Hva kan Wikipedia fortelle oss om kaffe?

Lars G. Johnsen
Magnus Breder Birkenes
Nasjonalbiblioteket
Wikipedia-akedemiet 2015

Nasjonalbiblioteket

- Wikepedia emneord og gruppering av innhold
- Strukturering av innhold
- Gruppering av artikler med innholdsord

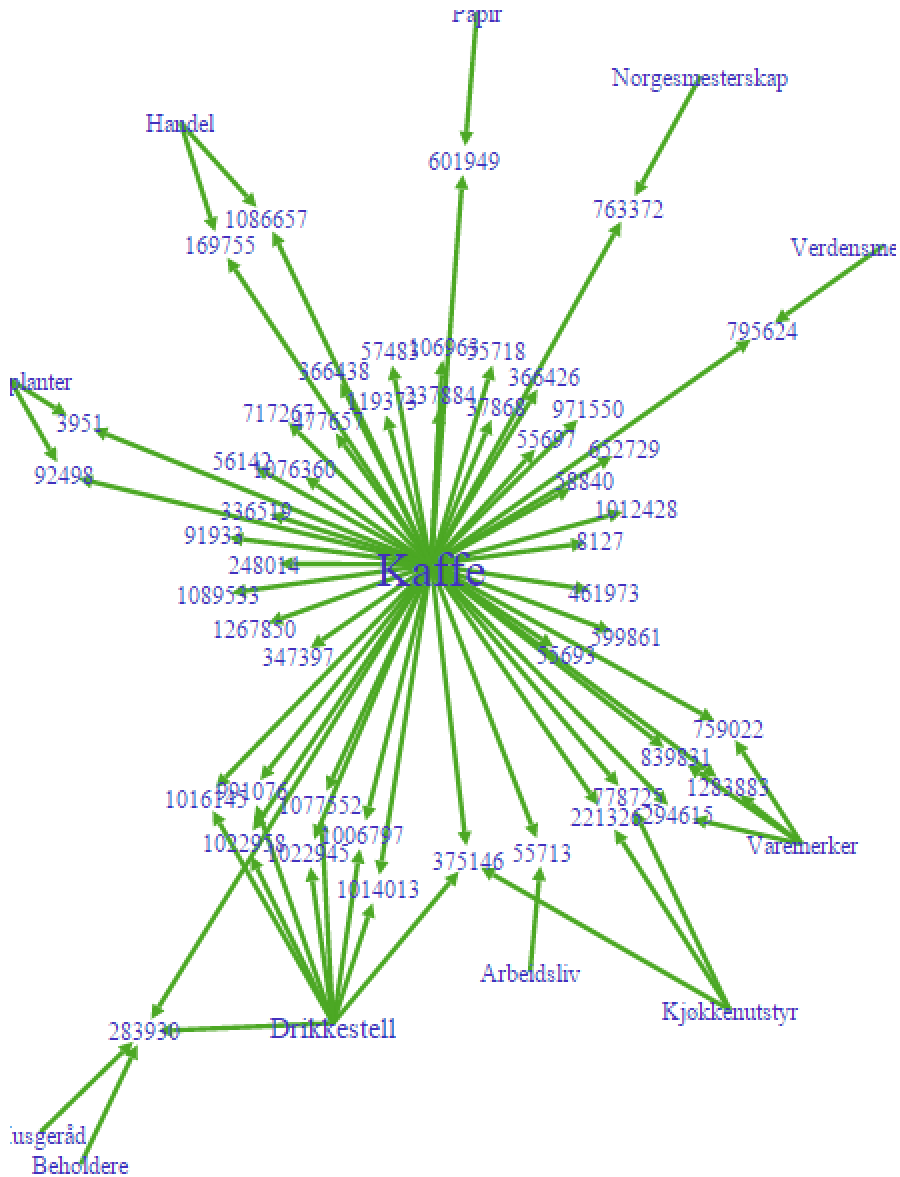
Relasjoner mellom innholds- og emneord

- Emner og tema
- Tekstklassifikasjon

Fra Wikipedia til data

- MySQL (wikipediadata)
- SQLite (arbeidsbase)
- Python (skripting)
- Gephi (visualisering)

F	Art.ID	Emne	Innhold
1	894284	Kortvinger	til
1	31563	Geologi	sies
1	26945	Rasisme	pasifister
1	189391	restkategori	to
1	407914	Bær	Vaccinium
1	478710	Tinget	etterkommes
1	526951	Kvinner	pensjonerte
1	187815	Menn	resulterte
1	1257701	Barokkomponister	Karl
1	947705	restkategori	på
1	410947	Menn	Trapezunt
1	134951	Göteborg	innbyggerantall



Papir
Norgesmesterskap
Handel
601949
1086657
169755
763372
795624
366438
366426
971550
planter
3951
652729
92498
1012428
248014
Kaffe
461973
1267850
599861
347397
759022
839831
1283883
1077552
1006797
221326
Varemerker
375146
55713
1014013
Arbeidsliv
Kjøkkenutstyr
283930
Drikkestell
Beholdere

4	USA	iranske
1	Trekirker	bakover
1	restkategori	SkyCargo
1	Tanagarer	lokalt
1	Litauen	rådet
1	Sprengstoff	jobb
1	Geofysikk	rombølger
1	Jesuittordenen	Juizhou
1	Portal:Færøyene/artikler	átti
2	Menn	verdenssensasjon
1	Tyskere	gjeldende
4	Dramakomedier	Frances

PMI (Pointwise Mutual Information)

Dividerer

Innholdsord i Emne/Emnetotalen

med

Innholdsordtotalen/Totalen

Og vekter med

Innholdsord i Emne

Kaffe


kaffen
starbucks
iskaffe
kaffekanner
kaffe
kaffepulveret
arabica
filteret
coffea
barista
robusta
nespresso
kokende
friele
café
kaffehuset
espresso
kokekaffe
spesialkaffe
löfbergs
bønnene
liberica
latte
beholderen
lait
baristakunst
coffee
kaffebønner
kaffetraktere
kjeldsberg
gevalia
kaffekjeler
kapslene
WordItOut

Ekteskap

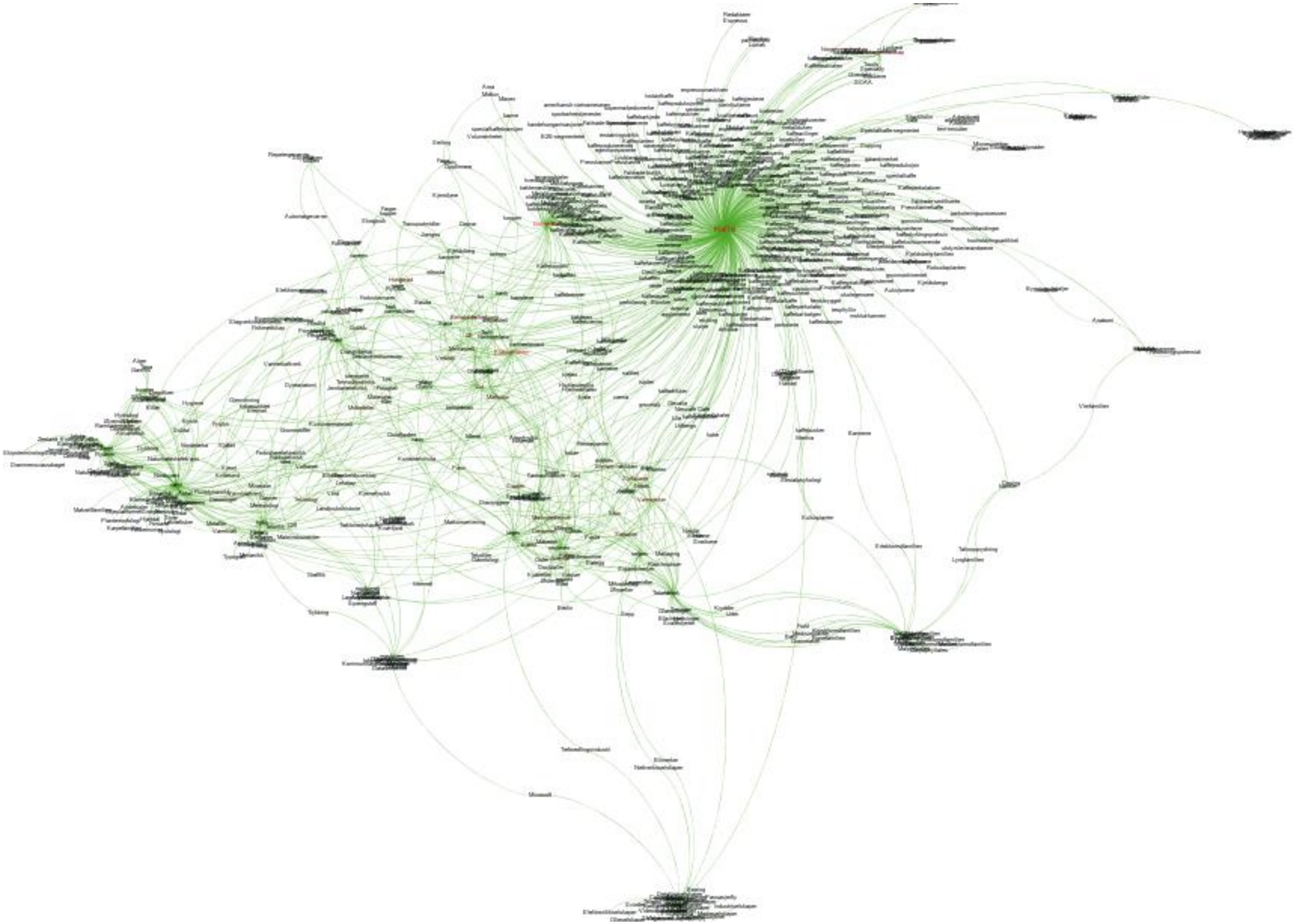

brudgommen ektefeller
gifte
bryllup ekteskap
likekjønnet bruden vielse
bryllupet gamos
vigsel parforhold ektevigsel
kvinne ektefellene inngått
forloveren inngå ekteskapsloven
gifter skilsmisse
hieros vigsler trolovelse
særeie felleseie inngåelse
polygami ektepakt
vigsleren ekteskapslov
ekteskapet likekjønnede
WordItOut

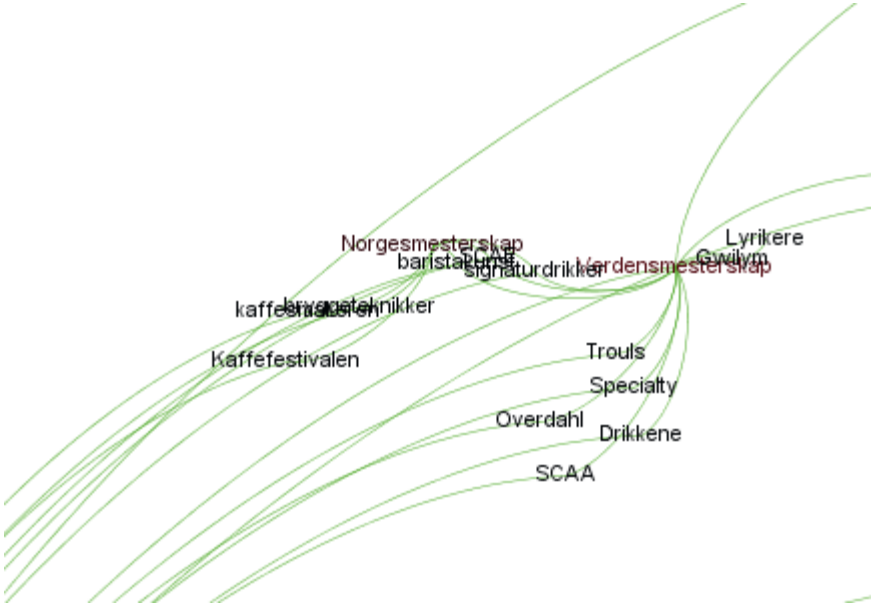
Bilmerker

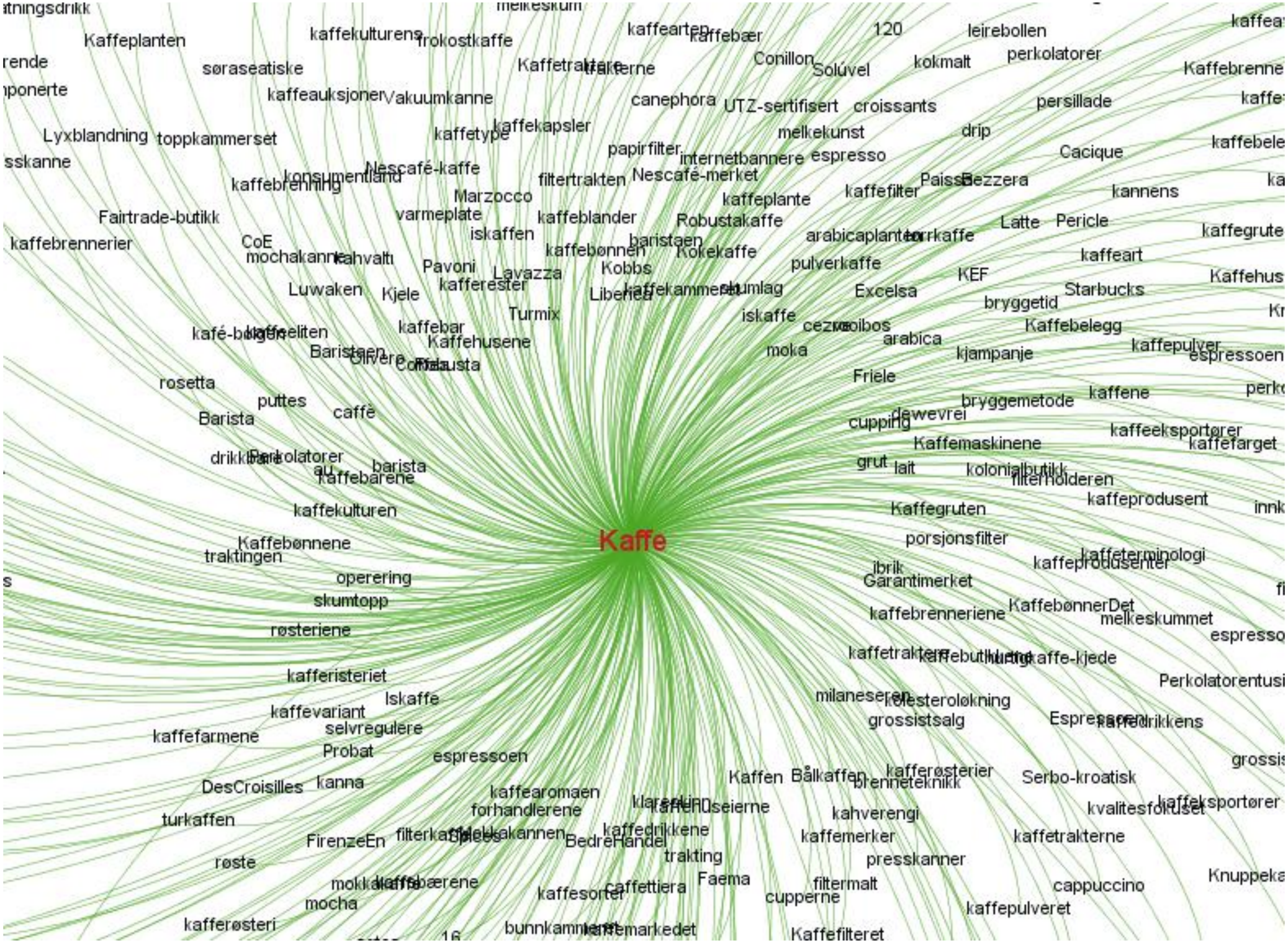

maserati sportsbiler
lastebiler
bilen
motors ford
selskapet
bil modell
company produksjon
fiat
rover daimler
lexus
fabrikken
buick merket
bilene
bilmerke
motorer
opel
motor biler
modeller
cadillac produserte
chrysler
modellen
modellene bilprodusent
WordItOut

Kaffe og undergrupper

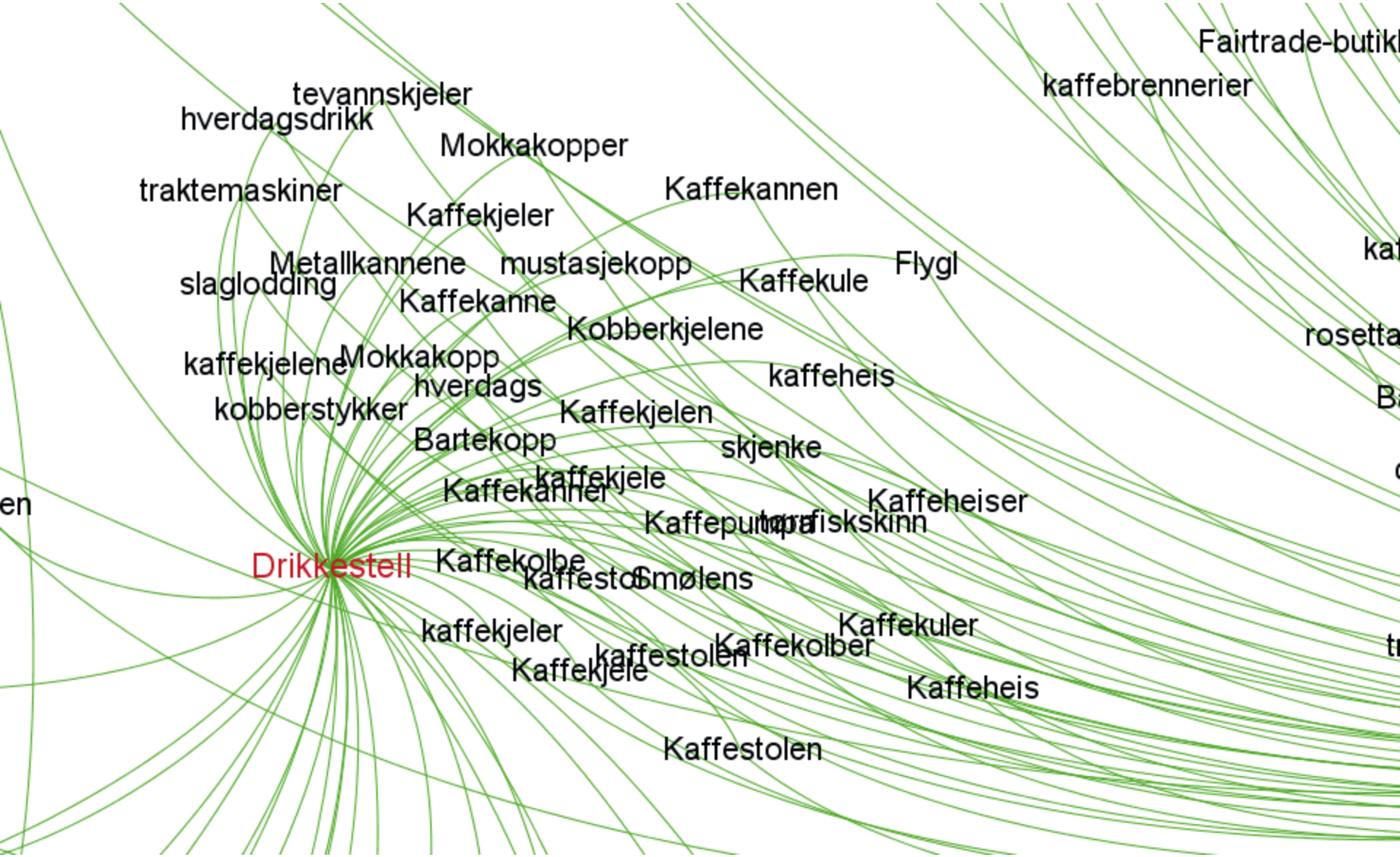
- Innholdsordene står også i tekster med forskjellige emneord




Norgesmesterskap
barista
signaturdrikker
Verdensmesterskap
Lyrikere
Gwilym
kaffe
Kaffefestivalen
Trouls
Specialty
Overdahl
Drikkene
SCAA



Kaffe
Kaffeplanten
søraseatiske
kaffeauksjoner
Vakuumkanne
Lyxblandning
toppkammerset
kaffetype
kaffekapsler
Fairtrade-butikk
kaffebrennerier
CoE
varmeplate
Marzocco
iskaffen
Pavoni
Lavazza
kafferester
Luwaken
Kjele
Turmix
kaffebar
Kaffehusene
rosetta
puttes
caffè
Barista
barista
kaffebarene
kaffekulturen
Kaffebønnene
traktingen
operering
skumtopp
rosteriene
kafferisteriet
Iskaffe
kaffevariant
selvregulere
kaffefarmene
Probat
espressoen
DesCroisilles
kanna
kaffearomaen
forhandlerene
turkaffen
FirenzeEn
røste
mocha
kafferøsteri
kaffesorter
Faema
cupperne
filtermalt
Kaffefilteret
bunnkammeret
kaffekulturens
frokostkaffe
kaffearten
kaffebær
120
leirebollen
Conillon
Solúvel
kokmalt
perkolatorer
canephora
UTZ-sertifisert
croissants
persillade
melkekunst
drip
papirfilter
internetbannere
espresso
Cacique
filtertrakten
Nescafé-merket
kaffeplante
kaffefilter
kannens
kaffeblander
Robustakaffe
Latte
Pericle
arabicaplanten
kaffegrute
Kokekaffe
kaffeart
Kobbs
pulverkaffe
KEF
Kaffehus
Excelsa
Starbucks
bryggetid
iskaffe
arabica
Kaffebelegg
kaffepulver
moka
kjampanje
espressoen
Friele
bryggemetode
kaffene
cupping
dewevrei
kaffeeksportører
Kaffemaskinene
kaffefarget
grut
lait
filterholderen
kaffeprodusent
Kaffegruten
porsjonsfilter
kaffeterminologi
ibrik
Garantimerket
melkeskummet
kaffebrenneriene
Perkolatorentusi
milaneseren
kolesterolnøkning
grossistsalg
kafferøsterier
brenneteknikk
Serbo-kroatisk
kahverengi
kaffeksportører
kaffemerker
kaffetrakterne
presskanner
cappuccino
kaffepulveret
Kaffen
trakting
Knuppeka

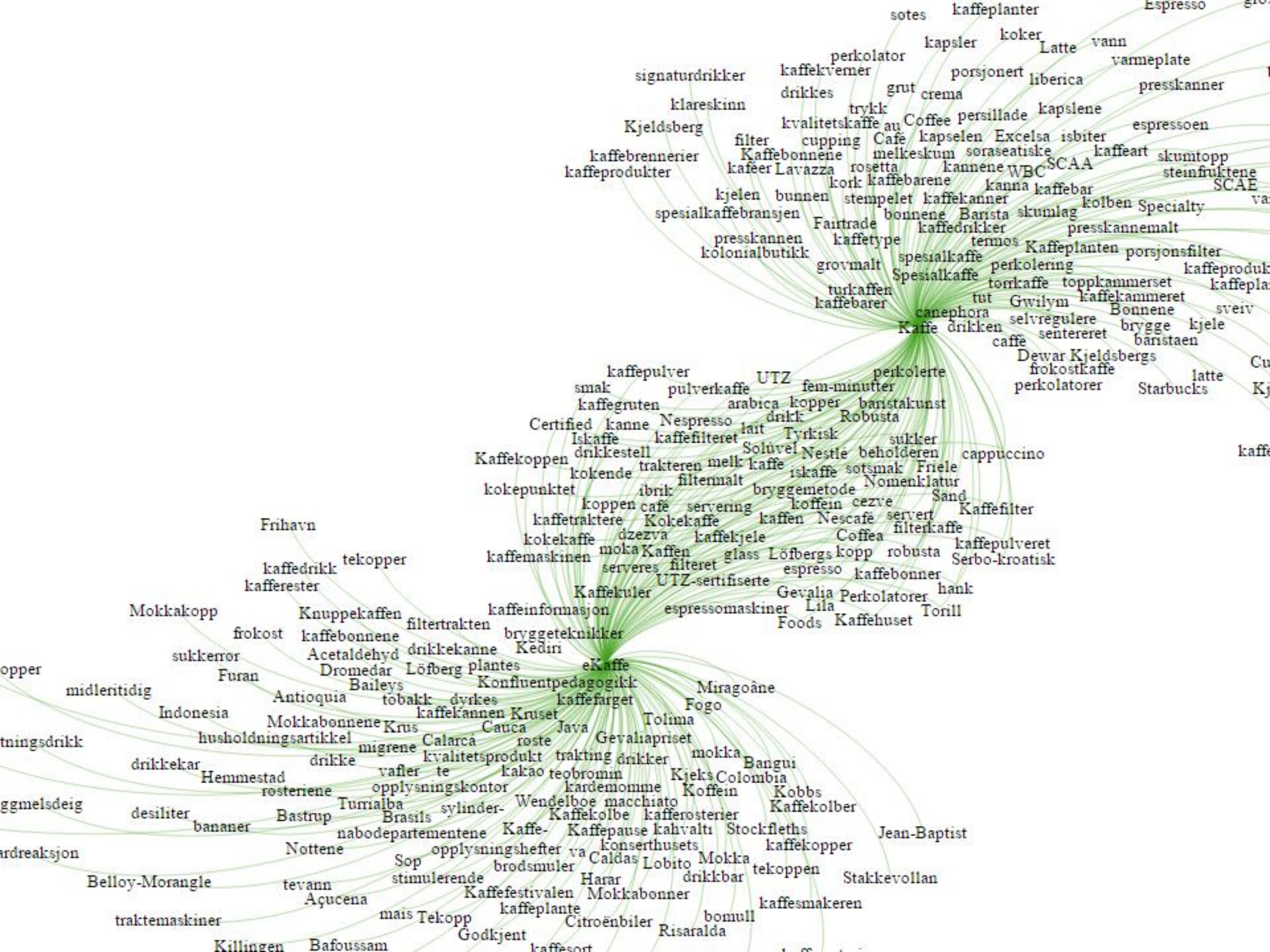

Fairtrade-butikk
kaffebrennerier
tevannskjeler
hverdagsdrikk
Mokkakopper
traktemaskiner
Kaffekannen
Kaffekjeler
Metallkannene
mustasjekopp
Flygl
slaglodding
Kaffekule
Kaffekanne
Kobberkjelene
rosetta
kaffekjelene
Mokkakopp
kaffeheis
hverdags
kobberstykker
Kaffekjelen
Bartekopp
skjenke
kaffekjele
Kaffekanner
Kaffeheiser
Kaffepumpa
Drikkestell
Kaffekolbe
Smølens
kaffekjeler
Kaffekuler
Kaffekolber
kaffestolen
Kaffekjele
Kaffeheis
Kaffestolen

Hva med gruppering på innholdsord?

- La «kaffe» være alle artikler som inneholder ordet «kaffe».

Assosiasjon

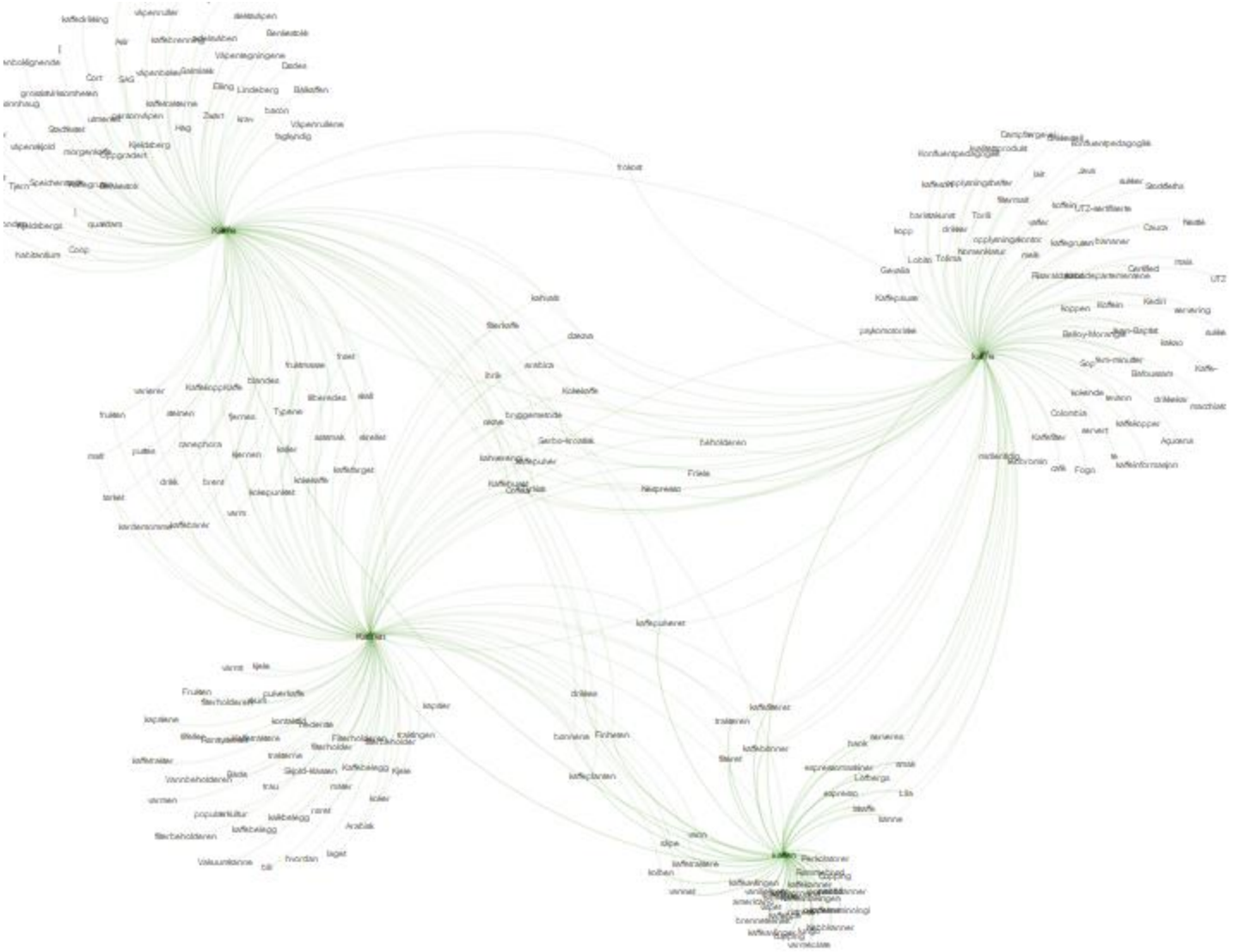
- For to ord «kaffe» og «kopp» i en artikkel
- Tell opp forekomster av «kaffe»
- Tell opp forekomster av «kopp»
- Ta produktet og del på størrelsen av artikkelen.
- Jo mindre artikkel, desto høyere score
- Jo fler «kaffe» eller «kopp», desto høyre score



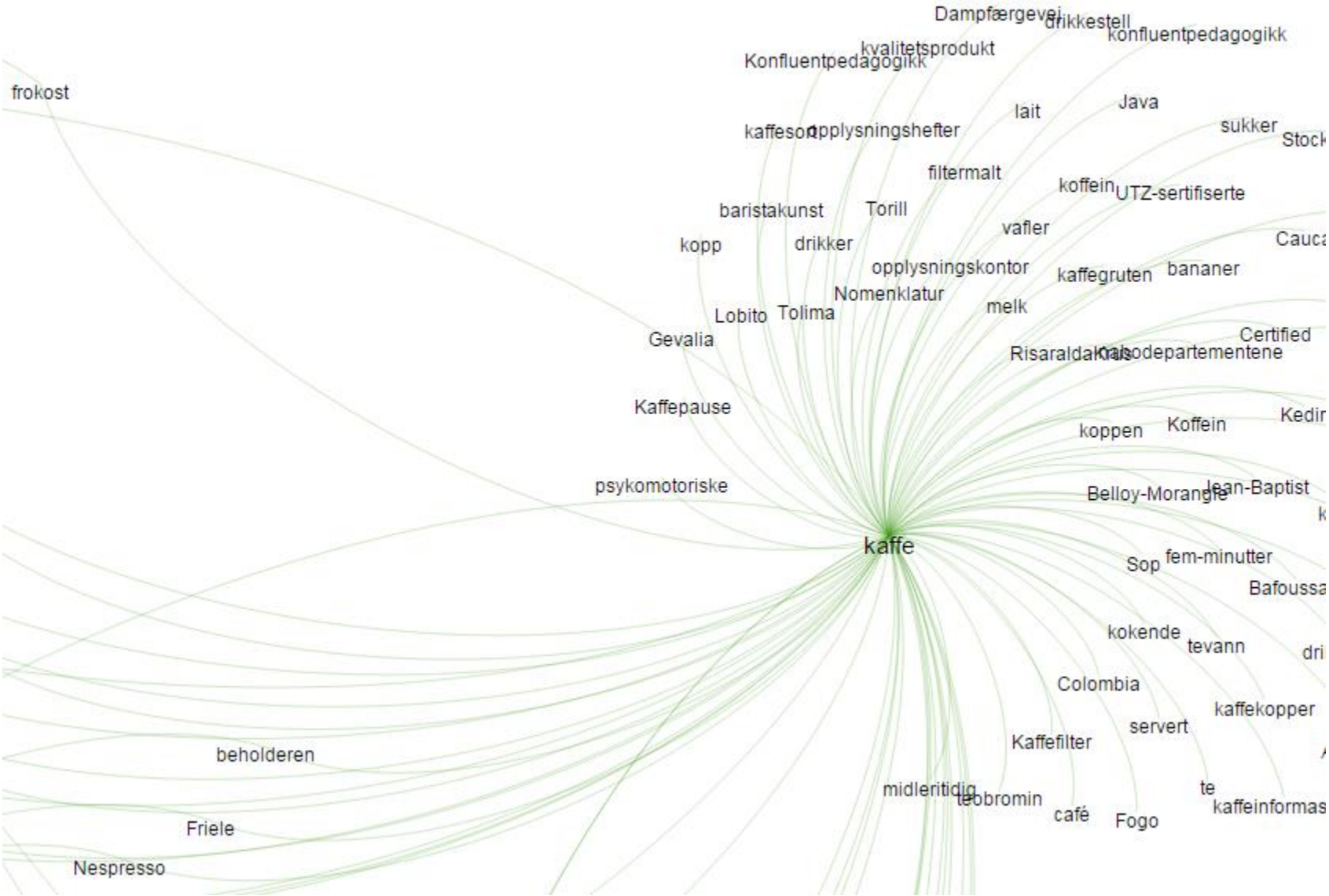
sotes
kaffeplanter
Espresso
kapsler
koker
Latte
vann
perkolator
varmeplate
signaturdrikker
kaffekverner
porsjonert
liberica
presskanner
drikkes
grut
crema
klareskinn
trykk
kvalitetskaffe
au
Coffee
persillade
kapslene
Kjeldsberg
espressoen
filter
cupping
Café
kapselen
Excelsa
isbiter
kaffebrennerier
Kaffebonnene
melkeskum
soraseatiske
kaffeart
skumtopp
kafeer
Lavazza
rosetta
kannene
WBC
SCAA
kaffeprodukter
steinfruktene
kork
kaffebarene
kanna
kaffebar
SCAE
kjelen
bunnen
stempelet
kaffekanner
kolben
Specialty
spesialkaffebransjen
bonnene
Barista
skumlag
Fairtrade
kaffedrikker
presskannemalt
presskannen
kaffetype
termos
Kaffeplanten
porsjonsfilter
kolonialbutikk
spesialkaffe
grovmalt
perkolering
Spesialkaffe
torrkaffe
toppkammerset
turkaffen
tut
Gwilym
kaffekammeret
kaffebarer
sveiv
canephora
Bonnene
selvregulere
Kaffe
drikken
brygge
kjele
sentereret
baristaen
caffè
Dewar
Kjeldsbergs
frokostkaffe
kaffepulver
perkolerte
latte
perkolatorer
UTZ
Starbucks
smak
pulverkaffe
fem-minutter
kaffegruten
arabica
kopper
baristakunst
Certified
kanne
Nespresso
drikk
Robusta
lait
Tyrkisk
Iskaffe
kaffefilteret
sukker
Solúvel
drikkestell
Nestlé
beholderen
cappuccino
Kaffekoppen
trakteren
melk
kaffe
sotsmak
Friele
kokende
iskaffe
filtermalt
Nomenklatur
kokepunktet
ibrik
bryggemetode
Sand
koppen
café
servering
koffein
cezve
Kaffefilter
kaffetraktere
Kokekaffe
kaffen
Nescafé
servert
filterkaffe
Frihavn
kokekaffe
dzezva
kaffekjele
Coffea
kaffepulveret
tekopper
kaffemaskinen
moka
Kaffen
glass
Löfbergs
kopp
robusta
Serbo-kroatisk
kaffedrikk
filteret
serveres
espresso
kaffebonner
kafferester
UTZ-sertifiserte
hank
Kaffekuler
Gevalia
Perkolatorer
Mokkakopp
Knuppekaffen
kaffeinformasjon
espressomaskiner
Lila
Torill
filtertrakten
Foods
Kaffehuset
frokost
kaffebonnene
bryggeteknikker
sukkerror
Acetaldehyd
drikkekanne
Kediri
opper
Dromedar
Löfberg
plantes
eKaffe
Furan
midleritidig
Baileys
Konfluentpedagogikk
Miragoâne
Antioquia
tobakk
dyrkes
kaffefarget
Fogo
Indonesia
kaffekannen
Kruset
Mokkabonnene
Krus
Cauca
Tolima
tningsdrikk
husholdningsartikkel
Calarcá
roste
Java
Gevaliapriset
migrene
kvalitetsprodukt
trakting
mokka
drikkekar
drikke
drikker
Bangui
vafler
te
kakao
teobromin
Hemmestad
Kjeks
Colombia
opplysningskontor
rosteriene
kardemomme
Kobbs
Koffein
ggmelsdeig
Turrialba
Wendelboe
macchiato
desiliter
sylinder-
Bastrup
Brasils
Kaffekolbe
kafferosterier
Kaffekolber
bananer
nabodepartementene
Kaffe-
Kaffepause
kahvaltı
Stockfleths
Jean-Baptist
Nottene
opplysningshefter
konserthusets
ardreaksjon
va
kaffekopper
Sop
Caldas
Lobito
Mokka
Belloy-Morangle
brodsmuler
tekoppen
Stakkevollan
tevann
stimulerende
Harar
drikkbar
Açucena
Kaffefestivalen
Mokkabonner
kaffesmakeren
kaffeplante
traktemaskiner
mais
Tekopp
bomull
Citroënbiler
Godkjent
Risaralda
Killingen
Bafoussam

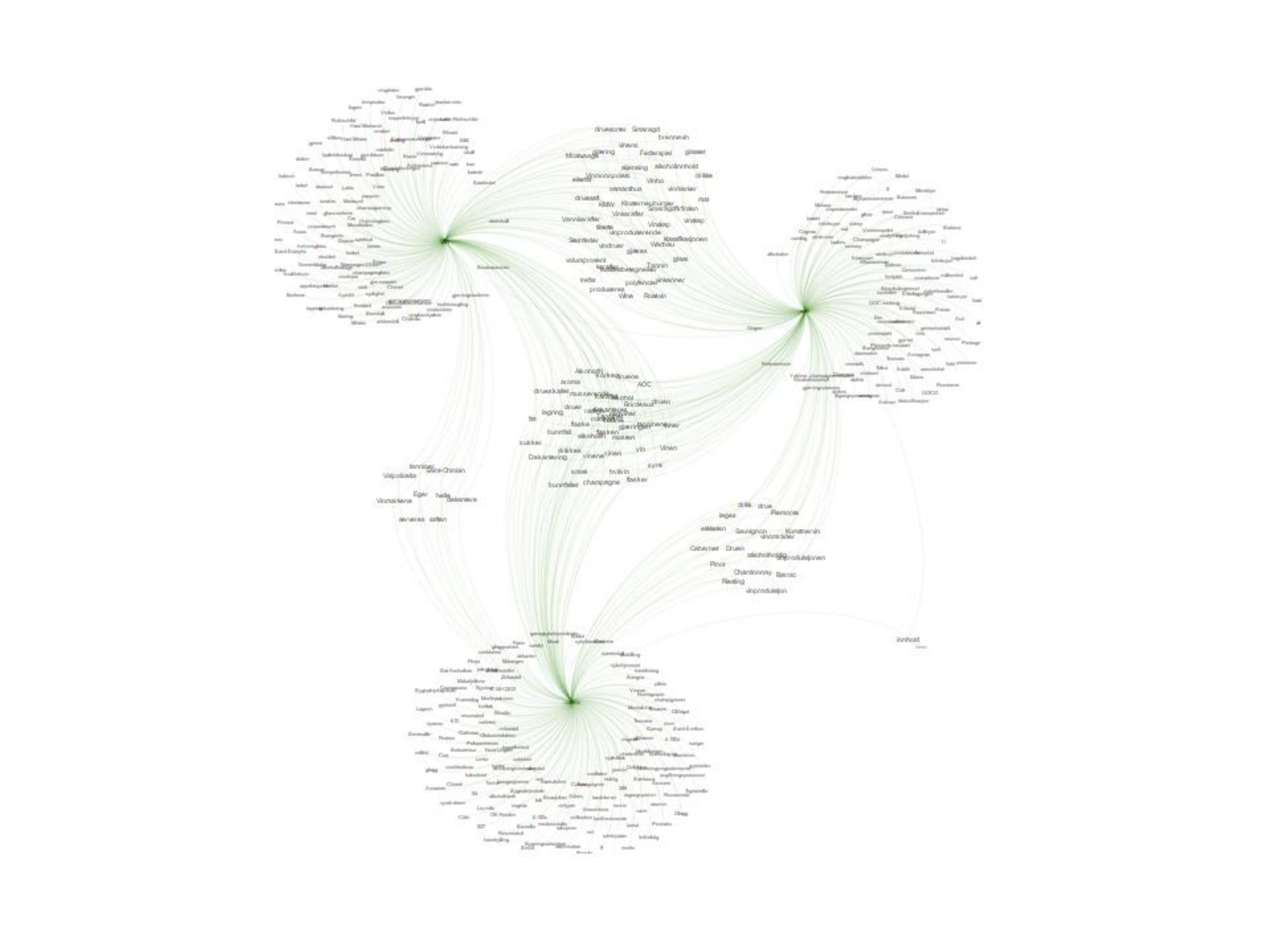
Leksem vs. ordformer

Leksem vs. ordformer



Leksem vs. ordformer




innhold