Covid19 politics / media twitter database

By Bruno Chaves
Public Governance working group
24-01-2022

Purpose of the presentation

- Present the raw data collected with minimum cleaning
  Appropriate cleaning depends on the analytical methods will be used.
- Make the data available and documented for research collaboration.
  - GitHub : https://github.com/chaves/acss_covid
  - Website : https://acss-covid.vercel.app/

  (*with programs and descriptive statistics but not data sources*)

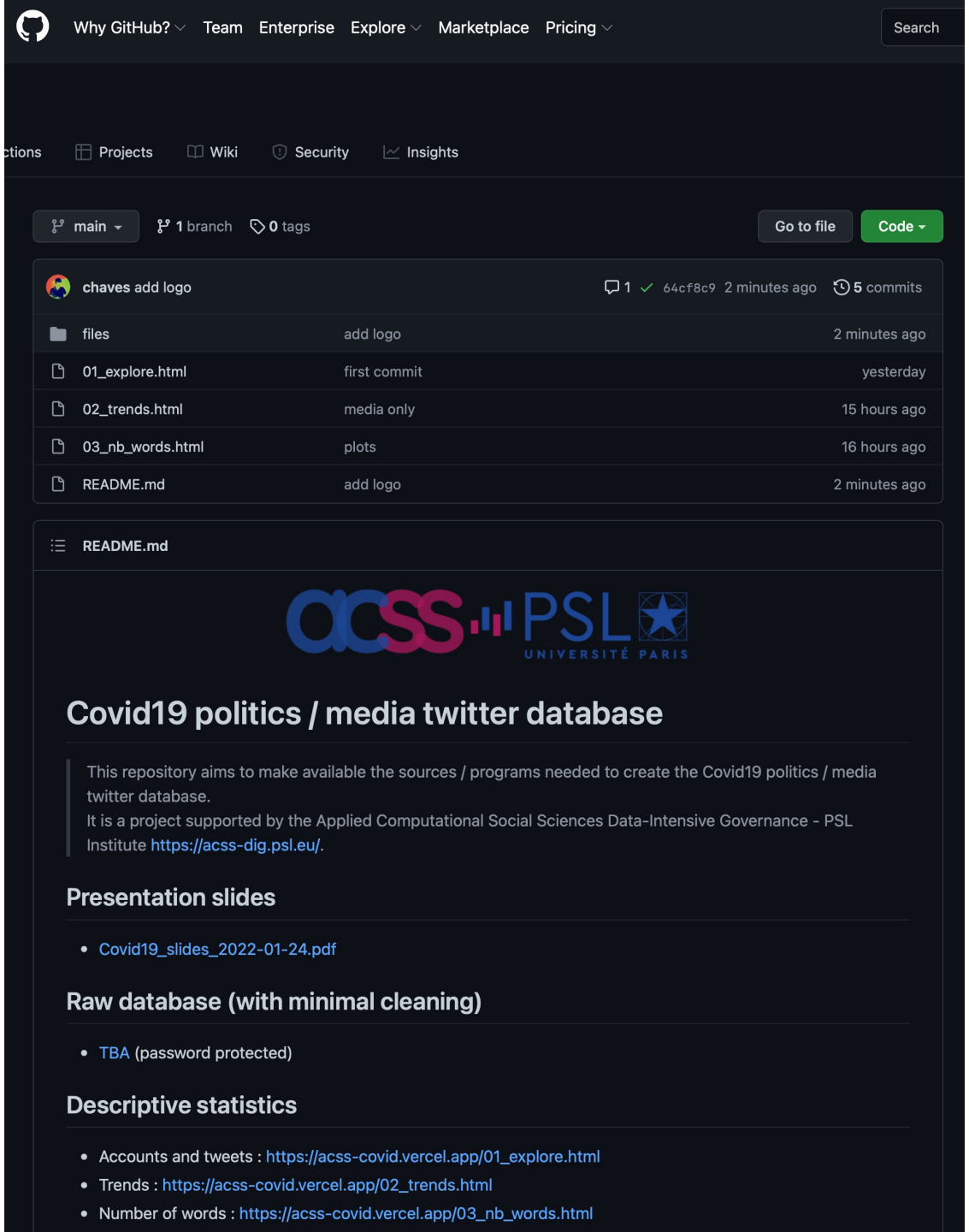
I had to restart the process from the beginning (except the scraping of the tweets) to :

- Assess the quality of the data
- Better understand what it contains. Indeed, the 3 other main contributors of the project have left: Abir, Nicolas and Antoine.

Note : I already analyzed this data (e.g. with STM) with a sample of the tweets with at least 10 words. This may be the subject of a future presentation.

GitHub repository

https://github.com/chaves/acss_covid



Round 1

- Explore the narration of politics v. experts (in health and economics) in the context of the pandemic in G20 countries.
- Data : **all tweets** by these accounts from **2019-12-01 to 2021-01-27** (*I scrapped this data*)
- BUT : it's very difficult to get a balanced sample :
  - Expert types (health v. economics and others) :
    - e.g. The topics of German tweets are more related to economic issues and French tweets to health issues. A stylized fact or a selection problem ?
  - Very contrasted countries (e.g. India, Russia, Brazil v. USA, UK, France)

Round 2

- Explore the narration of politics v. **media** (not experts) in the context of the pandemic.
- 12 countries : [USA, Germany, France, Sweden, Spain, UK, Italy, Canada, Netherlands, Australia, New Zealand, Poland]
  (we started with the G20 countries, than we reduced it to European and English speaking ones)
- Data : tweets by these accounts from **2019-12-01 to 2021-01-27**
  - Politics : all tweets (*scraping by Abir*)
  - Media : only the tweets related to Covid (*scraping by Nicolas*)
  - Total : **1 420 958 tweets**

Politics sampling

Official and personal accounts of ministers, ministries (*economics, heath but also defense, international relations, etc*.), presidents, and parties represented in the parliaments.

1. POLITICS_abir.docx
2. Covariables : covid_twitter_politics_accounts_v4_Antoine.xlsx

⇒ metadata based on the parties scores (parlgov database : https://www.parlgov.org/data-info/

Media sampling

We couldn’t find media that specializes in health and economics per country, so the solution is to focus on big media, each big media has a version/axe for heath and economics.

We found the Oxford (www.digitalnewsreport.org) website that is ranking the news media based on their study.

Keywords :

- **treatment_terms** = ['Vaccination', 'Test' ,'Testing' ,'Hydroxychloroquine', 'Reanimation', 'Spread', 'Vaccine', 'Vaccines', 'Spreading', 'Asymptomatic', 'Symptomatic', 'Masks']
- **virus_terms** = ['Coronavirus', 'Corona', 'COVID', 'Virus', 'Cluster', 'Endemic', 'Epidemic', 'Outbreak', 'Pandemic', 'Incubation', 'Health']
- **emergency_measures** = ['Quarantine', 'Distancing', 'Crisis', 'Emergency', 'lockdown', 'Lock', 'confinement', 'Immunity']
- **medical_institutions** = ['Hospitals', 'Hospital', 'Laboratory', 'Laboratories', 'Home']

Data gathering

- **Scraping** : **TWINT** (Twitter Intelligence Tool). Twint allows for scraping tweets from Twitter profiles without using Twitter's API.
- **Translation to (American*) English**:
  - Google gives us a trial of 300$ to try its services, among them the Translation API, but it's nothing for translating.
  - DeepL : a very good solution, but pricier than Google Translation API

⇒ I used Google. This is a very problematic issue because the identification of the language by Twitter is very bad, sometimes the tweets have words in several languages.

⇒ We can't translate all the tweets.

⇒ Solution: a new language detection module by Spacy that works well.

Note* : I usually convert British English words to American English (when using bag of words techniques)

Summary number of tweets

https://acss-covid.vercel.app/01_explore.html

1 420 958 tweets

Tweets by country

country	nb_tweets
GB	258133
ES	252169
IT	162398
US	136871
FR	127122
CA	126277
PL	120552
DE	99350
AU	71109
NL	31606
SW	20594
NZ	14777

Tweets by type

type	nb_tweets
media	879992
politics	540966

The number of tweets from New Zealand represents less than 6% of the tweets of UK

I noticed that we don’t have media accounts for New-Zealand, probably no data in the Oxford website

Summary by country / type

https://acss-covid.vercel.app/01_explore.html

Number of twitter accounts

country	nb_media	nb_politics
ES	15	36
IT	14	34
CA	14	34
PL	14	50
GB	13	50
US	13	16
DE	12	29
AU	12	35
FR	11	60
SW	11	25
NL	9	38

Number of tweets

country	nb_tweets_media	nb_tweets_politics
ES	171320	80849
GB	171130	87003
IT	117236	45162
US	104320	32551
CA	89110	37167
FR	82506	44616
DE	57126	42224
AU	55888	15221
PL	25833	94719
SW	4162	16432
NL	1361	30245

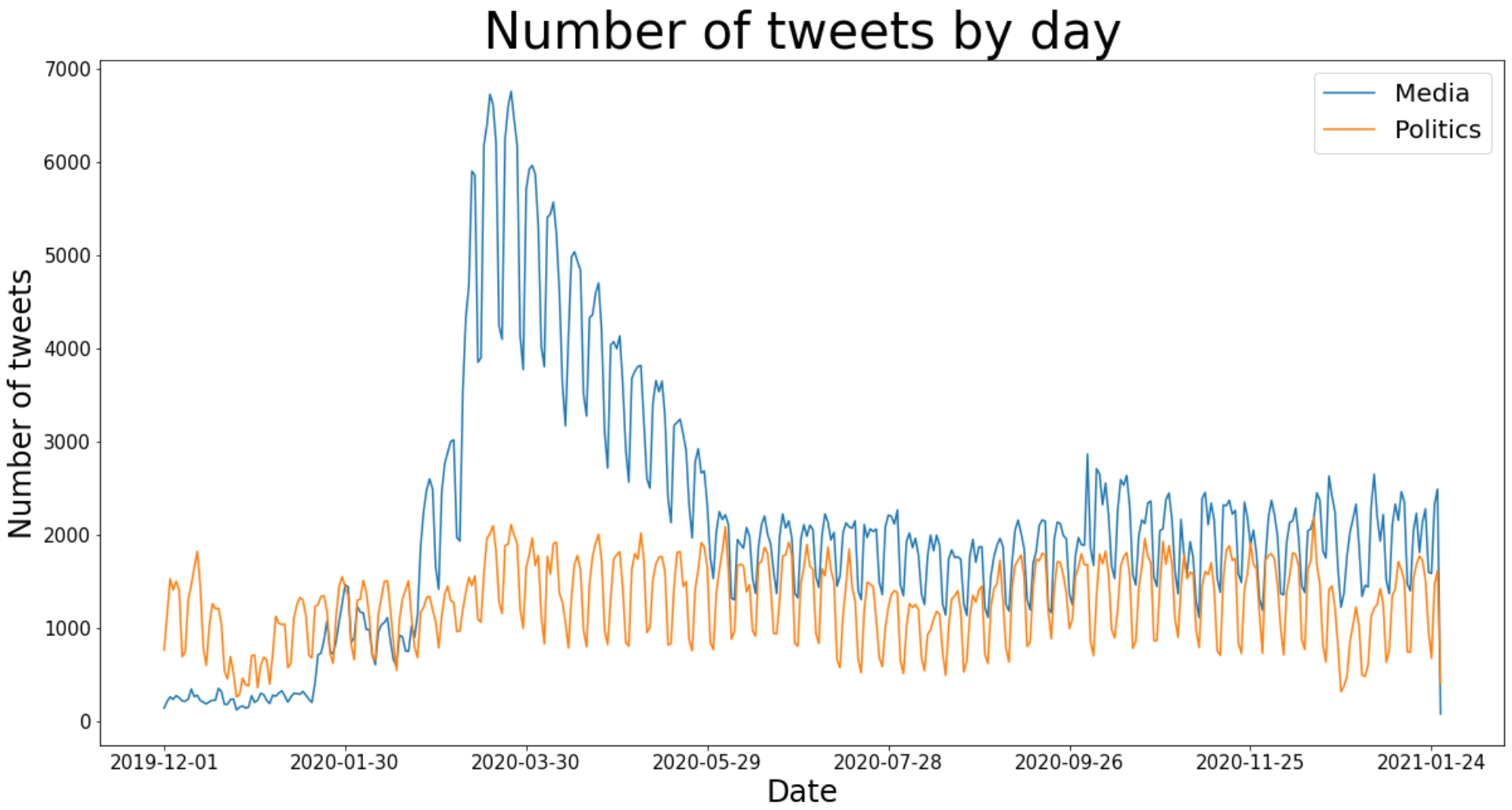
No media accounts for New-Zealand, so this row is dropped

Most active accounts

username	name	username_count	country	type
DailyMirror	Daily Mirror	32012	GB	media
Independent	The Independent	29259	GB	media
SkyNews	Sky News	28780	GB	media
repubblica	Repubblica	28285	IT	media
BFMTV	BFMTV	23408	FR	media
guardian	The Guardian	22645	GB	media
larazon_es	La Razón	22180	ES	media
CNN	CNN	18118	US	media
CTVNews	CTV News	18099	CA	media
20m	20minutos.es	17746	ES	media
ABC	ABC News	15487	US	media
eldiarioes	elDiario.es	15396	ES	media
TheSun	The Sun	14659	GB	media
thesnp	The SNP	14265	GB	politics
NBCNews	NBC News	14170	US	media
SkyTG24	Sky tg24	14125	IT	media
LaStampa	La Stampa	13684	IT	media
elperiodico	El Periódico	13305	ES	media
el_pais	EL PAÍS	12774	ES	media
franceinfo	franceinfo	12489	FR	media
konfederacja_	Konfederacja	12434	PL	politics
matteosalvinimi	Matteo Salvini	11538	IT	politics

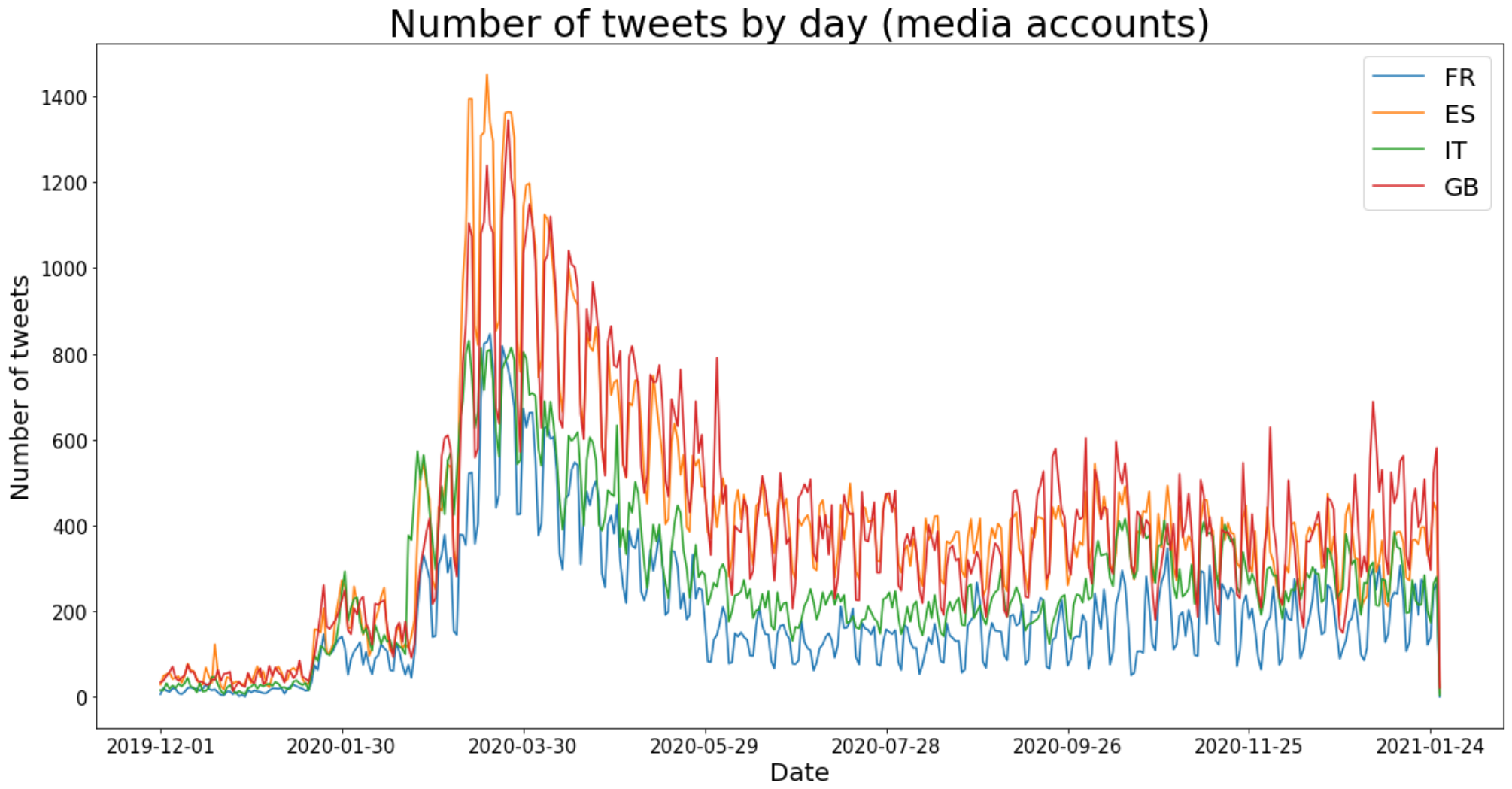
Number of tweets by day – type

https://acss-covid.vercel.app/02_trends.html

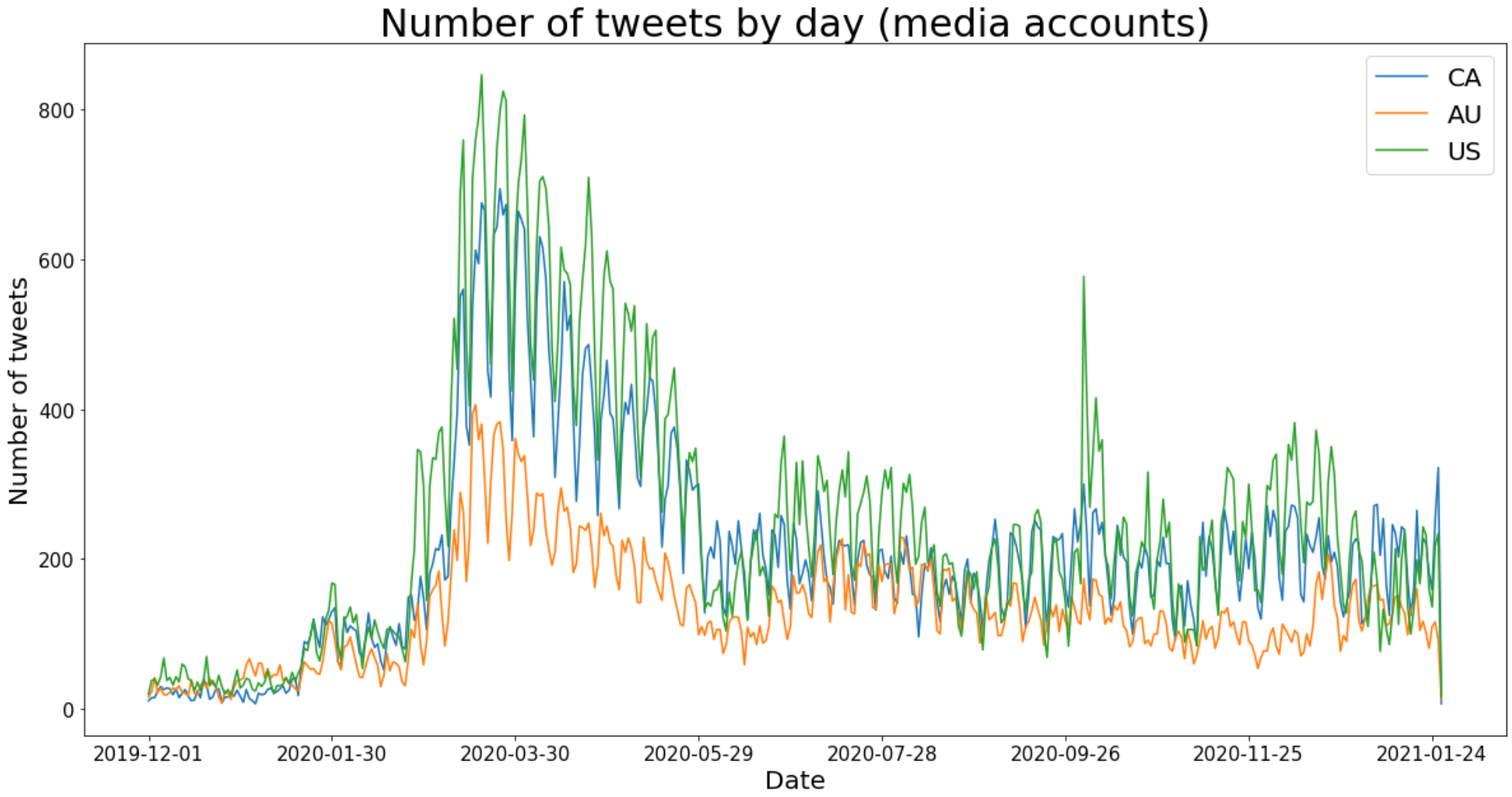


Media : number of tweets by day - country sample 1

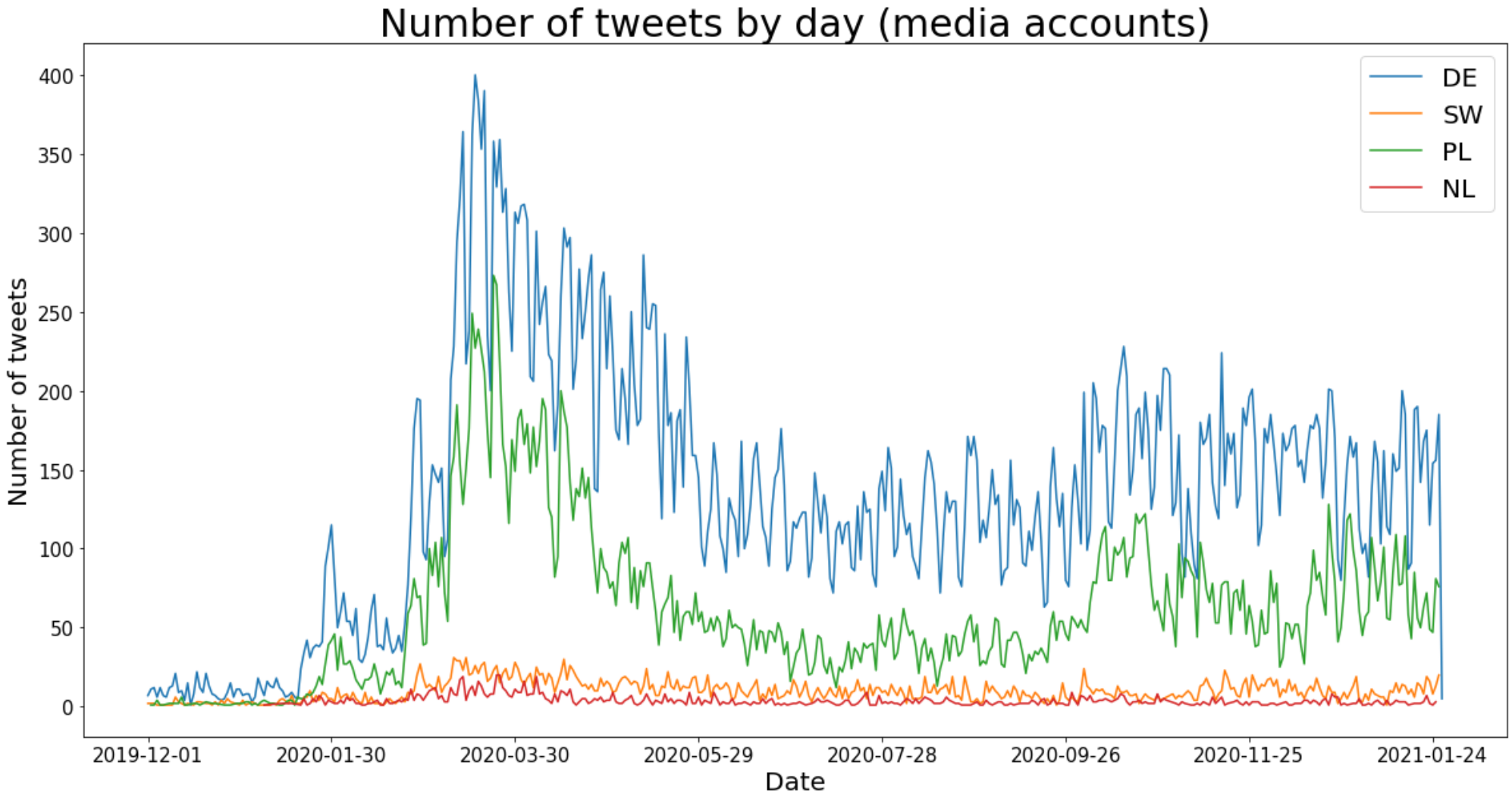
https://acss-covid.vercel.app/02_trends.html



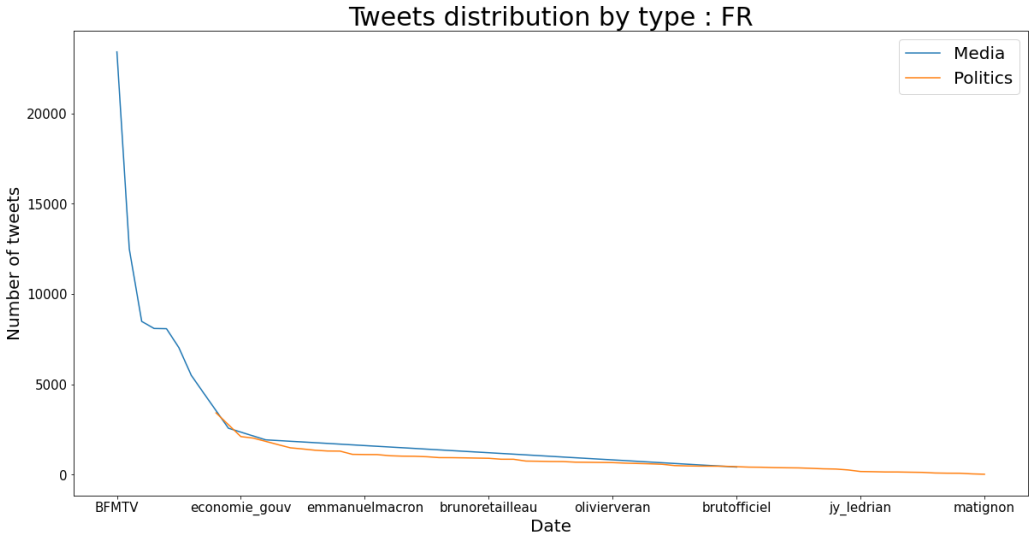
Media : number of tweets by day - country sample 2
https://acss-covid.vercel.app/02_trends.html

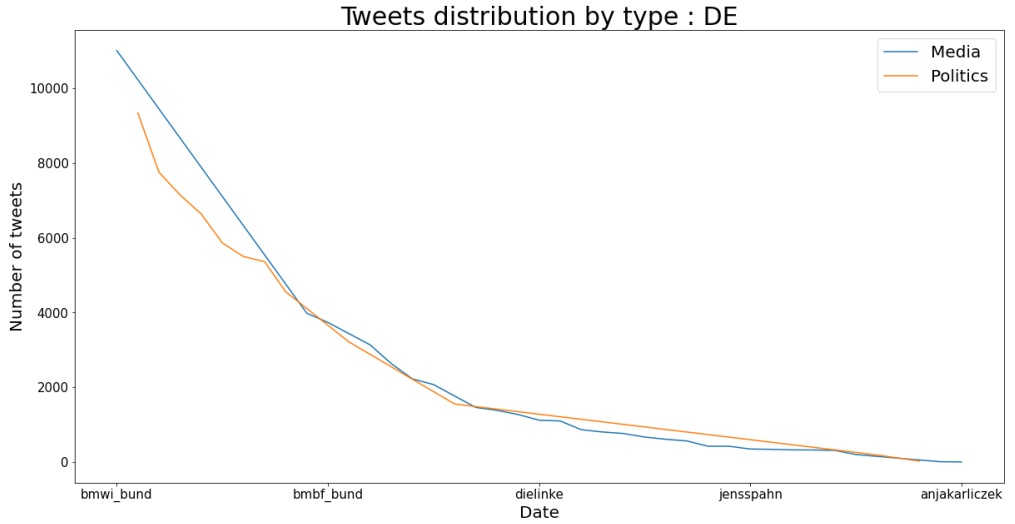


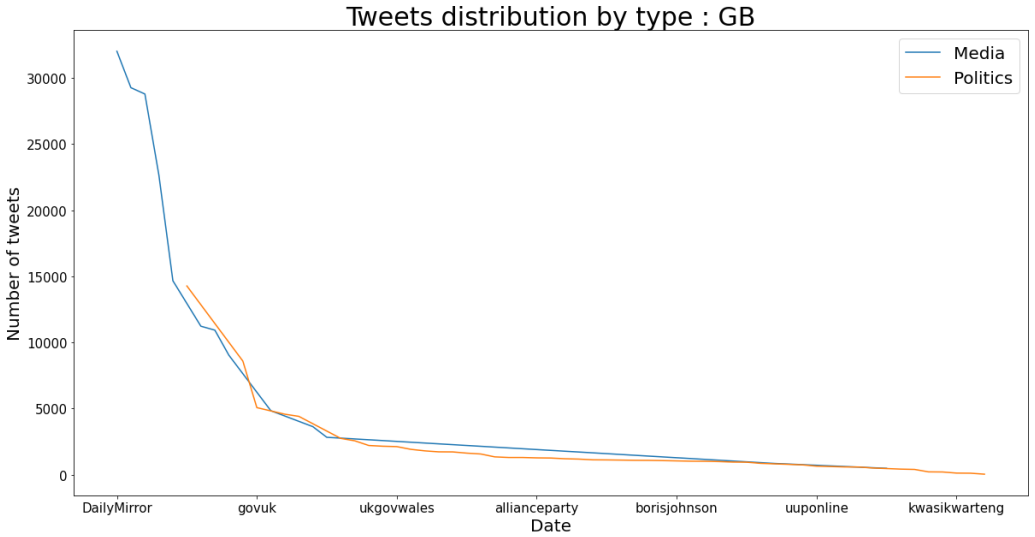
Media : number of tweets by day - country sample 3

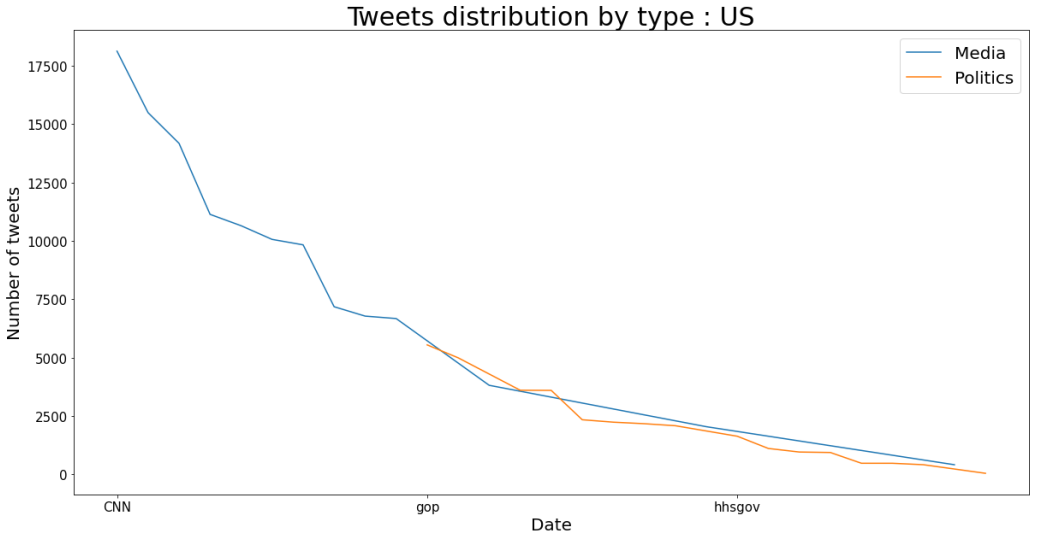


Tweets distribution by account type / country



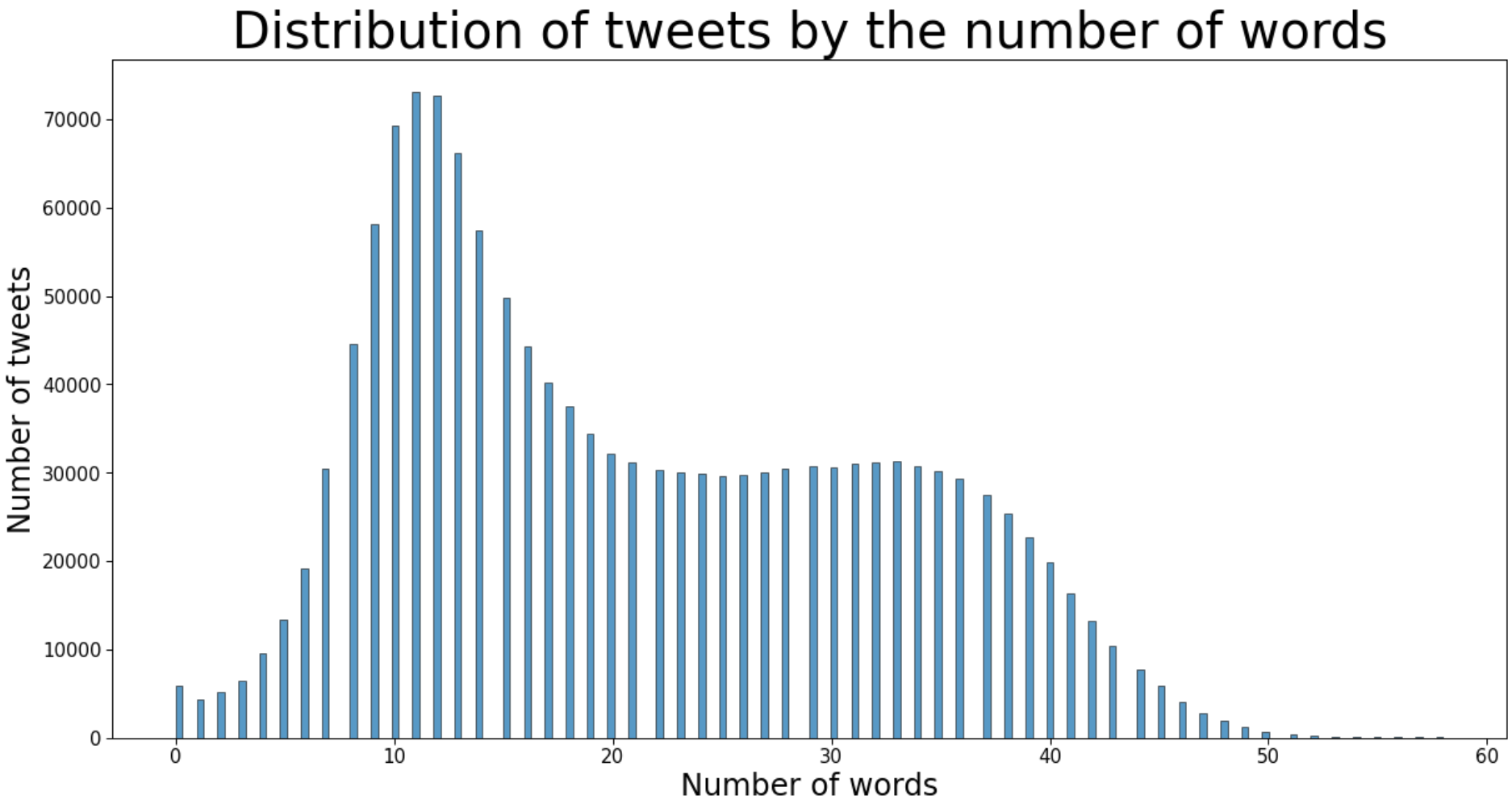






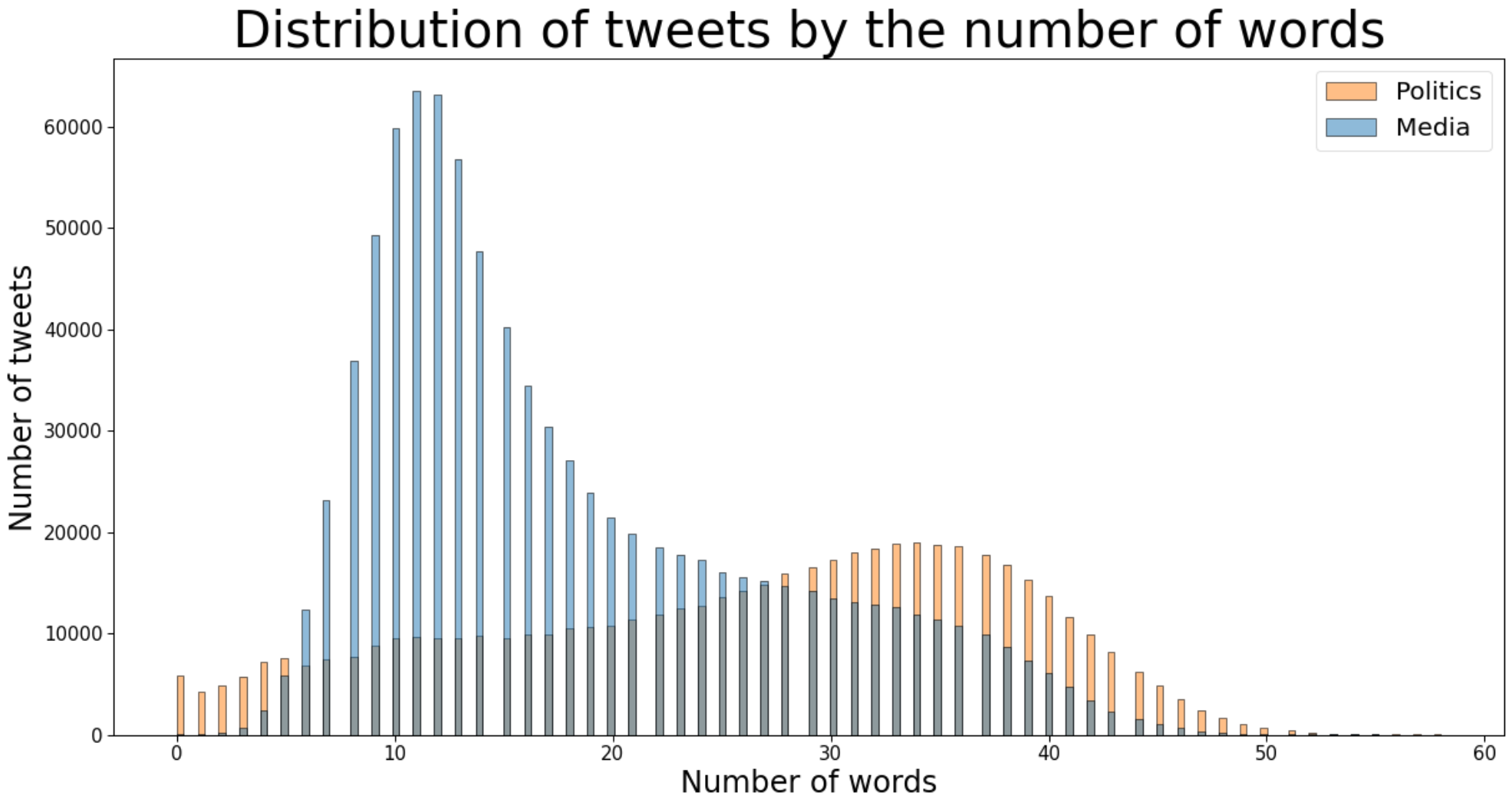
Distribution by words

https://acss-covid.vercel.app/03_nb_words.html



Distribution by words – account type

https://acss-covid.vercel.app/03_nb_words.html



Covariables for politics accounts

Adding co-variates is useful to improve the analysis. For example, in the context of STM (structural topic modeling).

Idea (Antoine?) : match the politics accounts with the parlgov database by their parties

- ***left_right*** : [0-10] → classifying parties by Ideological groupings left or right.
- ***family_name*** : [liberal, conservative, green, ...] → classifying parties by their philosophy.
- ***state_market*** : [0-10] → the more "market" the country is, the less it might be inclined to impose restrictions (low closure of shops, etc.)
- ***liberty_authority*** : [0-10] → the more the country has "liberty" the less it could be inclined to impose restrictions (lockdown, curfew, derogatory certificates, etc.)

Politics categories

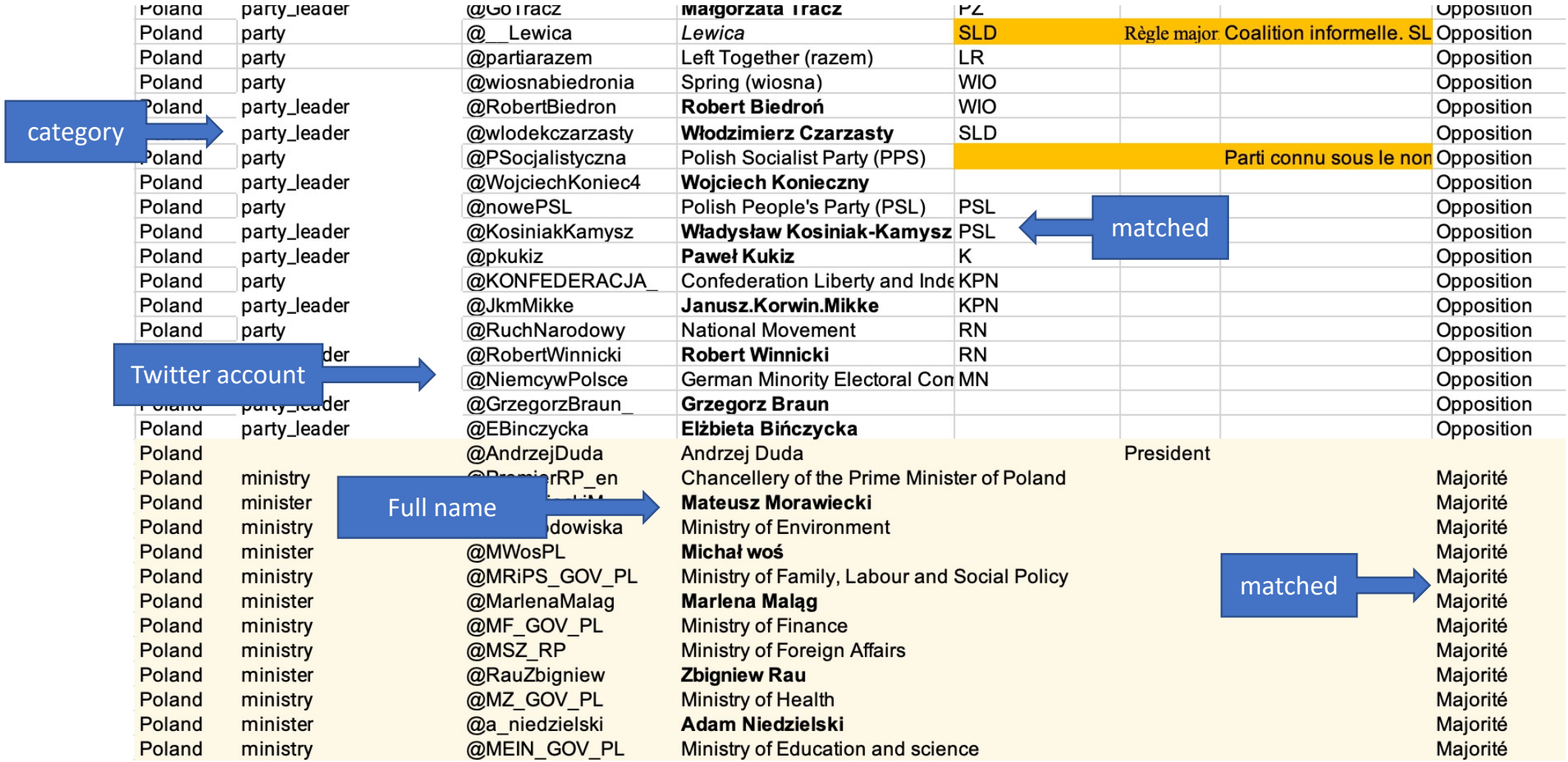
	category
party	131
party_leader	124
ministry	94
minister	87

country	category	count
Australia	minister	7
	ministry	9
	party	10
	party_leader	13
Canada	minister	8
	ministry	10
	party	9
	party_leader	9
France	minister	9
	ministry	9
	party	21
	party_leader	16
France	party	2
	party_leader	2
Germany	minister	7
	ministry	7
	party	7
	party_leader	9
Italy	minister	7
	ministry	7
	party	15
	party_leader	9
Netherlands	minister	6
	ministry	7
	party	13
	party_leader	12
New Zealand	minister	7
	ministry	5
	party	5
	party_leader	7
Poland	minister	10
	ministry	10
	party	15
	party_leader	16
Spain	minister	7
	ministry	7
	party	9
	party_leader	5
Sweden	minister	6
	ministry	6
	party	8
	party_leader	6
USA	minister	5
	ministry	5
	party	2
	party_leader	4
United Kingdom	minister	8
	ministry	12
	party	15
	party_leader	16

Issues (report by Abir and Antoine)

- USA does not exist in the Parlgov database, we tried to correct that but we didn't find any other complementary database ParlGov, that classifies parties on the same basis.
- Not all parties are in ParlGov, so Antoine helped in this step by working with the rule of majority, associating party groups of the parliaments.

Matching with the parlgov database



Poland	party_leader	@GoTracz	**Małgorzata Tracz**	PZ			Opposition
Poland	party	@__Lewica	*Lewica*	SLD	Règle major	Coalition informelle. SL	Opposition
Poland	party	@partiarazem	Left Together (razem)	LR			Opposition
Poland	party	@wiosnabiedronia	Spring (wiosna)	WIO			Opposition
Poland	party_leader	@RobertBiedron	**Robert Biedroń**	WIO			Opposition
Poland	party_leader	@wlodekczarzasty	**Włodzimierz Czarzasty**	SLD			Opposition
Poland	party	@PSocjalistyczna	Polish Socialist Party (PPS)			Parti connu sous le non	Opposition
Poland	party_leader	@WojciechKoniec4	**Wojciech Konieczny**				Opposition
Poland	party	@nowePSL	Polish People's Party (PSL)	PSL			Opposition
Poland	party_leader	@KosiniakKamysz	**Władysław Kosiniak-Kamysz**	PSL			Opposition
Poland	party_leader	@pkukiz	**Paweł Kukiz**	K			Opposition
Poland	party	@KONFEDERACJA_	Confederation Liberty and Inde	KPN			Opposition
Poland	party_leader	@JkmMikke	**Janusz.Korwin.Mikke**	KPN			Opposition
Poland	party	@RuchNarodowy	National Movement	RN			Opposition
Poland	party_leader	@RobertWinnicki	**Robert Winnicki**	RN			Opposition
Poland	party_leader	@NiemcywPolsce	German Minority Electoral Con	MN			Opposition
Poland	party_leader	@GrzegorzBraun_	**Grzegorz Braun**				Opposition
Poland	party_leader	@EBinczycka	**Elżbieta Bińczycka**				Opposition
Poland		@AndrzejDuda	Andrzej Duda		President		
Poland	ministry	@PremierRP_en	Chancellery of the Prime Minister of Poland				Majorité
Poland	minister	[illegible]	**Mateusz Morawiecki**				Majorité
Poland	ministry	[illegible]dowiska	Ministry of Environment				Majorité
Poland	minister	@MWosPL	**Michał woś**				Majorité
Poland	ministry	@MRiPS_GOV_PL	Ministry of Family, Labour and Social Policy				Majorité
Poland	minister	@MarlenaMalag	**Marlena Maląg**				Majorité
Poland	ministry	@MF_GOV_PL	Ministry of Finance				Majorité
Poland	ministry	@MSZ_RP	Ministry of Foreign Affairs				Majorité
Poland	minister	@RauZbigniew	**Zbigniew Rau**				Majorité
Poland	ministry	@MZ_GOV_PL	Ministry of Health				Majorité
Poland	minister	@a_niedzielski	**Adam Niedzielski**				Majorité
Poland	ministry	@MEIN_GOV_PL	Ministry of Education and science				Majorité

Data cleaning (round 1) – part 1

(for STM, other techniques may require a different cleaning)

- removed **links**, **emojis**, **mentions** and we **expanded the hashtags**. → *text_to_translate*
- removed **digits** and **punctuations**. → *text_bag_words*
- *number_words* : count number of words with size > 1 in *text_to_translate_no_digit*.
- the detected language by Twitter, in some cases it is **und**, 1.7% of the tweets (and that's when the tweets are composed of emojis, only links, attached words in form of hashtags, tweets with different languages and in case where twitter is unable to define the dominant language.) so these tweets are to delete
- We decided to translate tweets composed of more than 10 words and their language is not und:
- *to_keep* : 0 (don't translate) : <=10 words OR lang = und ; 1 (translate) : else

Data cleaning (round 1) – part 2

(for STM, other techniques may require a different cleaning)

- Removed all the noise tokens (parts of speech), usual stopwords and some additional: ['covid','covid19','corona','coronavirus','virus','sars’]
  (these words will appear in many topics and make their identification more difficult)
- Replaced British English words with American English (e.g. organization to organization)
- Created bi-grams (probabilistic with the python library Gensim)

Data cleaning (round 2)

Issues to be addressed in round 2:

- Twitter does a poor job in detecting the language. I noticed (ex-post) that quite many tweets with **und** can be translated and used.
- Translate tweets with less than 10 “useful” words ?
  Topic modeling techniques require at least a few words and translation is poor with not sufficient context since they use context to identify the semantics.
- Use the new Spacy feature to detect the languages to make additional checks

The data I send you

Right now, I send you the full data with minimal cleaning/translations for 2 reasons :

1. The right cleaning depends on the techniques to be used (*e.g. bag of words v. word embedding techniques*)
2. Cleaning means removing some tweets but we should first consider the quality of the samples. Very bad practice to change the sample after cleaning.

Issues to consider/discuss for a dream database

- USA not in the parlgov database. Do we need to use with it ? Do we have better alternatives ?
- No media for NZ
- Too few tweets for media in NL and SW ?
- Is our sample sufficiently balanced (both for politics and media) ?
  I would like someone to look in detail at the list of accounts to get a double checking
- Should we extend the initial period ? (quite costly)

I am open to any collaboration. This database has the potential to be great but it needs a team really invested in it.