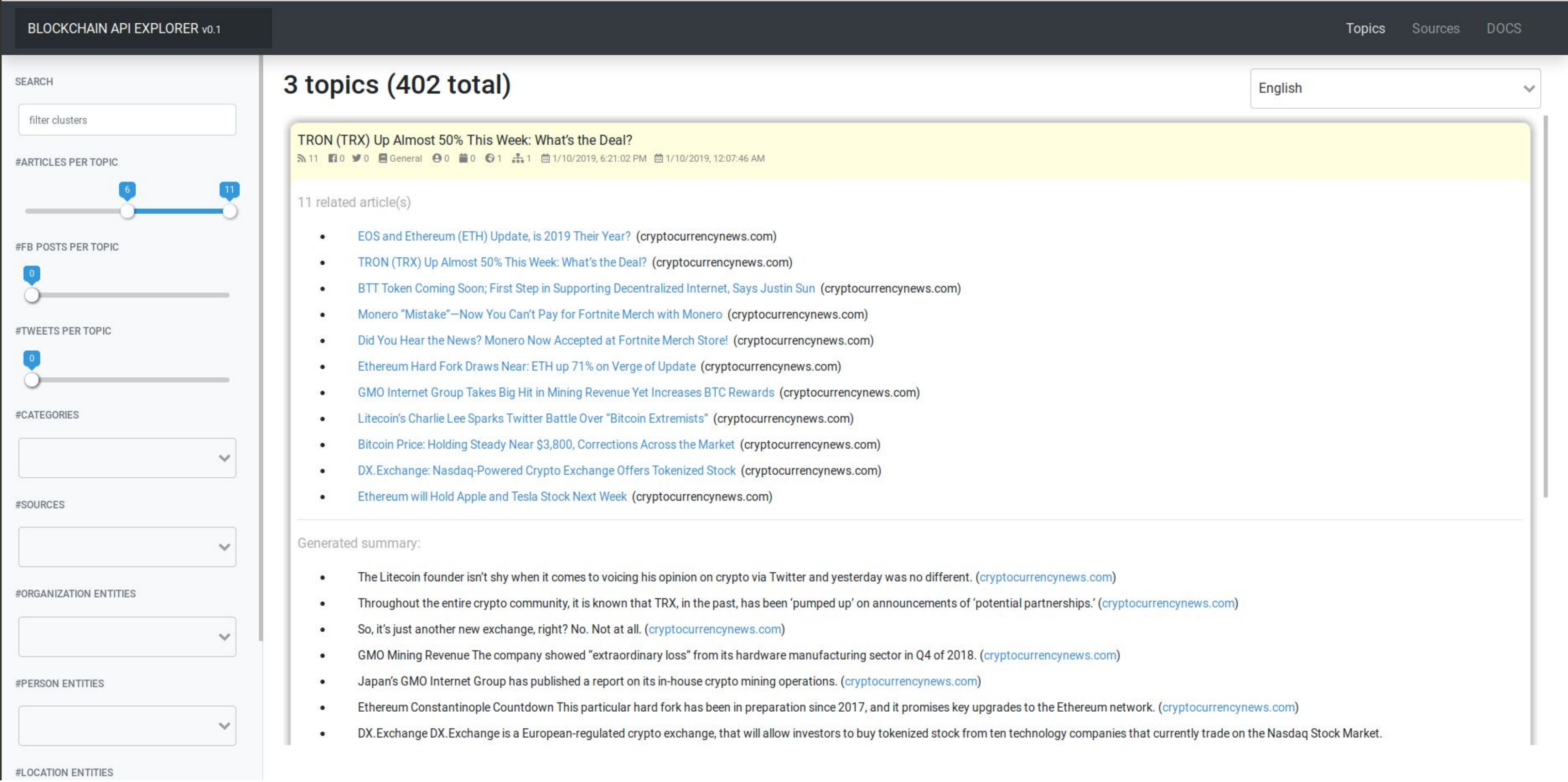
- Facebook Inc said on Wednesday it may have "unintentionally uploaded" email contacts of 1.5 million new users since May 2016, in what seems to be the latest privacy-related issue faced by the social media company. (reuters)
- discovery follows criticism of Facebook by security experts for a feature that asked new users for their email password as part of the sign-up process. (the guardian)
- In a statement, Facebook admitted collecting contact information from the digital address books of users who provided the social network with their email addresses and matching passwords. (washington times)
- Company says it has stopped using password verification feature that collected data Facebook has admitted to “unintentionally” uploading the address books of 1.5 million users without consent, and says it will delete the collected data and notify those affected.The (the guardian)
- As well as exposing users to potential security breaches, those who provided passwords found that, immediately after their email was verified, the site began “importing” contacts without asking for permission. (the guardian)

Extracted Entities:

Organization: Facebook

Wisconsin governor says he wants to renegotiate Foxconn contract

5 0 0 Technology 0 0 1 0 4/19/2019, 1:09:38 AM 4/18/2019, 5:11:47 AM



SEARCH

filter clusters

#ARTICLES PER TOPIC: 6 – 11

#FB POSTS PER TOPIC: 0

#TWEETS PER TOPIC: 0

#CATEGORIES

#SOURCES

#ORGANIZATION ENTITIES

#PERSON ENTITIES

#LOCATION ENTITIES

3 topics (402 total)

English

TRON (TRX) Up Almost 50% This Week: What's the Deal?

11 0 0 General 0 0 1 1 1/10/2019, 6:21:02 PM 1/10/2019, 12:07:46 AM

11 related article(s)

- EOS and Ethereum (ETH) Update, is 2019 Their Year? (cryptocurrencynews.com)
- TRON (TRX) Up Almost 50% This Week: What's the Deal? (cryptocurrencynews.com)
- BTT Token Coming Soon; First Step in Supporting Decentralized Internet, Says Justin Sun (cryptocurrencynews.com)
- Monero "Mistake"—Now You Can't Pay for Fortnite Merch with Monero (cryptocurrencynews.com)
- Did You Hear the News? Monero Now Accepted at Fortnite Merch Store! (cryptocurrencynews.com)
- Ethereum Hard Fork Draws Near: ETH up 71% on Verge of Update (cryptocurrencynews.com)
- GMO Internet Group Takes Big Hit in Mining Revenue Yet Increases BTC Rewards (cryptocurrencynews.com)
- Litecoin's Charlie Lee Sparks Twitter Battle Over "Bitcoin Extremists" (cryptocurrencynews.com)
- Bitcoin Price: Holding Steady Near $3,800, Corrections Across the Market (cryptocurrencynews.com)
- DX.Exchange: Nasdaq-Powered Crypto Exchange Offers Tokenized Stock (cryptocurrencynews.com)
- Ethereum will Hold Apple and Tesla Stock Next Week (cryptocurrencynews.com)

Generated summary:

- The Litecoin founder isn't shy when it comes to voicing his opinion on crypto via Twitter and yesterday was no different. (cryptocurrencynews.com)
- Throughout the entire crypto community, it is known that TRX, in the past, has been 'pumped up' on announcements of 'potential partnerships.' (cryptocurrencynews.com)
- So, it's just another new exchange, right? No. Not at all. (cryptocurrencynews.com)
- GMO Mining Revenue The company showed "extraordinary loss" from its hardware manufacturing sector in Q4 of 2018. (cryptocurrencynews.com)
- Japan's GMO Internet Group has published a report on its in-house crypto mining operations. (cryptocurrencynews.com)
- Ethereum Constantinople Countdown This particular hard fork has been in preparation since 2017, and it promises key upgrades to the Ethereum network. (cryptocurrencynews.com)
- DX.Exchange DX.Exchange is a European-regulated crypto exchange, that will allow investors to buy tokenized stock from ten technology companies that currently trade on the Nasdaq Stock Market.

Business Goals

Goals

Business goals

- improve the quality of the solutions our product offers
- allow NewSum technology to expand to new domains/markets

From a technical perspective

Measure and evaluate:

- the accuracy of candidate clustering components,
- the effectiveness (summary quality) of alternative summarization components
- the overall scalability of the system

Challenges

Business challenges

- Expansion to new markets should take domain specific characteristics into account as system parameters

- A product manager is not able to configure the product-related settings appropriate for each domain, so a semi-supervised process would be invaluable

From a technical perspective

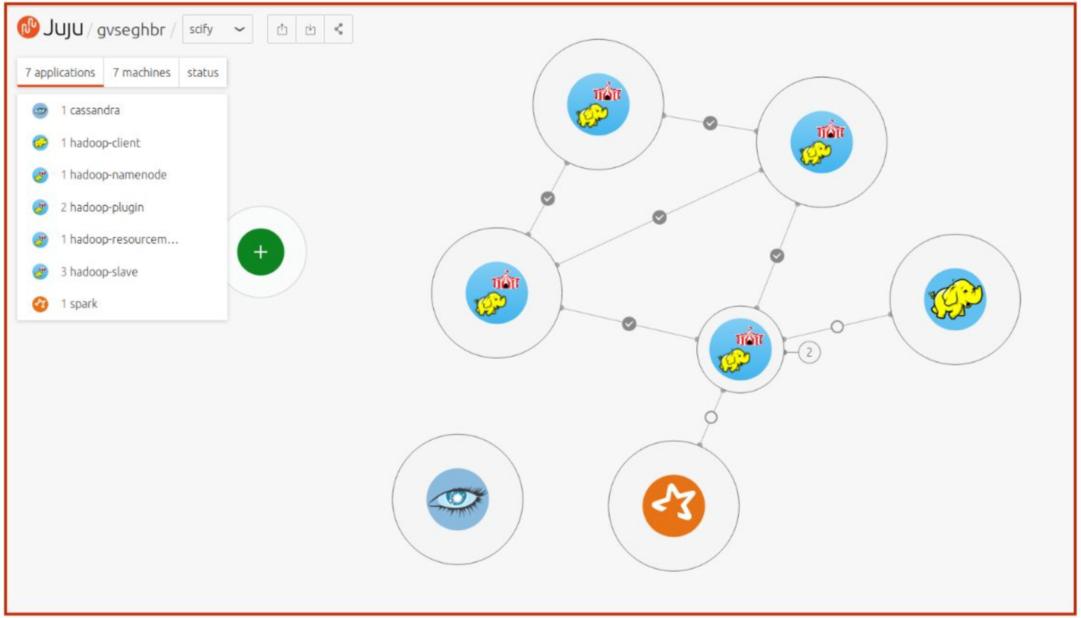
- Define a process for evaluating different clustering and summarization components

- Scale the algorithms to process thousand of sources/articles

The Experiments

Setup

Tengu testbed with the support of IMEC
Cassandra - Hadoop - Spark



Experiments

Experiment set 1

Goal:
Measure effectiveness of NewSum's candidate clustering implementations

Related datasets:
Multiling (articles with clustering information)
6GB database of news articles

Methodology:
Run clustering on 2 different clustering implementations and measure recall and precision.

Automatic evaluation for MultiLing dataset
Manual process for news articles dataset

Results

Selected the algorithm with higher precision & recall

Experiments

Experiment set 2

Goal:
Measure scalability

Related dataset:
6GB database of news articles

Methodology:
Run the clustering pipeline using as input a) the algorithm from experiment set 1 b) a variable number of articles.

Measure speed

Results

Increased 5 times the speed of the clustering pipeline!

Identified areas of improvement

Experiments

Experiment set 3

Goal:
Measure effectiveness of NewSum's candidate summarization implementations

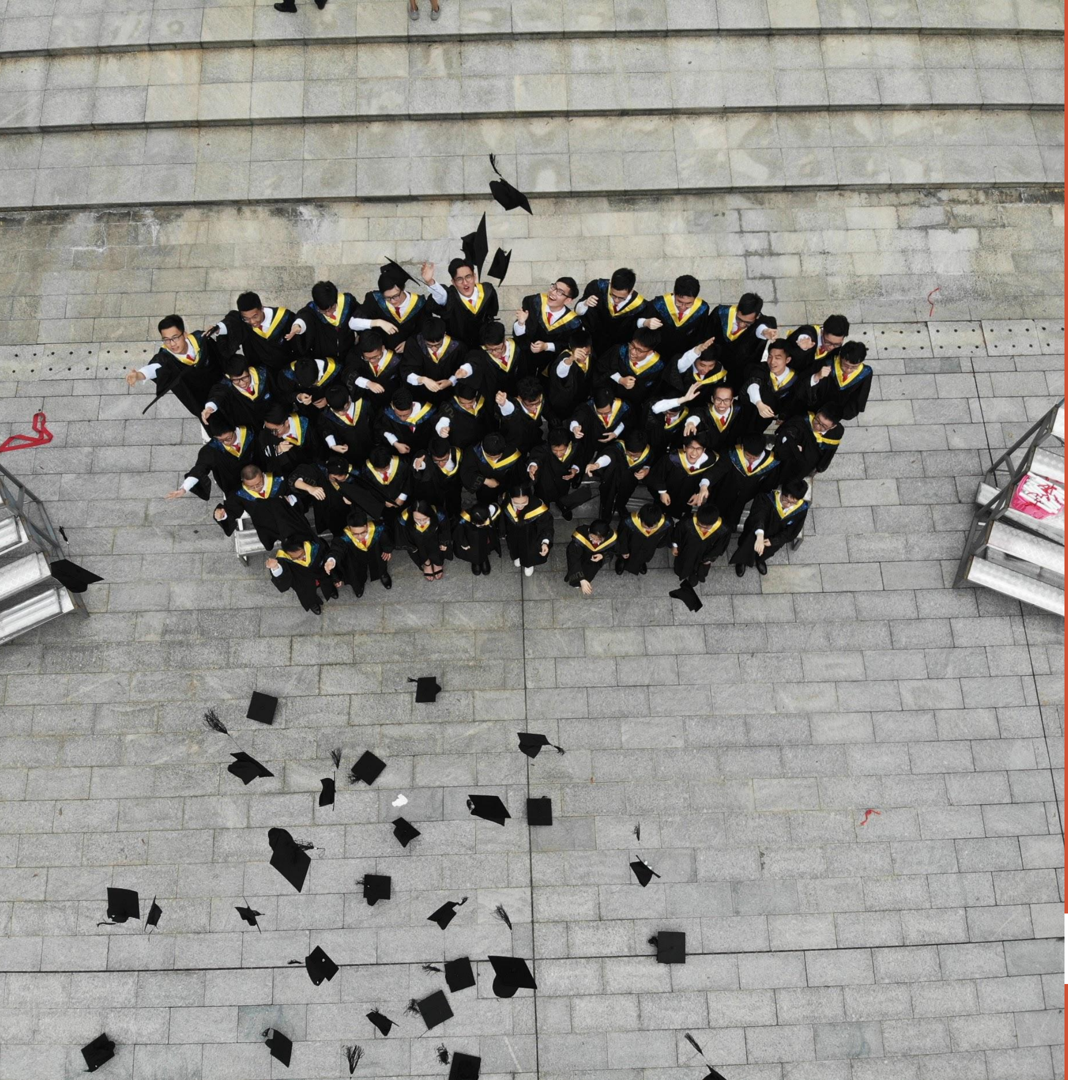
Related datasets:
6GB database of news articles

Methodology:
Run the summarization pipeline using as input a) configuration/parameter setting b) a number of clusters to be summarized.
Recall and precision were measured through a manual process.

Results:
Implemented/Identified the process for selecting the algorithm appropriate for each scenario

Results

Implemented/Identified the process for selecting the algorithm appropriate for each scenario



Conclusions

What we achieved

- Defined a process for evaluating clustering algorithms
- Defined a process for evaluating summarization components
- Increased 5 times the speed of the clustering pipeline!
- Measured scalability and identified bottlenecks

How Fed4Fire+ helped us

Patron’s support was crucial to the success of the experiments

How Fed4Fire+ helped us

Provided a quick way to start experimenting with big data without having to worry about the underlying technologies

How Fed4Fire+ helped us

Funding allowed us allocate time to implement the algorithms and analyze next steps

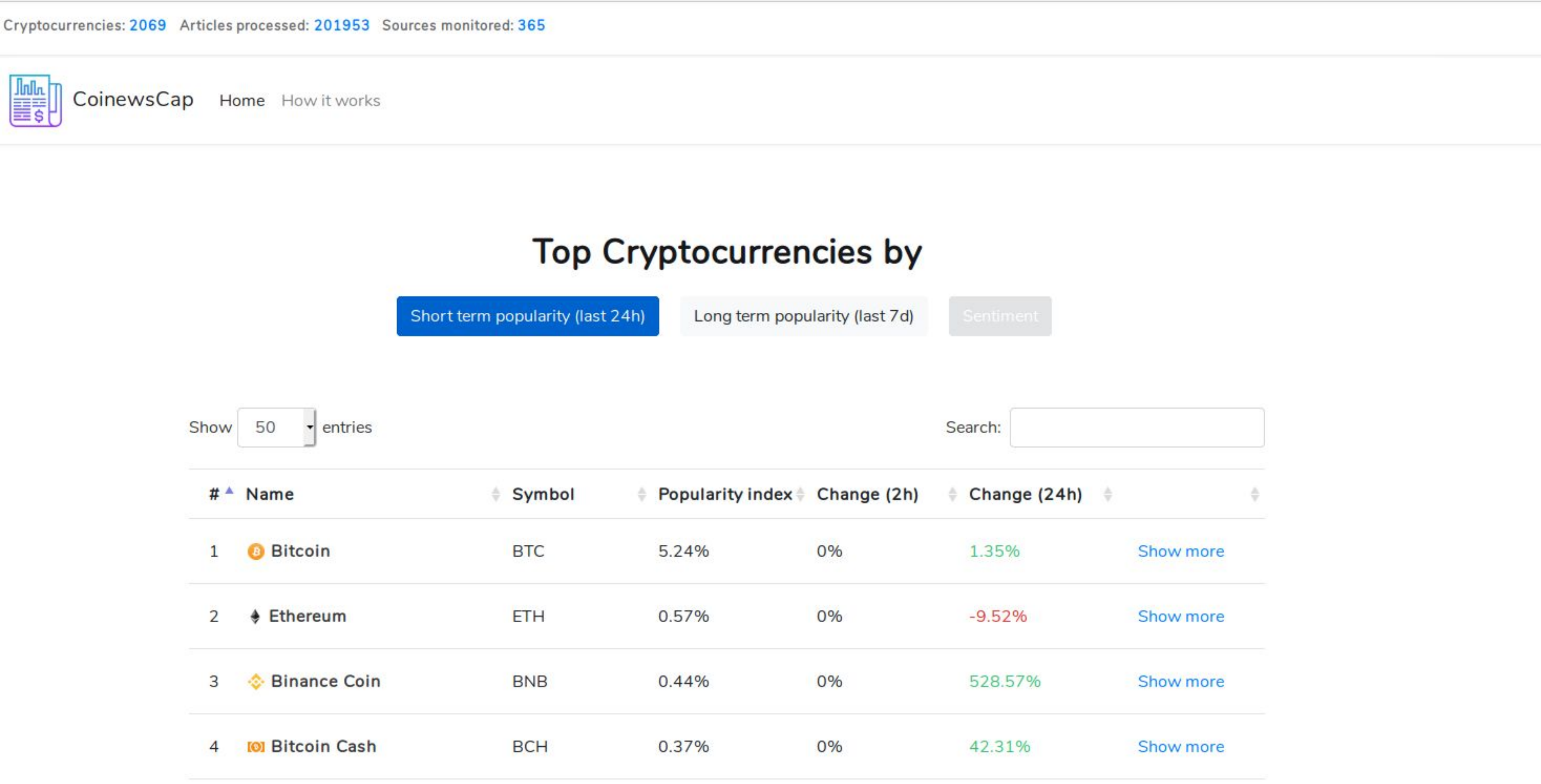
Next Steps

Next steps

Continue working on algorithm implementations
- Distributed N-gram graphs
- Improve clustering speed using blocking methodology

Automate the set up of a pipeline in a cloud environment to be used in production.

Release a domain specific product related to Blockchain news.


Cryptocurrencies: 2069 Articles processed: 201953 Sources monitored: 365
CoinewsCap
Home
How it works
Short term popularity (last 24h)
Long term popularity (last 7d)
Sentiment
Show 50 entries
Search:

Top Cryptocurrencies by

#	Name	Symbol	Popularity index	Change (2h)	Change (24h)	
1	Bitcoin	BTC	5.24%	0%	1.35%	Show more
2	Ethereum	ETH	0.57%	0%	-9.52%	Show more
3	Binance Coin	BNB	0.44%	0%	528.57%	Show more
4	Bitcoin Cash	BCH	0.37%	0%	42.31%	Show more

www.scify.org

WWW.FED4FIRE.EU

Co-funded by the European Union

Co-funded by the Swiss Confederation

This project has received funding from the European Union’s Horizon 2020 research and innovation programme, which is co-funded by the European Commission and the Swiss State Secretariat for Education, Research and Innovation, under grant agreement No 732638.