adn: African Digital News
Building a Multilingual Monitor Corpus

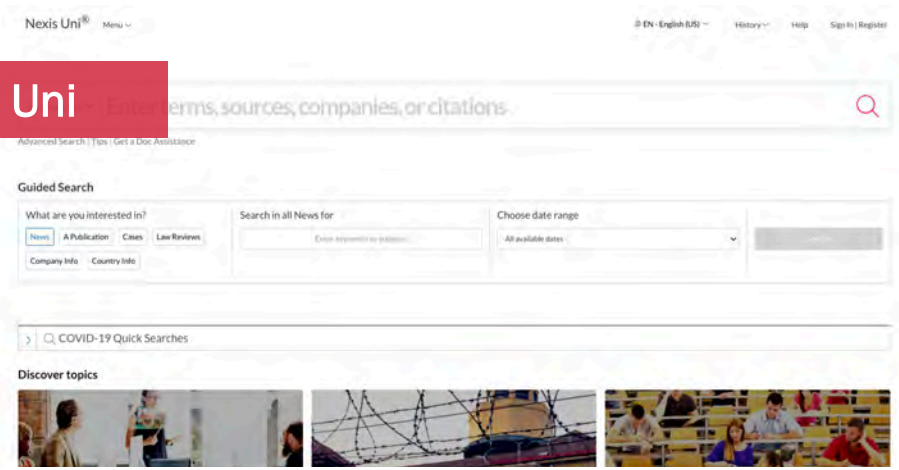
Dani Madrid-Morales
Peggy Lindner
Madhu Periyasamy
Braian Pita

April 19, 2021

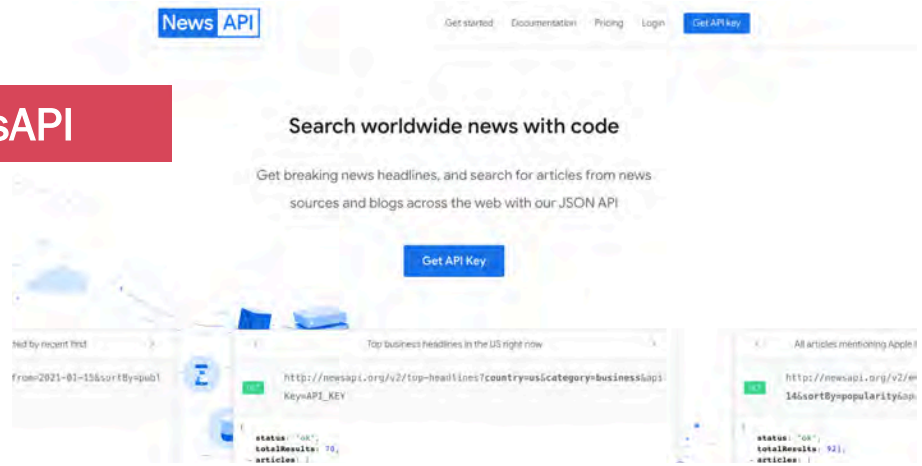
UNIVERSITY of HOUSTON

Limitations of existing archives/corpora

Nexis Uni



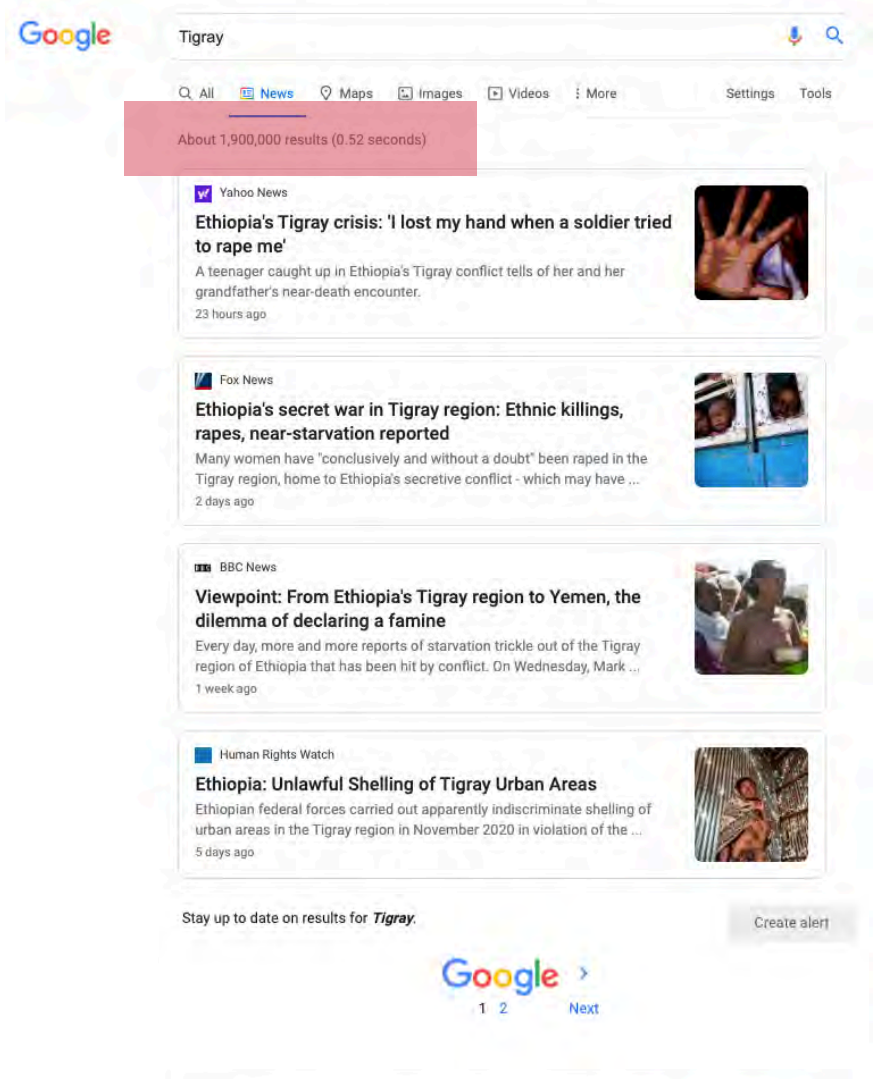
NewsAPI

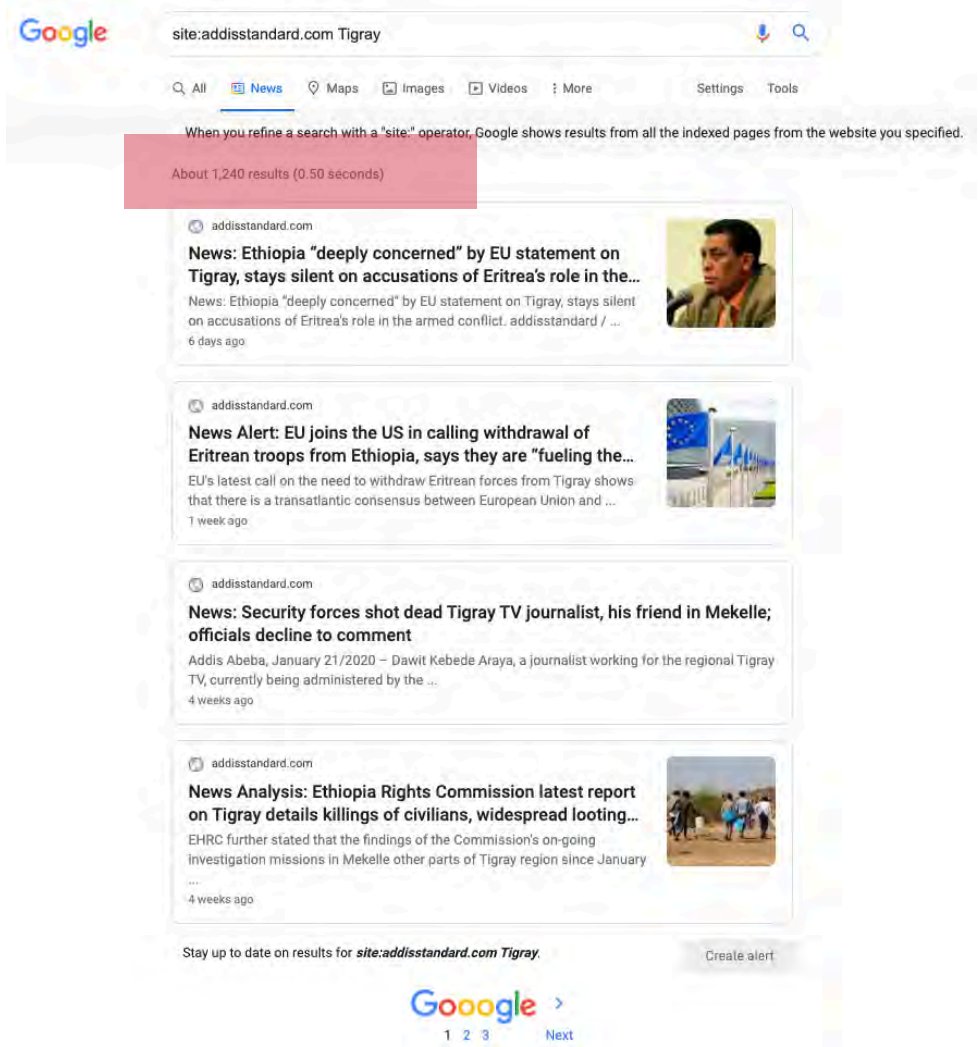


GDELT

NOW Corpus

Algorithmic biases

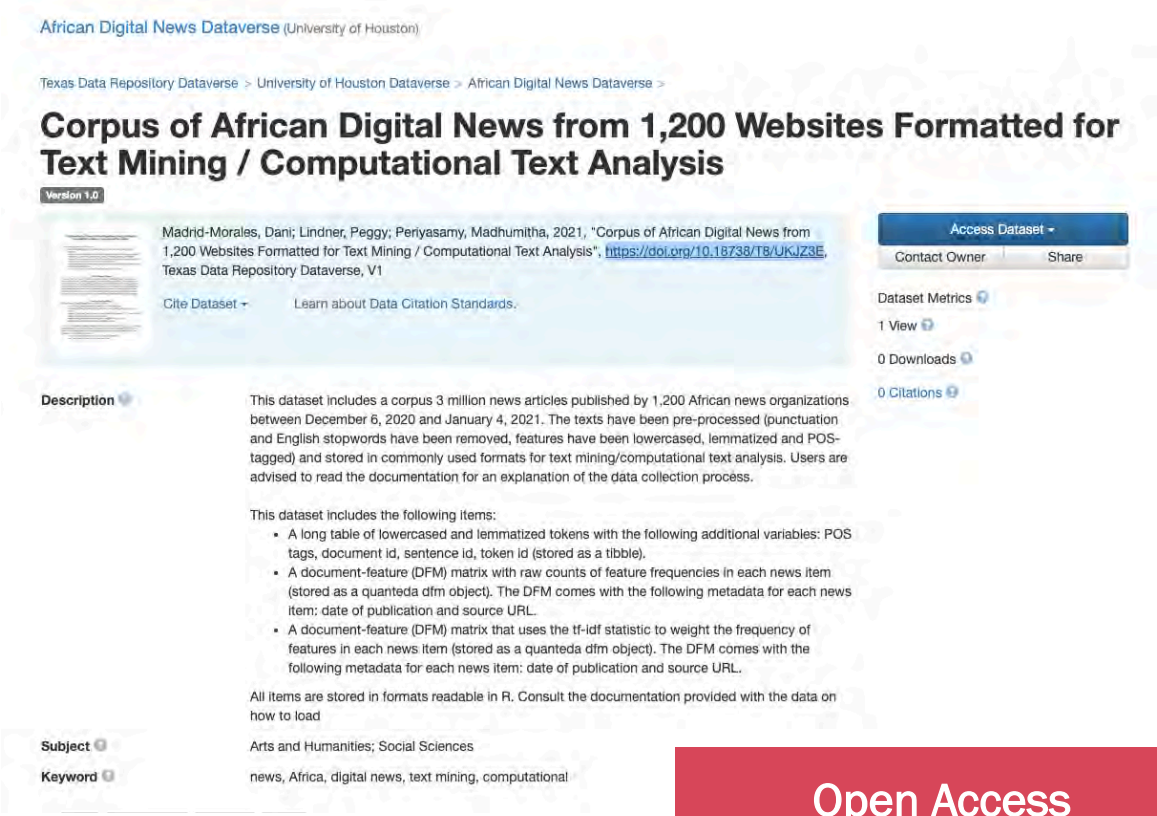




Africa Digital News Corpus

- In 2019 we started a pilot project to monitor news content on digital news in Africa in real time to create a **live corpus of news**.

- We built this using R and Python.



African Digital News Dataverse (University of Houston)

Texas Data Repository Dataverse > University of Houston Dataverse > African Digital News Dataverse >

Corpus of African Digital News from 1,200 Websites Formatted for Text Mining / Computational Text Analysis

Version 1.0

Madrid-Morales, Dani; Lindner, Peggy; Periyasamy, Madhumitha, 2021, "Corpus of African Digital News from 1,200 Websites Formatted for Text Mining / Computational Text Analysis", https://doi.org/10.18738/T8/UKJZ3E, Texas Data Repository Dataverse, V1

Cite Dataset · Learn about Data Citation Standards.

Access Dataset · Contact Owner · Share

Dataset Metrics
1 View
0 Downloads
0 Citations

Description

This dataset includes a corpus 3 million news articles published by 1,200 African news organizations between December 6, 2020 and January 4, 2021. The texts have been pre-processed (punctuation and English stopwords have been removed, features have been lowercased, lemmatized and POS-tagged) and stored in commonly used formats for text mining/computational text analysis. Users are advised to read the documentation for an explanation of the data collection process.

This dataset includes the following items:

- A long table of lowercased and lemmatized tokens with the following additional variables: POS tags, document id, sentence id, token id (stored as a tibble).
- A document-feature (DFM) matrix with raw counts of feature frequencies in each news item (stored as a quanteda dfm object). The DFM comes with the following metadata for each news item: date of publication and source URL.
- A document-feature (DFM) matrix that uses the tf-idf statistic to weight the frequency of features in each news item (stored as a quanteda dfm object). The DFM comes with the following metadata for each news item: date of publication and source URL.

All items are stored in formats readable in R. Consult the documentation provided with the data on how to load

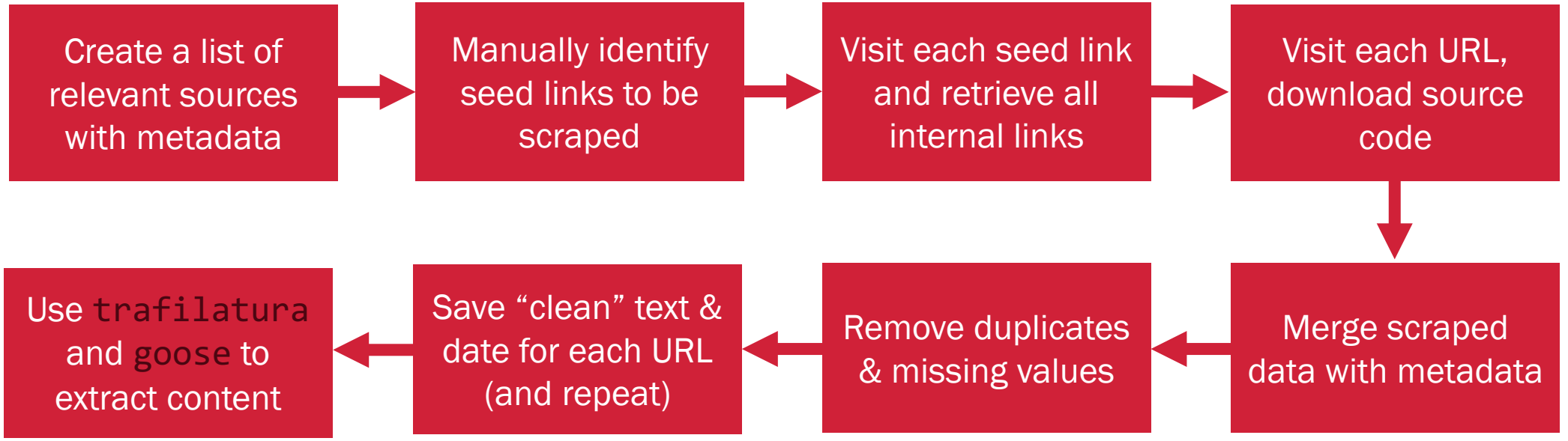
Subject Arts and Humanities; Social Sciences

Keyword news, Africa, digital news, text mining, computational

https://doi.org/10.18738/T8/UKJZ3E

Open Access
Open Source

Africa Digital News Corpus



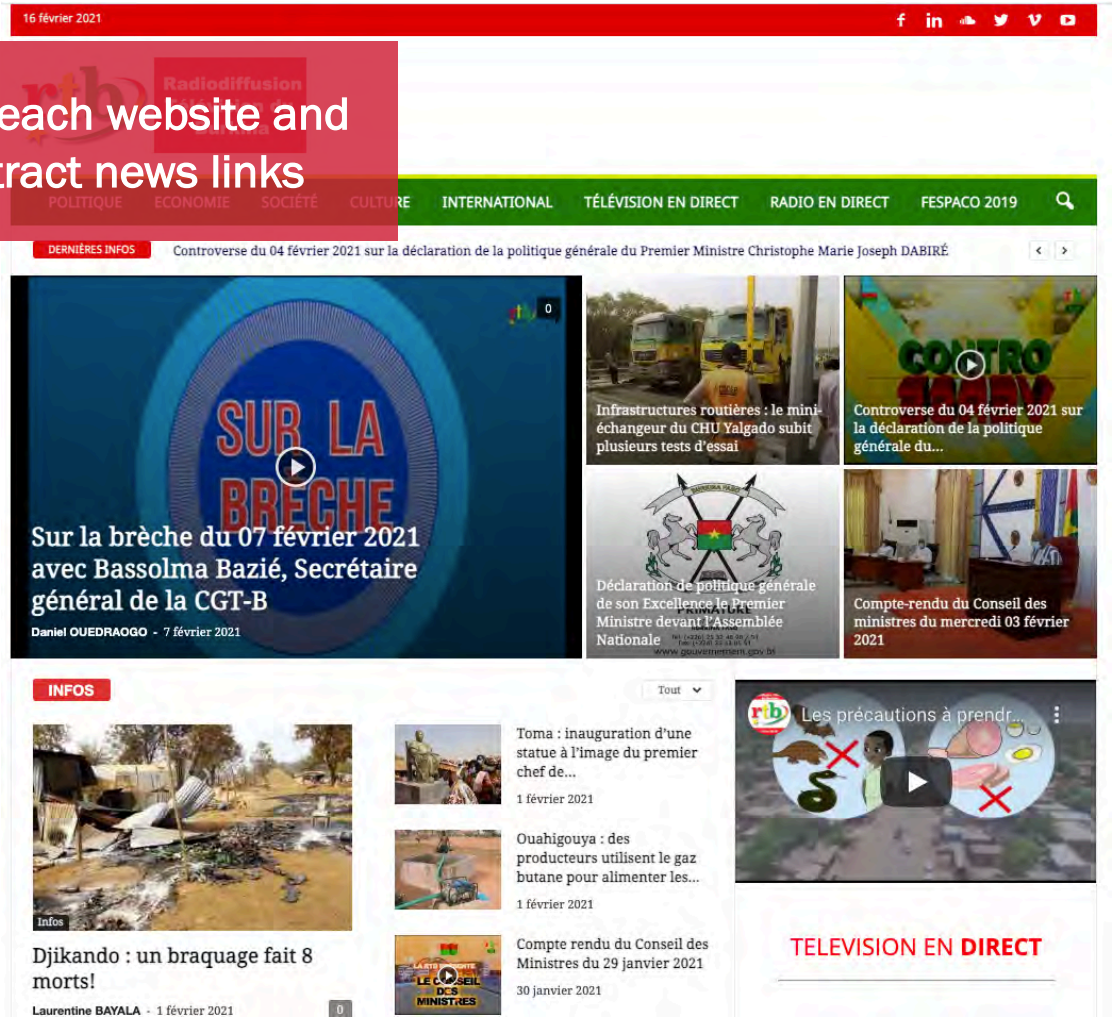
Process – Step 1: Seed Links

Create a list of pages you want to scrape

country_name	country_is	medium_name	alexa_sco	source	url	news_site
Burkina Faso	BF	http://www.aib.media	NA	xyzlinks	http://www.aib.media	TRUE
Burkina Faso	BF	http://www.aouaga.com	NA	xyzlinks	http://www.aouaga.com	TRUE
Burkina Faso	BF	http://www.aujourd8.net	NA	xyzlinks	http://www.aujourd8.net	TRUE
Burkina Faso	BF	http://www.burkina24.com	NA	xyzlinks	http://www.burkina24.com	TRUE
Burkina Faso	BF	http://www.burkinainfo.com	NA	xyzlinks	http://www.burkinainfo.com	TRUE
Burkina Faso	BF	http://www.faso-actu.net	NA	xyzlinks	http://www.faso-actu.net	TRUE
Burkina Faso	BF	http://www.fasopresse.net	NA	xyzlinks	http://www.fasopresse.net	TRUE
Burkina Faso	BF	http://www.fasozine.com	NA	xyzlinks	http://www.fasozine.com	TRUE
Burkina Faso	BF	http://www.leconomistedufaso.bf	NA	xyzlinks	http://www.leconomistedufaso.bf	TRUE
Burkina Faso	BF	http://www.lefaso.net	NA	xyzlinks	http://www.lefaso.net	TRUE
Burkina Faso	BF	http://www.lepays.bf	NA	xyzlinks	http://www.lepays.bf	TRUE
Burkina Faso	BF	http://www.lepays.bf	NA	xyzlinks	http://www.lepays.bf	TRUE
Burkina Faso	BF	http://www.lesaffairesbf.com	NA	xyzlinks	http://www.lesaffairesbf.com	TRUE
Burkina Faso	BF	http://www.lexpressdufaso-bf.com	NA	xyzlinks	http://www.lexpressdufaso-bf.com	TRUE
Burkina Faso	BF	http://www.lobservateur.bf	NA	xyzlinks	http://www.lobservateur.bf	TRUE
Burkina Faso	BF	http://www.lobservateur.bf	NA	xyzlinks	http://www.lobservateur.bf	TRUE
Burkina Faso	BF	http://www.monpulsar.com	NA	xyzlinks	http://www.monpulsar.com	TRUE
Burkina Faso	BF	http://www.mutationsbf.info	NA	xyzlinks	http://www.mutationsbf.info	TRUE
Burkina Faso	BF	http://www.omegabf.info	NA	xyzlinks	http://www.omegabf.info	TRUE
Burkina Faso	BF	http://www.ouagafm.bf	NA	xyzlinks	http://www.ouagafm.bf	TRUE
Burkina Faso	BF	http://www.radiob24.com	NA	xyzlinks	http://www.radiob24.com	TRUE
Burkina Faso	BF	http://www.rtb.bf	NA	xyzlinks	http://www.rtb.bf	TRUE
Burkina Faso	BF	http://www.savanefm.bf	NA	xyzlinks	http://www.savanefm.bf	TRUE
Burkina Faso	BF	http://www.sidwaya.info	NA	xyzlinks	http://www.sidwaya.info	TRUE
Burkina Faso	BF	http://www.sidwaya.info	NA	xyzlinks	http://www.sidwaya.info	TRUE
Burundi	BI	http://www.abpinfos.com	NA	xyzlinks	http://www.abpinfos.com	TRUE
Burundi	BI	http://www.arib.info	NA	xyzlinks	http://www.arib.info	TRUE
Burundi	BI	http://www.burundi24.wordpress.com	NA	xyzlinks	http://www.burundi24.wordpress.com	TRUE
Burundi	BI	http://www.burundi-agnews.org	NA	xyzlinks	http://www.burundi-agnews.org	TRUE
Burundi	BI	http://www.burundi-eco.com	NA	xyzlinks	http://www.burundi-eco.com	TRUE
Burundi	BI	http://www.burundi-forum.org	NA	xyzlinks	http://www.burundi-forum.org	TRUE
Burundi	BI	http://www.burundi-information.net	NA	xyzlinks	http://www.burundi-information.net	TRUE
Burundi	BI	http://www.fr.itaraburundi.com	NA	xyzlinks	http://www.fr.itaraburundi.com	TRUE
Burundi	BI	http://www.isanganiro.org	NA	xyzlinks	http://www.isanganiro.org	TRUE
Burundi	BI	http://www.itaraburundi.com	NA	xyzlinks	http://www.itaraburundi.com	TRUE
Burundi	BI	http://www.iwacu-burundi.org	NA	xyzlinks	http://www.iwacu-burundi.org	TRUE

Process – Step 2: Retrieve links

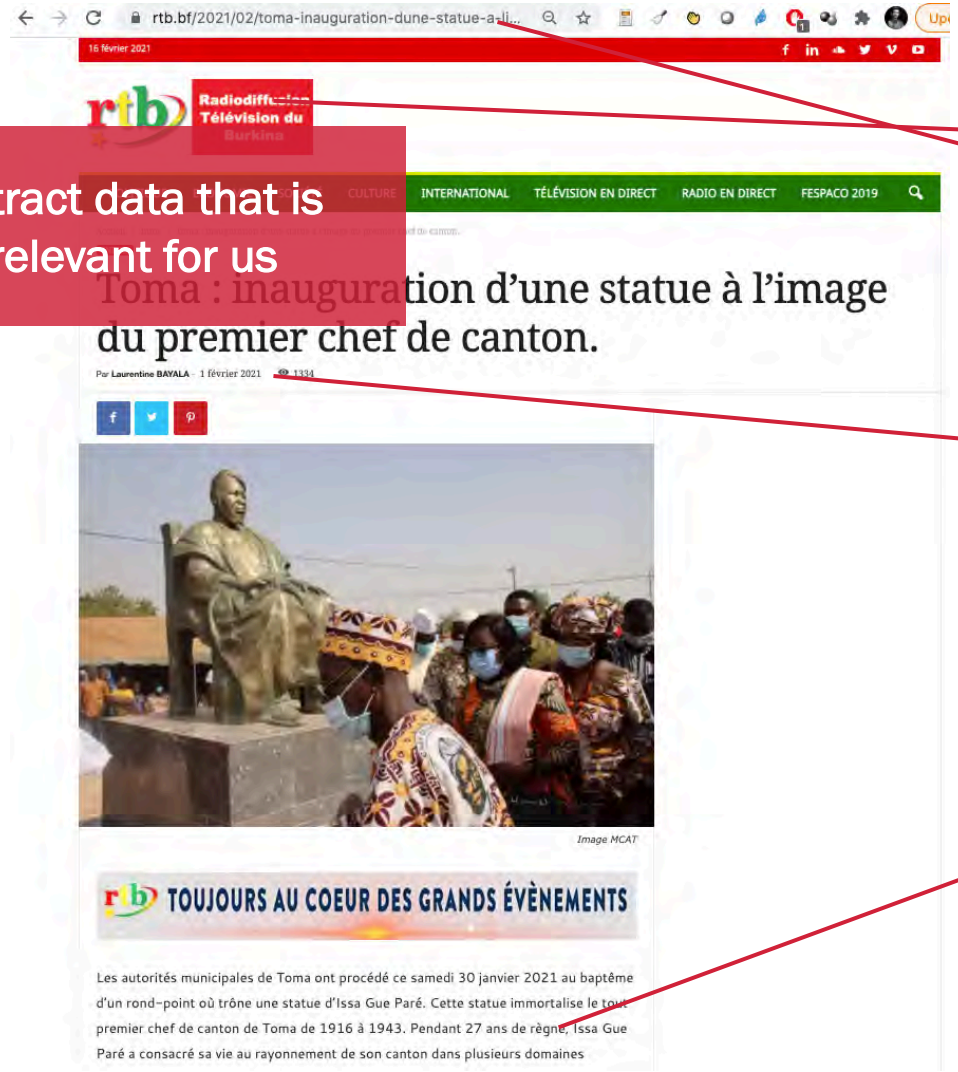
Visit each website and extract news links



Link	URL
Toma : inauguration d'une statue à l'image du premier chef de...	https://www.rtb.bf/2021/02/toma-inauguration-dune-statue-a-limage-du-premier-chef-de-canton/
Ouahigouya : des producteurs utilisent le gaz butane pour alimenter les...	https://www.rtb.bf/2021/02/ouahigouya-des-producteurs-utilisent-le-gaz-butane-pour-alimenter-les-motopompes/
Compte rendu du Conseil des Ministres du 29 janvier 2021	https://www.rtb.bf/2021/01/compte-rendu-du-conseil-des-ministres-du-29-janvier-2021/
Compte-rendu du conseil des ministres du vendredi 29 janvier 2021	https://www.rtb.bf/2021/01/compte-rendu-du-conseil-des-ministres-du-vendredi-29-janvier-2021/

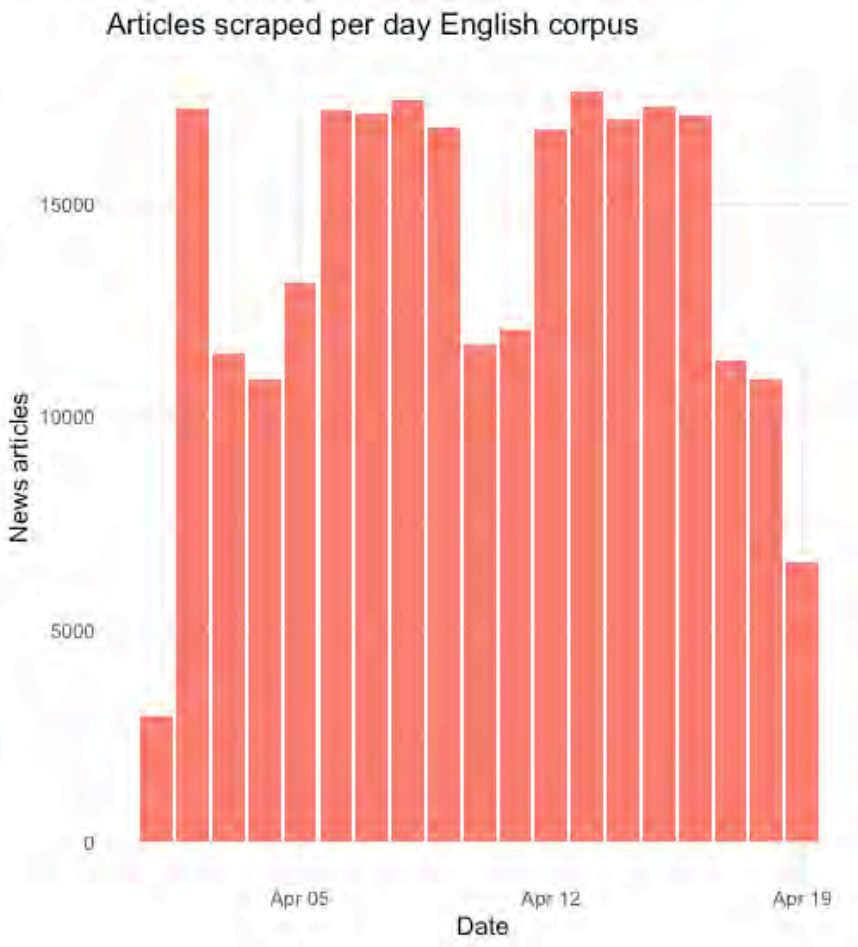
Process – Step 3: Extracting data

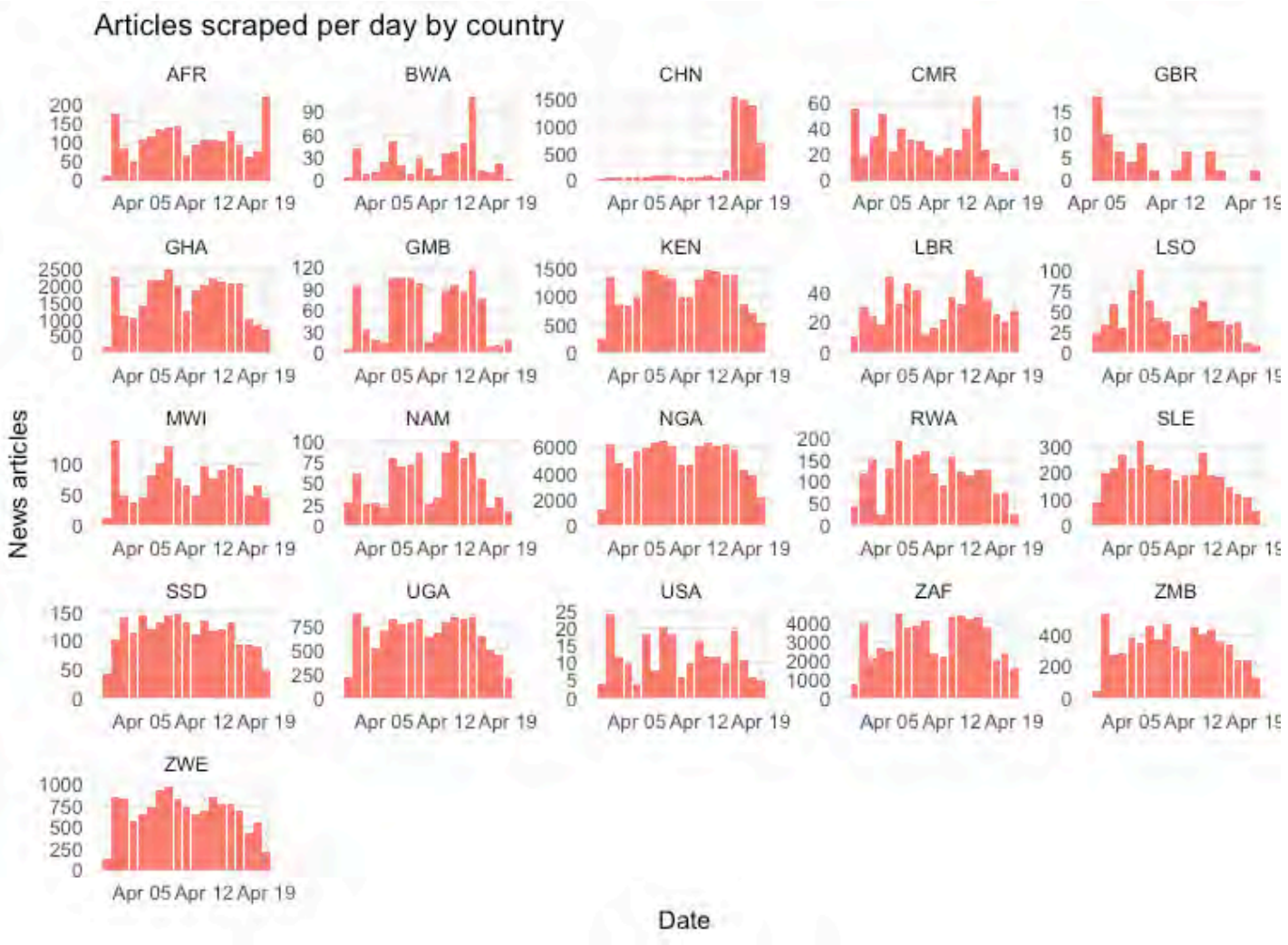
Extract data that is relevant for us



publish_date	source	country_iso	full_text	web_url
2020-04-07	abidjan.net	CI	Société Hamed Bakayoko déclaré positif au coronaviru...	https://news.abidjan.net/h/671218.html
2020-04-27	abidjan.net	CI	Société COVID 19: Un don en vivres et non vivres offe...	https://news.abidjan.net/h/672003.html
2020-04-20	abidjan.net	CI	Économie Covid-19: le directeur général de l'OMS peu...	https://news.abidjan.net/h/671694.html
2020-04-22	abidjan.net	CI	Économie Hausse du coût de la viande : les acteurs in...	https://news.abidjan.net/h/671803.html
2020-04-29	abidjan.net	CI	Société Secteur du tourisme et des loisirs : les résulta...	https://news.abidjan.net/h/672090.html
2020-03-21	abidjan.net	CI	Économie Covid-19: Les mesures urgentes prises par ...	https://news.abidjan.net/h/670625.html
2020-03-23	abidjan.net	CI	Société Journée mondiale de l'eau 2020 toujours d'ac...	https://news.abidjan.net/h/670673.html
2020-03-27	abidjan.net	CI	International New York, nouvel épicentre de la pandé...	https://news.abidjan.net/h/670803.html
2020-05-06	abidjan.net	CI	Société Ramadan : Hamed Bakayoko offre des vivres a...	https://news.abidjan.net/h/672361.html
2020-05-12	abidjan.net	CI	Société Coronavirus : Le Premier Ministre met la minis...	https://news.abidjan.net/h/672617.html
2020-05-14	abidjan.net	CI	Société Reprise des travaux de la Nationale A3 à Niak...	https://news.abidjan.net/h/672735.html
2020-05-27	abidjan.net	CI	Politique Présidentielle 2020: la révision de la liste éle...	https://news.abidjan.net/h/673528.html
2020-02-27	abidjan.net	CI	[redacted]	https://news.abidjan.net/h/669714.html
2020-04-03	abidjan.net	CI	Économie Covid- 19 Des producteurs de café-cacao r...	https://news.abidjan.net/h/671048.html
2020-04-05	abidjan.net	CI	Société Le préfet de Korhogo appelle la population à s...	https://news.abidjan.net/h/671111.html
2020-04-10	abidjan.net	CI	Région COVID-19 : Agboville lance la phase répressiv...	https://news.abidjan.net/h/671331.html
2020-04-12	abidjan.net	CI	Société Pâques 2020 : LEglise méthodiste plaide pour ...	https://news.abidjan.net/h/671409.html
2020-04-15	abidjan.net	CI	Art et Culture Covid 19: Des artistes et journalistes se...	https://news.abidjan.net/h/671480.html
2020-04-10	abidjan.net	CI	Société Coronavirus : Le centre de dépistage de Yopo ...	https://news.abidjan.net/h/671348.html
2020-04-04	abidjan.net	CI	Société Covid-19 : Sensibilisation à outrance à l'intéri...	https://news.abidjan.net/h/671076.html
2020-04-09	abidjan.net	CI	Société Covid-19: Voici les techniques utilisées par d...	https://news.abidjan.net/h/671283.html
2020-04-03	abidjan.net	CI	Société Treichville/ Lutte contre le Coronavirus : le co...	https://news.abidjan.net/h/671067.html
2020-04-15	abidjan.net	CI	Société Covid-19: lÉtat de Côte dIvoire proroge le dél...	https://news.abidjan.net/h/671500.html
2020-04-19	abidjan.net	CI	Société Bruno Koné fait des dons aux ressortissants d...	https://news.abidjan.net/h/671672.html
2020-03-06	abidjan.net	CI	Société Social: Report du gala de la Fondation Childre...	https://news.abidjan.net/h/670091.html
2020-03-18	abidjan.net	CI	Société Coronavirus: Latelier de formation des format...	https://news.abidjan.net/h/670503.html
2020-03-19	abidjan.net	CI	Politique Covid-19: La plateforme pro-Gbagbo, EDS, l...	https://news.abidjan.net/h/670561.html
2020-03-17	abidjan.net	CI	Société Crise du Coronavirus : Les Abidjanais vident l...	https://news.abidjan.net/h/670484.html
2020-03-27	abidjan.net	CI	Économie Impacts du Covid-19 dans lUEMOA : Avis r...	https://news.abidjan.net/h/670824.html
2020-03-26	abidjan.net	CI	Économie Covid-19/Abidjan : Le marché gouro ferme...	https://news.abidjan.net/h/670770.html
2020-05-06	abidjan.net	CI	Société Lutte contre le Covid-19: démarrage officiel d...	https://news.abidjan.net/h/672345.html
2020-05-04	abidjan.net	CI	Société Ramadan : 400 familles démunies soulagées à...	https://news.abidjan.net/h/672295.html
2020-05-10	abidjan.net	CI	Société Lutte contre le Covid-19: les commerçants du...	https://news.abidjan.net/h/672540.html
2020-05-13	abidjan.net	CI	Société Des enseignants musulmans offrent des kits a...	https://news.abidjan.net/h/672674.html
2020-05-12	abidjan.net	CI	Société Covid-19: des équipements sanitaires offerts ...	https://news.abidjan.net/h/672628.html
2020-05-20	abidjan.net	CI	Société Nuit du Qadr: Les musulmans ivoiriens invités...	https://news.abidjan.net/h/673090.html

Process – Step 3: Describing data





What does the code look like

```
##Function: Grab links from initial URLs##
get_links <- function(url){

  withTimeout({Sys.sleep(2)

    page <- read_html(url)

    web_link <- page %>%
      html_nodes("a") %>%
      html_attr("href")

    text_links <- page %>%
      html_nodes("a") %>%
      html_text()

    tibble(web_link,
           text_links,
           url)
  }, timeout = 10, onTimeout = "silent")
}
###
```

```
##Function: scrapes out text, date, and raw HTML data from a given link##
get_text_from_py <- function(url_link){

  withTimeout({
    traf <- import("trafilatura")
    requests <- import("requests")
    rand <- import("requests_random_user_agent")
    goose <- import("goose3")

    r <- traf$fetch_url(url_link)
    web_text <- traf$extract(r)
    web_raw1 <- r[1]

    web_date <- traf$metadata$extract_metadata(r) %>%
      str_extract("[0-9]{4}-[0-9]{2}-[0-9]{2}")

    g = goose$Goose({browser_user_agent = browser_agent})

    article <- g$extract(url = url_link)
    web_text2 <- article$cleaned_text
    web_date2 <- as.character(article$publish_datetime_utc)
    web_raw2 <- article$raw_html

    if (is.null(web_date2) == TRUE){
      web_date2 <- ""
    }

    if (is.null(web_text2) == TRUE){
      web_text2 <- ""
    }

    g$close()

    tibble(web_url = url_link,
           web_date = web_date,
           web_date2 = web_date2,
           web_sysdate = Sys.Date(),
           web_text = web_text,
           web_text2 = web_text2,
           web_raw1 = web_raw1,
           web_raw2 = web_raw2)
  }, timeout = 10, onTimeout = "silent")

}
```

Step 4: Processing the data (optional)

- Use spacy (or udpipe) to tokenize (or lemmatize) the text and POS (part of speech tag)

- We can then remove tokens that we don't "need" (stopwords, numbers, punctuation…)

Filter

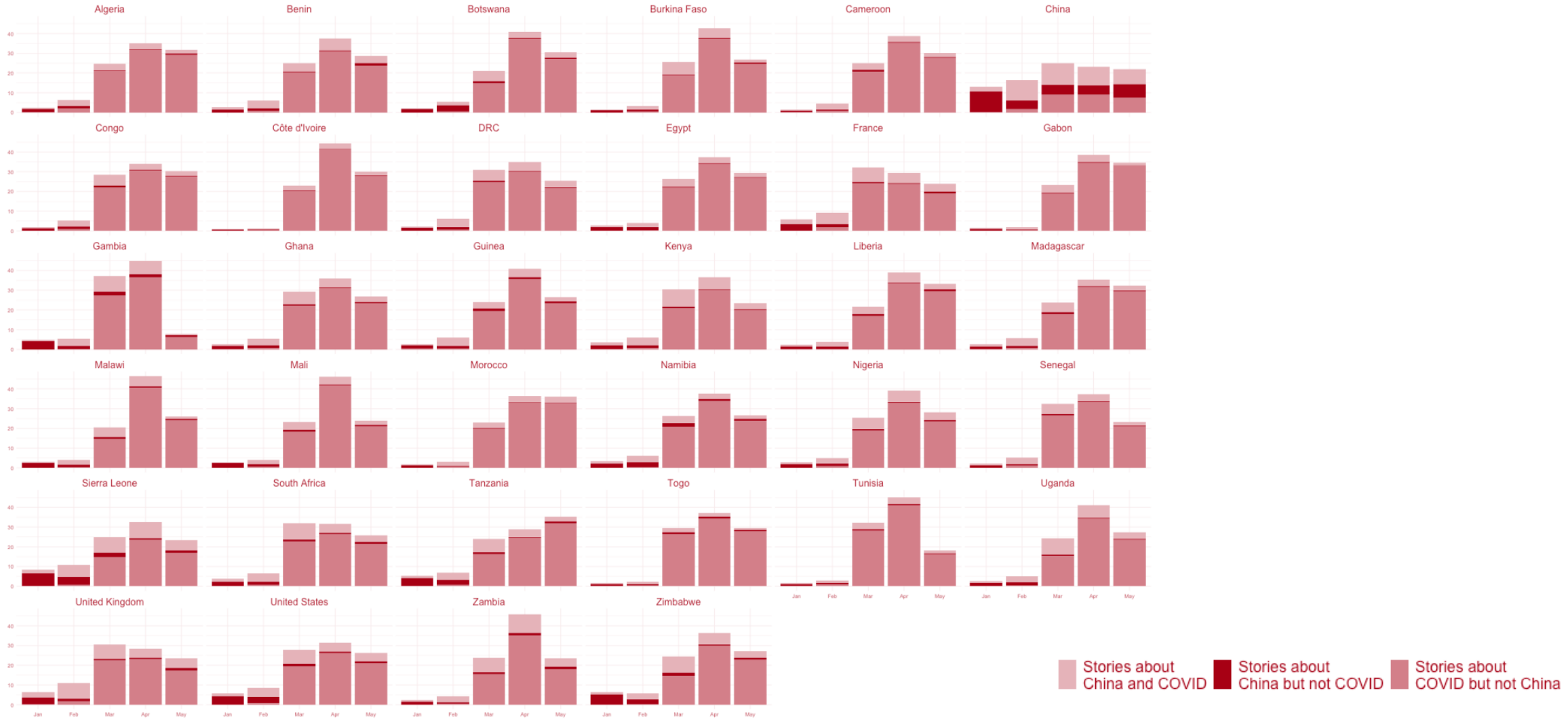
	doc_id	sentence_id	token_id	token	lemma	pos	entity
1	STMID_006086	1	1	When	when	ADV	
2	STMID_006086	1	2	Kem	Kem	PROPN	
3	STMID_006086	1	3	Senou	Senou	PROPN	
4	STMID_006086	1	4	Pavel	Pavel	PROPN	
5	STMID_006086	1	5	Daryl	Daryl	PROPN	
6	STMID_006086	1	6	,	,	PUNCT	
7	STMID_006086	1	7	a	a	DET	
8	STMID_006086	1	8	21-year	21-year	NUM	DATE_B
9	STMID_006086	1	9	-	-	PUNCT	DATE_I
10	STMID_006086	1	10	old	old	ADJ	DATE_I
11	STMID_006086	1	11	Cameroonian	cameroonian	ADJ	NORP_B
12	STMID_006086	1	12	student	student	NOUN	
13	STMID_006086	1	13	living	live	VERB	
14	STMID_006086	1	14	in	in	ADP	
15	STMID_006086	1	15	the	the	DET	
16	STMID_006086	1	16	Chinese	chinese	ADJ	NORP_B
17	STMID_006086	1	17	city	city	NOUN	
18	STMID_006086	1	18	of	of	ADP	
19	STMID_006086	1	19	Jingzhou	Jingzhou	PROPN	GPE_B
20	STMID_006086	1	20	,	,	PUNCT	

Applications

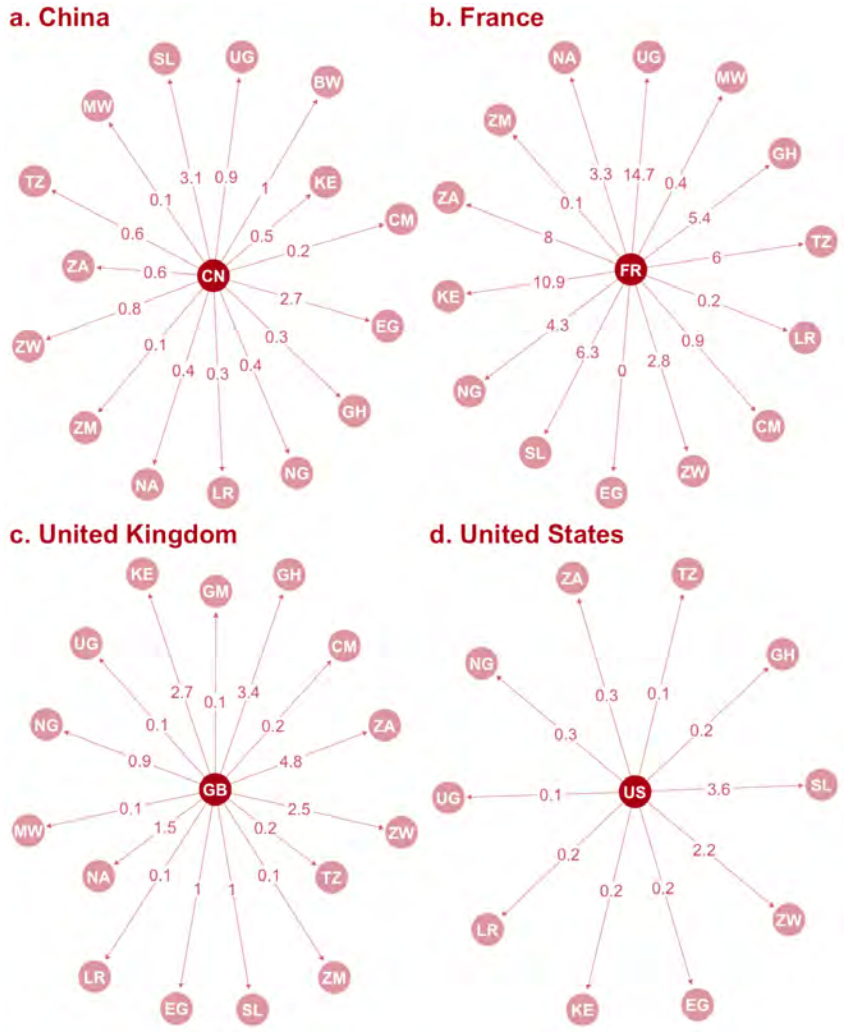
- This same process can be used to collect content from government websites, discussion fora, citizen journalism platforms…

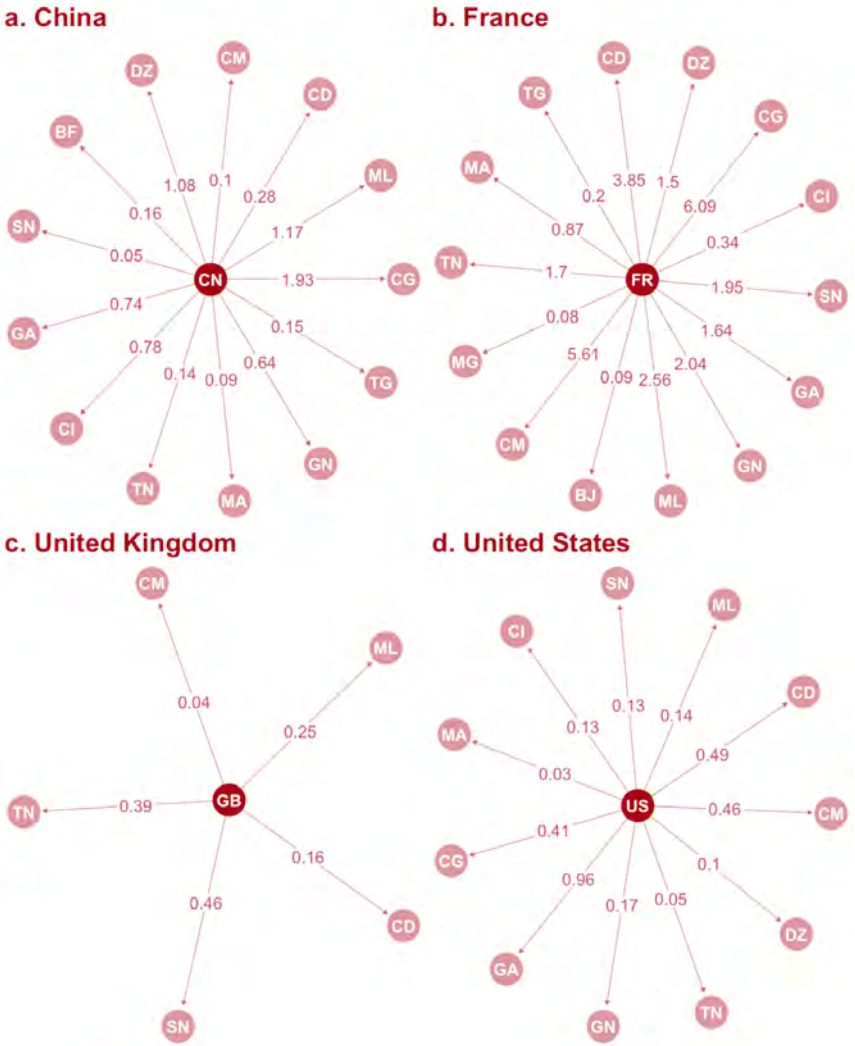
- With this type of data…
  - We can answer **descriptive questions** about content
  - We can answer questions about **linguistic patterns**
  - We can **compare sources** across geographic spaces
  - We can trace **longitudinal patterns**
  - We can **triangulate text data with fieldwork**

Application – News flows



Application – News flows





More information

- (Some of the) data and the code for this study is available in this repository:
  - https://github.com/danimadrid/africa_china_covid19

- There's more details on the process that I describe here in here:
  - https://doi.org/10.1080/23743670.2020.1820885

Limitations

- This approach would (in principle) exclude data that is behind **paywalls**; it is internet data heavy.

- These approaches are heavy on the **amount of bandwidth** that is required (cloud computing as alternative).

- Downtime of websites and other technical problems might limit how much data we can collect, so it is important to acknowledge **biases in our dataset**.

Next stages

- Pilot study with English text [2019]
- Beta version of multilingual database [current]
- Launch production stage [June 2021]
- Opening access to monitor corpus to users [September 2021]
- Data processing and insights [pending funding]

adn: African Digital News
Building a Multilingual Monitor Corpus

Dani Madrid-Morales, PhD
dmmorales2@uh.edu
@DMadrid_M
www.danimadrid.net

UNIVERSITY of HOUSTON