Corpus

- 24.787 pdf documents, representing 138,3 GB
- Period 1958 -1982, with documents in French, Dutch, German, Italian, Danish, English and Greek
- Tombstone metadata
- Conversion to .txt and split per language => 205.370 . txt files, representing 7,4 GB
- 835.717.292 words or 1.671.434 pages
- Usage of Topic Modeling and Word2Vec to create relevant metadata

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

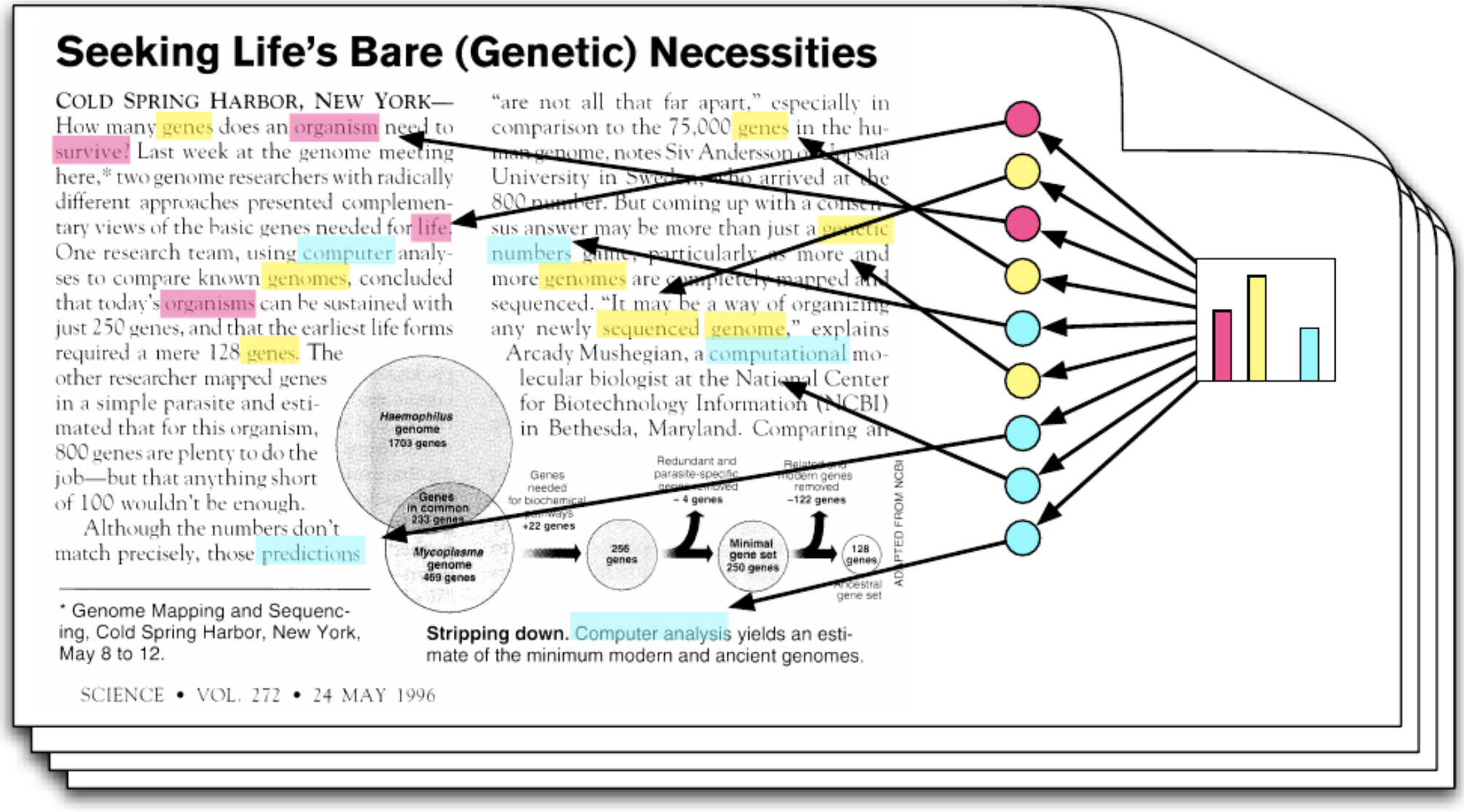
life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Topic proportions and assignments



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

David Blei (2012)

	A	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Topic Number	Salient tokens													
2	0	amounts	cereals	compensatory	products	rice	wheat	applicable	regulation	amount	eec	refund	falling	weight	referred
3	1	fund	decision	council	assistance	social	article	european	operations	february	training	regulation	persons	application	financial
4	2	article	member	provisions	states	paragraph	commission	council	directive	measures	proposal	laid	referred	state	provided
5	3	commission	council	committee	decision	european	community	member	states	draft	proposal	economic	opinion	concerning	treaty
6	4	securities	quantitative	textiles	authorities	arrangement	products	community	textile	limits	units	zealand	quotation	ciuts	agreement
7	5	directive	products	substances	product	annex	sea	medicinal	foodstuffs	acid	mediterranean	solution	health	laws	content
8	6	products	customs	weight	tariff	fruit	duties	ii	sugar	exceeding	falling	heading	common	content	net
9	7	tax	project	taxable	will	rate	amount	oil	services	taxation	system	supplies	production	excise	equipment
10	8	including	products	heading	articles	machines	parts	apparatus	metal	falling	ii	example	kind	paper	machinery
11	9	product	vehicles	tha	produot	agreement	motor	froa	bureaux	territory	produota	republic	based	including	nr
12	10	contract	credit	contracts	insurer	amount	guarantee	payment	insured	grouping	export	loss	works	bank	commercial
13	11	agreement	community	european	trade	economic	products	republic	negotiations	imports	article	council	cooperation	parties	commission
14	12	contracting	party	parties	agreement	agent	operating	executive	committee	task	agency	will	participants	article	operation
15	13	test	gas	pressure	directive	equipment	water	appliance	type	tests	measuring	inspection	requirements	appliances	conditions
16	14	increase	community	production	countries	prices	year	situation	will	market	world	total	level	price	increased
17	15	description	products	de	member	port	airport	fabrics	destination	european	communities	address	state	handicrafts	tariff
18	16	member	states	state	directive	law	national	undertakings	community	legal	provisions	rights	laws	territory	authorities
19	17	convention	protocol	agreement	article	treaty	council	force	states	commerce	accession	government	provisions	member	international
20	18	holdings	income	farm	labour	ua	ha	family	survey	pigs	land	alu	returning	poultry	paid
21	19	regulation	eec	article	regard	council	commission	community	european	products	amended	states	treaty	applicable	member
22	20	project	subsidy	aid	guidance	agricultural	regulation	eec	fund	granted	will	article	application	eaggf	construction
23	21	products	countries	community	tariff	states	preferences	member	islands	article	imports	quotas	duties	territories	commission
24	22	member	quota	states	community	tariff	share	state	shares	tons	reserve	metric	initial	commission	article
25	23	coal	steel	decision	ecsc	treaty	will	market	undertakings	article	commission	industry	aid	iron	production
26	24	fishing	fish	staff	community	fishery	species	annex	sea	allowance	servants	catch	officials	conditions	ii
27	25	products	originating	protocol	manufacture	certificate	article	goods	movement	customs	community	exceed	product	joint	conditions
28	26	tho	bo	cf	tha	ir	ar	ii	ia	ho	io	en	ti	cr	er
29	27	amounts	meat	compensatory	beef	animals	veal	eggs	amount	regulation	monetary	bovine	eec	weight	poultry
30	28	price	prices	market	year	marketing	intervention	sugar	production	wine	community	fixed	products	common	quality
31	29	aid	milk	food	powder	skimmed	tons	countries	community	metric	programme	commission	supply	council	butteroil
32	30	european	data	projects	industry	programme	community	development	processing	will	aircraft	statistics	study	national	market
33	31	social	workers	employment	training	work	benefits	working	women	health	labour	action	migrant	security	family
34	32	tender	intervention	invitation	butter	tenders	export	agency	decision	market	minimum	levy	article	price	tenderer

README | EU_en_Pres1-50 | EU_en_Pres1-100 | EU_en_Pres2-100 | EU_en_Pres2-50 | EU_en_Pres3-100 | EU_en_Pres3-50

Content language:
(en) English

Simple search

Advanced search

Browse

- Browse the subject-oriented version

Download

- By domain
- Permuted alphabetical
- Multilingual list
- Alphabetical index
- EuroVoc SKOS/RDF
- EuroVoc XML

Your proposals

- Contribute
- New approved concepts

barley

RDF/XML

60 AGRI-FOODSTUFFS
MT 6006 plant product
BT1 cereals
RT malt [6026]

URI *http://eurovoc.europa.eu/2193*

Has Exact Match
Barley (AGROVOC)
Barley (ECLAS)
Barley (STW)
barley (UNBIS)

LANGUAGE EQUIVALENTS

BG	ечемик
ES	cebada
CS	ječmen
DA	byg
DE	Gerste
ET	oder
EL	κριθάρι
EN	**barley**
FR	orge
GA	barley *(under translation)*
HR	ječam
IT	orzo
LV	mieži
LT	miežiai
HU	árpa
MT	xgħir
NL	gerst
PL	jęczmień
PT	cevada
RO	orz
SK	jačmeň
SL	ječmen
FI	ohra
SV	korn
SR	јечам
MK	јачмен
SQ	elb

Human input - topic	EuroVoc URI	Label	Salient tokens		
coal and steel industry aid	http://eurovoc.europa.eu/852	ECSC aid	coal	steel	aid
ree trade	http://eurovoc.europa.eu/3584	trade restriction	tax	trade	duty
ducation	http://eurovoc.europa.eu/668	education	education	training	social
mployment policies	http://eurovoc.europa.eu/217190	employment policy	employment	workers	work
nergy production	http://eurovoc.europa.eu/2715	energy production	energy	oil	coal
ood trade	http://eurovoc.europa.eu/2446	food policy	gatt	tariff	cheese
gricultural aid	http://eurovoc.europa.eu/2965	aid to agriculture	aid	agricultural	measures

ECSC aid

RDF/XML | TURTLE | CONTRIBUTE | MAP

16 ECONOMICS

MT 1606 economic policy
BT1 EU aid
BT2 economic support
RT ECSC loan [1021]

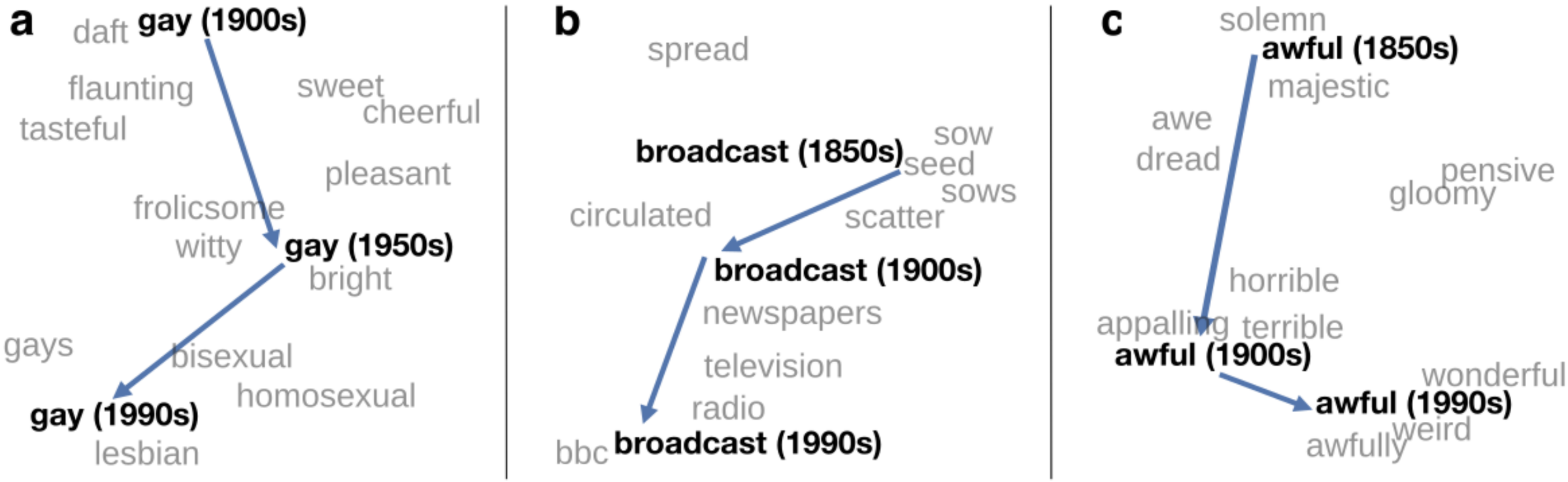
URI *http://eurovoc.europa.eu/852*

LANGUAGE EQUIVALENTS

BG	помощ на Европейско обединение за въглища и стомана
ES	ayuda CECA
CS	pomoc Evropského společenství uhlí a oceli
DA	EKSF-støtte
DE	EGKS-Beihilfe
ET	ESTÜ abi
EL	ενισχύσεις ΕΚΑΧ
EN	**ECSC aid**
FR	aide CECA
GA	ECSC aid *(under translation)*
HR	potpora EZUČ-a
IT	aiuto CECA
LV	EOTK palīdzība
LT	EAPB pagalba
HU	ESZAK-segély
MT	għajnuna KEFA
NL	EGKS-steun
PL	pomoc EWWS
PT	auxílio CECA
RO	ajutor CECO
SK	pomoc ESUO
SL	pomoč ESPJ
FI	EHTY:n tuki
SV	EKSG-stöd

Topic labeling

- Hulpus et al (2013) & Allahyaria and Kochuta (2015) use the graph structure of DBPedia to rank the different label candidates
- But - graph structure of DBPedia as a knowledge structure is not terribly coherent …
- Our approach : use pre-trained Word2Vec to exclude in an iterative manner the “outliers” from the tokens of a topic and match the remaining token with Eurovoc



https://nlp.stanford.edu/projects/histwords/

```
>>> word_vectors.doesnt_match("amounts meat compensatory beef animals veal eggs amount regulation monetary bovine eec weight poultry conditions prices competent imports authorities
".split())
'eec'
>>> word_vectors.doesnt_match("amounts meat compensatory beef animals veal eggs amount regulation monetary bovine weight poultry conditions prices competent imports authorities".sp
lit())
'competent'
>>> word_vectors.doesnt_match("amounts meat compensatory beef animals veal eggs amount regulation monetary bovine weight poultry conditions prices imports authorities".split())
'compensatory'
>>> word_vectors.doesnt_match("amounts meat beef animals veal eggs amount regulation monetary bovine weight poultry conditions prices imports authorities".split())
'regulation'
>>> word_vectors.doesnt_match("amounts meat beef animals veal eggs amount monetary bovine weight poultry conditions prices imports authorities".split())
'conditions'
>>> word_vectors.doesnt_match("amounts meat beef animals veal eggs amount monetary bovine weight poultry prices imports authorities".split())
'authorities'
>>> word_vectors.doesnt_match("amounts meat beef animals veal eggs amount monetary bovine weight poultry prices imports".split())
'weight'
>>> word_vectors.doesnt_match("amounts meat beef animals veal eggs amount monetary bovine poultry prices imports".split())
'monetary'
>>> word_vectors.doesnt_match("amounts meat beef animals veal eggs amount bovine poultry prices imports".split())
'amount'
>>> word_vectors.doesnt_match("amounts meat beef animals veal eggs bovine poultry prices imports".split())
'amounts'
>>> word_vectors.doesnt_match("meat beef animals veal eggs bovine poultry prices imports".split())
'prices'
>>> word_vectors.doesnt_match("meat beef animals veal eggs bovine poultry imports".split())
'imports'
>>> word_vectors.doesnt_match("meat beef animals veal eggs bovine poultry".split())
'animals'
>>> word_vectors.doesnt_match("meat beef veal eggs bovine poultry".split())
'bovine'
>>> word_vectors.doesnt_match("meat beef veal eggs poultry".split())
'eggs'
>>> word_vectors.doesnt_match("meat beef veal poultry".split())
'poultry'
>>> word_vectors.doesnt_match("meat beef veal".split())
'veal'
>>> word_vectors.doesnt_match("meat beef".split())
'beef'
>>> word_vectors.doesnt_match("meat".split())
'meat'
>>>
```

Topic label from Word2Vec	Topic Number	Salient tokens								
barley	0	amounts	cereals	compensatory	products	rice	wheat	applicable	regulation	amount
December	1	fund	decision	council	assistance	social	article	european	operations	february
council	2	article	member	provisions	states	paragraph	commission	council	directive	measures
committee	3	commission	council	committee	decision	european	community	member	states	draft
ciuts/textile	4	securities	quantitative	textiles	authorities	arrangement	products	community	textile	limits
labelling	5	directive	products	substances	product	annex	sea	medicinal	foodstuffs	acid
sugar	6	products	customs	weight	tariff	fruit	duties	ii	sugar	exceeding
tax	7	tax	project	taxable	will	rate	amount	oil	services	taxation
apparatus	8	including	products	heading	articles	machines	parts	apparatus	metal	falling
german/productversions	9	product	vehicles	tha	produot	agreement	motor	froa	bureaux	territory
guarantee	10	contract	credit	contracts	insurer	amount	guarantee	payment	insured	grouping
agreements	11	agreement	community	european	trade	economic	products	republic	negotiations	imports
party	12	contracting	party	parties	agreement	agent	operating	executive	committee	task
appliances	13	test	gas	pressure	directive	equipment	water	appliance	type	tests
prices	14	increase	community	production	countries	prices	year	situation	will	market
handicrafts	15	description	products	de	member	port	airport	fabrics	destination	european
laws	16	member	states	state	directive	law	national	undertakings	community	legal
ratification	17	convention	protocol	agreement	article	treaty	council	force	states	commerce
labour	18	holdings	income	farm	labour	ua	ha	family	survey	pigs
council	19	regulation	eec	article	regard	council	commission	community	european	products
eaggf/project	20	project	subsidy	aid	guidance	agricultural	regulation	eec	fund	granted
states	21	products	countries	community	tariff	states	preferences	member	islands	article
tariff	22	member	quota	states	community	tariff	share	state	shares	tons
ecsc/market	23	coal	steel	decision	ecsc	treaty	will	market	undertakings	article
fishery	24	fishing	fish	staff	community	fishery	species	annex	sea	allowance
products	25	products	originating	protocol	manufacture	certificate	article	goods	movement	customs
/	26	tho	bo	cf	tha	ir	ar	ii	ia	ho
meat	27	amounts	meat	compensatory	beef	animals	veal	eggs	amount	regulation
prices	28	price	prices	market	year	marketing	intervention	sugar	production	wine
programme/butteroil	29	aid	milk	food	powder	skimmed	tons	countries	community	metric
programme	30	european	data	projects	industry	programme	community	development	processing	will
labour/programme	31	social	workers	employment	training	work	benefits	working	women	health
tender	32	tender	intervention	invitation	butter	tenders	export	agency	decision	market
oil	33	energy	oil	community	power	supply	electricity	production	nuclear	stations
import	34	goods	customs	duties	processing	authorities	import	transit	origin	country
programme	35	pollution	environment	water	protection	environmental	waste	quality	states	action
vehicles	36	type	approval	vehicle	directive	eec	vehicles	member	requirements	light
wood	37	seed	paper	products	community	imports	ceilings	commission	member	wood
appropriations	38	expenditure	financial	budget	appropriations	year	appropriation	staff	article	commission
vehicles	39	transport	road	traffic	carriage	system	goods	vehicles	regulation	inland
food	40	country	recipient	community	european	agreement	economic	delivery	aid	goods
policies	41	economic	policy	member	community	monetary	states	policies	measures	will
agriculture	42	regional	areas	regions	agricultural	development	directive	farming	region	population
prices	43	agreement	market	products	article	undertakings	prices	competition	commission	trade
development	44	countries	developing	states	cooperation	community	trade	acp	aid	associated
proposals	45	community	will	commission	development	action	problems	general	international	policy
programme/jrc	46	programme	nuclear	materials	project	material	technical	reactor	energy	activities
luxembourg/nederland	47	kingdom	ireland	united	member	italy	states	rate	italian	netherlands
fibre	48	woven	cotton	fabrics	fibres	man	articles	flax	knitted	textile
companies	49	company	board	european	companies	group	employees	article	management	supervisory

Content language:
(en) English

Simple search

Advanced search

Browse

- Browse the subject-oriented version

Download

- By domain
- Permuted alphabetical
- Multilingual list
- Alphabetical index
- EuroVoc SKOS/RDF
- EuroVoc XML

Your proposals

- Contribute
- New approved concepts

barley

RDF/XML

60 AGRI-FOODSTUFFS
MT 6006 plant product
BT1 cereals
RT malt [6026]

URI *http://eurovoc.europa.eu/2193*

Has Exact Match
Barley (AGROVOC)
Barley (ECLAS)
Barley (STW)
barley (UNBIS)

LANGUAGE EQUIVALENTS

BG	ечемик
ES	cebada
CS	ječmen
DA	byg
DE	Gerste
ET	oder
EL	κριθάρι
EN	**barley**
FR	orge
GA	barley *(under translation)*
HR	ječam
IT	orzo
LV	mieži
LT	miežiai
HU	árpa
MT	xgħir
NL	gerst
PL	jęczmień
PT	cevada
RO	orz
SK	jačmeň
SL	ječmen
FI	ohra
SV	korn
SR	јечам
MK	јачмен
SQ	elb

Content language:
(en) English

Simple search

council

- Advanced search

Browse

- Browse the subject-oriented version

Download

- By domain
- Permuted alphabetical
- Multilingual list
- Alphabetical index
- EuroVoc SKOS/RDF
- EuroVoc XML

Your proposals

- Contribute
- New approved concepts

Search results (52 concepts)

MT 0416 electoral procedure and voting

election

local election UF ***council*** *election*

MT 0436 executive power and public service

executive body

Council of Ministers

public administration

local government UF *county* ***council***

local government UF *town* ***council***

MT 0806 international affairs

international instrument

European convention UF *convention of the* ***Council*** *of Europe*

MT 1006 EU institutions and European civil service

EU institution

Council configuration

Council of the European Union UF ***Council*** *of European Ministers*

Council of the European Union UF ***Council*** *of the European Communities*

Council of the European Union UF ***Council*** *of the Union*

Council of the European Union UF *EC* ***Council***

Council of the European Union UF *EU* ***Council***

Council of the European Union UF *European Union* ***Council***

EU **Council** Presidency UF *EC* ***Council*** *Presidency*

EU **Council** committee UF *EC* ***Council*** *committee*

Ecofin UF *Ecofin* ***Council***

Ecofin UF *Economic and Financial Affairs* ***Council***

European **Council**

President of the European **Council**

Start dating !

- Automation with machine learning methods is inevitable - we need it !
- But … develop a nuanced view by experimenting on your own
- Best cheap and cheerful dating partners :
    - Topic Modeling = > LDA
    - Word embeddings => Word2Vec

Relevant links

- Paper on the usage of Topic Modeling applied on EC's archival holdings
- Vellino & Alberts paper on auto-classification of email
- NARA report on the automation of Records Management
- Topic Modeling tutorial for historians
- Topic Modeling tool Simon Hengchen