Tracing conceptual change in messy data (2): self-reliance as boon and bane

© Joris van Eijnatten

Universiteit Utrecht

Faculty of Humanities

My question

- which popular concepts of “Europe” do newspapers reveal?

OR

- has football always been more important than Brussels?

AND

- how can we trace concepts of “Europe” in messily digitized newspapers?

My question

Three approaches:

1. The mundane
   - e.g. crossword puzzels, weather forecasts
2. Visions of/for the future
   - e.g. political ideals, philosophical views
3. Popular competition
   - e.g. Miss Europe, Eurovision Song Contest, Football

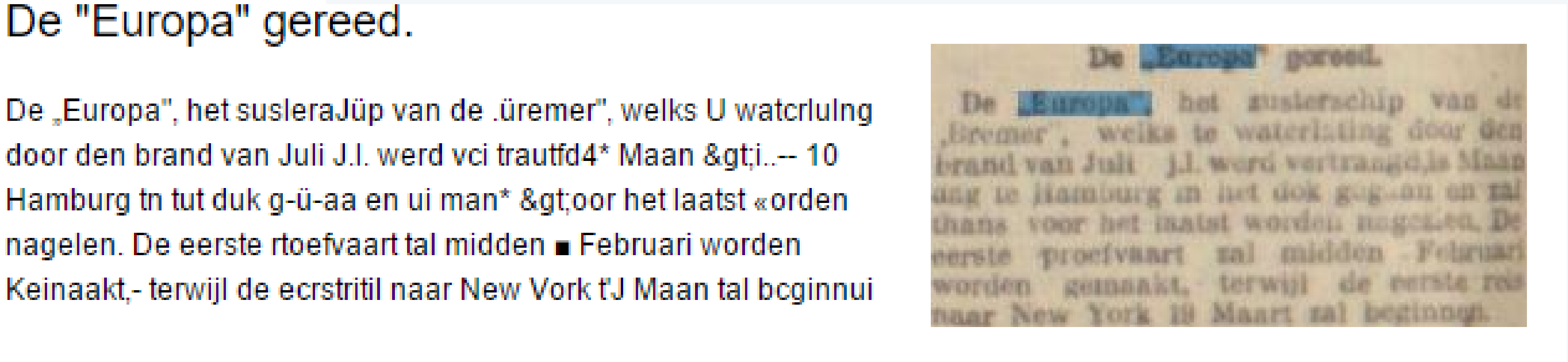
What I need

- A handy toolbox
  - to trace conceptual change
  - in ± big data
  - of not so very good quality
  - over a longer period of time
  - in more than one language

- comparative analysis over time and space

Data

- “messy data”



De "Europa" gereed.

De „Europa", het susleraJüp van de .üremer", welks U watcrlulng door den brand van Juli J.l. werd vci trautfd4* Maan &gt;i..-- 10 Hamburg tn tut duk g-ü-aa en ui man* &gt;oor het laatst «orden nagelen. De eerste rtoefvaart tal midden ■ Februari worden Keinaakt,- terwijl de ecrstritil naar New Vork t'J Maan tal bcginnui

De „Europa" gereed.

De „Europa", het zusterschip van de
„Bremer", welks te waterlating door den
brand van Juli j.l. werd vertraagd, is Maan
ing te Hamburg in het dok gegaan en zal
thans voor het laatst worden nagezien. De
eerste proefvaart zal midden Februari
worden gemaakt, terwijl de eerste reis
naar New York 19 Maart zal beginnen.

susleraJüp = <zusterschip>
.üremer“ = <“Bremer”>
U watcrlulng = <te waterlating>
vci trautfd4* = <vertraagd is>

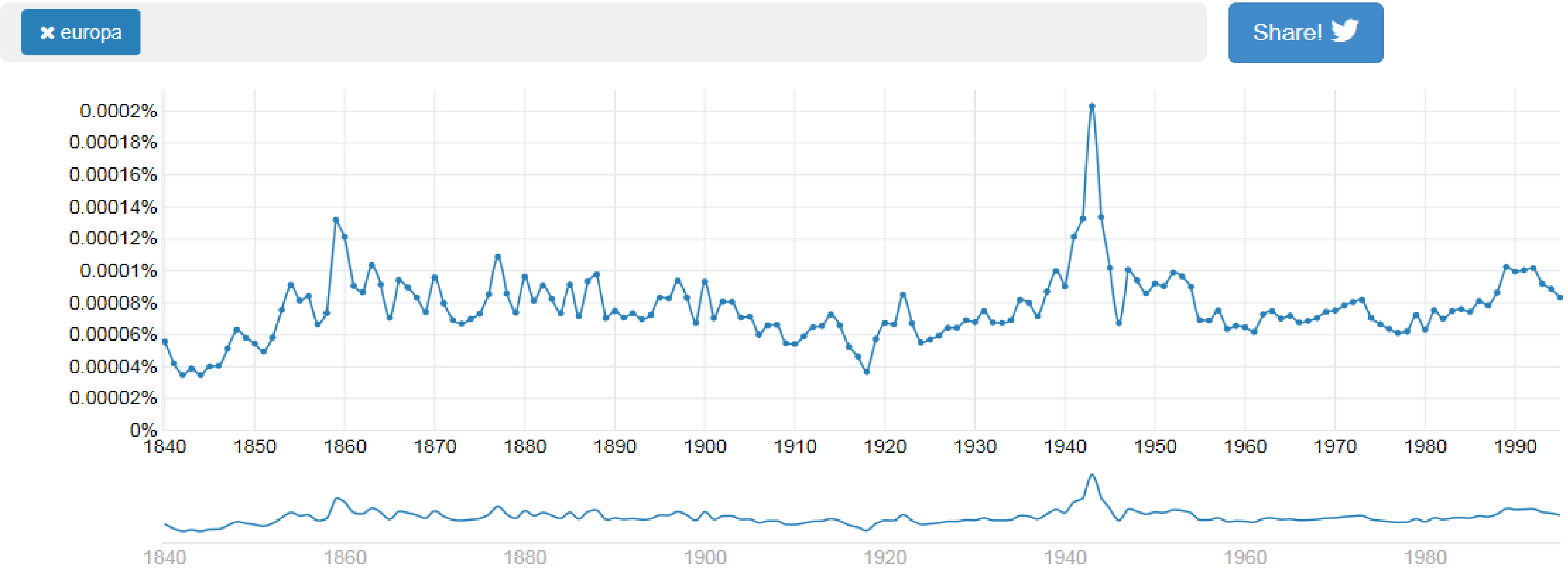
Rotterdamsch nieuwsblad, 08-01-1930

Toolbox, January 2016

- which accessible & robust tools do we actually have?

nGrams (e.g. Delpher)

> insightful but too simple & too rigid

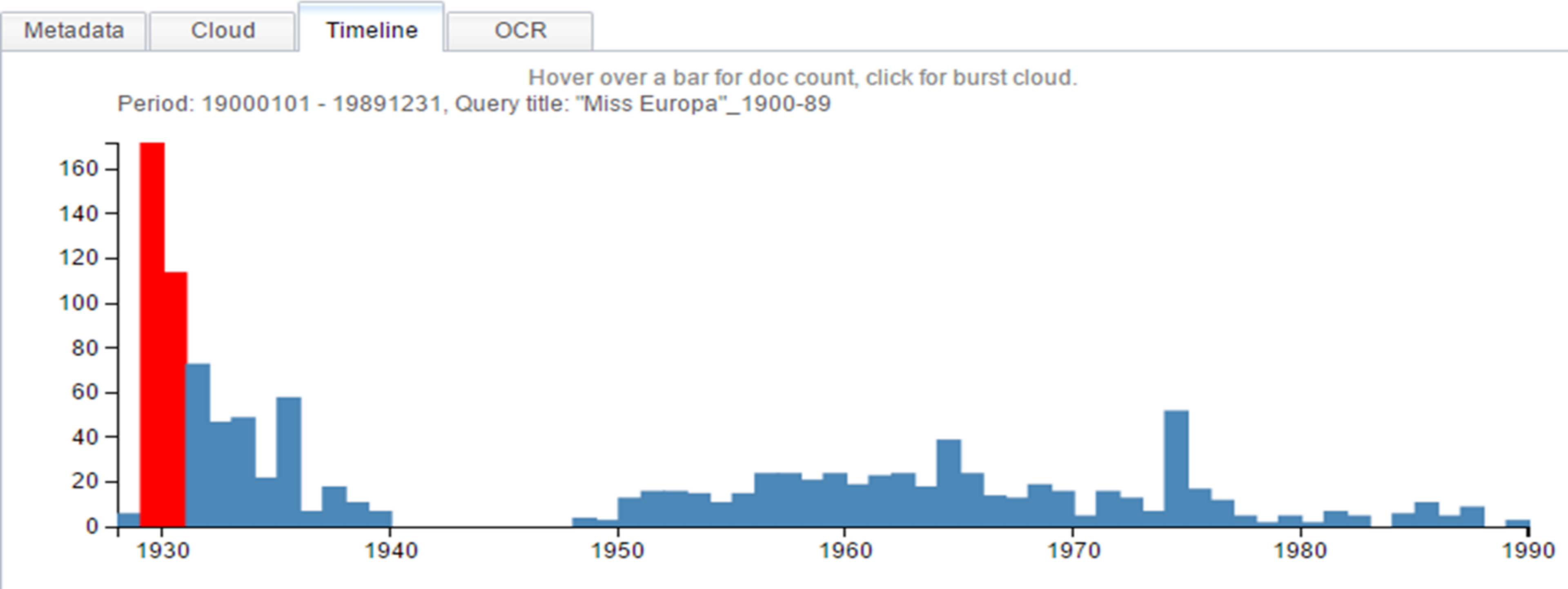


Toolbox, January 2016

- which accessible & robust tools do we actually have?

semantic text-mining (e.g. Texcavator)

> temporal dimension, proven, but restricted to KB corpus



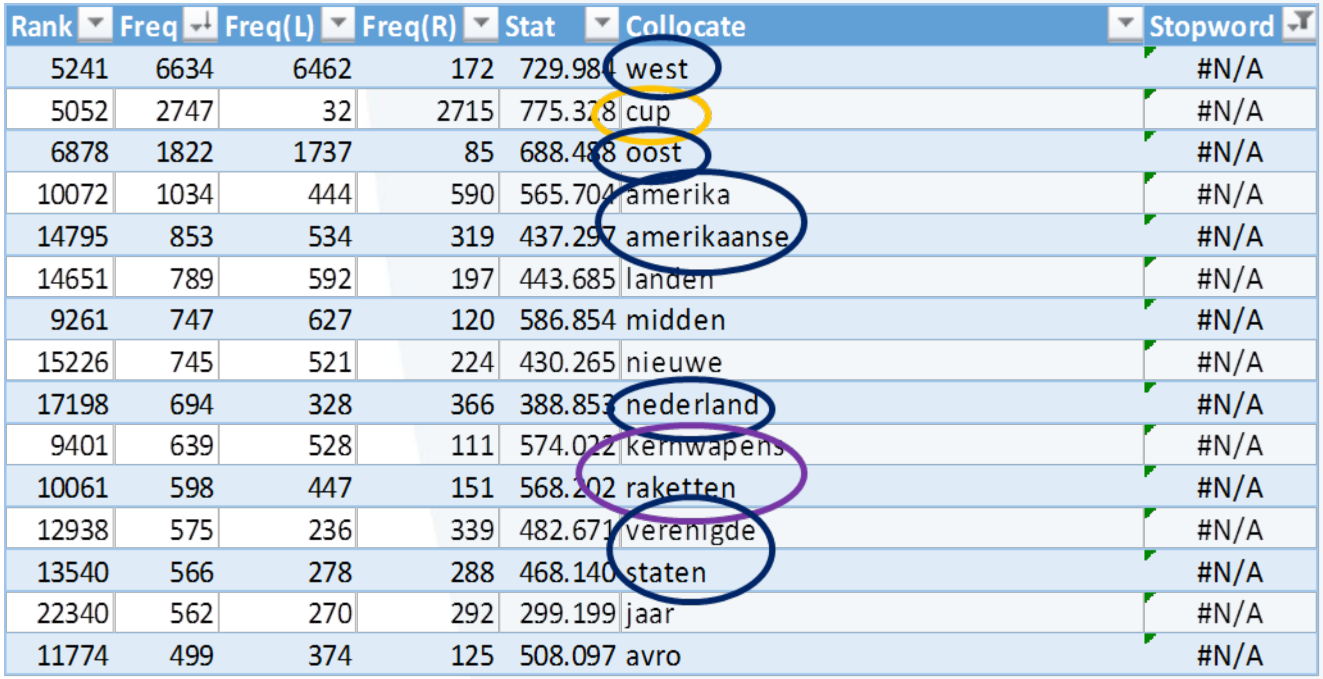
Hover over a bar for doc count, click for burst cloud.

Period: 19000101 - 19891231, Query title: "Miss Europa"_1900-89

Toolbox, January 2016

- which accessible & robust tools do we actually have?

corpus linguistics (e.g. Antconc)
> proven, but no temporal dimension



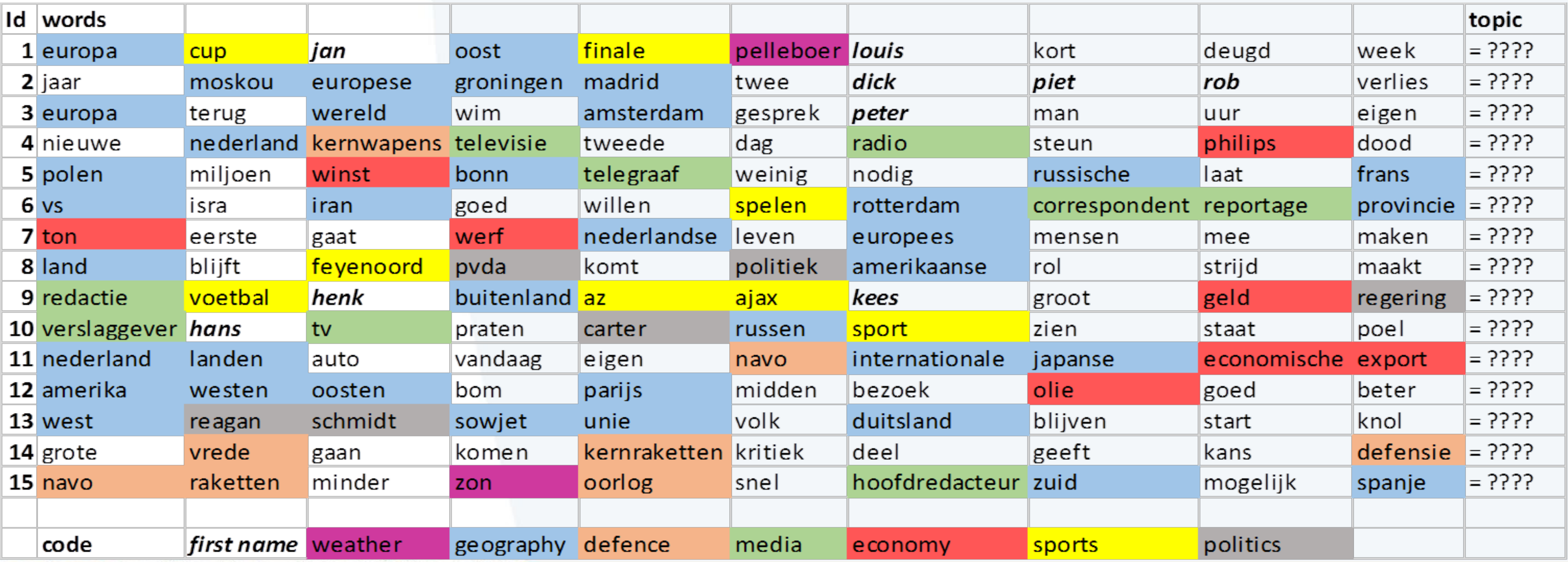
Rank	Freq	Freq(L)	Freq(R)	Stat	Collocate	Stopword
5241	6634	6462	172	729.984	west	#N/A
5052	2747	32	2715	775.328	cup	#N/A
6878	1822	1737	85	688.488	oost	#N/A
10072	1034	444	590	565.704	amerika	#N/A
14795	853	534	319	437.297	amerikaanse	#N/A
14651	789	592	197	443.685	landen	#N/A
9261	747	627	120	586.854	midden	#N/A
15226	745	521	224	430.265	nieuwe	#N/A
17198	694	328	366	388.853	nederland	#N/A
9401	639	528	111	574.022	kernwapens	#N/A
10061	598	447	151	568.202	raketten	#N/A
12938	575	236	339	482.671	verenigde	#N/A
13540	566	278	288	468.140	staten	#N/A
22340	562	270	292	299.199	jaar	#N/A
11774	499	374	125	508.097	avro	#N/A

Toolbox, January 2016

- which accessible & robust tools do we actually have?

topic modelling (e.g. Mallet)

> proven, but no temporal dimension



Id	words										topic
1	europa	cup	*jan*	oost	finale	pelleboer	*louis*	kort	deugd	week	= ????
2	jaar	moskou	europese	groningen	madrid	twee	*dick*	*piet*	*rob*	verlies	= ????
3	europa	terug	wereld	wim	amsterdam	gesprek	*peter*	man	uur	eigen	= ????
4	nieuwe	nederland	kernwapens	televisie	tweede	dag	radio	steun	philips	dood	= ????
5	polen	miljoen	winst	bonn	telegraaf	weinig	nodig	russische	laat	frans	= ????
6	vs	isra	iran	goed	willen	spelen	rotterdam	correspondent	reportage	provincie	= ????
7	ton	eerste	gaat	werf	nederlandse	leven	europees	mensen	mee	maken	= ????
8	land	blijft	feyenoord	pvda	komt	politiek	amerikaanse	rol	strijd	maakt	= ????
9	redactie	voetbal	*henk*	buitenland	az	ajax	*kees*	groot	geld	regering	= ????
10	verslaggever	*hans*	tv	praten	carter	russen	sport	zien	staat	poel	= ????
11	nederland	landen	auto	vandaag	eigen	navo	internationale	japanse	economische	export	= ????
12	amerika	westen	oosten	bom	parijs	midden	bezoek	olie	goed	beter	= ????
13	west	reagan	schmidt	sowjet	unie	volk	duitsland	blijven	start	knol	= ????
14	grote	vrede	gaan	komen	kernraketten	kritiek	deel	geeft	kans	defensie	= ????
15	navo	raketten	minder	zon	oorlog	snel	hoofdredacteur	zuid	mogelijk	spanje	= ????
	code	*first name*	weather	geography	defence	media	economy	sports	politics		

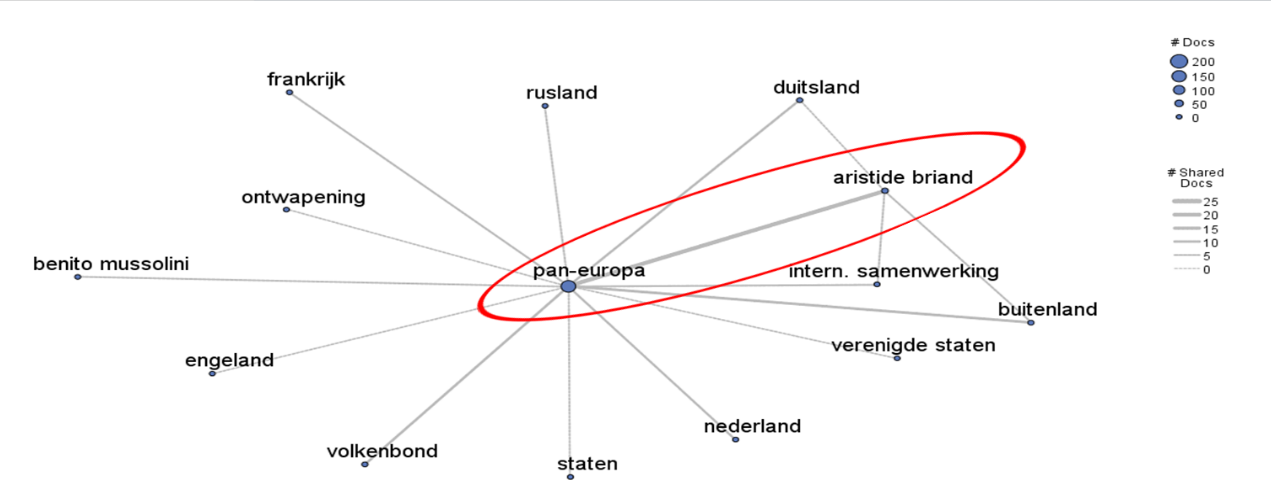
Toolbox, January 2016

- which accessible & robust tools do we actually have?

text analytics package (e.g. SPSS Modeler)

> proven, but no temporal dimension & black box

Category web of <pan-Europa>, article titles, N = 42,712 docs
Dutch territorial newspapers 1930-31

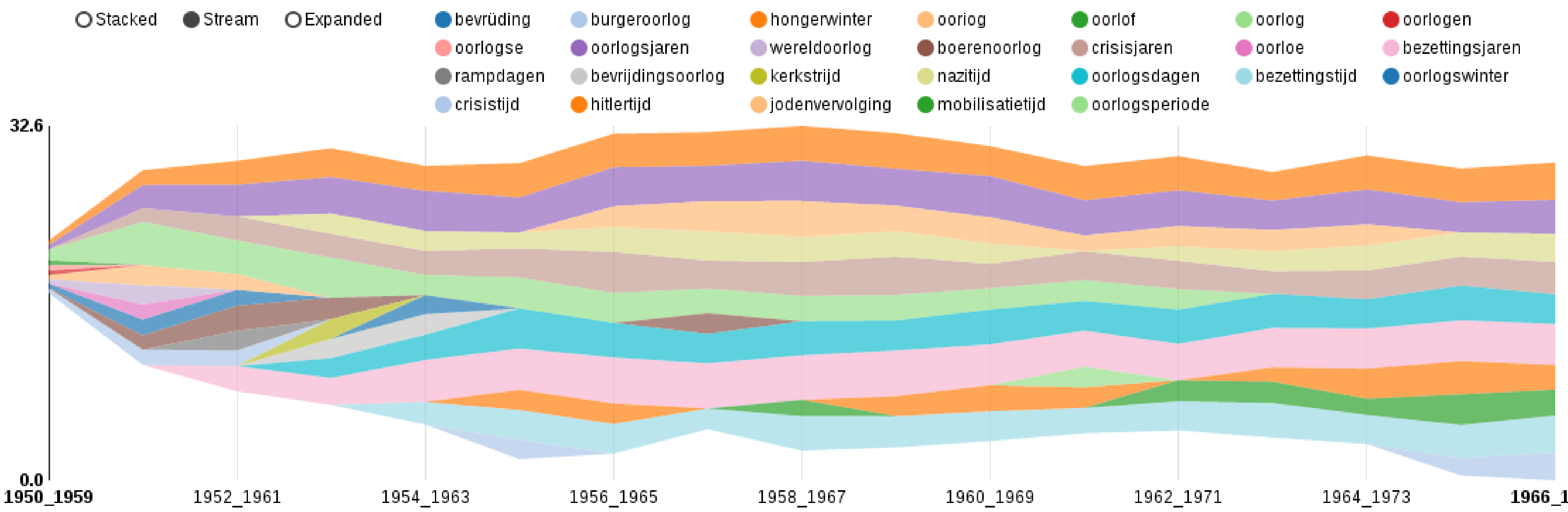


Toolbox, January 2016

- which accessible & robust tools do we actually have?

vector-space modelling (e.g. ShiCo)

> temporal dimension, unproven & black box



Case: analysing the mundane

assignment

- “determine the frequency of locations mentioned in weather forecasts and plot them on a dynamic, time-based graph”

hypothesis

- “weather forecasts offer insight into the geographical dimensions of a popular concept of Europe”

Weather forecasts

WEER IN EUROPA
Door het zich naar Zuid west-lerland terugtrekken van een hoge-drttkgebied stroomt minder koude lucht van de Oceaan West- en Midden-Europa binnen. Koud blijft het nog in het zuiden, met temperaturen in Spanje en aan de lUvlèra van vijf tot tien graden. Ook in het Alpengebied blijft het nog winters met enige sneeuwval, maar later zal de zachtere lucht verder naar Midden-Europa doordringen.

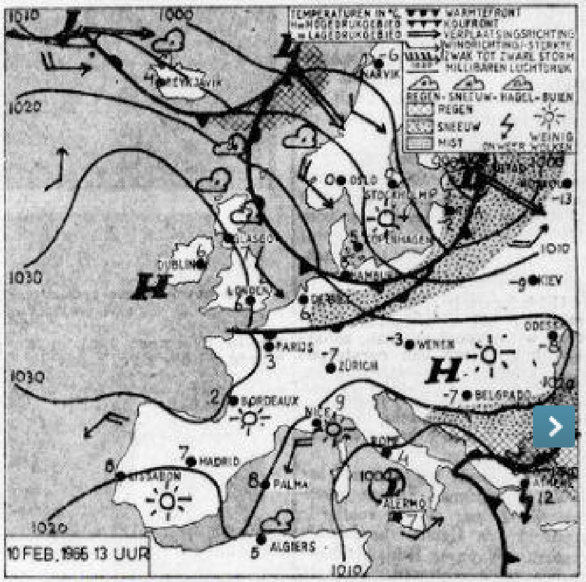
Zachter dan normaal

Van onze weerkundige correspondent.
Amsterdam, donderdag.

Er is vrijwel geen kans, dat in de eerstkomende dagen opnieuw koude polaire luchtmassa's naar Nederland zullen stromen, zoals in het begin van deze week het geval was. We zijn nu in een situatie terechtgekomen, waarbij de wind in hoofdzaak uit west tot noordwest blijft waaien

en dit betekent, dat het weer z
ter dan normaal voor deze tijd
het jaar zal zijn. Overdag temp
turen rond 6 graden, 's nachts g
vorst.

In de noordwestelijke stromin
komen kleine storingen mee. D
om zal er vandaag over het a
meen weer veel bewolking zijn
enkele meest lichte buien.



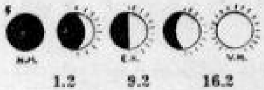
10 FEB. 1966 13 UUR

WEER IN EUROPA
Door het zich naar Zuidwest-Ierland terugtrekken van een hoge-drukgebied stroomt minder koude lucht van de Oceaan West- en Midden-Europa binnen.
Koud blijft het nog in het zuiden, met temperaturen in Spanje en aan de Rivièra van vijf tot tien graden. Ook in het Alpengebied blijft het nog winters met enige sneeuwval, maar later zal de zachtere lucht verder naar Midden-Europa doordringen.

HET WEER VAN GISTEREN

Station	weer	Max. temp.	Neerslag 7-19 uur
Den Helder	zw. bew.	6	0
Ypenburg	geh. bew.	6	0
Vlissingen	geh. bew.	6	0.1
Eelde	zw. bew.	7	0.3
De Bilt	l. regen	6	1
Twente	regen	6	0.1
Eindhoven	regenbui	6	0.3
Z.-Limburg	mist	3	1
Helsinki	onbewolkt	—3	0
Stockholm	l. bew.	2	0
Oslo	l. bew.	3	0
Kopenhagen	onbewolkt	6	0
Aberdeen	zw. bew.	8	0
Londen	geh. bew.	7	0
Amsterdam	regenbui	6	0.1
Brussel	motregen	5	2
Luxemburg	sneeuw	—	—
Parijs	geh. bew.	3	0
Bordeaux	l. bew.	6	0
Grenoble	onbewolkt	1	0
Nice	onbewolkt	9	0
Berlijn	zw. bew.	2	1
Frankfort	motsneeuw	—1	0.1
München	geh. bew.	—3	0
Zürich	geh. bew.	—3	0
Genève	onbewolkt	0	0
Locarno	onbewolkt	9	0
Wenen	geh. bew.	—2	0
Innsbruck	geh. bew.	—3	0
Belgrado	onbewolkt	—5	0
Rome	onbewolkt	6	0
Barcelona	onbewolkt	8	0
Madrid	onbewolkt	8	0
Majorca	onbewolkt	8	0
Lissabon	onbewolkt	11	0

VRIJDAG:
Zon op 8.01, onder 17.48. Maan onder 5.33, op 13.11. Voertuigen licht op van 18.03 tot 7.44.



Problem

no immediately available tool could help me to perform this simple assignment

so I tried a little
self-reliance

Weather forecasts: methodology

- at some point in your research, write down your methodology
- it helps you to better understand what you have been doing

- weather forecasts required a 4-step methodology

Methodology: Step 1

1. Extract dataset of weather forecasts
    1. Extract per decade from the KB terrabyte all articles and advertisements published between 1880 and 1990 containing the word “temperatu*”
    2. Reduce the dataset to records that are very probably weather forecasts, using the query “isstartstring(article_dc_title,"weerber"), or isstartstring(article_dc_title,"weerkundig"), or hasmidstring(article_dc_title,"ivierberi"), or hasmidstring(article_dc_title,"emperatu"), or hasmidstring(article_dc_title,"eerberi"), or hasmidstring(article_dc_title,"eerkun"), or hasmidstring(article_dc_title,"eerover"), or hasmidstring(article_dc_title,"eersge"), or hasmidstring(article_dc_title,"eerstoe"), or hasmidstring(article_dc_title,"eerverw"), or hasmidstring(article_dc_title,"meteo"), or hasmidstring(text_content,"weerb"), or hasmidstring(text_content,"weerk"), or hasmidstring(text_content,"weerover"), or hasmidstring(text_content,"weersge"), or hasmidstring(text_content,"weerstoe"), or hasmidstring(text_content,"weerverw"), or hasmidstring(text_content,"meteo"), or hasmidstring(text_content,"vorst"), or hasmidstring(text_content,"bewolk"), or hasmidstring(text_content,"buie") , or hasmidstring(text_content,"dooi") , or hasmidstring(text_content,"droog"), or hasmidstring(text_content,"graden"), or hasmidstring(text_content,"hitte"), or hasmidstring(text_content,"visser"), or hasmidstring(text_content,"koud"), or hasmidstring(text_content,"neerslag"), or hasmidstring(text_content,"onweer"), or hasmidstring(text_content,"opklaring"), or hasmidstring(text_content,"regen"), or hasmidstring(text_content,"sneeuw"), or hasmidstring(text_content,"storm"), or hasmidstring(text_content,"vries"), or hasmidstring(text_content,"warm"), or hasmidstring(text_content,"wind"), or hasmidstring(text_content,"wolke"), or hasmidstring(text_content,"zonn"), or hasmidstring(text_content,"Bilt"), or hasmidstring(text_content,"Bildt"), or hasmidstring(text_content,"storm"), or hasmidstring(text_content,"uchtdruk"), or hasmidstring(text_content,"arometer"), or hasmidstring(text_content,"Barom"), or hasmidstring(text_content,"depress")”
    3. Further reduce the dataset to records that are exclusively weather forecasts
    4. Divide dataset into subsets based on newspaper titles (different newspapers have different OCRs)

Methodology: Step 2

2. Get a list of place names and further refine dataset
    5. Differentiate between
        - towns/cities: Leuven, Sint Job in ’t Goor, Silly, Corny
        - countries: Ireland, Germany, Lichtenstein
        - regions within a country: Friesland, Bavaria
        - regions across countries: Lapland, the Alps
    6. Generate a basic list of place names to further refine the dataset
    7. Remove all records that mention only De Bilt (“d[\s\S]*?bilt”)
    8. Get list of names for each category (see 5):
        - e.g. github.com/David-Haim /CountriesToCities**JSON**

Methodology: Step 3

3. Determine frequencies of place names in datasets

    9. Create a list of name variants using regular expressions
        - Zürich = z[uüe]{1,2}rich|[uüe]{1,2}rich|zÃ¼rich|z[lt]irich|zurieh
        - West Germany = west[\s\-]?deuts.{1,5}d|west[\-\s]?du[\s]?it[s\s]{0,2}land
    10. Determine frequencies of place names in data set
    11. Use and augment a list of geographical stopwords (= false place names, n= 3,135)
        - “alexander”, “hjo”, “zomergem”
    12. Use and augment list of mistaken identities
        - “middelburg; Belgium”, “los angeles; Spain”, “china; Russia”
    13. Repeat 10 through 12

Methodology: Step 4

4. Normalize and plot on map
    14. Normalize frequencies of place names over the total number of records per dataset
    15. Obtain coordinates for place names (lat & long)
    16. Plot place names and frequencies on map
    17. Take a very long vacation

Methodology: which tools?

1. Texcavator
2. SPSS Modeler
3. Excel (manual)
4. SPSS Modeler
5. Excel
6. Python 3 script
7. Python 3 script
8. Browser
9. Regex editor
10. Python 3 script
11. Python 3 script
12. Python 3 script
13. Python 3 script
14. Python 3 script
15. Hamster maps
16. Carto DB / Google Fusion Tables
17. EasyJet

```
import pandas as pd
import numpy as np

# get input
name_file = pd.read_csv('OriginalNames_Regex.csv', sep=(';'), names=['Concept', 'Concept+', 'Category',
'Country', 'Regex', 'Latitude', 'Longitude'], encoding='ISO-8859-1')
name_pat = (name_file['Regex'])

prompt1 = input('Please insert the name of the TRUE data file you want to process \n')
prompt2 = input('Please insert the name of your RAW output frequency file \n')
weather_file = pd.read_csv(prompt1, sep=(';'), names=('_id', 'paper_dc_date', 'paper_dc_language',
'paper_dc_title', 'paper_dc_publisher', 'paper_dcterms_issued','paper_dcterms_spatial',
'paper_dcterms_temporal', 'paper_dcx_issuenumber','article_dc_subject','article_dc_title', 'text_content'),
encoding='ISO-8859-1')

#generate frequencies of regex names
content = pd.Series(weather_file['text_content'])
name_dict = {}
for name in name_pat:
    name_dict[name] = 0
result_list = []
for item in name_dict:
    result = content.str.contains(item, case=False, regex=True).sum()
    if result == 0:
        pass
    elif result > 0:
        result_list.append([item, result])
```

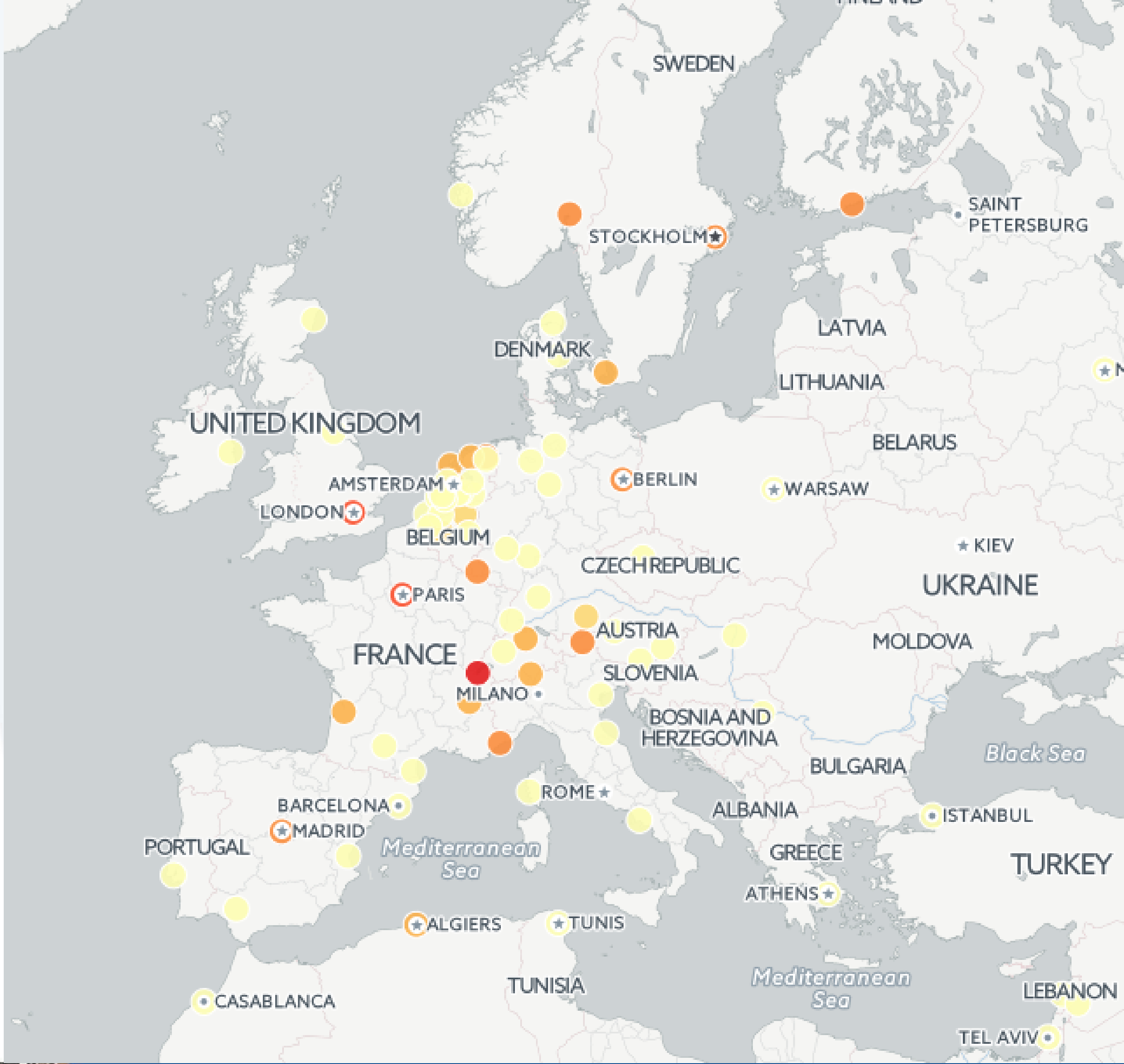
city/country	category	latitude	longitude	frequency50-59	norm_50-59
amsterdam; Netherlands	Town	52.370.215	4.895.167	788	690.018
geneva; Switzerland	Town	46.204.390	6.143.157	554	485.114
de bilt; Netherlands	Town	52.109.271	5.180.967	185	161.996
paris; France	Town	48.856.614	2.352.221	135	118.214
london; United Kingdom	Town	51.507.350	-0.127758	130	113.835
oslo; Norway	Town	59.913.868.800	10.752.245.400	117	102.452
groningen; Netherlands	Town	53.219.383	6.566.501	113	98.949
berlin; Germany	Town	52.520.006	13.404.953	105	91.944
stockholm; Sweden	Town	59.329.323.500	18.068.580.800	93	81.436
helsinki; Finland	Town	60.169.855.700	24.938.379.000	91	79.685
luxembourg; Luxembourg	Town	49.611.621	6.131.934	90	78.809
innsbruck; Austria	Town	47.269.212.400	11.404.102.400	75	65.674
madrid; Spain	Town	40.416.775.400	-3.703.790.200	75	65.674
nice; France	Town	43.710.172.800	7.261.953.200	75	65.674
zurich; Switzerland	Town	47.376.886	8.541.694	60	52.539
copenhagen; Denmark	Town	55.676.096.800	12.568.337.100	59	51.664
den helder; Netherlands	Town	52.956.280	4.760.797	59	51.664
locarno; Switzerland	Town	46.166.998	8.794.264	57	49.912
algiers; Algeria	Town	36.753.770	3.058.792	57	49.912
leeuwarden; Netherlands	Town	53.201.233.400	5.799.913.300	56	49.037
grenoble; France	Town	45.188.529	5.724.523	53	46.41
bordeaux; France	Town	44.837.789	-0.579179	48	42.032
eindhoven; Netherlands	Town	51.441.641	5.469.722	43	37.653
brussels; Belgium	Town	50.850.339	4.351.710	42	36.778

Normalised frequency of towns in weather forecasts

De Telegraaf

1950-1959

n = 1,142



Lessons learnt, September 2016

self-reliance: boon or bane?

PROS or CONS

- programming is pleasant (great fun!)
- method is mandatory (write up your stuff)
- numbers are needed (you can’t avoid statistics)
- autonomy is arduous (time, time, time)
- halting is hard (where or when to draw the line?)

Lessons learnt, September 2016

there isn’t really very much out there (yet)

- the available tools are great but they are only a first step

remain in control

- insight into tools is crucial but involves blood, sweat, tears, pain, suffering, anguish, despair, depression... and time

help is expensive

- if you are poor, be self-reliant

get results

- you won’t be taken seriously otherwise

small is better

- work with specific tools in a circumscribed area

innovation often sucks

- funding institutions need to think about their language