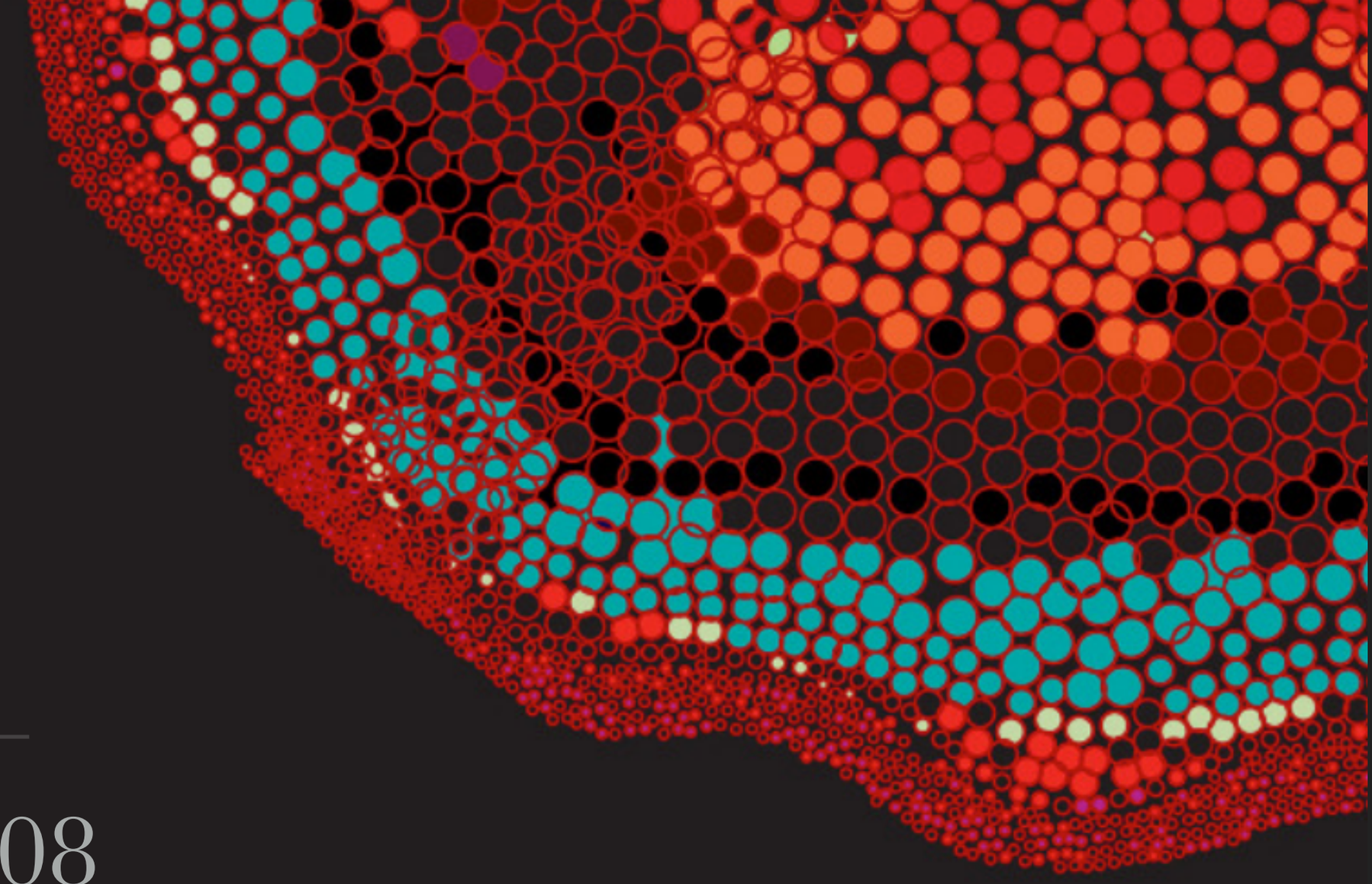


LECTURE 08

Text Visualization

they
puppy
is
her
shake
frantic
men
sad
blood
whisper
me
from
who
do
still
think
leg
soar
delirious
then
luscious
have
you
live
has
drunk
these
spray
have
finger
need
ful
s
an
breast
delicate
to
not
sleep
there
take
as
and
pole
your
must
let
boy
through
could
honey
sausage
death
produce
cook
this
language
some
above
moon
behind
head
of
out
my
said
like
size
ship
smooth
read
mean
rock
egg
blow
tongue
stare
drool
petal
lust
ask
juice
like
friend
am
the
skin
day
bed

Text And Document

- We have huge resources of text information:
  - Libraries
  - Email archives
  - WWW documents
  - Twitter feeds

Can we visualize these resources?

Levels Of Text Representations

- Lexical Levels
  - Transforming a string of characters into a sequence of atomic entities called *tokens* (e.g., characters, words, phrases)
- Syntactic Level
  - Identifying and tagging each token's function.
- Semantic level
  - Extraction of meaning and relationship between pieces of knowledge derived from the structure identified in the syntactic level.

Google Search I'm Feeling Lucky

Google offered in: 日本語

Vector Space Model

The Vector Space Model

There is a great deal of controversy about the safety of genetically engineered foods. Advocates of biotechnology often say that the risks are overblown. ‘‘There have been 25,000 trials of genetically modified crops in the world, now, and not a single incident, or anything dangerous in these releases,’’ said a spokesman for Adventa Holdings, a UK biotech firm. During the 2000 presidential campaign, then-candidate George W. Bush said that ‘‘study after study has shown no evidence of danger.’’ And Clinton Administration Agriculture Secretary Dan Glickman said that ‘‘test after rigorous scientific test’’ had proven the safety of genetically engineered products.

The Vector Space Model

- The paragraph contains 98 string tokens, 74 terms, and 48 terms when stop words are removed. Here is a sample of the term vector that would be generated by the pseudocode:

genetically	said	safety	engineered	study	test	great	deal	controversy	foods
3	3	2	2	2	2	1	1	1	1

id	men	entered	bank	charlotte	missiles	masks	aryan	guns	witnesses	reported	silver	suv	august
seg1.txt	0.239441	0	0.153457	0.195243	0	0.237029	0	0.195243	0.237029	0.140004	0.195243	0.237029	0
seg13.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg14.txt	0	0.192197	0	0	0	0	0	0	0	0	0	0	0.172681
seg15.txt	0	0	0	0	0	0	0	0	0	0	0	0	0.149652
seg16.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg17.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg18.txt	0	0.158432	0	0	0	0	0	0	0	0	0	0	0
seg19.txt	0	0	0	0.197255	0	0	0	0	0	0.141447	0	0	0.155038
seg2.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg20.txt	0	0.234323	0	0	0	0	0	0	0	0	0	0	0
seg21.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg22.txt	0	0	0	0	0.139629	0	0.127389	0	0	0	0	0	0
seg23.txt	0	0	0	0	0	0	0	0	0	0.180656	0	0	0
seg24.txt	0	0	0	0	0	0	0.117966	0	0	0.117966	0	0	0
seg25.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg26.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg27.txt	0	0	0.235418	0	0	0	0.214781	0	0	0	0	0	0
seg28.txt	0	0	0	0	0.151753	0	0	0	0	0	0	0	0
seg29.txt	0	0	0	0	0	0	0.129852	0	0	0	0	0	0.142329
seg3.txt	0	0	0	0	0.18432	0	0	0	0	0	0	0	0
seg30.txt	0.078262	0	0	0	0	0	0	0	0	0	0	0	0
seg31.txt	0	0	0.213409	0	0	0	0.194701	0	0	0	0	0	0
seg32.txt	0	0	0	0	0	0	0	0	0	0	0	0	0

Tasks Of Text Analysis

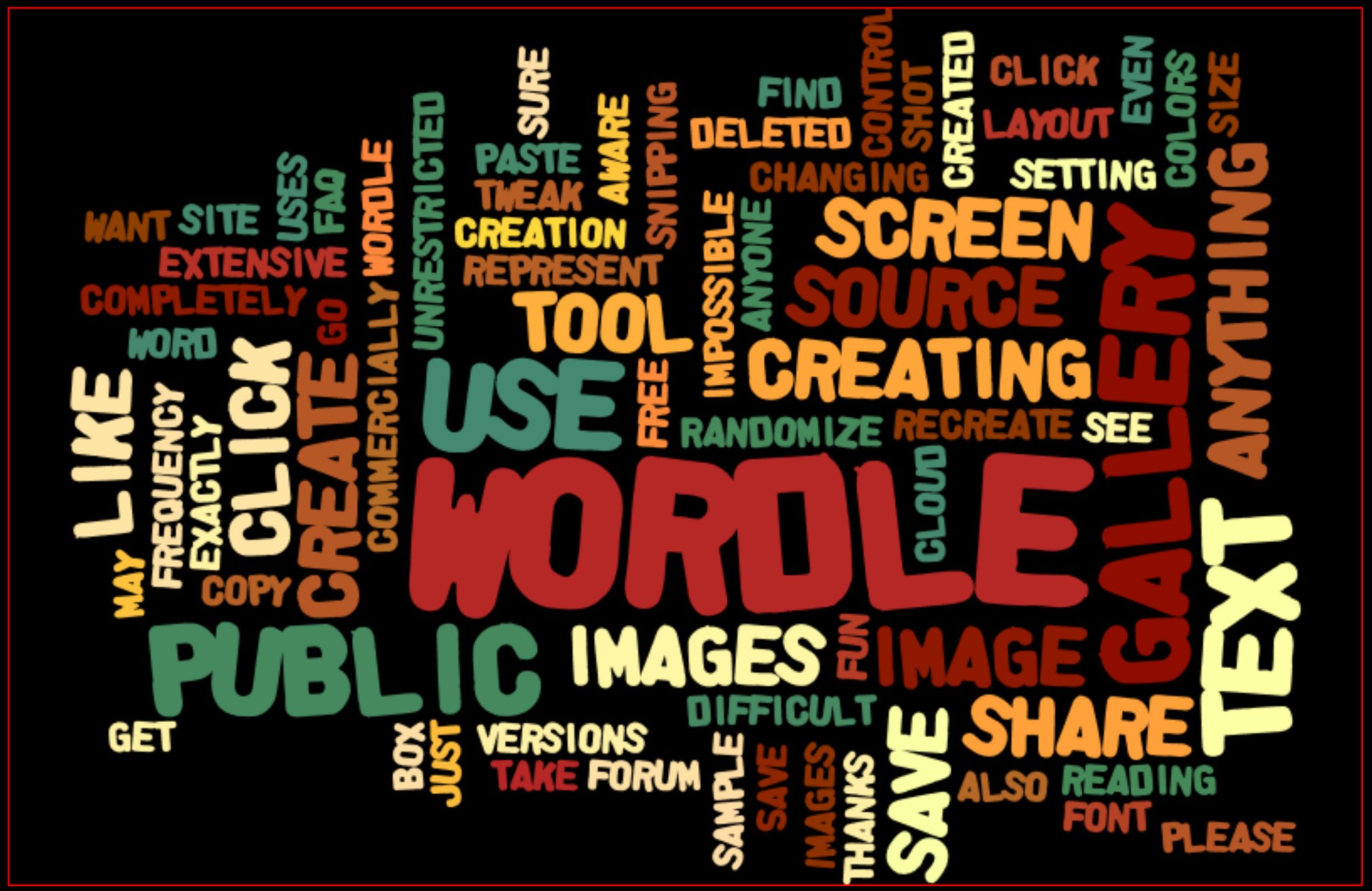
- Which documents are similar to a specific one?
- Which documents are relevant to a given collection?
- Which documents are relevant to a given search query?
- Finding patterns or structures
- Finding a document's main theme

Single Document

Tag Clouds


author biotechnology build concerned contained crops danger detected diet dr earthsave eating
engineered extra firm foods found ge genetically incident labeled life monitoring
monsanto ph press prevent products proven releases researcher risks rissler safety save
sequence shown soybeans stephens study surprised test think trials unfortunately validity vegetarian wall wild world
0s 60s 70s 80s 90s acoustic albums i own alternative alternative metal alternative
rock ambient american anime atmospheric avant-garde awesome beautiful black metal blues blues rock
ritish britpop brutal death metal canadian celtic chill chillout christian classic classic rock classical comedy
ool country cover dance dark ambient darkwave death metal disco doom metal downtempo drum and bass dub eas
stening ebm electro electronic electronica emo experimental favorite favorite songs
avorites favourite favourites female female vocalist female vocalists finnish folk folk metal folk rock french
un funk german goth gothic gothic metal gothic rock grindcore grunge guitar hard rock hardcore heavy
metal hip hop hip-hop house idm indie indie pop indie rock industrial industrial metal

Wordle



http://www.wordle.net/

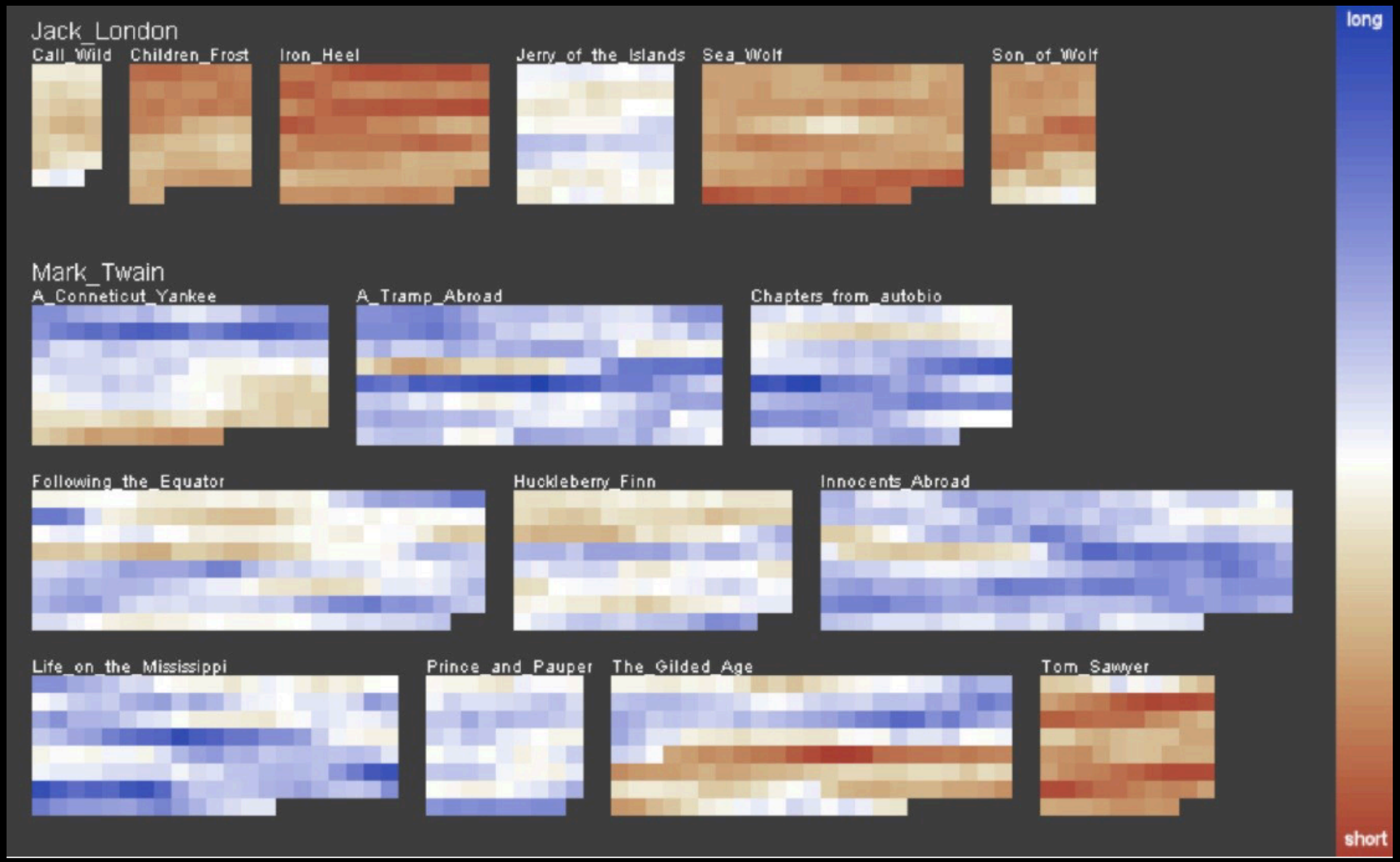
Word Tree

17 hits

- **the**
  - **most** part you wouldn't know [if you were eating them] but the point being that you wouldn't need to know .
  - **point** being that you wouldn't need to know .
  - **safety of genetically engineered**
    - foods .
    - products .
  - **risks** are overblown .
  - **world**
    - , now , and not a single incident , or anything dangerous in these releases , " said a spokesman for adventa holdings , a uk
    - [conari press , 2550 ninth st .
  - **2000** presidential campaign , then - candidate george w .
  - **case** ?
  - **union** of concerned scientists , dr .
  - **epa** , she is one of the nation's leading authorities on the environmental risks of genetically engineered foods .
  - **nation's** leading authorities on the environmental risks of genetically engineered foods .
  - **environmental** risks of genetically engineered foods .
  - **trials** and studies .
  - **wall** street journal found that 16 of 20 vegetarian foods labeled as being " free " of genetically engineered products actually contained ge soybeans .
  - **validity** of all their safety tests ?
  - **author** of diet for a new america and founder of earthsave international .

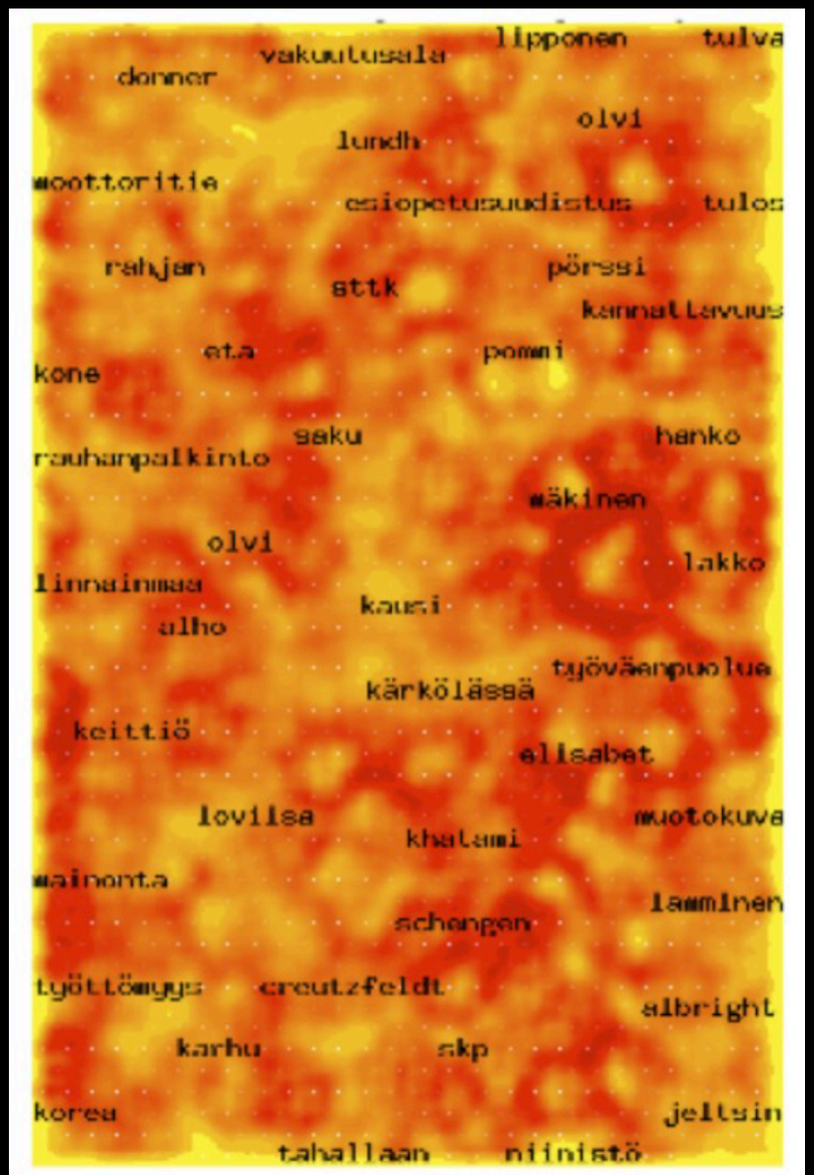

Enter phrase #1:
why doesn't he
130,000K
ask me out
60,800K
want a relationship
43,400K
like me quiz
2,010K
want to marry me
77,000K
call me
163,000K
like me
222,000K
call
94,000K
want me
92,300K
love me
22,900K
love me anymore
trust me
talk to me
just leave
leave
32,900K
88,200K
124,000K
57,400K
64,300K
17,300K
15,600K
20,000K
31,500K
115,000K
Enter phrase #2:
why doesn't she

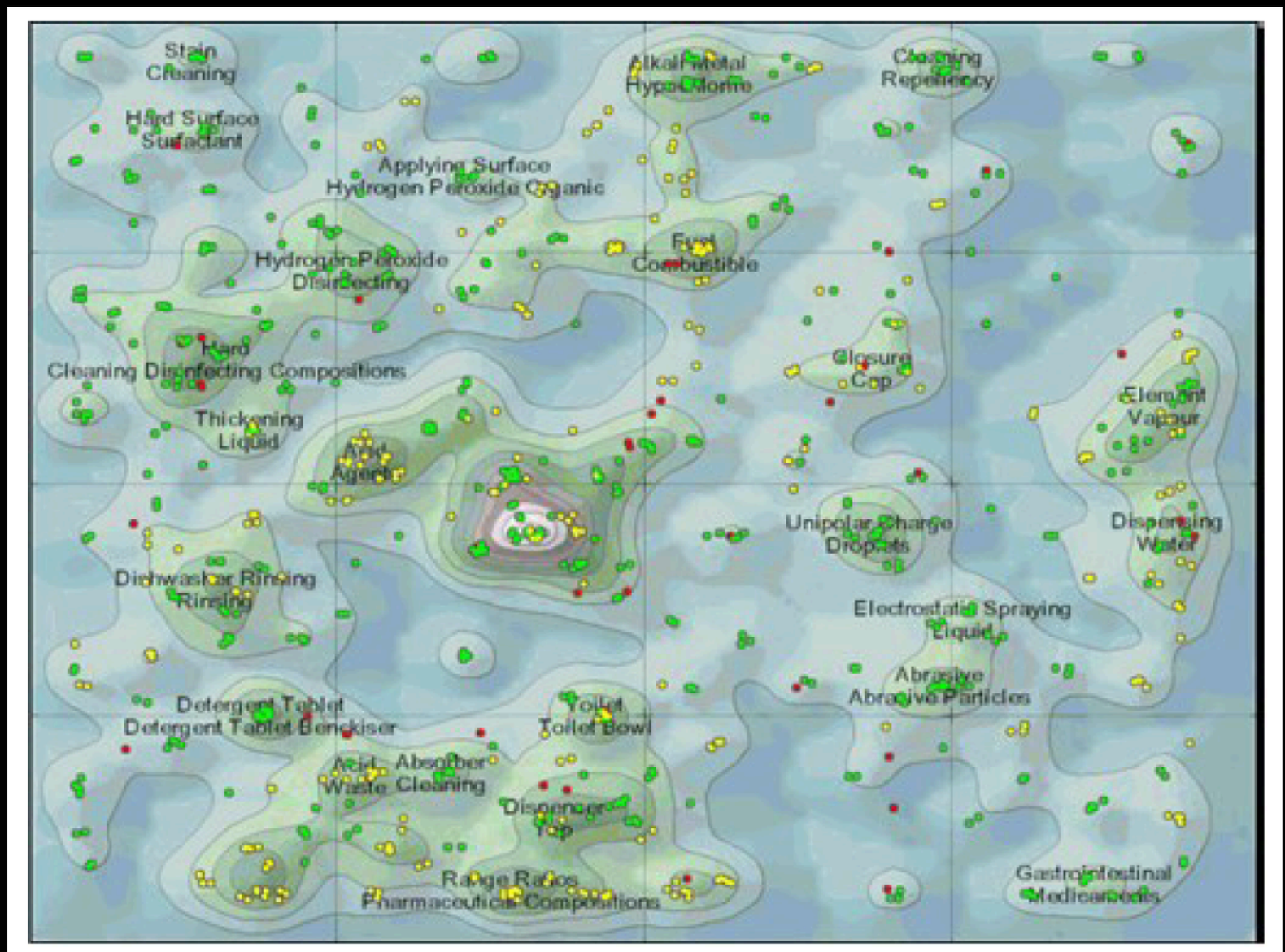
Literature Fingerprinting



Sentence Length

Document Collection



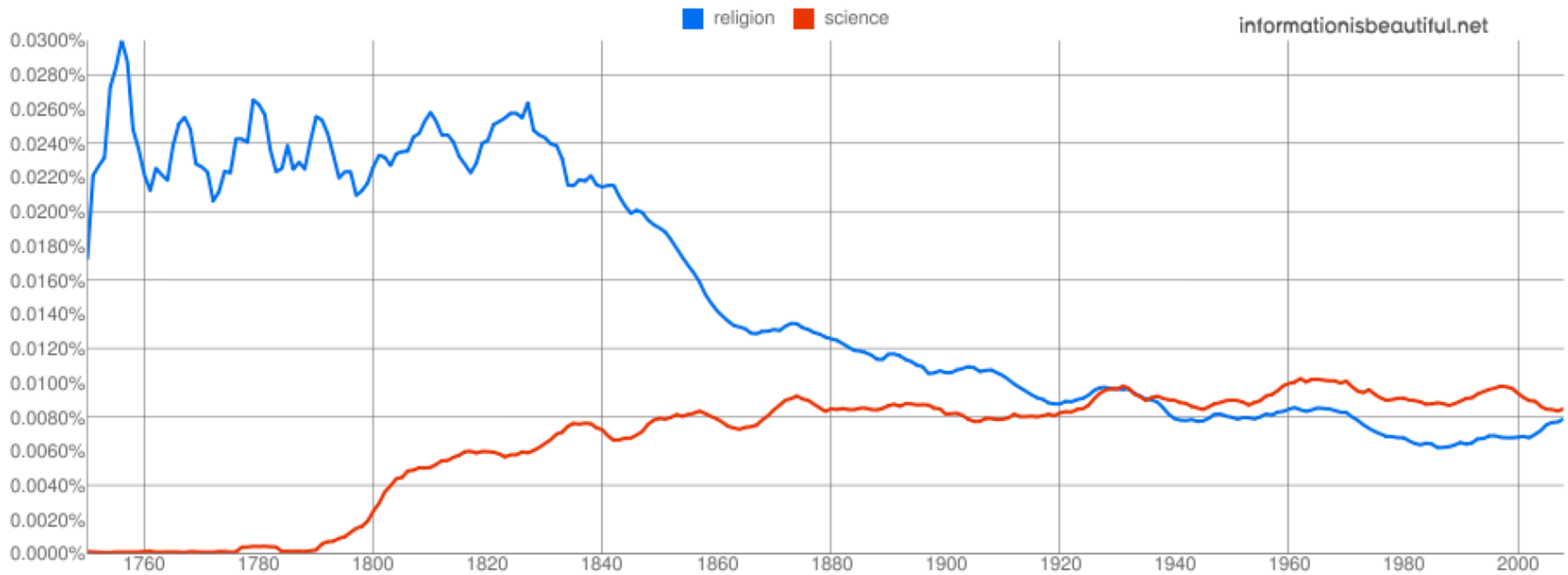


The labels show the topical areas, and color represents the number of documents, with light areas containing more.

Joined Analysis

Google NGram

https://books.google.com/ngrams


religion
science
informationisbeautiful.net
0.0300%
0.0280%
0.0260%
0.0240%
0.0220%
0.0200%
0.0180%
0.0160%
0.0140%
0.0120%
0.0100%
0.0080%
0.0060%
0.0040%
0.0020%
0.0000%
1760
1780
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000