Data and Network Analysis on Wikipedia Outline Computer Science & Artificial Intelligence

Introduction to Digital Humanities Final Project
Professor Moses Boudourides
Joonha Yu

Goals:

- Visualize network patterns within pages originating from two outlines
- Identify existing communities in the network
- Identify major topics within pages originating from two outlines
- Find relationships between words used in Wikipedia pages

1. Wikipedia Networks of Hyperlinks and Topic Modeling of Pages Originating from the *Outline of computer science* and the *Outline of artificial intelligence*

```
excluded=['International Standard Book Number',
        'International Standard Serial Number',
        'JSTOR',
        'Library of Congress Control Number',
        "Digital object identifier",
        "Integrated Authority File",
        "PubMed Identifier",
        "PubMed Central",
        "OCLC",
        "Wayback Machine",
        "ArXiv",
        "Bibcode",
        "ACM Computing Classification System",
        "Academic Press",
        "Website",
        "World Wide Web",
        "BioRxiv",
        "CiteSeerX",
        "Telecommunication network",
        "Web sites",
        "Daylight saving time",
        "International Standard Name Identifier",
        "Système universitaire de documentation",
        ...
        "Midpeninsula Free University",
        "California",
        "Random House",
        "Associated Press",
        "Computer History Museum",
        "Venture capitalist",
        "List of programmers",
        "List of computer scientists",
        "AFIPS",
        "Electronic publication",
        "Time zone",
        "Scientific American"
         ]
```

Collection Process:

- Used python module wikipedia to collect data
- Two outline pages:
  - “Outline of computer science”
  - “Outline of artificial intelligence”
- Excluded 86 links non-relevant to topics
- Verify collected links

Verification Process

- PEAS – Performance, Environment, Actuators, Sensors
- Percept (artificial intelligence) –
- Perceptual Computing –
- Rule-based system –
- Self-management (computer science) –
- Soft computing –

[nonexistent page shown in red]

[redirect page]

Collection Result:

- “Outline of Computer Science” has 234 functioning hyperlinks
- “Outline of Artificial Intelligence” has 407 functioning hyperlinks
- Two outlines have 26 hyperlinks in common
- In total, there are 615 unique hyperlinks

The directed graph of hyperlinked Wikipedia pages originating from two outlines:
'Outline of computer science and Outline of artificial intelligence'

Graph Diagnostics:

The graph has 477 nodes and 2199 edges
The graph is a simple graph
The graph is an unweighted graph
The graph is a directed graph
The graph is not a bipartite graph
The graph is not a tree
The graph is not strongly connected and it has 429 strongly connected components
The graph is a weakly connected graph
The graph has no isolates
The density of the graph is 0.010
The transitivity of the graph is 0.287

2. Wikipedia Networks of Reciprocating Hyperlinks of Pages Originating from the *Outline of computer science* and the *Outline of artificial intelligence*

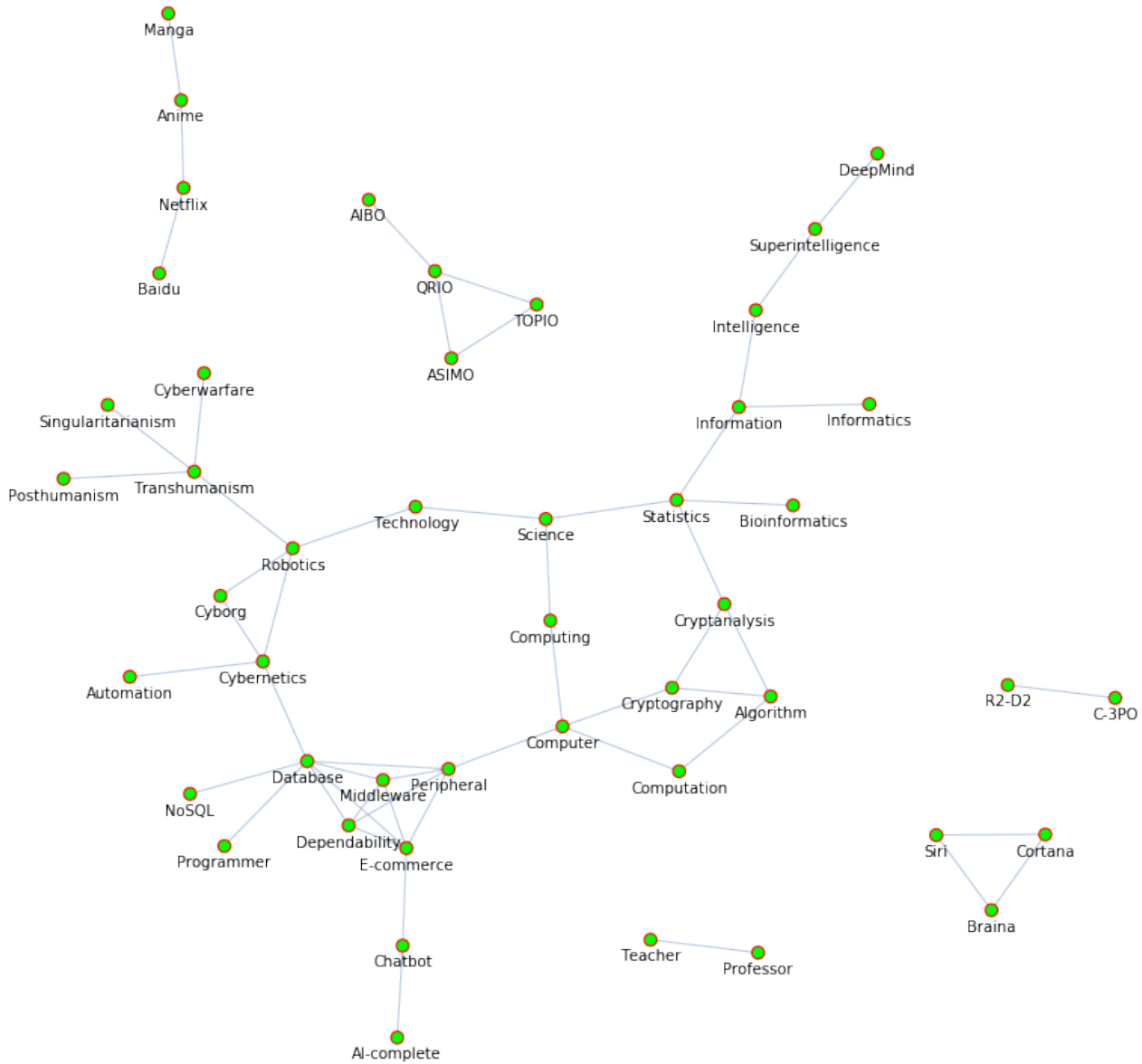
Reciprocation Analysis:

The graph has 2199 hyperlinks.
Among those, there are 108 reciprocating hyperlinks.

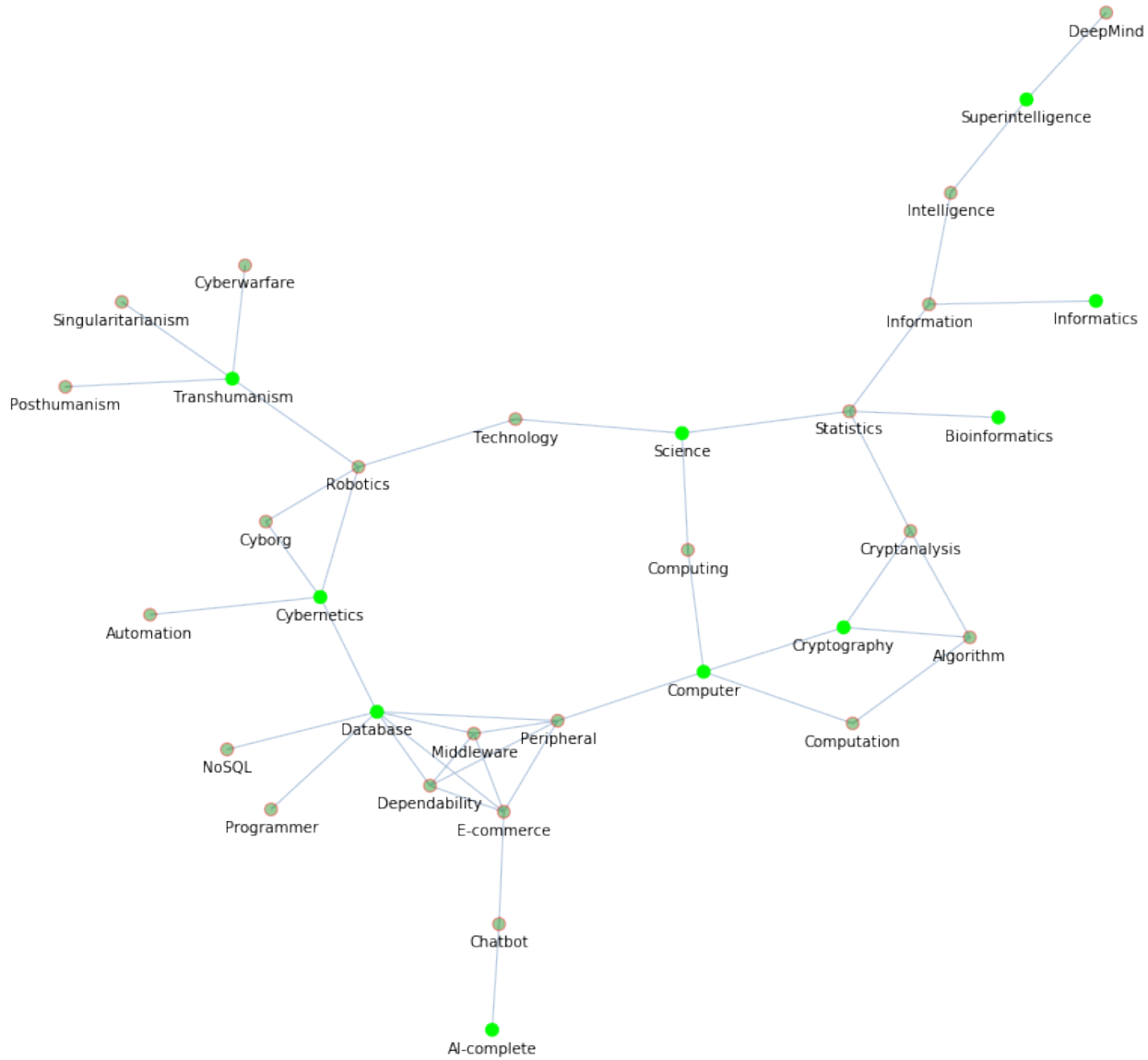
Graph Diagnostics:

The graph has 47 nodes and 54 edges
The graph is a simple graph
The graph is an unweighted graph
The graph is an undirected graph
The graph is not a bipartite graph
The graph is not a tree
The graph is a disconnected graph and it has 6 connected components
The largest connected component of this graph has 32 nodes and 42 edges
The graph has no isolates
The density of the graph is 0.050
The transitivity of the graph is 0.365
The diameter of the graph is 11

The network of reciprocated hyperlinks of Wikipedia pages
originating from 'Outline of computer science and Outline of artificial intelligence'

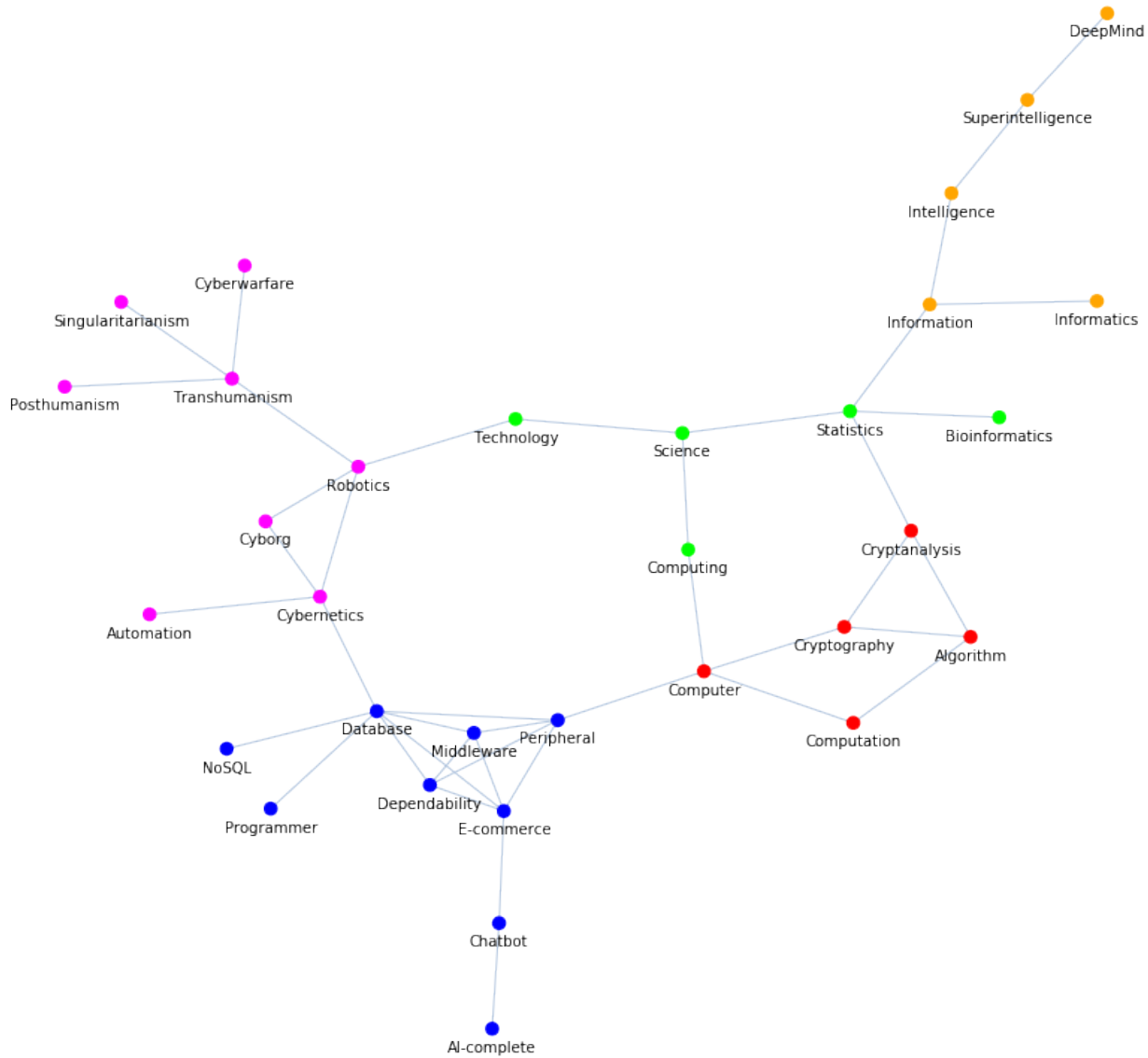


The largest connected component of the
network of reciprocated hyperlinks of Wikipedia pages
originating from 'Outline of computer science and Outline of artificial intelligence'

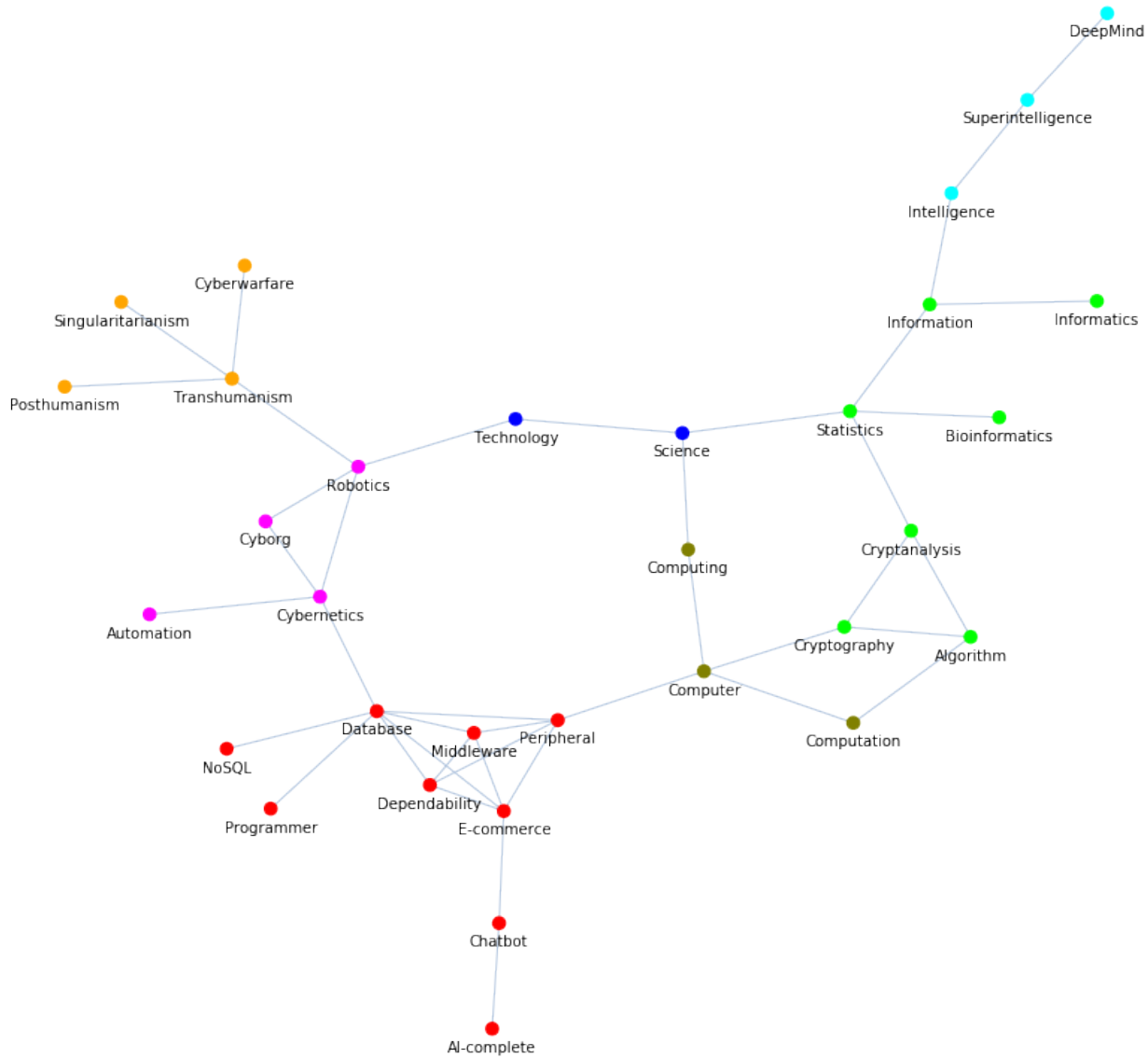


Community Analysis

The largest connected component of the
network of reciprocated hyperlinks of Wikipedia pages
originating from 'Outline of computer science and Outline of artificial intelligence'
colored in 5 Louvain communities



The largest connected component of the
network of reciprocated hyperlinks of Wikipedia pages
originating from 'Outline of computer science and Outline of artificial intelligence'
colored in 7 Fluid communities



3. Topic Modeling of Pages Originating from the *Outline of computer science* and the *Outline of artificial intelligence*

Previous Collection Result:

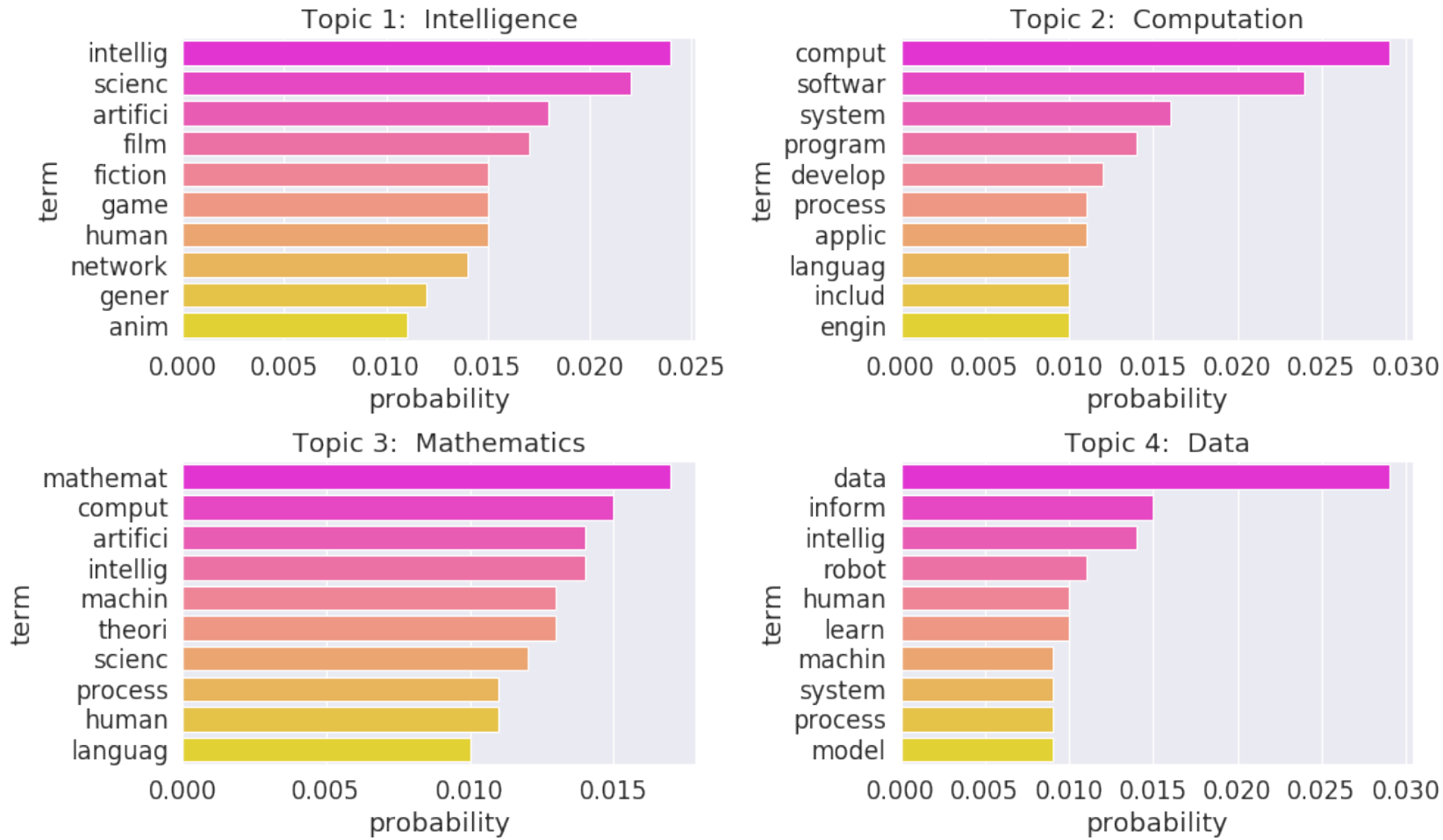
- “Outline of Computer Science” has 234 functioning hyperlinks
- “Outline of Artificial Intelligence” has 407 functioning hyperlinks
- Two outlines have 26 hyperlinks in common
- In total, there are 615 unique hyperlinks

Preparing for Topic Modeling:

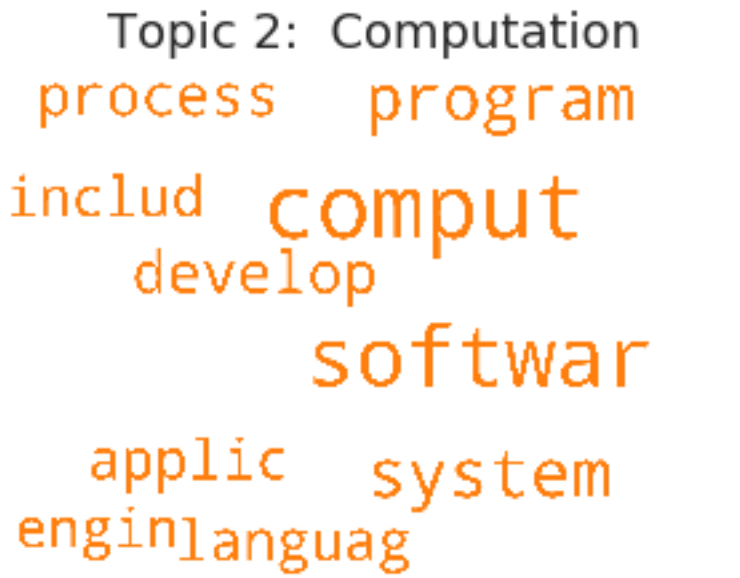
- Collect Wikipedia summaries from 615 pages.
- Tokenize 11890 words from the summaries.
- Preprocess by lemmatize stemming -> 6237 words.
- To focus on relevant words, filter vocabs which appear at least on 20 pages -> 60 words

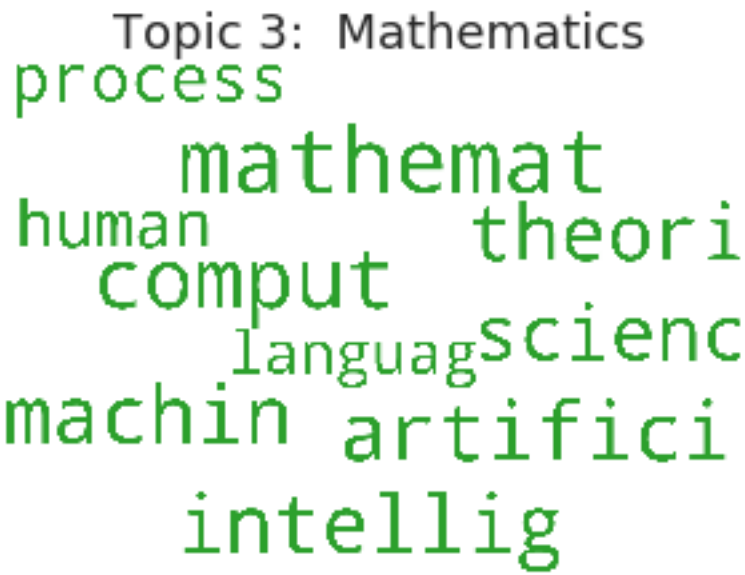
Used nltk library

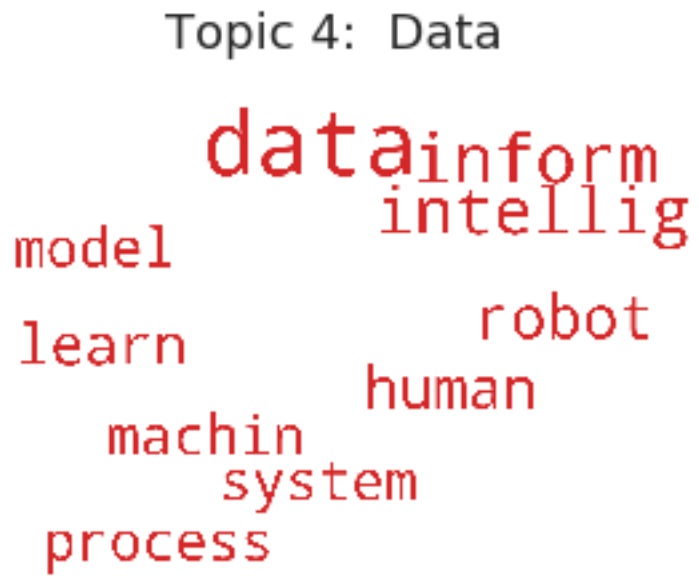
Topic Modeling (TM) of the 615 Wikipedia hyperlinks of 'Outline of computer science and Outline of artificial intelligence'



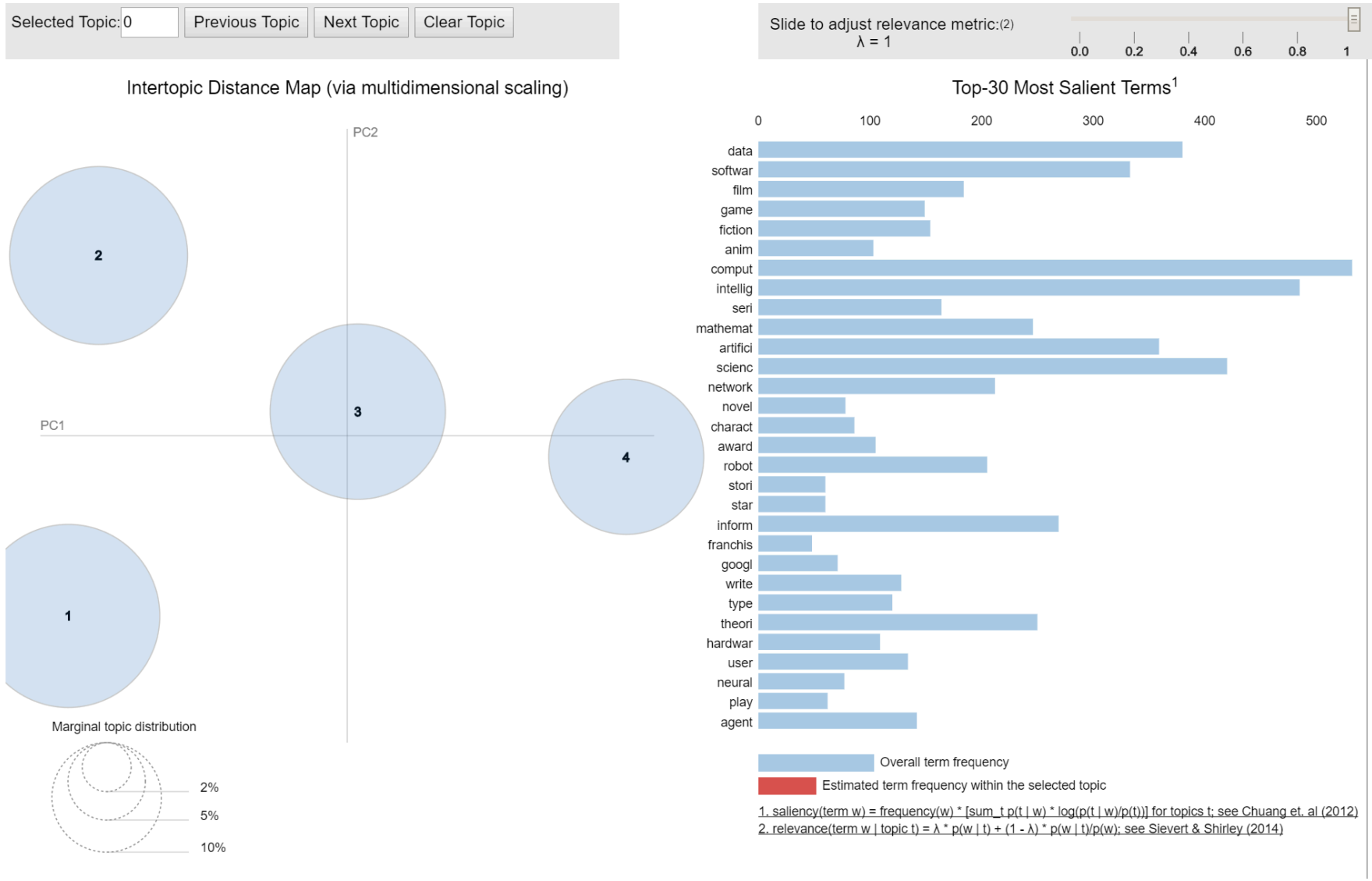

Topic 1: Intelligence
scienc
game
network
film
intellig
anim
human
fiction
gener
artifici


Topic 2: Computation
process
program
includ
comput
develop
softwar
applic
system
engin
languag

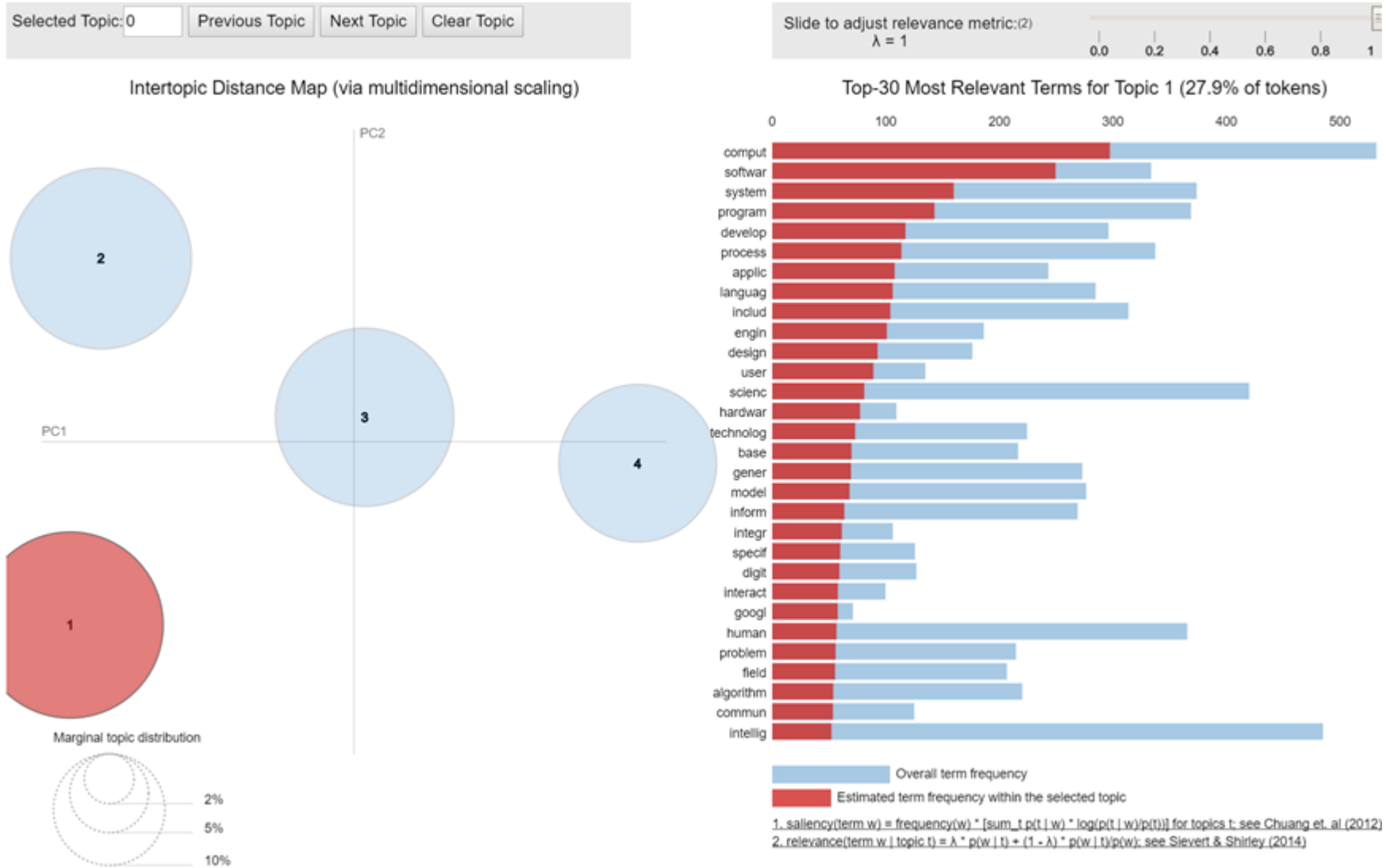

Topic 3: Mathematics
process
mathemat
human
theori
comput
languag
scienc
machin
artifici
intellig


Topic 4: Data
data
inform
intellig
model
robot
learn
human
machin
system
process

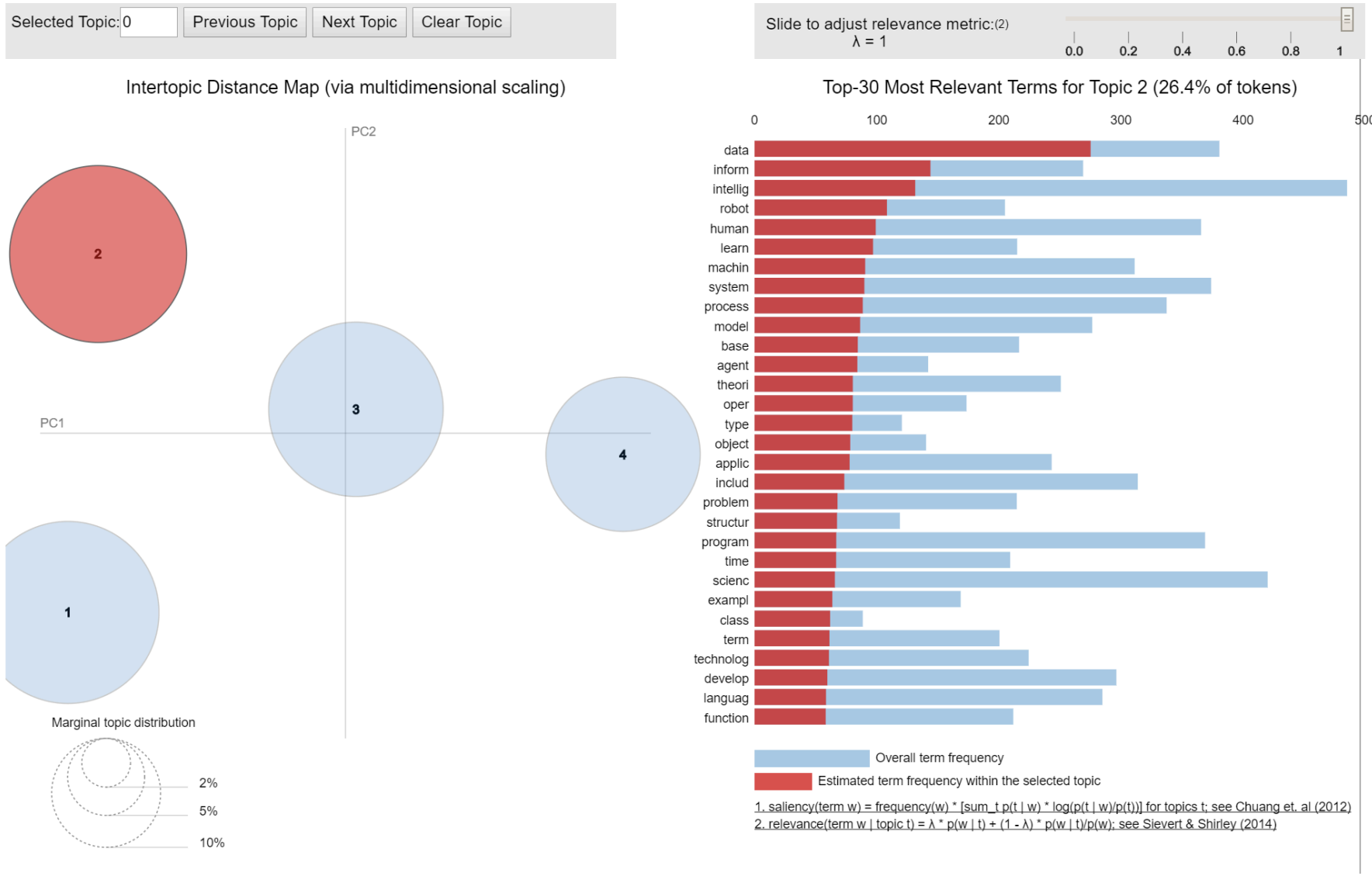
Out[29]:


Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
Marginal topic distribution
2%
5%
10%
Top-30 Most Salient Terms1
0
100
200
300
400
500
data
softwar
film
game
fiction
anim
comput
intellig
seri
mathemat
artifici
scienc
network
novel
charact
award
robot
stori
star
inform
franchis
googl
write
type
theori
hardwar
user
neural
play
agent
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

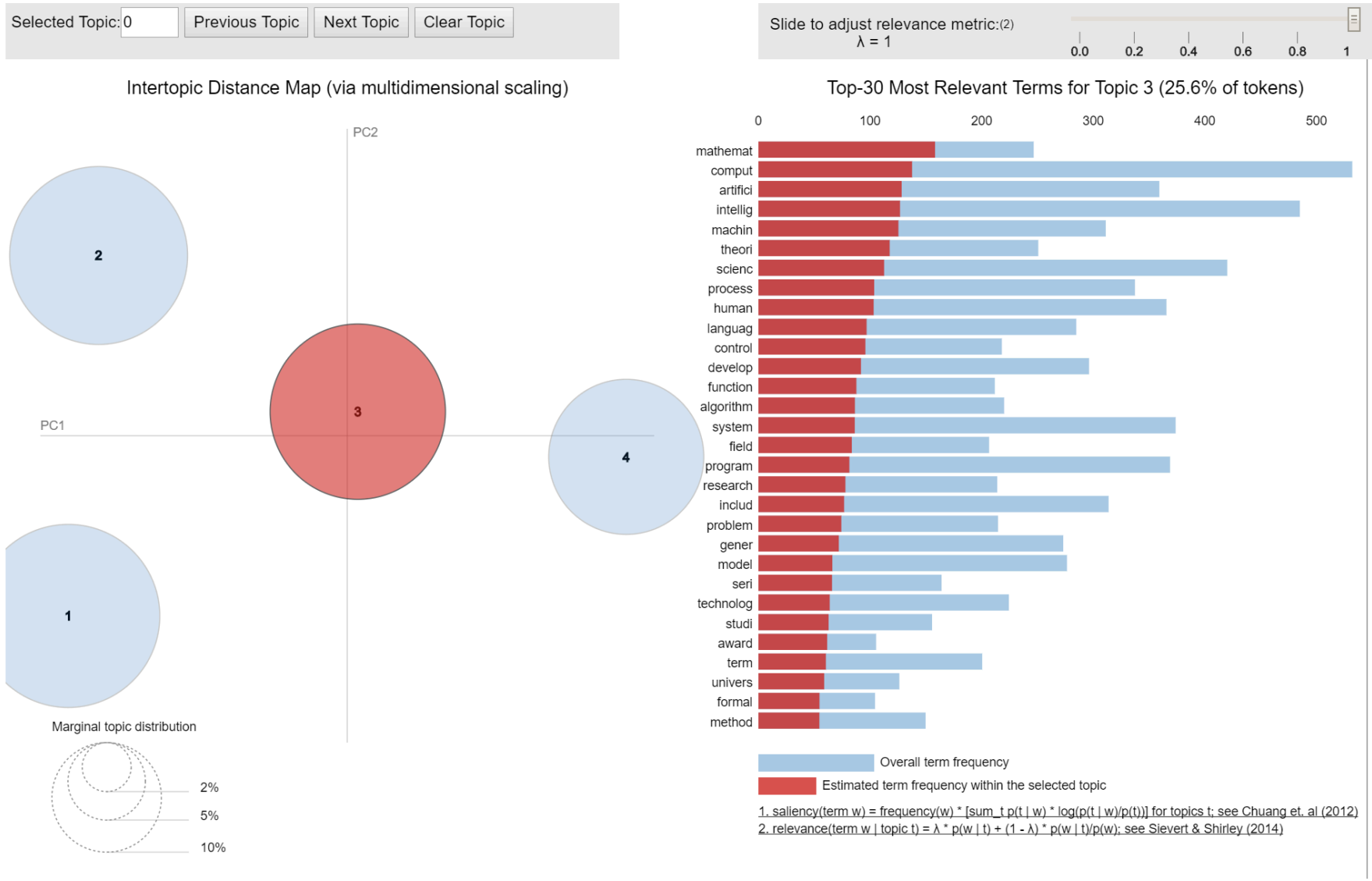
Out[29]:


Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (27.9% of tokens)
0
100
200
300
400
500
comput
softwar
system
program
develop
process
applic
languag
includ
engin
design
user
scienc
hardwar
technolog
base
gener
model
inform
integr
specif
digit
interact
googl
human
problem
field
algorithm
commun
intellig
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

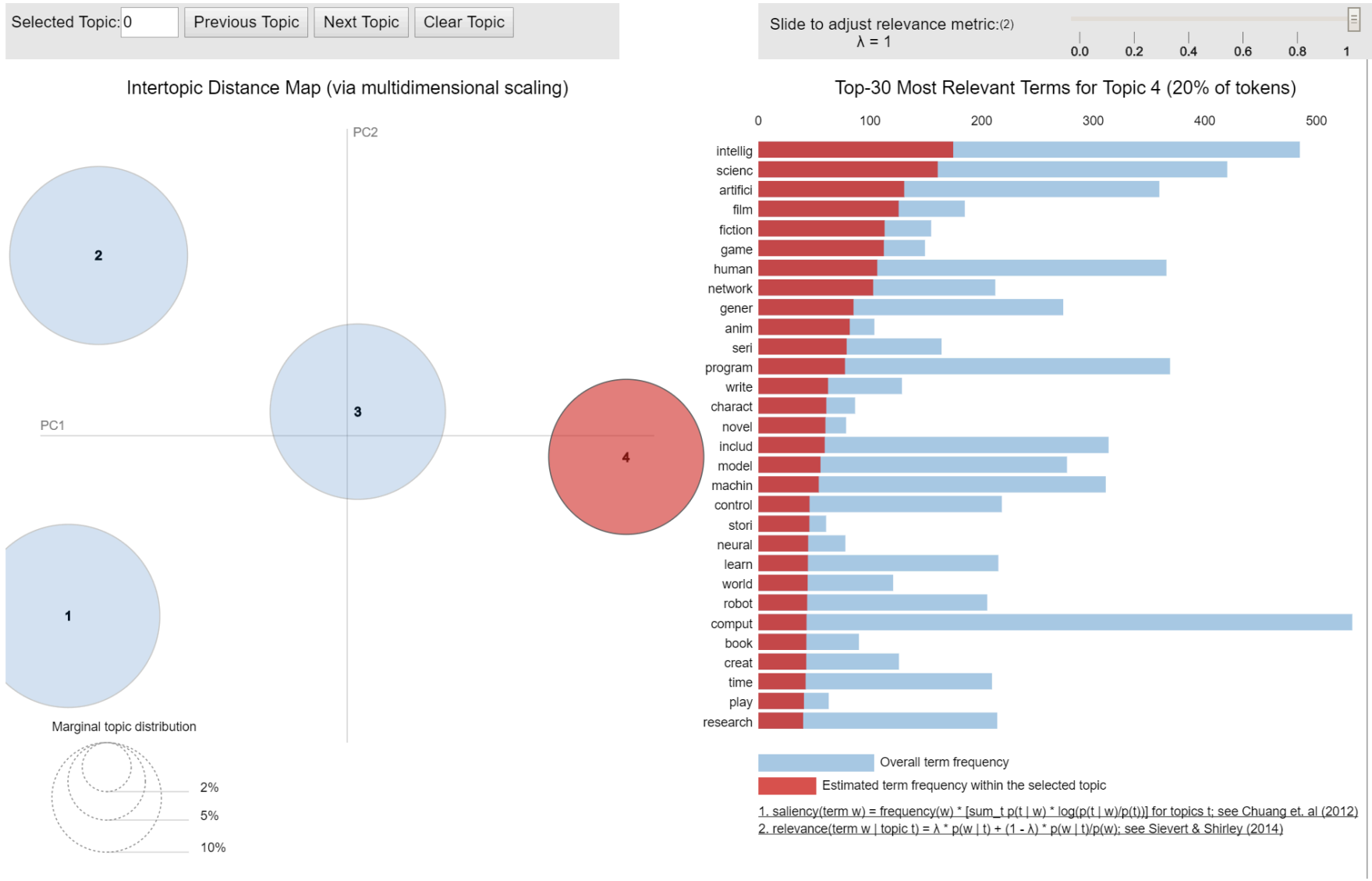
Out[29]:


Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 2 (26.4% of tokens)
0
100
200
300
400
500
data
inform
intellig
robot
human
learn
machin
system
process
model
base
agent
theori
oper
type
object
applic
includ
problem
structur
program
time
scienc
exampl
class
term
technolog
develop
languag
function
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Out[29]:


Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 3 (25.6% of tokens)
0
100
200
300
400
500
mathemat
comput
artifici
intellig
machin
theori
scienc
process
human
languag
control
develop
function
algorithm
system
field
program
research
includ
problem
gener
model
seri
technolog
studi
award
term
univers
formal
method
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Out[29]:


Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 4 (20% of tokens)
0
100
200
300
400
500
intellig
scienc
artifici
film
fiction
game
human
network
gener
anim
seri
program
write
charact
novel
includ
model
machin
control
stori
neural
learn
world
robot
comput
book
creat
time
play
research
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

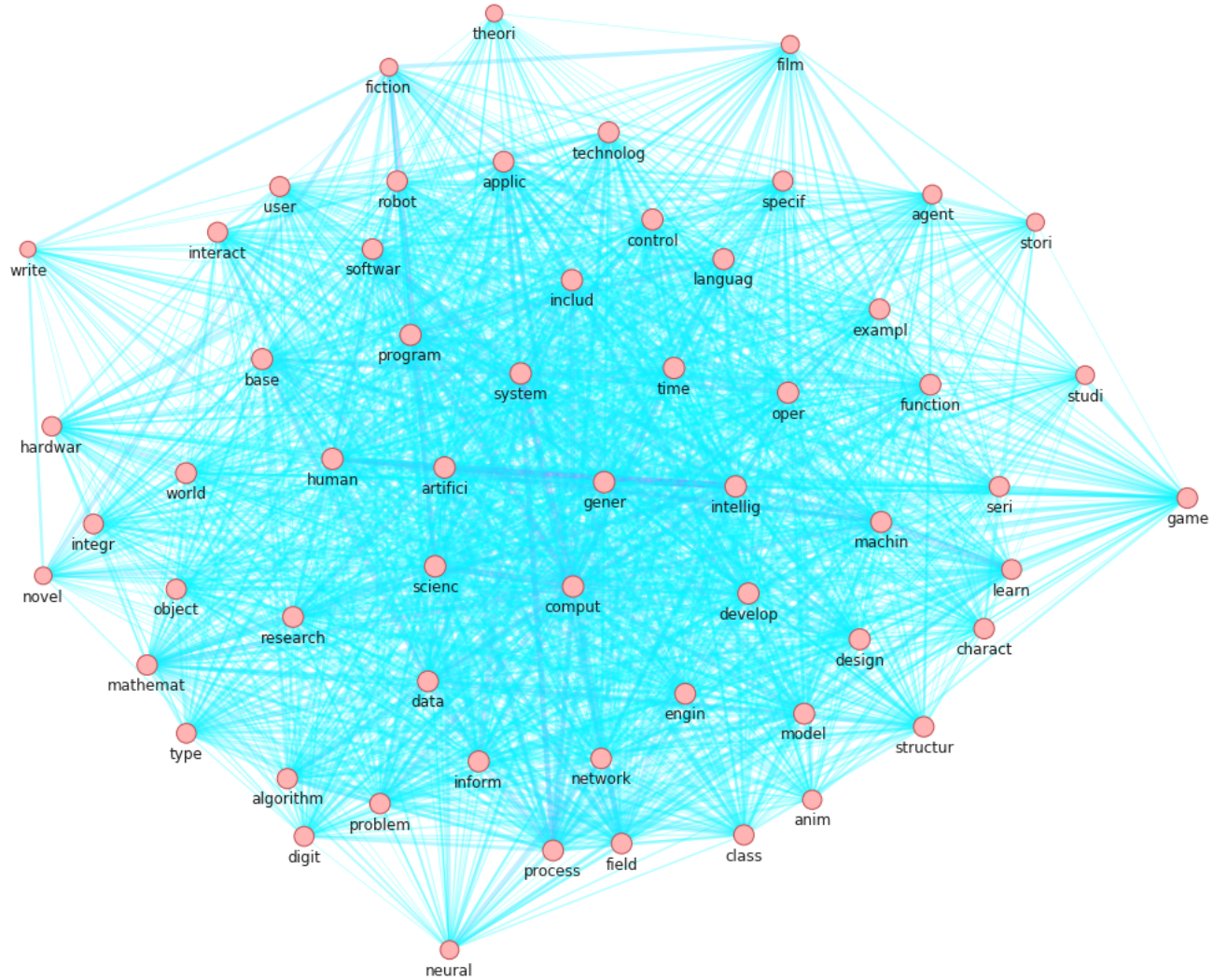
Visualization of Sententially Co-occurrent terms

['agent',
 'algorithm',
 'anim',
 'applic',
 'artifici',
 'base',
 'charact',
 'class',
 'comput',
 'control',
 'data',
 'design',
 'develop',
 …
'problem',
 'process',
 'program',
 'research',
 'robot',
 'scienc',
 'seri',
 'softwar',
 'specif',
 'stori',
 'structur',
 'studi',
 'system',
 'technolog',
 'theori',
 'time',
 'type',
 'user',
 'world',
 'write']

Preparing for Visualization:

- Among 615 nodes, in total of 60 words without aliases cooccur in at least 20 pages.
- Sentiment scores (polarity score) divided by their average. -1 to 1

The graph of sententially co-occurrent TM terms
in the 615 Wikipedia hyperlinks of 'Outline of computer science and Outline of artificial intelligence'
weighted over their average sentiment score



Graph Result:

- The graph is weighted
- The graph is a connected graph
- It has 59 nodes and 1536 edges.
- The term that does not cooccur is ‘googl’
- The density of this graph is 0.898

Possible Future Works:

Applying the similar analysis to different Wikipedia outlines
Applying the similar analysis on other wiki sources
Finding more efficient ways to conduct analysis

Thank You