

Monitoring Displacement Crises with Python: A Humanitarian Project by Data for Democracy



George Richardson



georgerichardson.net



[georgerichardson](https://github.com/georgerichardson)

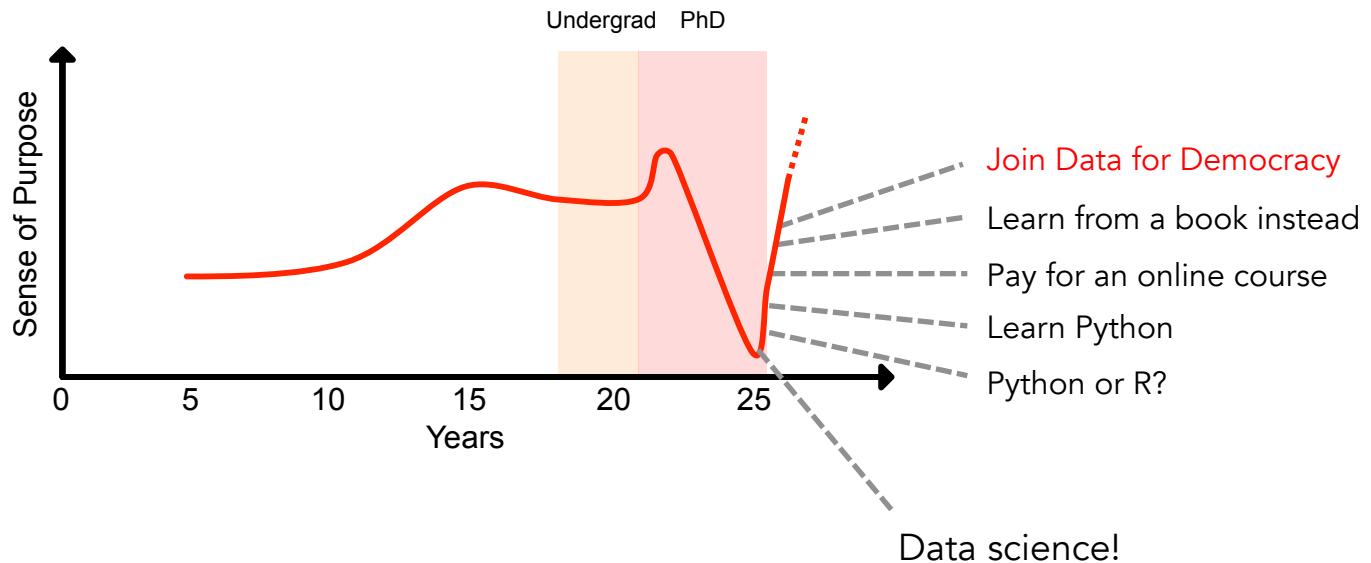


utopiadispatch.com



[@g_r_richardson](https://twitter.com/@g_r_richardson)

A Life in Data

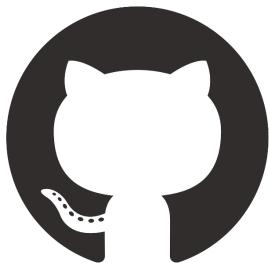


Data for Democracy

What is D4D?



=



+



github.com/data4democracy

datafordemocracy.slack.com



What is D4D?



gerrymandering

election transparency

government expenditures

hate speech

fact checking

open source tools

immigration

A Humanitarian Project



The graphic features a stylized Earth globe at the bottom, showing continents in black against a white background. A series of black vertical bars of increasing height rise from the left side of the globe, representing data points or events. Above the globe, the word "#IDTECT" is written in large, bold, blue capital letters. Below it, the text "INTERNAL DISPLACEMENT EVENT TAGGING, EXTRACTION AND CLUSTERING TOOL" is written in a smaller, blue sans-serif font. In the bottom right corner of the graphic area, the iDMC logo and the unite IDEAS logo are placed side-by-side.

<https://unite.un.org/ideas/content/idetect>

Monitoring Displacement

Internally Displaced Persons

An internally displaced person (IDP) is **someone who is forced to flee his or her home but who remains within his or her country's borders**. They are often referred to as refugees, although they do not fall within the legal definitions of a refugee.

https://en.wikipedia.org/wiki/Internally_displaced_person

- Armed conflicts
- General violence
- Human rights abuses
- Slow onset disasters
- Sudden onset disasters
- Development investments

IDPs - The Numbers

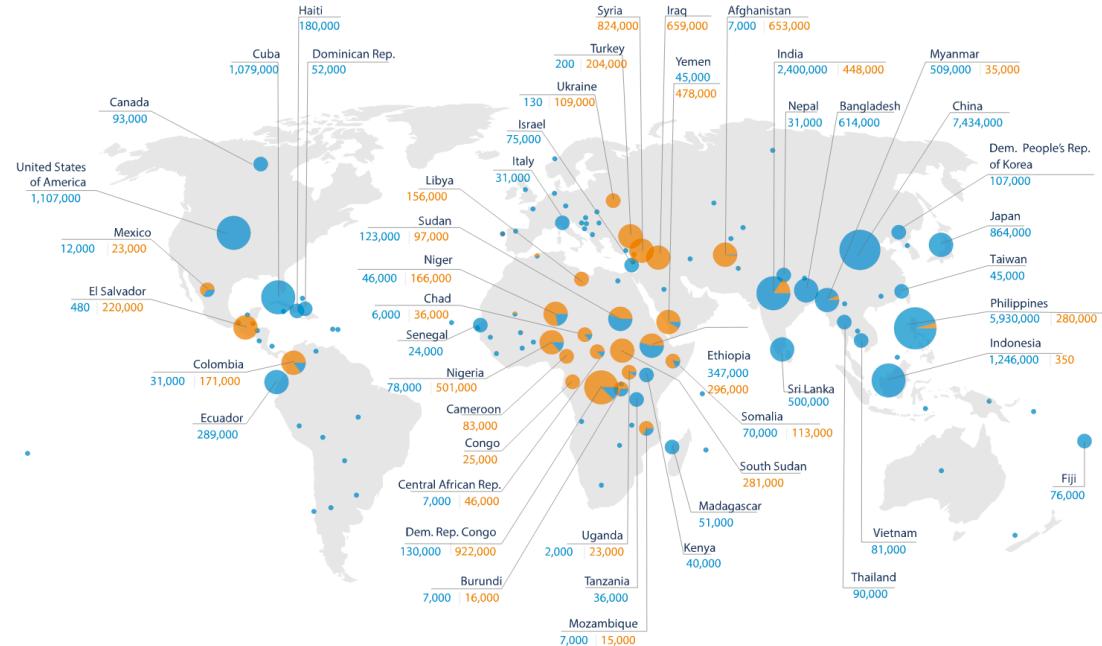
Total Number of IDPs
in December 2016

40.3m

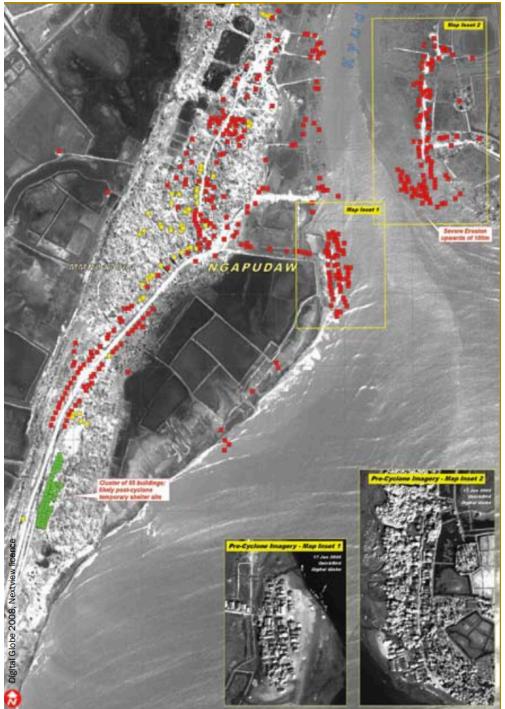
New Displacements in 2016

Conflict & Violence 6.9m

Disaster 24.2m



Monitoring IDPs - The Current Approach



<http://www.fmreview.org/sites/fmr/files/FMRdownloads/en/climatechange/bjorgo-pisano-lyons-heisig.pdf>

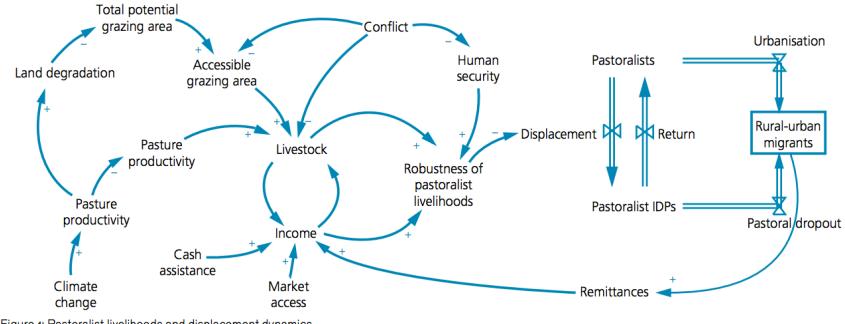
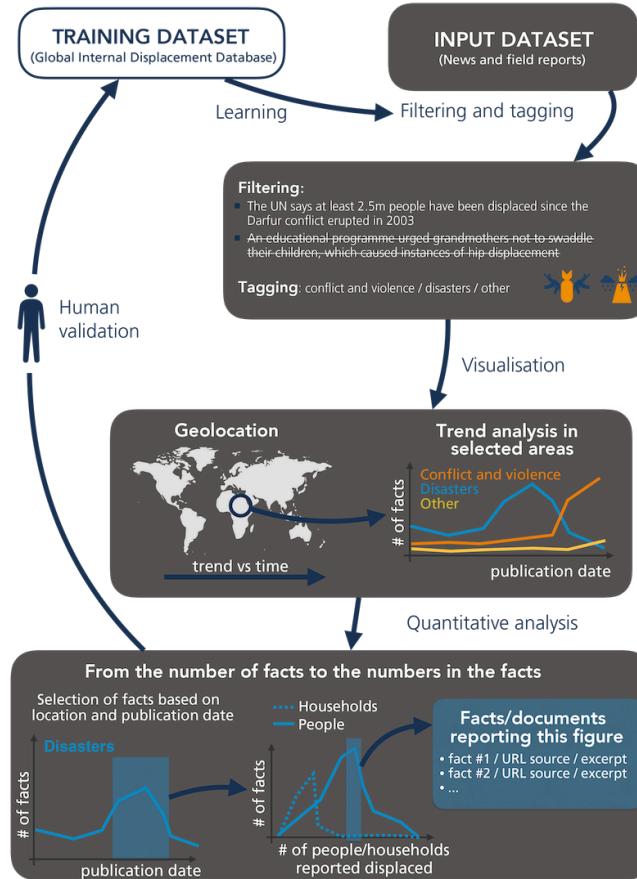


Figure 1: Pastoralist livelihoods and displacement dynamics

IDMC 2014 Annual Report

The screenshot displays the ReliefWeb homepage with a specific focus on the GIEWS Country Brief for El Salvador. The brief is dated 30-June-2017 and is attributed to the Food and Agriculture Organization of the United Nations. It highlights ongoing drought conditions in El Salvador and Guatemala. The page features a navigation bar with links to Home, Updates, Countries, Disasters, Topics, Organizations, Jobs, and Training, along with social media sharing options. A search bar is also present at the top.



Data Collection:
Scrape content from incoming URLs

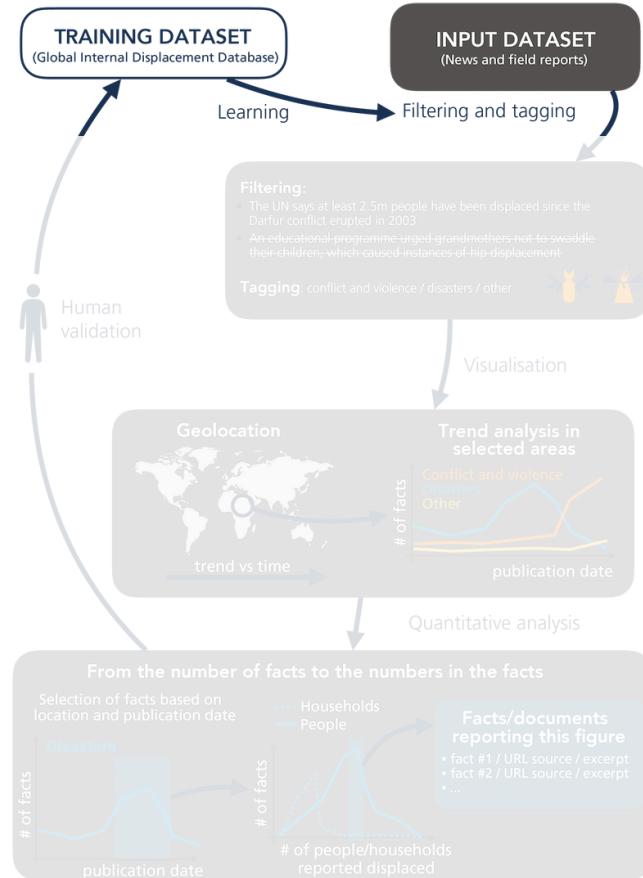
Classification:
Relevant or not relevant?
Conflict and violence or disaster?

Information Extraction:
Displacement figure, location, date, reporting term, reporting unit

Our Solution

Part I – Content Retrieval

Human Verification:
Accuracy and reliability
scoring. Model retraining.

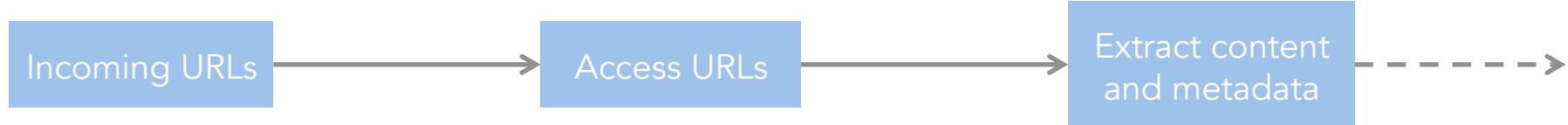


Data Collection:
Scrape content from incoming URLs

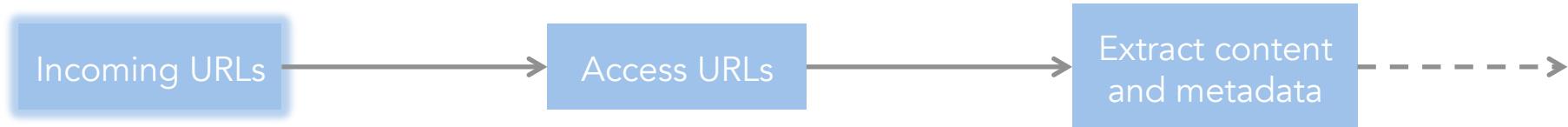
Classification:
Relevant or not relevant?
Conflict and violence or disaster?

Information Extraction:
Displacement figure, location, date,
reporting term, reporting unit

Scraping

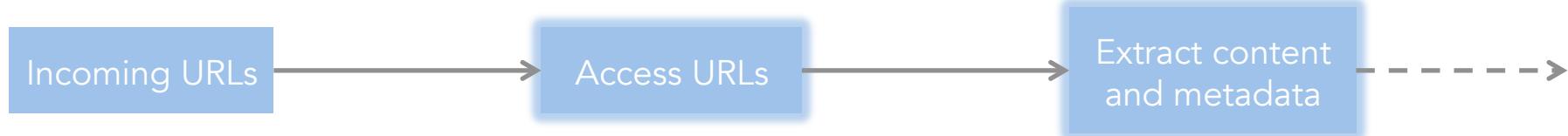


Scraping



SQLAlchemy

Scraping



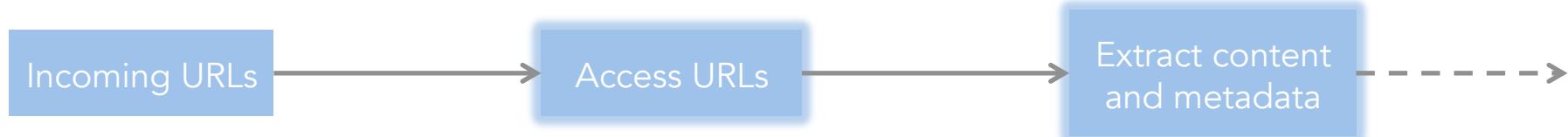
reliefweb

— **Algeria Sun** —
We never sleep

RAPID CITY
Journal



Scraping



reliefweb

— Algeria Sun —
We never sleep

RAPID CITY
Journal



```
$ pip install newspaper3k
```



Newspaper: Article scraping & curation

Release v0.1.2. ([Installation](#)).

Inspired by [requests](#) for its simplicity and powered by [lxml](#) for its speed.

Newspaper – Content Grabbing for Lazy People

```
In [1]: import newspaper
```

```
In [2]: url = "http://utopiadispatch.com/episodes/episode-1-utopia-or-dispatch/"  
a = newspaper.Article(url)  
a.download()  
a.parse()
```

```
In [3]: a.title
```

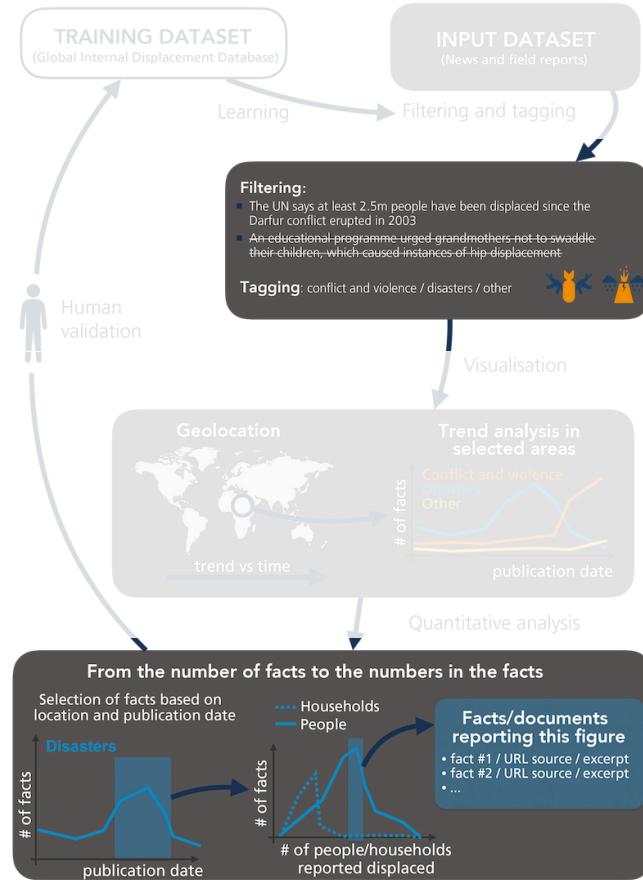
```
Out[3]: 'Episode 1 - Utopia or Dispatch'
```

```
In [4]: a.text
```

```
Out[4]: 'In the first episode of Utopia Dispatch we take a lighthearted tour of Thomas More's "Utopia", the book that coined the term 500 years ago. We see how the ideas hold up today, and how they might set the direction for our search for the perfect world.\n\nWe also bring you some good news, with The Good, The Best, and The Bestest.'
```

Part II – Article Processing

Human Verification:
Accuracy and reliability
scoring. Model retraining.



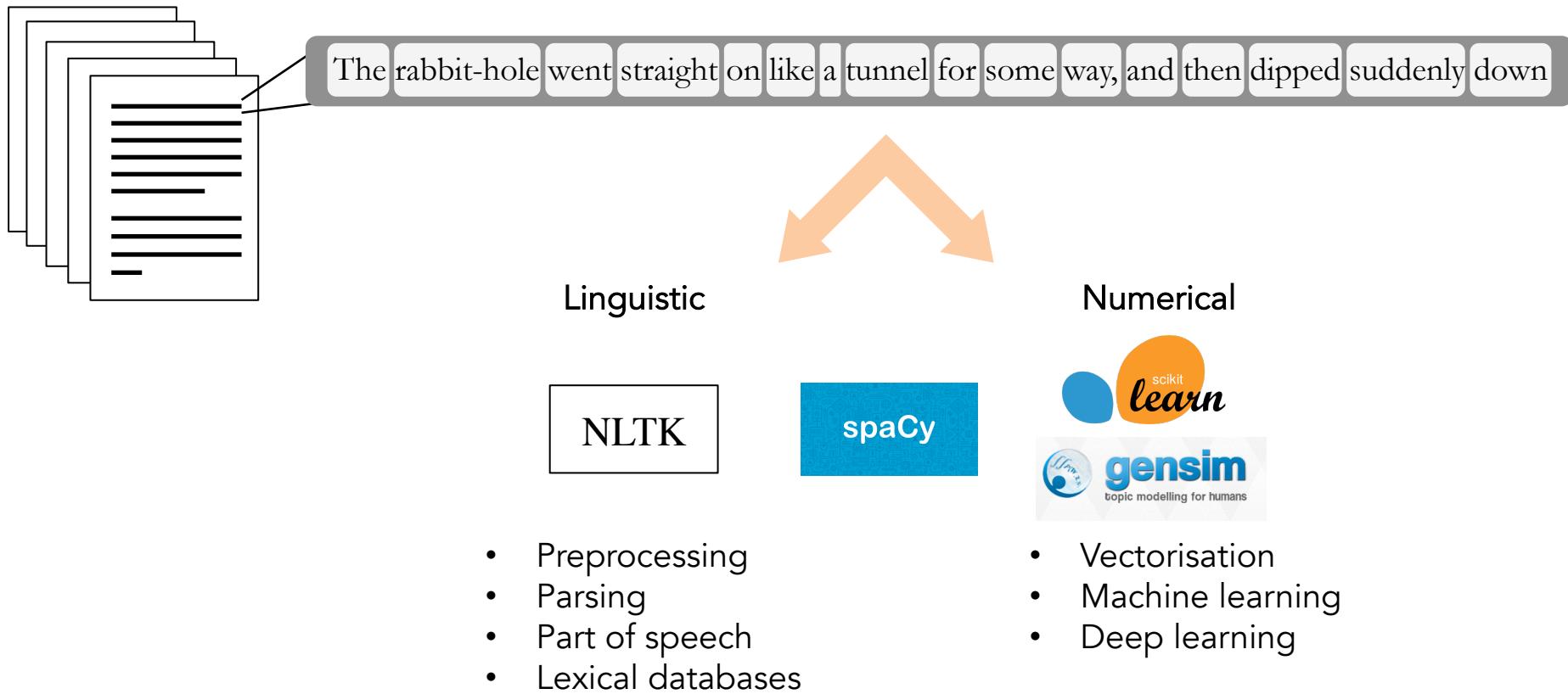
Data Collection:
Scrape content from incoming URLs

Classification:
Relevant or not relevant?
Conflict and violence or disaster?

Information Extraction:
Displacement figure, location, date,
reporting term, reporting unit

“NLP is a way for computers to analyze, understand, and derive meaning from human language in a smart and useful way.”

Matt Kiser, Algorithmia



Content Confusion

The storm killed more than 500 people in Haiti and at least 23 in the U.S. — nearly half of them in North Carolina. At least three people were missing. The full extent of the disaster in North Carolina was still unclear, but it appeared that thousands of homes were damaged, and more were in danger of flooding.

...

A levee in Lumberton appeared to fail overnight, but officials later concluded that floodwaters had flowed around it. About 1,500 people had to be rescued early Monday. Most of them were in knee-deep water, but some fled to rooftops as the brown waters swirled around them.

Content Confusion

The **storm** killed more than 500 people in Haiti and at least 23 in the U.S. — nearly half of them in **North Carolina**. At least three people were missing. The full extent of the **disaster** in **North Carolina** was still unclear, but it appeared that thousands of homes were damaged, and more were in danger of **flooding**.

...

A levee in **Lumberton** appeared to fail overnight, but officials later concluded that floodwaters had flowed around it. About **1,500 people** had to be **rescued** early **Monday**. Most of them were in knee-deep **water**, but some fled to rooftops as the brown **waters** swirled around them.

Continuous Content Confusion

...NDRF Director General O P Singh told PTI that while five teams are being airlifted for immediate deployment from its base in Odisha to Uttar Pradesh, the rest five are being picked by choppers from Bathinda in Punjab and will be sent to Bihar.

...

Singh said the teams will be in addition to the 56 such contingents which are undertaking flood combat operations in these two states, besides Rajasthan, Madhya Pradesh and Uttarakhand.

...

"So far, the NDRF teams have evacuated more than 26,400 people from various flood-prone areas in the country this monsoon season.

...

In Bihar on Monday, NDRF teams evacuated 3,400 people from Didarganj, 580 from Bakhtiyarpur, 545 from Danapur, 380 from Chhapra, 355 from Vaishali and 15 from Maner in Patna. The NDRF said 11 flood rescue teams rescued 275 people from Ballia, 275 from Varanasi and 325 from Chitrakoot in UP on yesterday. Nearly 150 marooned people were shifted to safer places from Rewa district in Madhya Pradesh on Sunday.

Continuous Content Confusion

...NDRF Director General O P Singh told PTI that while five teams are being airlifted for immediate **deployment** from its **base** in Odisha to Uttar Pradesh, the rest five are being picked by **choppers** from Bathinda in Punjab and will be sent to Bihar.

...

Singh said the teams will be in addition to the 56 such contingents which are undertaking **flood combat** operations in these two states, besides **Rajasthan**, **Madhya Pradesh** and **Uttarakhand**.

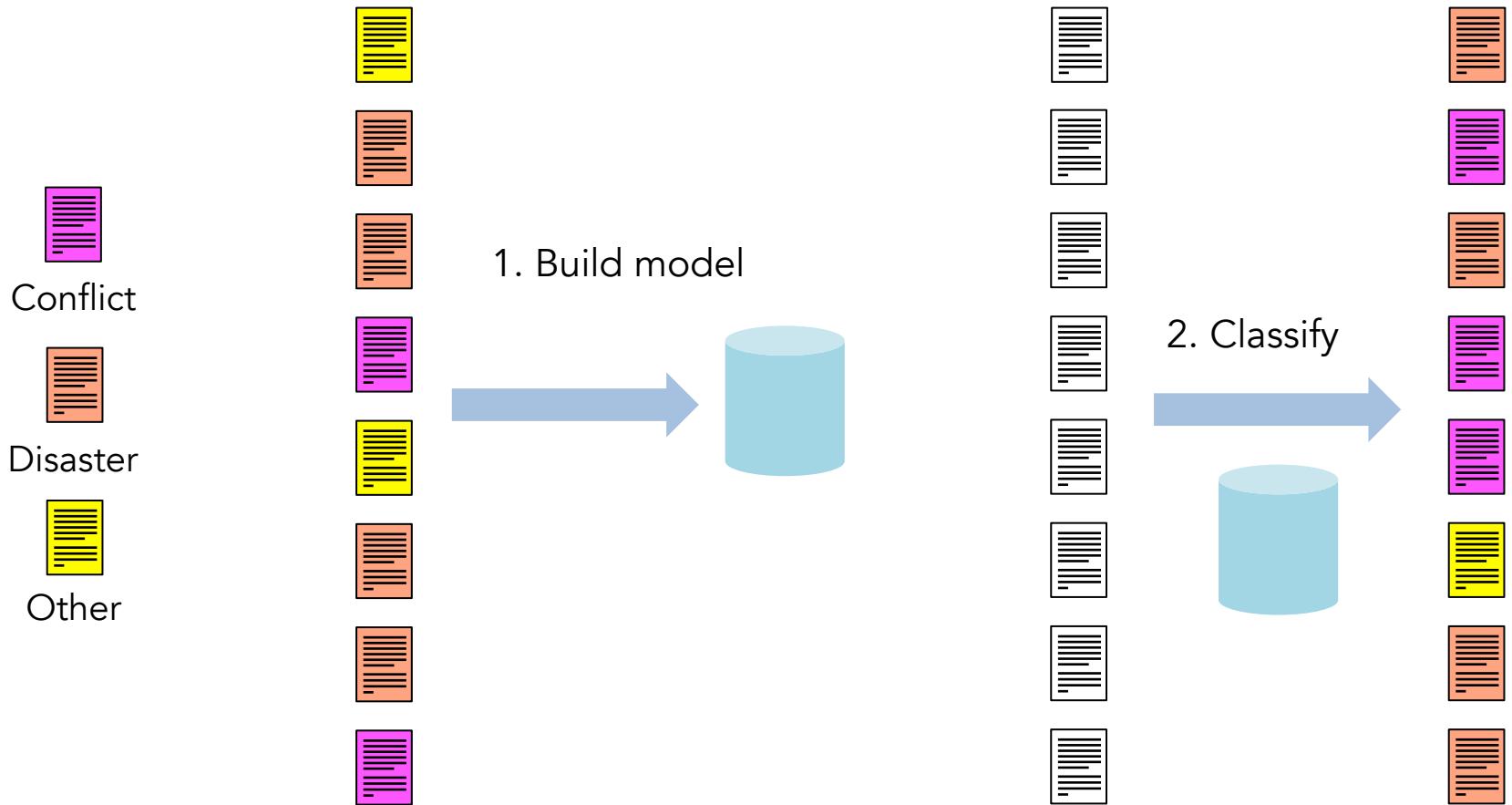
...

"So far, the NDRF teams have **evacuated** more than **26,400 people** from various **flood-prone** areas in the country this **monsoon** season.

...

In **Bihar** on **Monday**, NDRF teams **evacuated 3,400 people** from **Didarganj**, 580 from **Bakhtiyarpur**, 545 from **Danapur**, 380 from **Chhapra**, 355 from **Vaishali** and 15 from **Maner** in **Patna**. The NDRF said 11 flood rescue teams rescued **275 people** from **Ballia**, **275** from **Varanasi** and **325** from **Chitrakoot** in UP on yesterday. Nearly **150 marooned people** were **shifted to safer places** from **Rewa district** in **Madhya Pradesh** on **Sunday**.

Classification



Keyword Approach

```
conflict_tokens = ['war', 'conflict', 'military', 'ceasefire', 'terrorism',
                   'fighting', 'militia', 'rebels', 'violence', 'violent',
                   'clash', 'insurgent', 'besiege', 'bomb', 'gun', 'combat',
                   'siege', 'battle', 'battleground', 'explode', 'explosive',
                   'peace', 'truce', 'airstrike']
disaster_tokens = ['flood', 'wildfire', 'fire', 'earthquake', 'mudslide',
                   'landslide', 'washed', 'hurricane', 'storm', 'rain',
                   'rainfall', 'river', 'sea', 'disaster', 'volcano', 'typhoon',
                   'blaze', 'tremor', 'drought', 'disease', 'malnutrition',
                   'virus', 'health', 'tornado', 'forest', 'snow']
```

	Precision	Recall	f1
Conflict	0.86	0.69	0.77
Disaster	0.90	0.89	0.89
Other	0.58	0.71	0.64
Avg	0.82	0.80	0.81

Article Classification

Term Frequency Matrix

	country	evacuated	fighting	flood	...	houses	storm
Document 1	2	1	2			1	
Document 2	1	2		2			
Document 3	2						3

Term Frequency-Inverse Document Frequency Matrix (tf-idf)

	country	evacuated	fighting	flood	...	houses	storm
Document 1	0.12	0.21	0.51			0.02	
Document 2	0.10	0.25		0.38			
Document 3	0.12						0.47

Article Classification



Dimensionality Reduction (SVD)



Topic Matrix (Latent Semantic Indexing) ($k \sim 100$)

	topic 0	topic 1	topic 2	topic 3	...	topic k
Document 1	0.12	0.93	0.38	0.07		
Document 2	0.34	0.25	0.30	0.46		
Document 3	0.01	0.14	0.87	0.22		

topic 0: 0.7 * flooding + 0.2 * flee + 0.14 * water + ...

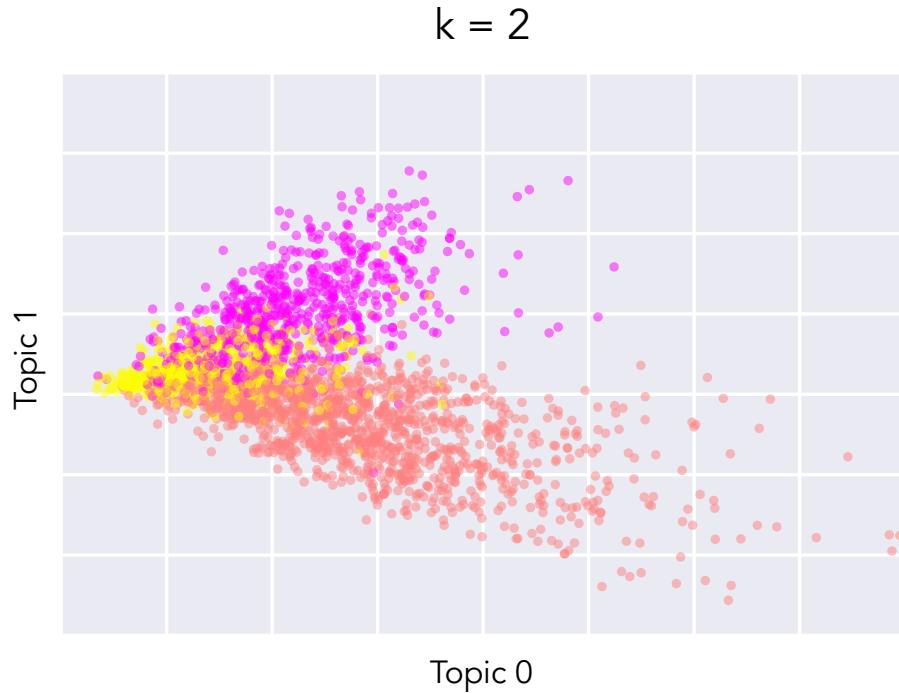
topic 1: 0.6 * vehicles + 0.3 * combat + 0.14 * nation + ...

topic 2: 0.4 * evacuated + 0.2 * people + 0.2 * houses + ...

...

Article Classification

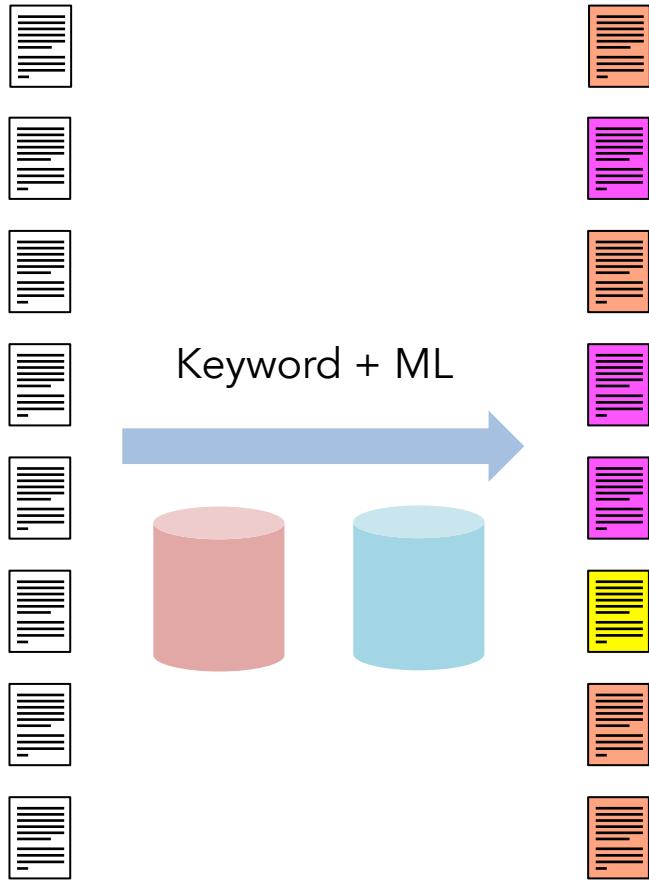
-  Conflict
-  Disaster
-  Other



$k = 100$ with Random Forest

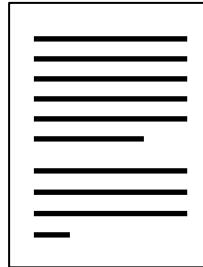
	Precision	Recall	f1
Conflict	0.77	0.8	0.79
Disaster	0.91	0.97	0.94
Other	0.73	0.57	0.64
Avg	0.84	0.85	0.84

Article Classification



	Precision	Recall	f1
Conflict	0.98	0.97	0.98
Disaster	0.98	0.99	0.99
Other	0.97	0.95	0.96
Avg	0.98	0.98	0.98

Information Extraction



- Date
- Location
- Displacement Figure
- Reporting Term
- Reporting Unit

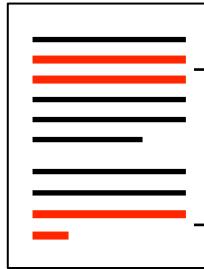
Reporting Terms

- displaced
- evacuated
- forced to flee
- homeless
- in relief camp
- sheltered
- relocated
- destroyed housing
- partially destroyed housing
- uninhabitable housing

Reporting Units

- people
- persons
- individuals
- children
- inhabitants
- residents
- migrants
- families
- households
- houses
- homes

Articles to Reports



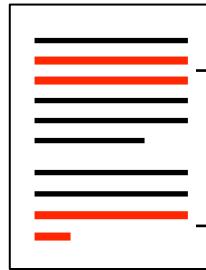
on **Monday**, NDRF teams **evacuated 3,400 people** from **Didarganj**

NDRF teams have **evacuated** more than **26,400 people** from various **flood**-prone areas

Report = Key word co-occurrences:

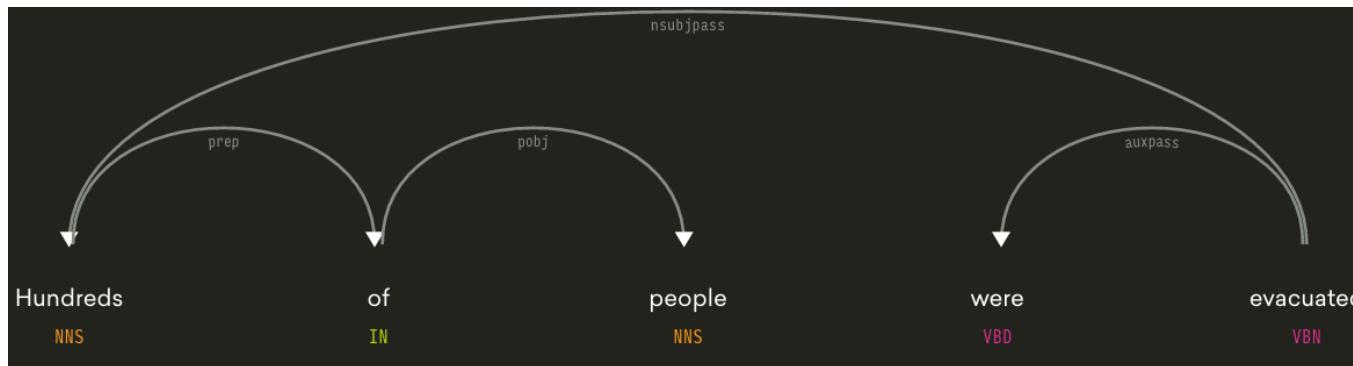
e.g. "people" + "evacuated"

Information Extraction



on **Monday**, NDRF teams **evacuated** **3,400 people** from **Didarganj**

NDRF teams have **evacuated** more than **26,400 people** from various **flood-prone** areas



Reporting Term (displaced, evacuated, ...) -> verb

Reporting Unit (people, houses, ...) -> object or subject of the verb

Entity Recognition

spaCy + displaCy

More than 200 CARDINAL people where evacuated from
their homes in GPE last week DATE .

2,000 citizens were moved to safety

Reporting Terms

- displaced
- evacuated
- forced to flee
- homeless
- in relief camp
- sheltered
- relocated
- destroyed housing
- partially destroyed housing
- uninhabitable housing

Reporting Units

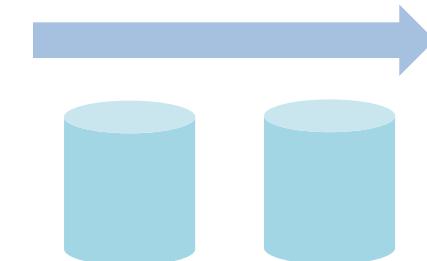
- people
- persons
- individuals
- children
- inhabitants
- residents
- migrants
- families
- households
- houses
- homes

Information Extraction



- Word2Vec Models
- Word Vectorisation
- Bigrams

	a1	a2	a3	a4	a5	...	an
r0							
r1							
r3							
r4							
r5							
...							
rm							

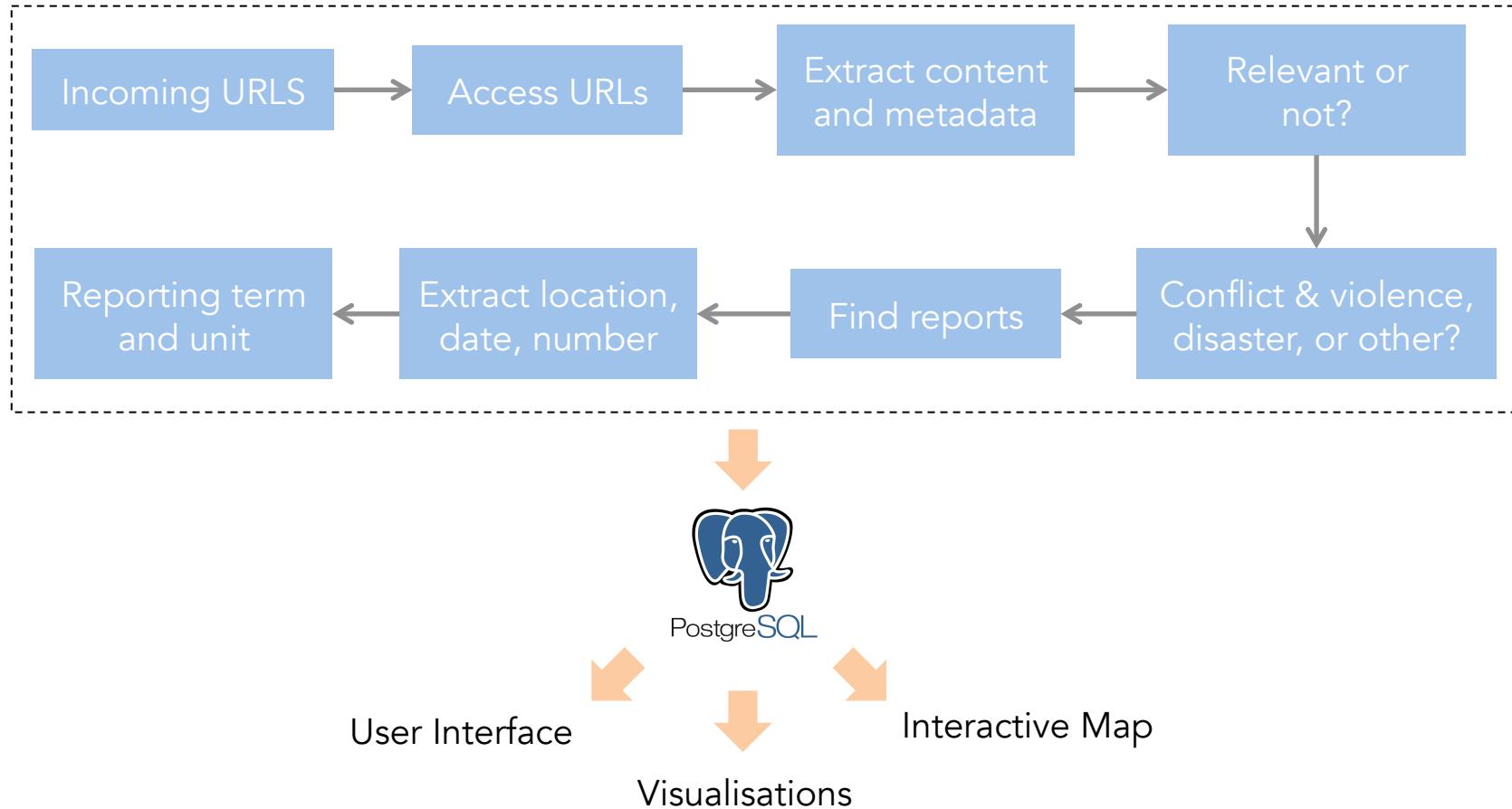


Rules + ML



F1-score Reporting Unit: 0.73 → 0.93
F1-score Reporting Term: 0.68 → 0.71

The Whole Picture



The Result

Welcome to the United Nations. It's your world.

UNITED NATIONS MEETINGS COVERAGE AND PRESS RELEASES

PRESS RELEASE

PI/2207
22 JUNE 2017

Data for Democracy Wins Unite Ideas #IDTECT Data Challenge to Monitor Worldwide Patterns of Internal Displacement

NEW YORK, 22 June (Office of Information and Communications Technology) — The United Nations announced today that Data for Democracy (D4D) has won the [Unite Ideas](#) Internal Displacement Event Tagging and Extraction Clustering Tool (#IDTECT) data challenge.

Data for Democracy is an inclusive community-driven initiative for data scientists and technologists to volunteer and collaborate on projects that make a positive impact on society. The D4D #IDTECT solution was built by a team of volunteers from around the world, including Aneel Nazareth, George Richardson, Simon Bedford, Wendy Mak, James Allen, Yane Frenski, Domingo Hui, Charles Neiswender, Daniel Forsyth, Joshua Arnold and Alex Rich.

"Open-source developers are proving to be an invaluable resource to the United Nations and our partners"

"a machine learning approach is better suited for a non-static context like internal displacement as the themes and topics can be updated (online learning)"

What Next?

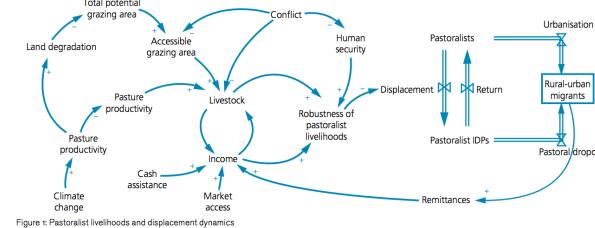
21st Century Displacement Monitoring



Satellite Image Analysis



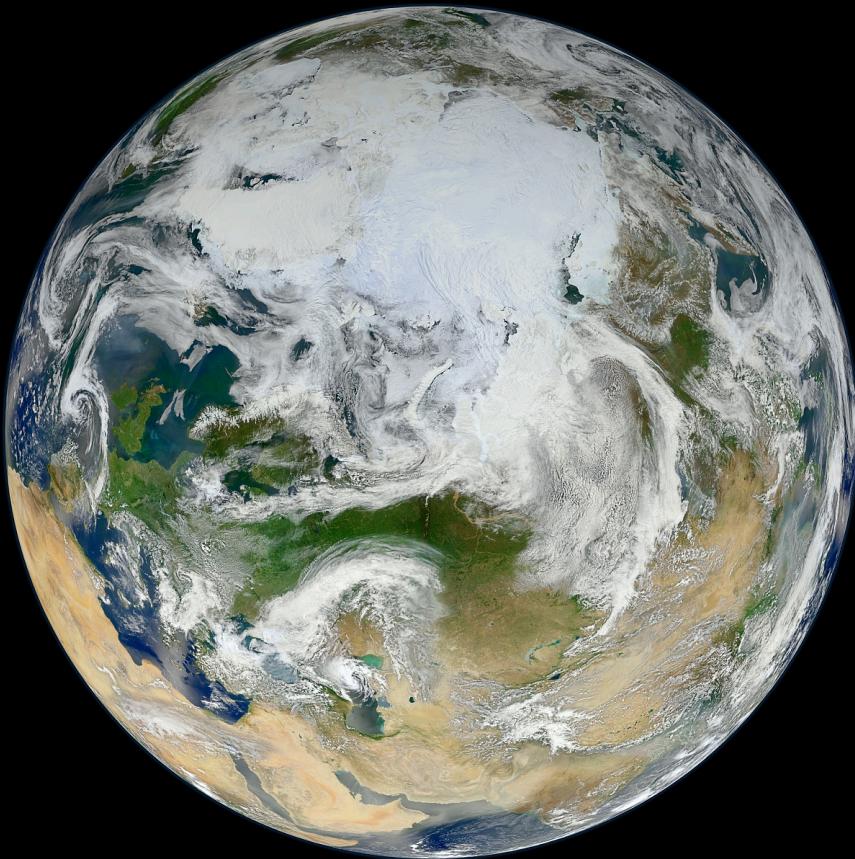
Risk Modelling



Human NLP
Human/Computer NLP

A screenshot of the reliefweb website. The header includes the reliefweb logo, navigation links for LANS, BLOG, MOBILE, ABOUT US, HELP, LOGIN/REGISTER, and social media icons. The main content area displays a 'GIEWS Country Brief: El Salvador 30-June-2017' report from the Food and Agriculture Organization of the United Nations. The report summary states: 'Cereal production in 2017 anticipated to remain at last year's high level' and 'Cereal imports forecast to sharply decline in 2017/18 marketing year'. The page also features a map of Central America with a focus on El Salvador and a news and press releases section.

Reflections



https://eoimages.gsfc.nasa.gov/images/imagerecords/78000/78349/arctic_vir_2012147_lrg.jpg



Form an Alliance

- Domain expertise
- External partnerships



Launch Something

- Do something
- Really anything



Have Fun!

- Interest area
- Awesome people



- Loose collective
- No rules, just guidelines
- Open source model applied to civic good



DATA FOR DEMOCRACY

- ★ Simon Bedford
- ★ Aneel Nazareth
- ★ Wendy Mak
- ★ James Allen
- ★ Yane Frenski
- ★ Domingo Hui
- ★ Charles Neiswender
- ★ Daniel Forsyth
- ★ Joshua Arnold
- ★ Alex Rich
- ★ Everyone else!



Get Involved



Data for Democracy Meetup
Mondays 6:30pm @ Galvanize



Seattle Social Good Series
Civic Mixer

- Meet community organisations looking for tech partnerships

Sat July 22nd 10am – 4pm
@ General Assembly

openseattle.org/civic-tech-mixer