T-61.6060 Presentation

Tag and document relations

Aki Saarinen, Laszlo Kozma, Stevan Keraudy

26.2.2008

Objectives

- Grouping of similar documents together
- Grouping of similar tags together

Outline of the methodology

1. Fetching of data and saving it to XML-files
2. Preprocessing + feature extraction with a Python script
3. (Dimension reduction)
4. Visualization with Self-Organizing Maps
5. Tag prediction with various methods

Data source: Newsvine.com

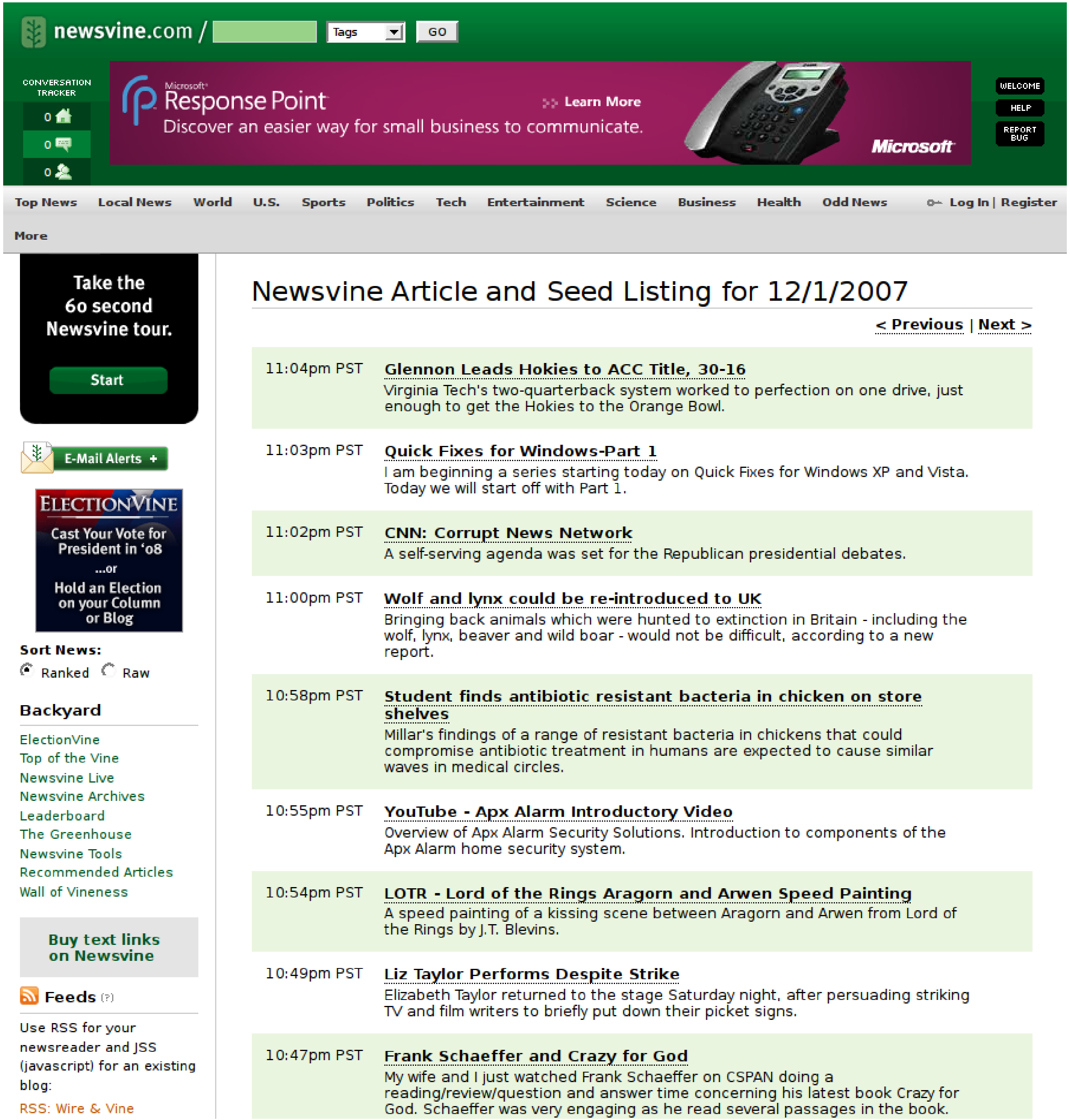
- Newsvine.com is a community news website
- Created in 2005
- Based in Seattle
- Owned by msnbc.com
- Articles can be written by anyone
- Any kind of topic can be covered
- Around 1000 articles per day

1. Data Fetching

- Retrieve all articles from December 2007
- Script to retrieve data written in Java
- Store data into XML files
- 1 XML file per article

Fetching algorithm

```
For each day
    Go to page that lists all articles
    Get the list of articles
    For each article
        Get meta information
        Go to the article page
        Get tags, category and text
        If article not empty (to reduce spam)
            Write information to XML file
```



Newsvine Article and Seed Listing for 12/1/2007

< Previous | Next >

11:04pm PST **Glennon Leads Hokies to ACC Title, 30-16**
Virginia Tech's two-quarterback system worked to perfection on one drive, just enough to get the Hokies to the Orange Bowl.

11:03pm PST **Quick Fixes for Windows-Part 1**
I am beginning a series starting today on Quick Fixes for Windows XP and Vista. Today we will start off with Part 1.

11:02pm PST **CNN: Corrupt News Network**
A self-serving agenda was set for the Republican presidential debates.

11:00pm PST **Wolf and lynx could be re-introduced to UK**
Bringing back animals which were hunted to extinction in Britain - including the wolf, lynx, beaver and wild boar - would not be difficult, according to a new report.

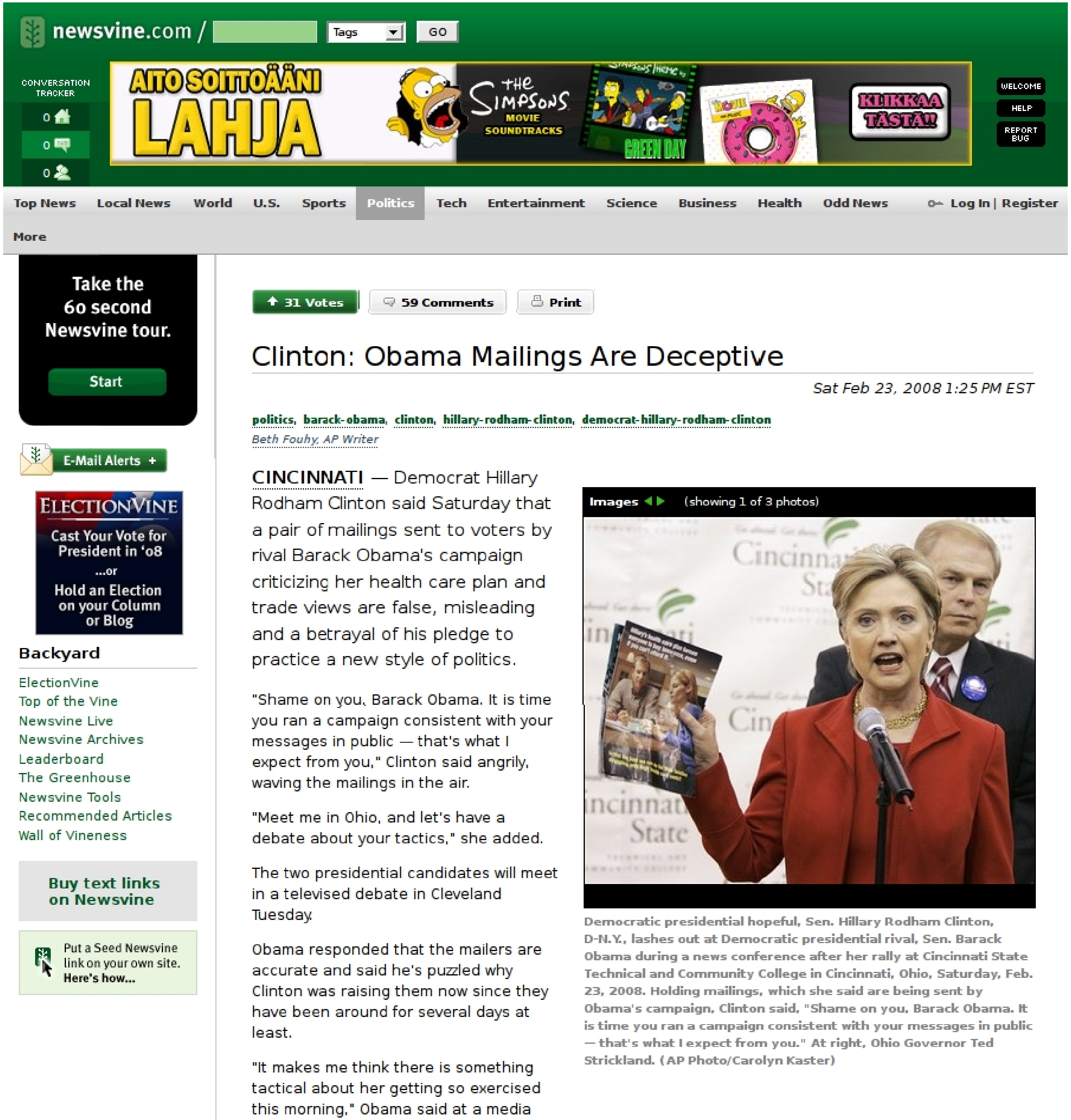
10:58pm PST **Student finds antibiotic resistant bacteria in chicken on store shelves**
Millar's findings of a range of resistant bacteria in chickens that could compromise antibiotic treatment in humans are expected to cause similar waves in medical circles.

10:55pm PST **YouTube - Apx Alarm Introductory Video**
Overview of Apx Alarm Security Solutions. Introduction to components of the Apx Alarm home security system.

10:54pm PST **LOTR - Lord of the Rings Aragorn and Arwen Speed Painting**
A speed painting of a kissing scene between Aragorn and Arwen from Lord of the Rings by J.T. Blevins.

10:49pm PST **Liz Taylor Performs Despite Strike**
Elizabeth Taylor returned to the stage Saturday night, after persuading striking TV and film writers to briefly put down their picket signs.

10:47pm PST **Frank Schaeffer and Crazy for God**
My wife and I just watched Frank Schaeffer on CSPAN doing a reading/review/question and answer time concerning his latest book Crazy for God. Schaeffer was very engaging as he read several passages in the book.



Clinton: Obama Mailings Are Deceptive

Sat Feb 23, 2008 1:25 PM EST

politics, barack-obama, clinton, hillary-rodham-clinton, democrat-hillary-rodham-clinton

Beth Fouhy, AP Writer

CINCINNATI — Democrat Hillary Rodham Clinton said Saturday that a pair of mailings sent to voters by rival Barack Obama's campaign criticizing her health care plan and trade views are false, misleading and a betrayal of his pledge to practice a new style of politics.

"Shame on you, Barack Obama. It is time you ran a campaign consistent with your messages in public — that's what I expect from you," Clinton said angrily, waving the mailings in the air.

"Meet me in Ohio, and let's have a debate about your tactics," she added.

The two presidential candidates will meet in a televised debate in Cleveland Tuesday.

Obama responded that the mailers are accurate and said he's puzzled why Clinton was raising them now since they have been around for several days at least.

"It makes me think there is something tactical about her getting so exercised this morning," Obama said at a media

Democratic presidential hopeful, Sen. Hillary Rodham Clinton, D-N.Y., lashes out at Democratic presidential rival, Sen. Barack Obama during a news conference after her rally at Cincinnati State Technical and Community College in Cincinnati, Ohio, Saturday, Feb. 23, 2008. Holding mailings, which she said are being sent by Obama's campaign, Clinton said, "Shame on you, Barack Obama. It is time you ran a campaign consistent with your messages in public — that's what I expect from you." At right, Ohio Governor Ted Strickland. (AP Photo/Carolyn Kaster)

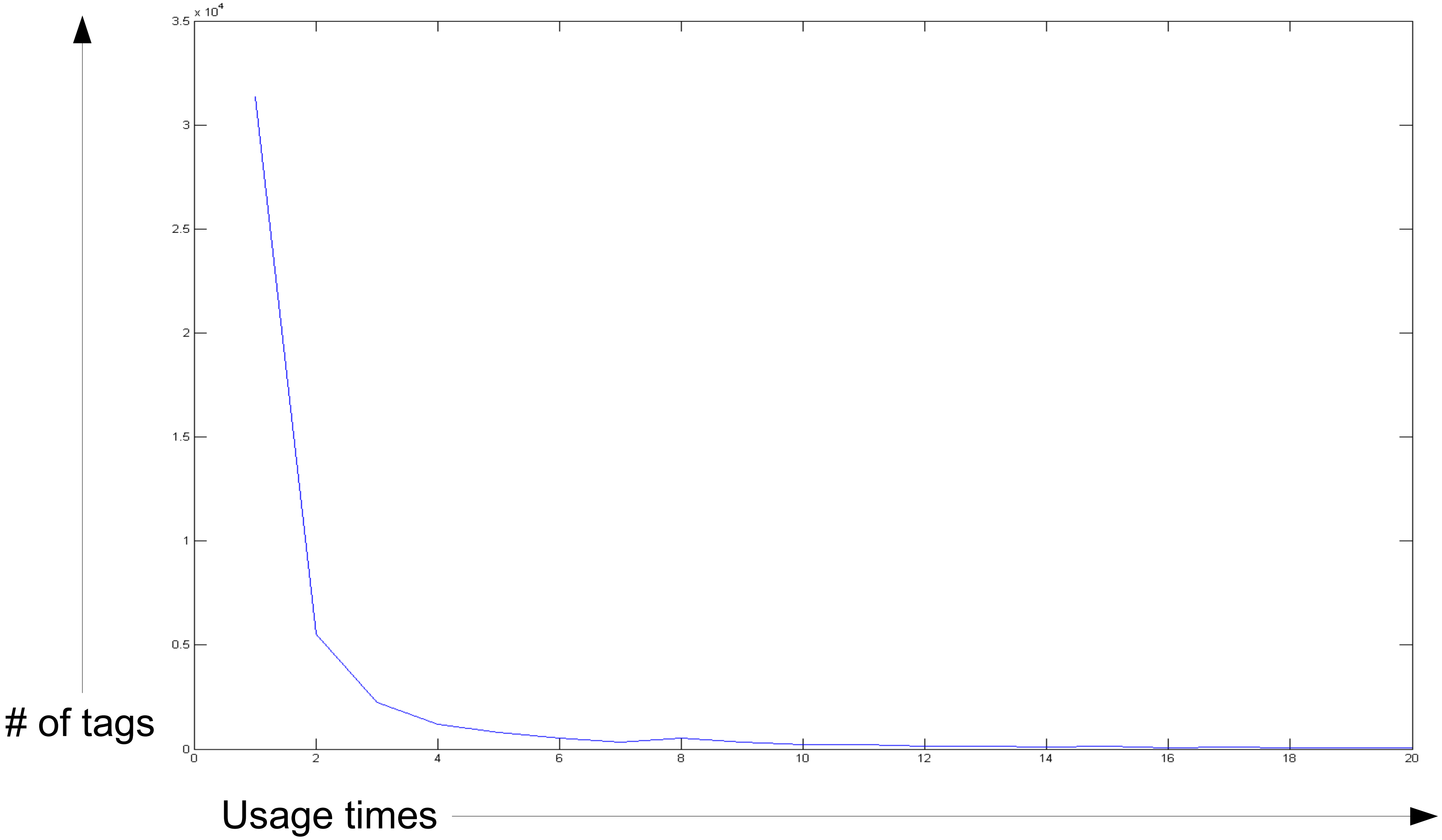
Information retrieved

- For each document, we retrieved:
    - Date and time
    - Title
    - Summary
    - Tags
    - Category
    - Text
- We basically only used the tags for analysis, other information could have been used too but we decided to leave it out because of a relatively tight schedule

Figures about the collected data

- **Newsvine.com**
  - 31 days: 2007 December 1st to 31st
  - 27 779 documents -> average 896 docs/day
  - 45 039 unique tags
  - 31 360 tags used only once

Tag histogram for Newsvine



Figures about the collected data

- As a curiosity, we also fetched data from Slashdot, even though most of the analysis was done for Newsvine
- **Slashdot.org**
  - 6 months (2007-07 to 2007-12)
  - 4017 documents -> average 22 docs / day
  - 5123 unique tags
  - 4052 tags used only once

  - *Free tip: don't fetch ~1000 pages/day from a single machine or you will get banned :)*

2. Preprocessing and feature extraction

- Read data from XML-files
  - Drop tags which are used less than N times
  - Ignore option for the most general tags (categories)

- Generate 2 matrices to be processed with Matlab
  - Tags used in documents (docs=rows, tags=columns)
  - Tags occurring together (tags=rows, tags=columns)

Generated matrices

- Document-tag (=tags used in documents)

	Gw-bush	Elections	Cooking	Internet	Web2.0	...
20071201-1	1	1	0	0	0	
20071201-2	0	0	0	1	1	
...						

- Tag-tag (=tags occurring together)

	Gw-bush	Elections	Cooking	Internet	Web2.0	...
Gw-bush	600	100	1	2	0	
Elections	100	520	0	10	0	
Cooking	1	0	80	0	0	
Internet	2	10	0	400	200	
Web2.0	0	0	0	200	350	
...						

Normalization for tag-tag-matrix

- Original matrix

	Gw-bush	Elections	Cooking	Internet	Web2.0	...
Gw-bush	600	100	1	2	0	
Elections	100	520	0	10	0	
Cooking	1	0	80	0	0	
Internet	2	10	0	400	200	
Web2.0	0	0	0	200	350	
...						

- Each cell divided by total count of matching tag

	Gw-bush	Elections	Cooking	Internet	Web2.0	...
Gw-bush	600/600	100/600	1/600	2/600	0	
Elections	100/520	520/520	0	10/520	0	
Cooking	1/80	0	80/80	0	0	
Internet	2/400	10/400	0	400/400	200/400	
Web2.0	0	0	0	200/350	350/350	
...						

3. Dimension reduction

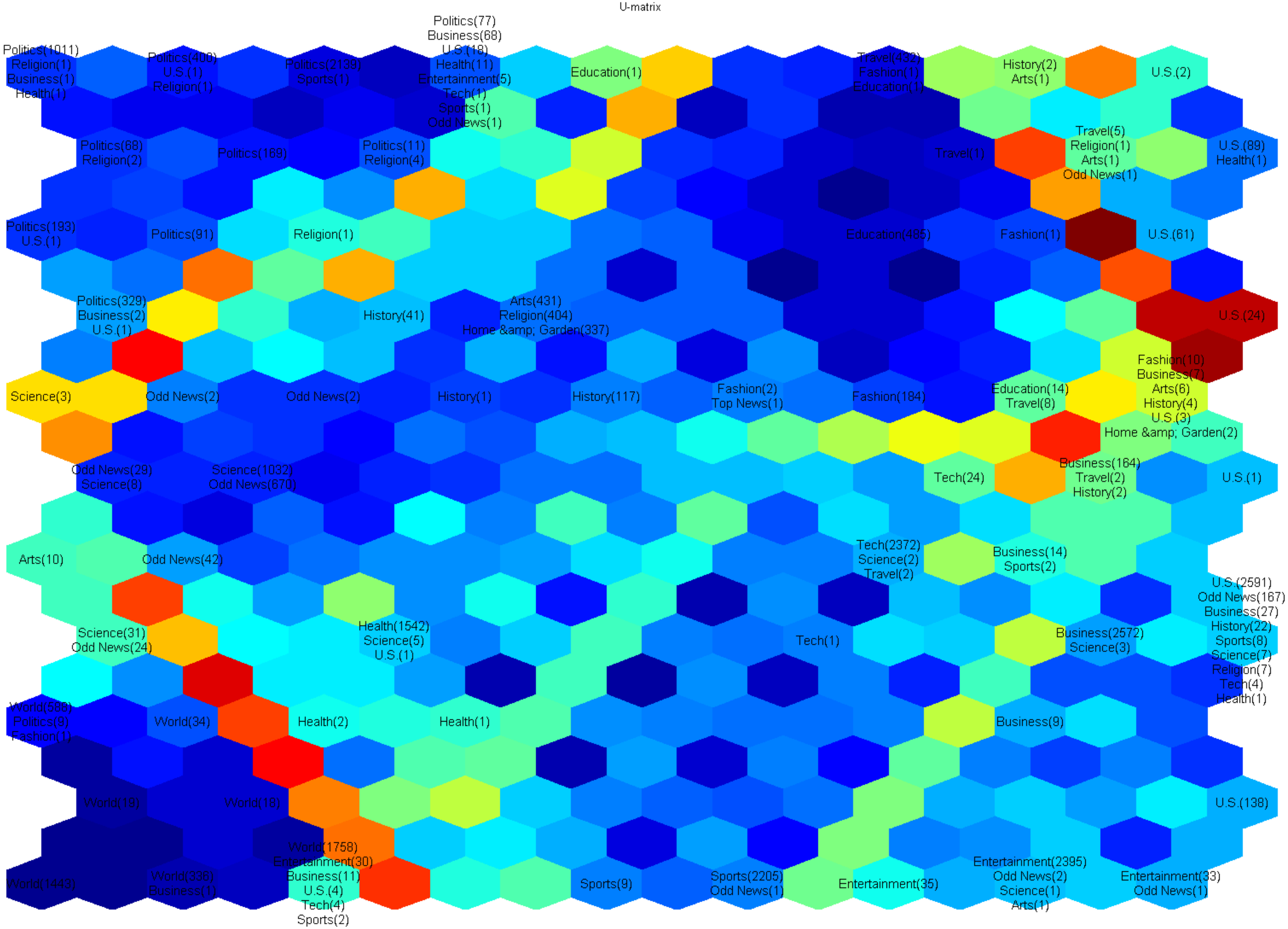
- Generated document-tag-matrices are quite big
  - Document-tag-matrix size for whole Dec 2007 is 27779 x 45039
  - Need for dimension reduction
- Achieved using Principle Component Analysis (PCA)
  - We used for example top-100 or top-200 components
  - SOM could do it too, but takes a very long time to process, PCA is convenient
  - Even PCA is not doable for the whole month, half a month used

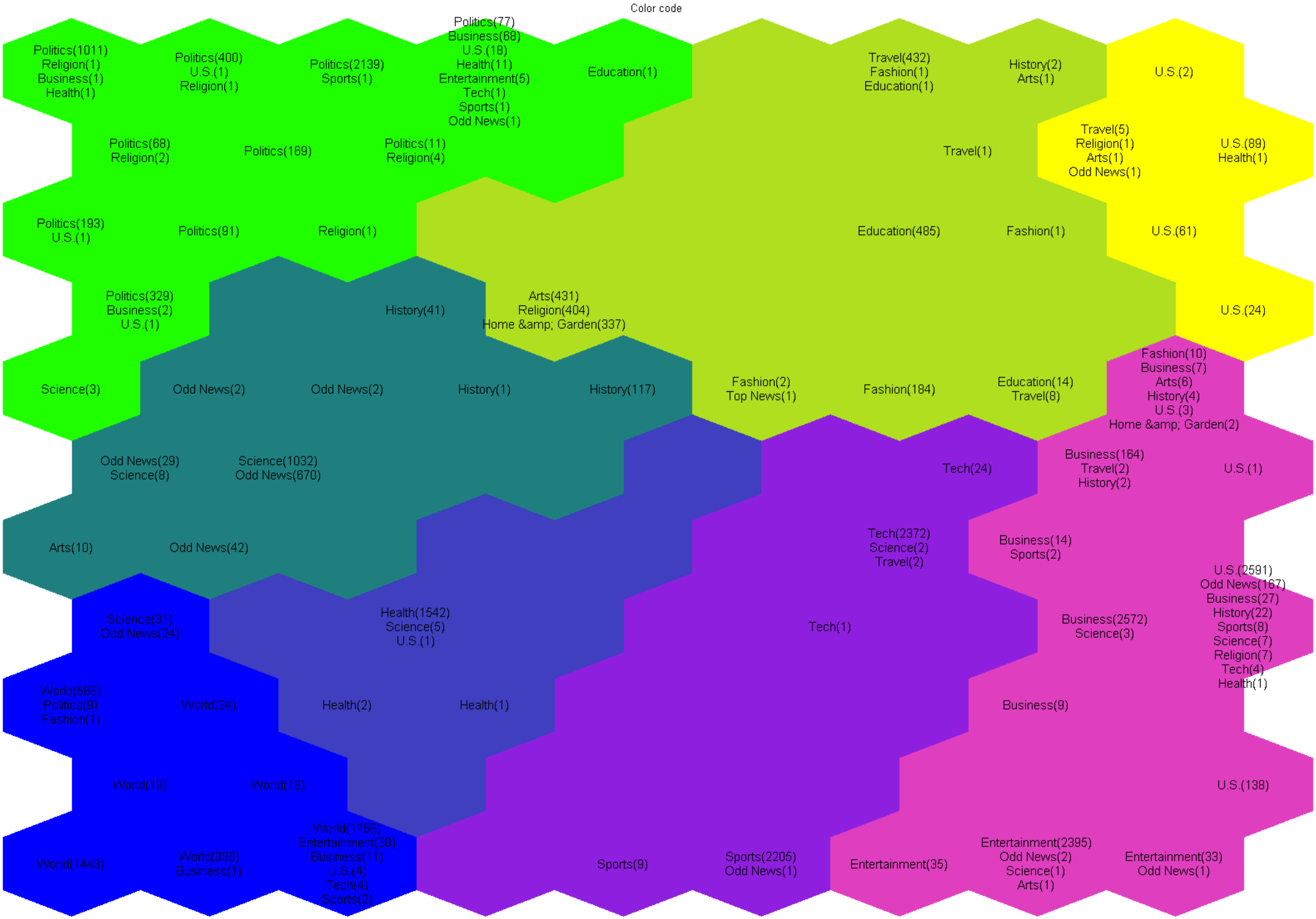
4. Visualization with SOMs

- Self-Organizing Maps
  - Artificial neural networks that can map multidimensional data to (usually) 2 dimensions
  - Map seeks to preserve topological properties of the input space
  - Results can be inspected visually
  - U-matrices and k-means clustering shown

SOM #1

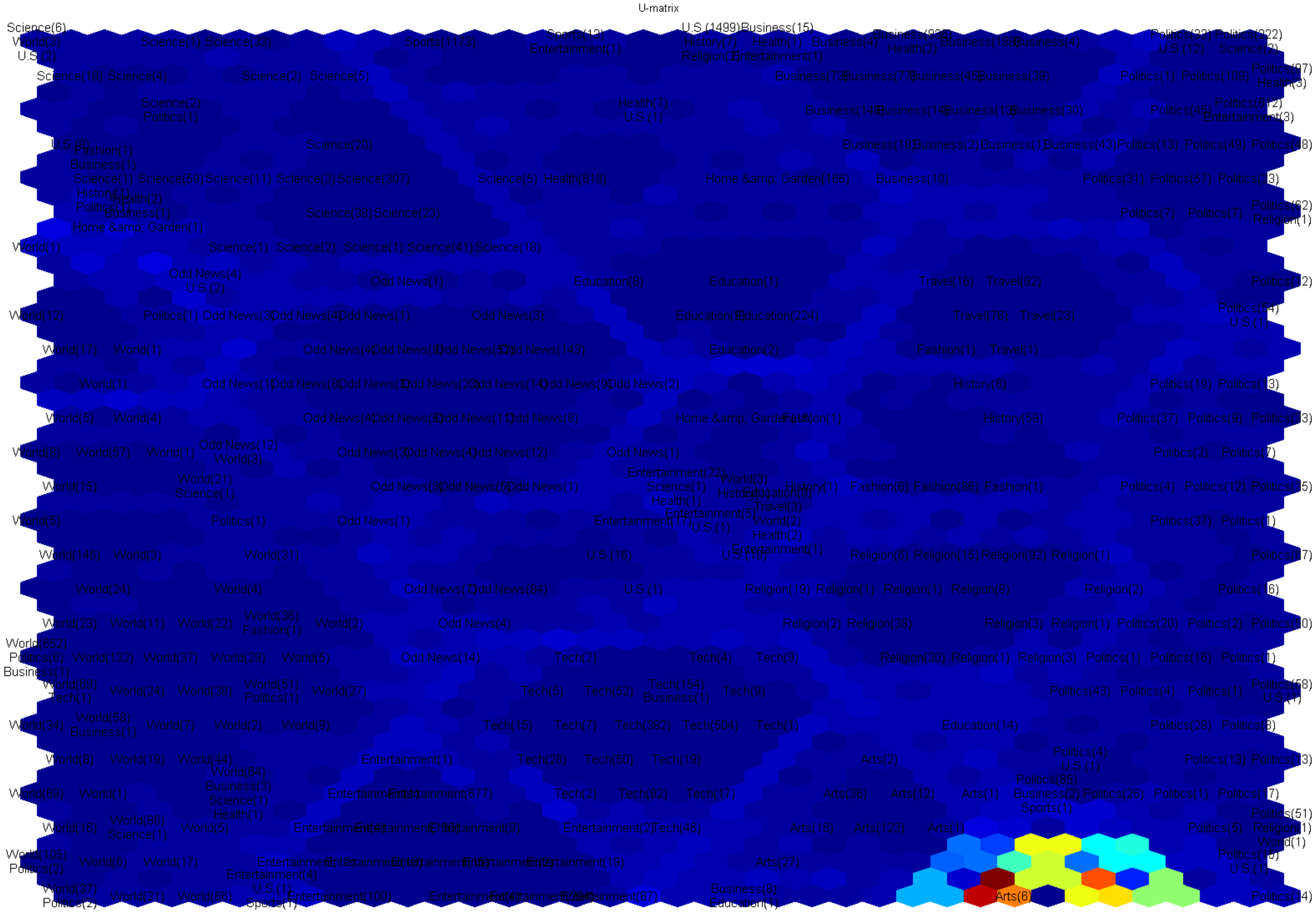
- Data from Newsvine.com
  - Whole December 2007
- Document-tag-matrix used as input to SOM
  - Only tags which occur >= 40 times are used
  - No ignore list (category tags are there, too)
  - 27779 documents, 423 tags
  - No need to do PCA (matrix size: 27779 x 423)
  - Labeled with categories from the site

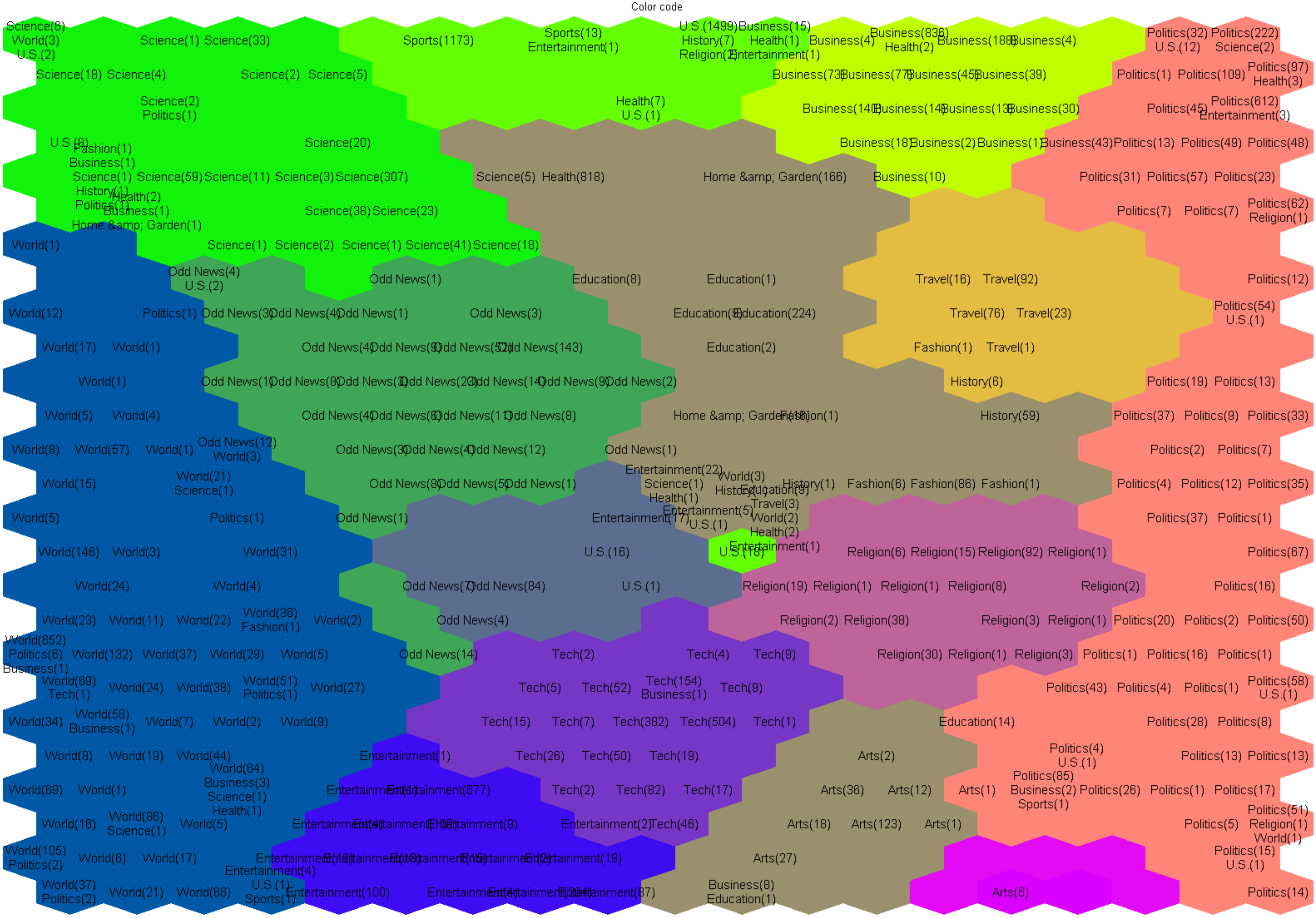

U-matrix
Politics(1011)
Religion(1)
Business(1)
Health(1)
Politics(400)
U.S.(1)
Religion(1)
Politics(2138)
Sports(1)
Politics(77)
Business(68)
U.S.(18)
Health(11)
Entertainment(5)
Tech(1)
Sports(1)
Odd News(1)
Education(1)
Travel(432)
Fashion(1)
Education(1)
History(2)
Arts(1)
U.S.(2)
Politics(68)
Religion(2)
Politics(169)
Politics(11)
Religion(4)
Travel(1)
Travel(5)
Religion(1)
Arts(1)
Odd News(1)
U.S.(89)
Health(1)
Politics(193)
U.S.(1)
Politics(91)
Religion(1)
Education(485)
Fashion(1)
U.S.(61)
Politics(329)
Business(2)
U.S.(1)
History(41)
Arts(431)
Religion(404)
Home &amp; Garden(337)
U.S.(24)
Fashion(10)
Business(7)
Arts(6)
History(4)
U.S.(3)
Home &amp; Garden(2)
Science(3)
Odd News(2)
Odd News(2)
History(1)
History(117)
Fashion(2)
Top News(1)
Fashion(184)
Education(14)
Travel(8)
Odd News(29)
Science(8)
Science(1032)
Odd News(670)
Tech(24)
Business(164)
Travel(2)
History(2)
U.S.(1)
Arts(10)
Odd News(42)
Tech(2372)
Science(2)
Travel(2)
Business(14)
Sports(2)
U.S.(2591)
Odd News(167)
Business(27)
History(22)
Sports(8)
Science(7)
Religion(7)
Tech(4)
Health(1)
Science(31)
Odd News(24)
Health(1542)
Science(5)
U.S.(1)
Tech(1)
Business(2572)
Science(3)
World(586)
Politics(9)
Fashion(1)
World(34)
Health(2)
Health(1)
Business(9)
World(19)
World(18)
U.S.(138)
World(1758)
Entertainment(30)
Business(11)
U.S.(4)
Tech(4)
Sports(2)
World(1443)
World(338)
Business(1)
Sports(9)
Sports(2205)
Odd News(1)
Entertainment(35)
Entertainment(2395)
Odd News(2)
Science(1)
Arts(1)
Entertainment(33)
Odd News(1)


Color code
Politics(1011) Religion(1) Business(1) Health(1)
Politics(400) U.S.(1) Religion(1)
Politics(2139) Sports(1)
Politics(77) Business(68) U.S.(18) Health(11) Entertainment(5) Tech(1) Sports(1) Odd News(1)
Education(1)
Travel(432) Fashion(1) Education(1)
History(2) Arts(1)
U.S.(2)
Politics(68) Religion(2)
Politics(169)
Politics(11) Religion(4)
Travel(1)
Travel(5) Religion(1) Arts(1) Odd News(1)
U.S.(89) Health(1)
Politics(193) U.S.(1)
Politics(91)
Religion(1)
Education(485)
Fashion(1)
U.S.(61)
Politics(329) Business(2) U.S.(1)
History(41)
Arts(431) Religion(404) Home &amp; Garden(337)
U.S.(24)
Science(3)
Odd News(2)
Odd News(2)
History(1)
History(117)
Fashion(2) Top News(1)
Fashion(184)
Education(14) Travel(8)
Fashion(10) Business(7) Arts(6) History(4) U.S.(3) Home &amp; Garden(2)
Odd News(29) Science(8)
Science(1032) Odd News(670)
Tech(24)
Business(164) Travel(2) History(2)
U.S.(1)
Arts(10)
Odd News(42)
Tech(2372) Science(2) Travel(2)
Business(14) Sports(2)
U.S.(2591) Odd News(167) Business(27) History(22) Sports(8) Science(7) Religion(7) Tech(4) Health(1)
Science(31) Odd News(24)
Health(1542) Science(5) U.S.(1)
Tech(1)
Business(2572) Science(3)
World(588) Politics(9) Fashion(1)
World(34)
Health(2)
Health(1)
Business(9)
World(19)
World(18)
U.S.(138)
World(1443)
World(336) Business(1)
World(1758) Entertainment(30) Business(11) U.S.(4) Tech(4) Sports(2)
Sports(9)
Sports(2205) Odd News(1)
Entertainment(35)
Entertainment(2395) Odd News(2) Science(1) Arts(1)
Entertainment(33) Odd News(1)

SOM #2

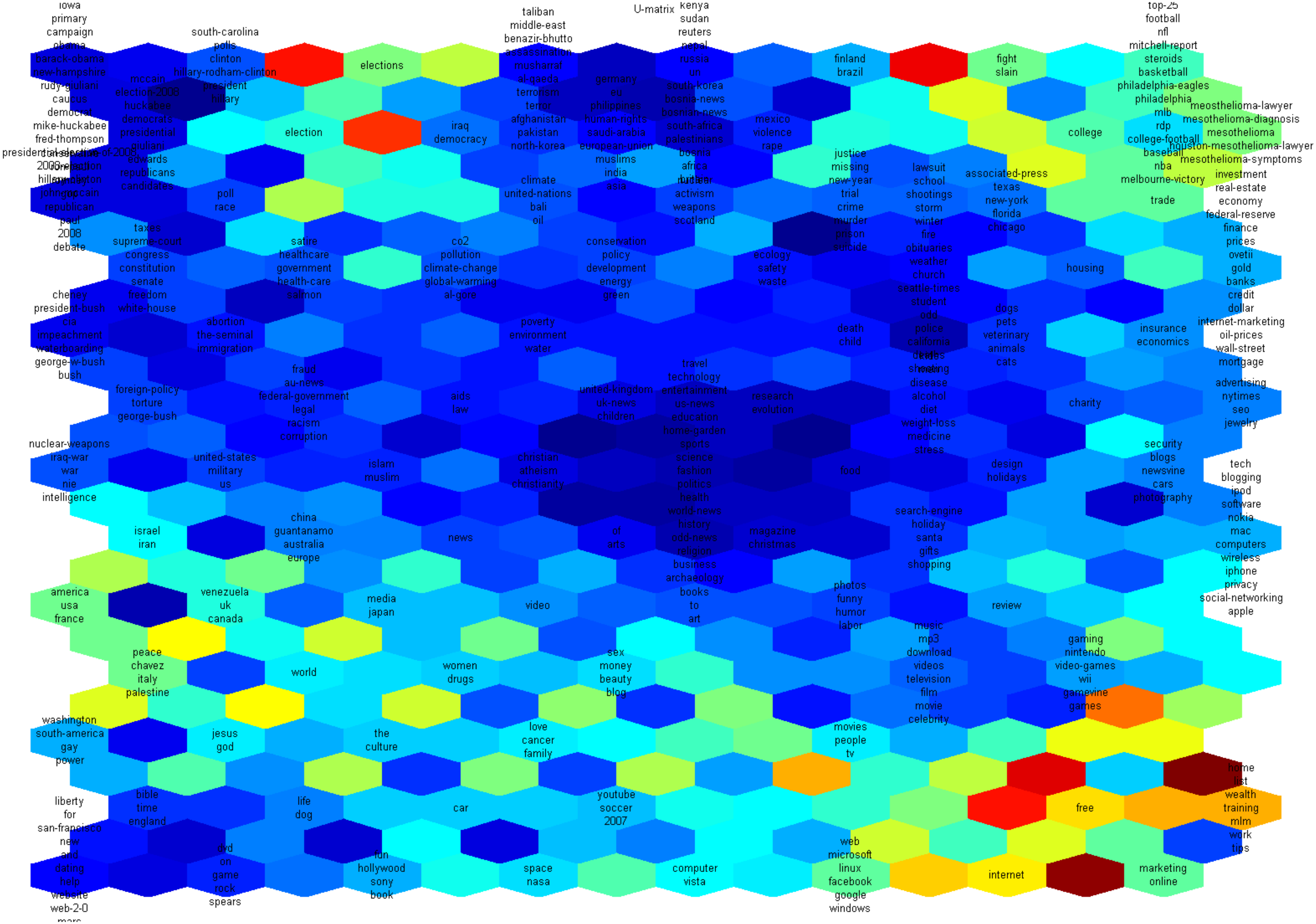
- Data from Newsvine.com
  - December 1st to December 15th
- Document-tag-matrix used as input to SOM
  - Only tags which occur more than 3 times are used
  - No ignore list (category tags are there, too)
  - 14679 documents, 4951 tags
  - Need to reduce with PCA, top-100 components used
  - Labeled with categories from the site

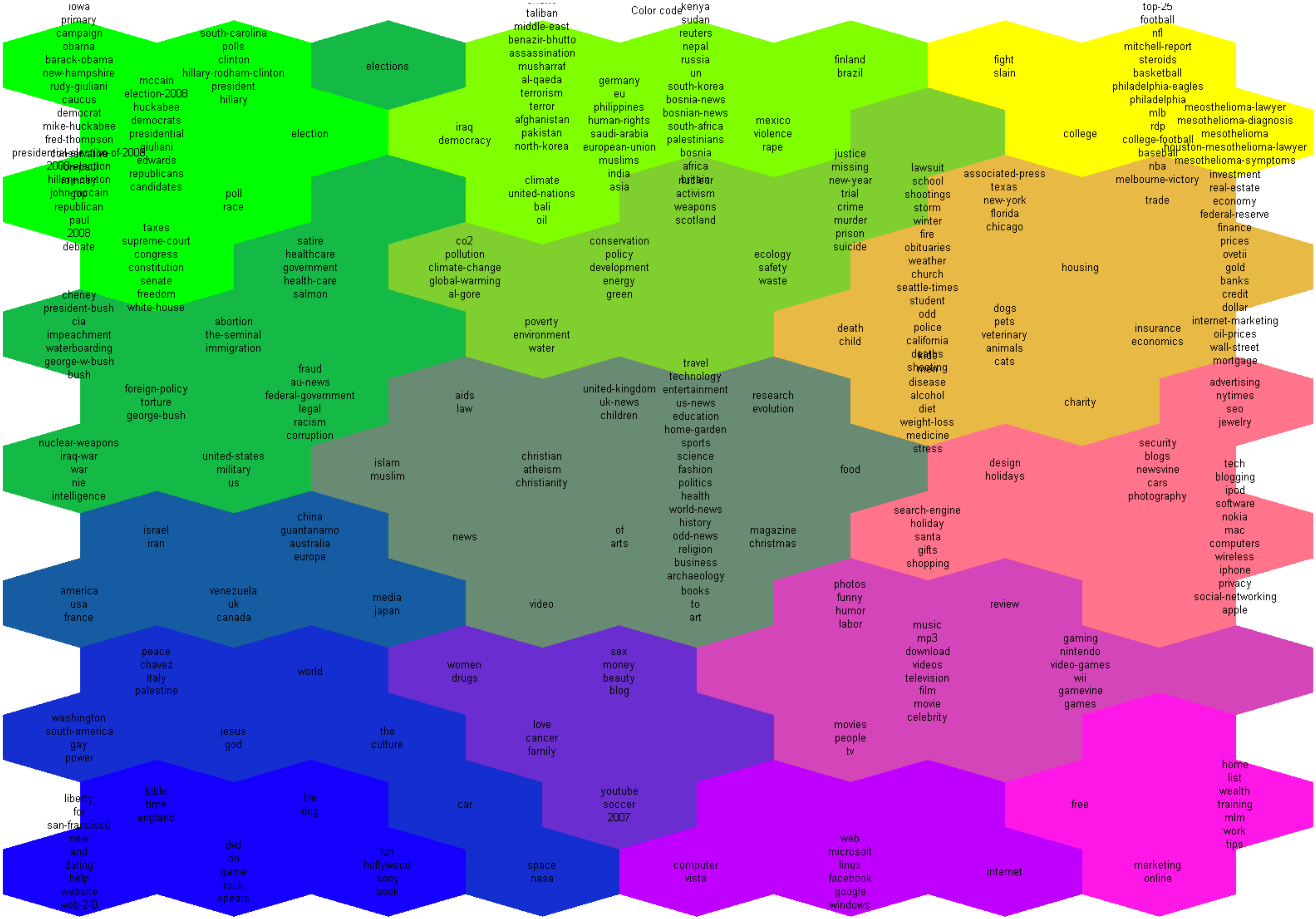

U-matrix
Science(6) World(3) U.S.(2)
Science(1) Science(33)
Sports(1173)
Sports(13) Entertainment(1)
U.S.(1499) Business(15) History(7) Health(1) Religion(2) Entertainment(1)
Business(4)
Business(326) Health(2)
Business(186) Business(4)
Politics(32) U.S.(12)
Politics(222) Science(2)
Science(18) Science(4)
Science(2) Science(5)
Business(73) Business(77) Business(45) Business(39)
Politics(1) Politics(109)
Politics(87) Health(3)
Science(2) Politics(1)
Health(7) U.S.(1)
Business(14) Business(14) Business(13) Business(30)
Politics(46)
Politics(612) Entertainment(3)
U.S.(8) Fashion(1) Business(1)
Science(20)
Business(18) Business(2) Business(1) Business(43) Politics(13) Politics(49) Politics(8)
Science(1) Science(69) Science(11) Science(3) Science(307)
Science(5) Health(818)
Home &amp; Garden(166)
Business(10)
Politics(31) Politics(57) Politics(3)
History(1) Politics(1) Health(2) Business(1)
Science(39) Science(23)
Politics(7) Politics(7)
Politics(62) Religion(1)
Home &amp; Garden(1)
World(1)
Science(1) Science(2) Science(1) Science(41) Science(18)
Odd News(4) U.S.(2)
Odd News(1)
Education(8)
Education(1)
Travel(18) Travel(92)
Politics(12)
World(12)
Politics(1) Odd News(3) Odd News(4) Odd News(1)
Odd News(3)
Education(6) Education(224)
Travel(76) Travel(28)
Politics(54) U.S.(1)
World(17) World(1)
Odd News(4) Odd News(9) Odd News(52) Odd News(143)
Education(2)
Fashion(1) Travel(1)
World(1)
Odd News(1) Odd News(8) Odd News(2) Odd News(23) Odd News(14) Odd News(9) Odd News(2)
History(6)
Politics(19) Politics(13)
World(5) World(4)
Odd News(4) Odd News(6) Odd News(11) Odd News(8)
Home &amp; Garden(8) Religion(1)
History(59)
Politics(37) Politics(9) Politics(33)
World(8) World(57) World(1)
Odd News(12) World(3)
Odd News(3) Odd News(4) Odd News(12)
Odd News(1)
Politics(2) Politics(7)
World(15)
World(21) Science(1)
Odd News(8) Odd News(5) Odd News(1)
Entertainment(22) Science(1) Health(1)
World(3) History(1) Education(3) Travel(3)
History(1)
Fashion(6) Fashion(86) Fashion(1)
Politics(4) Politics(12) Politics(35)
World(5)
Politics(1)
Odd News(1)
Entertainment(17)
Entertainment(5) U.S.(1)
World(2) Health(2)
Politics(37) Politics(1)
World(146) World(3)
World(31)
U.S.(16)
U.S.(18) Entertainment(1)
Religion(6) Religion(15) Religion(92) Religion(1)
Politics(67)
World(24)
World(4)
Odd News(7) Odd News(84)
U.S.(1)
Religion(19) Religion(1) Religion(1) Religion(8)
Religion(2)
Politics(16)
World(23) World(11) World(22)
World(38) Fashion(1)
World(2)
Odd News(4)
Religion(2) Religion(38)
Religion(3) Religion(1) Politics(20) Politics(2) Politics(10)
World(652) Politics(6) Business(1)
World(132) World(37) World(29) World(5)
Odd News(14)
Tech(2)
Tech(4) Tech(9)
Religion(30) Religion(1) Religion(3) Politics(1) Politics(16) Politics(1)
World(63) Tech(1)
World(24) World(38)
World(51) Politics(1)
World(27)
Tech(5) Tech(52)
Tech(154) Business(1)
Tech(9)
Politics(43) Politics(4) Politics(1)
Politics(58) U.S.(1)
World(34)
World(58) Business(1)
World(7) World(2) World(8)
Tech(15) Tech(7) Tech(382) Tech(504) Tech(1)
Education(14)
Politics(28) Politics(8)
World(8) World(19) World(44)
Entertainment(1)
Tech(26) Tech(50) Tech(19)
Arts(2)
Politics(4) U.S.(1)
Politics(19) Politics(13)
World(69) World(1)
World(64) Business(3) Science(1) Health(1)
Entertainment(1) Entertainment(677)
Tech(2) Tech(82) Tech(17)
Arts(38) Arts(12) Arts(1)
Politics(85) Business(2) Sports(1)
Politics(26) Politics(1) Politics(17)
World(16)
World(86) Science(1)
World(5)
Entertainment(4) Entertainment(18) Entertainment(9)
Entertainment(2) Tech(46)
Arts(18) Arts(123) Arts(1)
Politics(51) Politics(5) Religion(1) World(1)
World(105) Politics(2)
World(8) World(17)
Entertainment(4)
Entertainment(2) Entertainment(3) Entertainment(15) Entertainment(9) Entertainment(19)
Arts(27)
Politics(19) U.S.(1)
World(37) Politics(2)
World(21) World(66)
U.S.(1) Sports(1)
Entertainment(100)
Entertainment(4) Entertainment(5) Entertainment(87)
Business(8) Education(1)
Arts(6)
Politics(14)


Color code
Science(6) World(3) U.S.(2)
Science(1)
Science(33)
Sports(1173)
Sports(13) Entertainment(1)
U.S.(1499) History(7) Religion(2)
Business(15) Health(1) Entertainment(1)
Business(4)
Business(836) Health(2)
Business(188)
Business(4)
Politics(32) U.S.(12)
Politics(222) Science(2)
Science(18)
Science(4)
Science(2)
Science(5)
Business(73)
Business(77)
Business(45)
Business(39)
Politics(1)
Politics(109)
Politics(97) Health(3)
Science(2) Politics(1)
Health(7) U.S.(1)
Business(140)
Business(14)
Business(13)
Business(30)
Politics(45)
Politics(612) Entertainment(3)
U.S.(8)
Fashion(1) Business(1)
Science(20)
Business(18)
Business(2)
Business(1)
Business(43)
Politics(13)
Politics(49)
Politics(48)
Science(1)
Science(59)
Science(11)
Science(3)
Science(307)
Science(5)
Health(818)
Home &amp; Garden(166)
Business(10)
Politics(31)
Politics(57)
Politics(23)
History(1) Politics(1)
Health(2) Business(1)
Home &amp; Garden(1)
Science(38)
Science(23)
Politics(7)
Politics(7)
Politics(62) Religion(1)
World(1)
Science(1)
Science(2)
Science(1)
Science(41)
Science(18)
Odd News(4) U.S.(2)
Odd News(1)
Education(8)
Education(1)
Travel(16)
Travel(92)
Politics(12)
World(12)
Politics(1)
Odd News(3)
Odd News(4)
Odd News(1)
Odd News(3)
Education(9)
Education(224)
Travel(76)
Travel(23)
Politics(54) U.S.(1)
World(17)
World(1)
Odd News(4)
Odd News(9)
Odd News(52)
Odd News(143)
Education(2)
Fashion(1)
Travel(1)
World(1)
Odd News(1)
Odd News(8)
Odd News(3)
Odd News(23)
Odd News(14)
Odd News(9)
Odd News(2)
History(6)
Politics(19)
Politics(13)
World(5)
World(4)
Odd News(4)
Odd News(6)
Odd News(11)
Odd News(8)
Home &amp; Garden(18)
Fashion(1)
History(59)
Politics(37)
Politics(9)
Politics(33)
World(8)
World(57)
World(1)
Odd News(12) World(3)
Odd News(3)
Odd News(4)
Odd News(12)
Odd News(1)
Politics(2)
Politics(7)
World(15)
World(21) Science(1)
Odd News(8)
Odd News(5)
Odd News(1)
Entertainment(22) Science(1) Health(1) Entertainment(5) U.S.(1)
World(3) History(1) Education(9) Travel(3) World(2) Health(2) Entertainment(1)
History(1)
Fashion(6)
Fashion(86)
Fashion(1)
Politics(4)
Politics(12)
Politics(35)
World(5)
Politics(1)
Odd News(1)
Entertainment(17)
Politics(37)
Politics(1)
World(146)
World(3)
World(31)
U.S.(16)
U.S.(16)
Religion(6)
Religion(15)
Religion(92)
Religion(1)
Politics(67)
World(24)
World(4)
Odd News(7)
Odd News(84)
U.S.(1)
Religion(19)
Religion(1)
Religion(1)
Religion(8)
Religion(2)
Politics(16)
World(23)
World(11)
World(22)
World(36) Fashion(1)
World(2)
Odd News(4)
Religion(2)
Religion(38)
Religion(3)
Religion(1)
Politics(20)
Politics(2)
Politics(50)
World(652) Politics(6) Business(1)
World(132)
World(37)
World(29)
World(5)
Odd News(14)
Tech(2)
Tech(4)
Tech(9)
Religion(30)
Religion(1)
Religion(3)
Politics(1)
Politics(16)
Politics(1)
World(69) Tech(1)
World(24)
World(38)
World(51) Politics(1)
World(27)
Tech(5)
Tech(52)
Tech(154) Business(1)
Tech(9)
Politics(43)
Politics(4)
Politics(1)
Politics(58) U.S.(1)
World(34)
World(58) Business(1)
World(7)
World(2)
World(9)
Tech(15)
Tech(7)
Tech(382)
Tech(504)
Tech(1)
Education(14)
Politics(28)
Politics(8)
World(8)
World(19)
World(44)
Entertainment(1)
Tech(26)
Tech(50)
Tech(19)
Arts(2)
Politics(4) U.S.(1)
Politics(13)
Politics(13)
World(69)
World(1)
World(64) Business(3) Science(1) Health(1)
Entertainment(1)
Entertainment(677)
Tech(2)
Tech(82)
Tech(17)
Arts(36)
Arts(12)
Arts(1)
Politics(85) Business(2) Sports(1)
Politics(26)
Politics(1)
Politics(17)
World(16)
World(86) Science(1)
World(5)
Entertainment(4)
Entertainment(18)
Entertainment(9)
Entertainment(2)
Tech(46)
Arts(18)
Arts(123)
Arts(1)
Politics(5)
Politics(51) Religion(1) World(1)
World(105) Politics(2)
World(6)
World(17)
Entertainment(12)
Entertainment(12)
Entertainment(15)
Entertainment(2)
Entertainment(19)
Entertainment(4)
Arts(27)
Politics(15) U.S.(1)
World(37) Politics(2)
World(21)
World(66)
U.S.(1) Sports(1)
Entertainment(100)
Entertainment(4)
Entertainment(204)
Entertainment(87)
Business(8) Education(1)
Arts(6)
Politics(14)

SOM #3

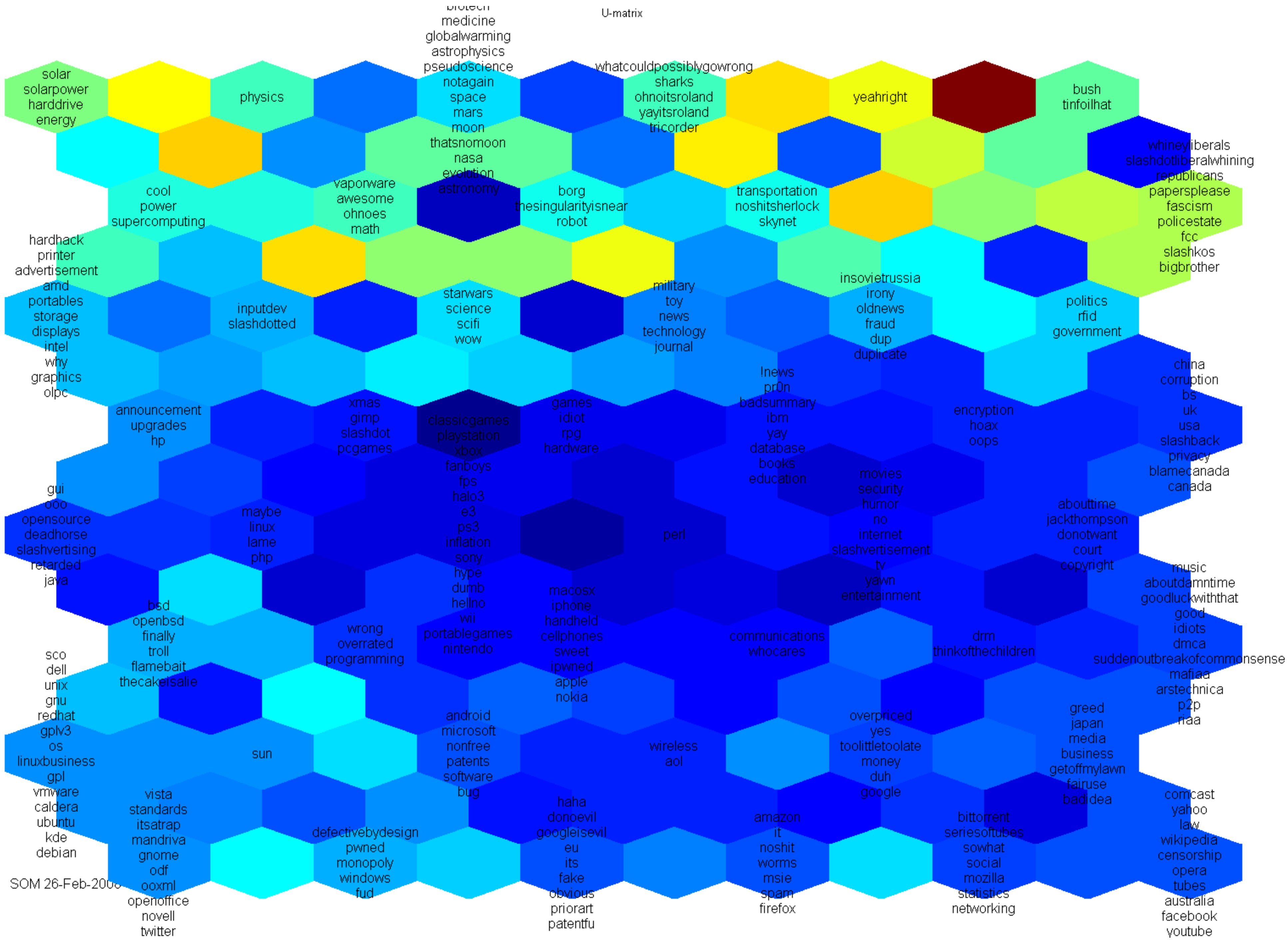
- Data from Newsvine.com
  - Whole December 2007
- Normalized tag-tag-matrix used as input to SOM
  - Only tags which occur >= 40 times are used
  - No ignore list (category tags are there, too)
  - 423 tags
  - No need to do PCA (matrix size: 423 x 423)

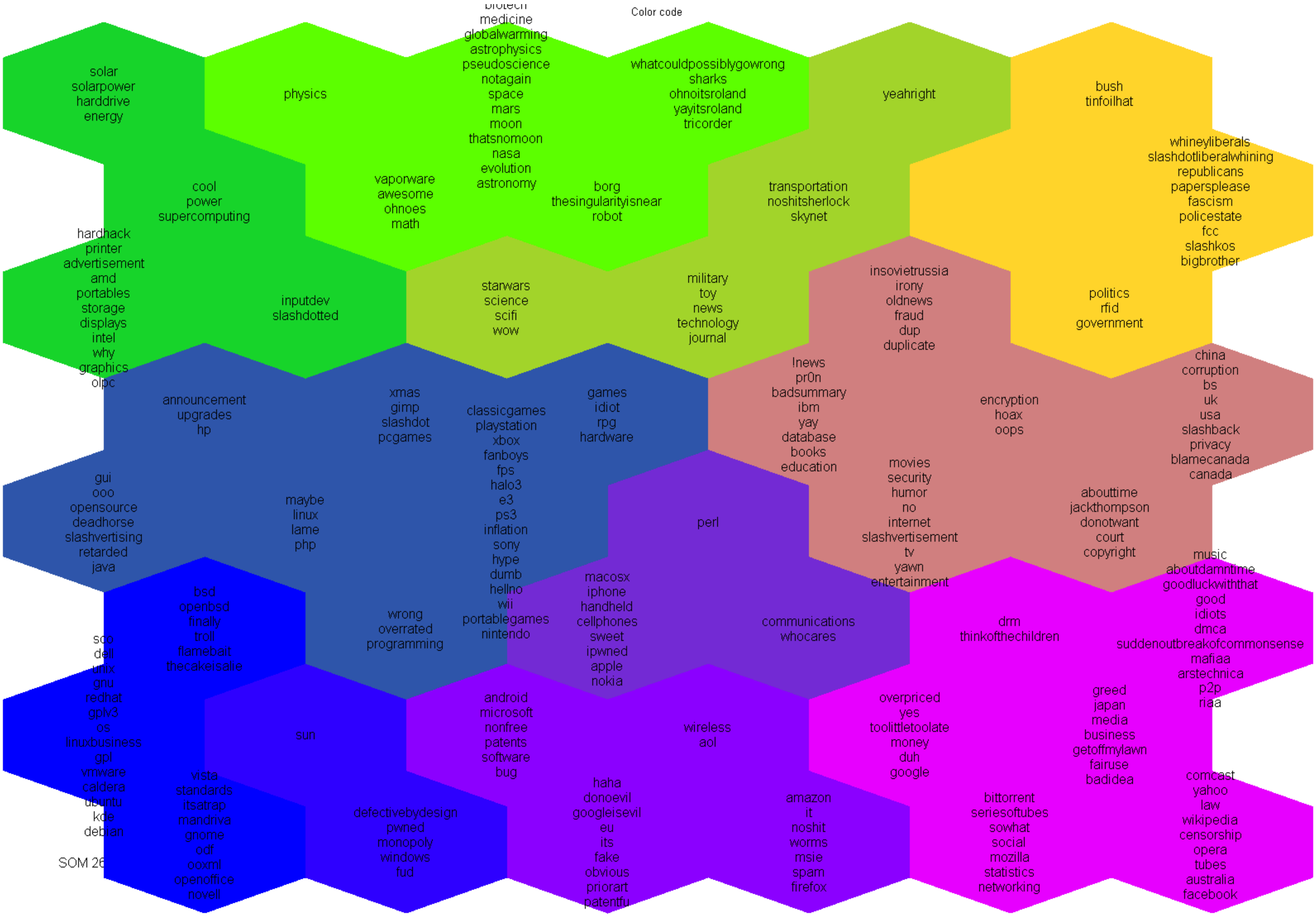

U-matrix
iowa
primary
campaign
obama
barack-obama
new-hampshire
rudy-giuliani
caucus
democrat
mike-huckabee
fred-thompson
presidential-election-of-2008
2008-election
hillary-clinton
john-mccain
republican
paul
2008
debate
mccain
election-2008
huckabee
democrats
presidential
giuliani
edwards
republicans
candidates
south-carolina
polls
clinton
hillary-rodham-clinton
president
hillary
elections
election
iraq
democracy
poll
race
taliban
middle-east
benazir-bhutto
assassination
musharraf
al-qaeda
terrorism
terror
afghanistan
pakistan
north-korea
germany
eu
philippines
human-rights
saudi-arabia
european-union
muslims
india
asia
kenya
sudan
reuters
nepal
russia
un
south-korea
bosnia-news
bosnian-news
south-africa
palestinians
bosnia
africa
darfur
activism
weapons
scotland
mexico
violence
rape
finland
brazil
justice
missing
new-year
trial
crime
murder
prison
suicide
fight
slain
lawsuit
school
shootings
storm
winter
fire
obituaries
weather
church
seattle-times
student
odd
police
california
deaths
shooting
disease
alcohol
diet
weight-loss
medicine
stress
associated-press
texas
new-york
florida
chicago
college
top-25
football
nfl
mitchell-report
steroids
basketball
philadelphia-eagles
philadelphia
mlb
rdp
college-football
baseball
nba
melbourne-victory
trade
meosthelioma-lawyer
mesothelioma-diagnosis
mesothelioma
houston-mesothelioma-lawyer
mesothelioma-symptoms
investment
real-estate
economy
federal-reserve
finance
prices
ovetii
gold
banks
credit
dollar
internet-marketing
oil-prices
wall-street
mortgage
advertising
nytimes
seo
jewelry
climate
united-nations
bali
oil
taxes
supreme-court
congress
constitution
senate
freedom
white-house
satire
healthcare
government
health-care
salmon
co2
pollution
climate-change
global-warming
al-gore
conservation
policy
development
energy
green
ecology
safety
waste
housing
cheney
president-bush
cia
impeachment
waterboarding
george-w-bush
bush
abortion
the-seminal
immigration
poverty
environment
water
death
child
dogs
pets
veterinary
animals
cats
insurance
economics
fraud
au-news
federal-government
legal
racism
corruption
foreign-policy
torture
george-bush
aids
law
united-kingdom
uk-news
children
travel
technology
entertainment
us-news
education
home-garden
sports
science
fashion
politics
health
world-news
history
odd-news
religion
business
archaeology
books
to
art
research
evolution
charity
nuclear-weapons
iraq-war
war
nie
intelligence
united-states
military
us
islam
muslim
christian
atheism
christianity
food
design
holidays
security
blogs
newsvine
cars
photography
tech
blogging
ipod
software
nokia
mac
computers
wireless
iphone
privacy
social-networking
apple
search-engine
holiday
santa
gifts
shopping
china
guantanamo
australia
europe
israel
iran
news
of
arts
magazine
christmas
america
usa
france
venezuela
uk
canada
media
japan
video
photos
funny
humor
labor
review
music
mp3
download
videos
television
film
movie
celebrity
gaming
nintendo
video-games
wii
gamevine
games
peace
chavez
italy
palestine
world
women
drugs
sex
money
beauty
blog
washington
south-america
gay
power
jesus
god
the
culture
love
cancer
family
movies
people
tv
home
list
wealth
training
mlm
work
tips
liberty
for
san-francisco
new
and
dating
help
website
web-2-0
mars
bible
time
england
life
dog
car
youtube
soccer
2007
dvd
on
game
rock
spears
fun
hollywood
sony
book
space
nasa
computer
vista
web
microsoft
linux
facebook
google
windows
internet
free
marketing
online


Color code
iowa
primary
campaign
obama
barack-obama
new-hampshire
rudy-giuliani
caucus
democrat
mike-huckabee
fred-thompson
presidential-election-of-2008
south-carolina
polls
clinton
hillary-rodham-clinton
president
hillary
mccain
election-2008
huckabee
democrats
presidential
giuliani
edwards
republicans
candidates
elections
election
taliban
middle-east
benazir-bhutto
assassination
musharraf
al-qaeda
terrorism
terror
afghanistan
pakistan
north-korea
iraq
democracy
germany
eu
philippines
human-rights
saudi-arabia
european-union
muslims
india
asia
kenya
sudan
reuters
nepal
russia
un
south-korea
bosnia-news
bosnian-news
south-africa
palestinians
bosnia
africa
britain
nuclear
activism
weapons
scotland
mexico
violence
rape
finland
brazil
fight
slain
college
top-25
football
nfl
mitchell-report
steroids
basketball
philadelphia-eagles
philadelphia
mlb
rdp
college-football
baseball
nba
melbourne-victory
meosthelioma-lawyer
mesothelioma-diagnosis
mesothelioma
houston-mesothelioma-lawyer
mesothelioma-symptoms
investment
real-estate
economy
federal-reserve
finance
prices
ovetii
gold
banks
credit
dollar
internet-marketing
oil-prices
wall-street
mortgage
trade
2008-election
hillary-clinton
john-mccain
republican
paul
2008
debate
poll
race
taxes
supreme-court
congress
constitution
senate
freedom
white-house
satire
healthcare
government
health-care
salmon
climate
united-nations
bali
oil
co2
pollution
climate-change
global-warming
al-gore
conservation
policy
development
energy
green
ecology
safety
waste
justice
missing
new-year
trial
crime
murder
prison
suicide
lawsuit
school
shootings
storm
winter
fire
obituaries
weather
church
seattle-times
student
odd
police
california
deaths
kids
shooting
health
disease
alcohol
diet
weight-loss
medicine
stress
associated-press
texas
new-york
florida
chicago
housing
dogs
pets
veterinary
animals
cats
death
child
insurance
economics
charity
cheney
president-bush
cia
impeachment
waterboarding
george-w-bush
bush
abortion
the-seminal
immigration
poverty
environment
water
fraud
au-news
federal-government
legal
racism
corruption
foreign-policy
torture
george-bush
aids
law
united-kingdom
uk-news
children
travel
technology
entertainment
us-news
education
home-garden
sports
science
fashion
politics
health
world-news
history
odd-news
religion
business
archaeology
books
to
art
research
evolution
advertising
nytimes
seo
jewelry
nuclear-weapons
iraq-war
war
nie
intelligence
united-states
military
us
islam
muslim
christian
atheism
christianity
food
design
holidays
security
blogs
newsvine
cars
photography
tech
blogging
ipod
software
nokia
mac
computers
wireless
iphone
privacy
social-networking
apple
israel
iran
china
guantanamo
australia
europe
news
of
arts
magazine
christmas
search-engine
holiday
santa
gifts
shopping
america
usa
france
venezuela
uk
canada
media
japan
video
photos
funny
humor
labor
review
music
mp3
download
videos
television
film
movie
celebrity
gaming
nintendo
video-games
wii
gamevine
games
peace
chavez
italy
palestine
world
women
drugs
sex
money
beauty
blog
washington
south-america
gay
power
jesus
god
the
culture
love
cancer
family
movies
people
tv
home
list
wealth
training
mlm
work
tips
liberty
for
san-francisco
new
and
dating
help
website
web-2-0
bible
time
england
life
dog
car
youtube
soccer
2007
free
dvd
on
game
rock
spears
fun
hollywood
sony
book
space
nasa
computer
vista
web
microsoft
linux
facebook
google
windows
internet
marketing
online

SOM #4

- Data from Slashdot.org
  - From July 2007 to December 2007
- Normalized tag-tag-matrix used as input to SOM
  - Only tags which occur >= 5 times are used
  - No ignores for categories
  - 268 tags
  - No need to do PCA


U-matrix
biotech
medicine
globalwarming
astrophysics
pseudoscience
notagain
space
mars
moon
thatsnomoon
nasa
evolution
astronomy
solar
solarpower
harddrive
energy
physics
whatcouldpossiblygowrong
sharks
ohnoitsroland
yayitsroland
tricorder
yeahright
bush
tinfoilhat
whineyliberals
slashdotliberalwhining
republicans
papersplease
fascism
policestate
fcc
slashkos
bigbrother
cool
power
supercomputing
vaporware
awesome
ohnoes
math
borg
thesingularityisnear
robot
transportation
noshitsherlock
skynet
hardhack
printer
advertisement
amd
portables
storage
displays
intel
why
graphics
olpc
inputdev
slashdotted
starwars
science
scifi
wow
military
toy
news
technology
journal
insovietrussia
irony
oldnews
fraud
dup
duplicate
politics
rfid
government
!news
pr0n
badsummary
ibm
yay
database
books
education
china
corruption
bs
uk
usa
slashback
privacy
blamecanada
canada
announcement
upgrades
hp
xmas
gimp
slashdot
pcgames
classicgames
playstation
xbox
fanboys
fps
halo3
e3
ps3
inflation
sony
hype
dumb
hellno
wii
portablegames
nintendo
games
idiot
rpg
hardware
encryption
hoax
oops
movies
security
humor
no
internet
slashvertisement
tv
yawn
entertainment
gui
ooo
opensource
deadhorse
slashvertising
retarded
java
maybe
linux
lame
php
perl
abouttime
jackthompson
donotwant
court
copyright
music
aboutdamntime
goodluckwiththat
good
idiots
dmca
suddenoutbreakofcommonsense
mafiaa
arstechnica
p2p
riaa
bsd
openbsd
finally
troll
flamebait
thecakeisalie
wrong
overrated
programming
macosx
iphone
handheld
cellphones
sweet
ipwned
apple
nokia
communications
whocares
drm
thinkofthechildren
sco
dell
unix
gnu
redhat
gplv3
os
linuxbusiness
gpl
vmware
caldera
ubuntu
kde
debian
sun
android
microsoft
nonfree
patents
software
bug
wireless
aol
overpriced
yes
toolittletoolate
money
duh
google
greed
japan
media
business
getoffmylawn
fairuse
badidea
SOM 26-Feb-2008
vista
standards
itsatrap
mandriva
gnome
odf
ooxml
openoffice
novell
twitter
defectivebydesign
pwned
monopoly
windows
fud
haha
donoevil
googleisevil
eu
its
fake
obvious
priorart
patentfu
amazon
it
noshit
worms
msie
spam
firefox
bittorrent
seriesoftubes
sowhat
social
mozilla
statistics
networking
comcast
yahoo
law
wikipedia
censorship
opera
tubes
australia
facebook
youtube


Color code
biotech medicine globalwarming astrophysics pseudoscience notagain space mars moon thatsnomoon nasa evolution astronomy
solar solarpower harddrive energy
physics
whatcouldpossiblygowrong sharks ohnoitsroland yayitsroland tricorder
yeahright
bush tinfoilhat
cool power supercomputing
vaporware awesome ohnoes math
borg thesingularityisnear robot
transportation noshitsherlock skynet
whineyliberals slashdotliberalwhining republicans papersplease fascism policestate fcc slashkos bigbrother
hardhack printer advertisement amd portables storage displays intel why graphics olpc
inputdev slashdotted
starwars science scifi wow
military toy news technology journal
insovietrussia irony oldnews fraud dup duplicate
politics rfid government
china corruption bs uk usa slashback privacy blamecanada canada
announcement upgrades hp
xmas gimp slashdot pcgames
classicgames playstation xbox fanboys fps halo3 e3 ps3 inflation sony hype dumb hellno wii portablegames nintendo
games idiot rpg hardware
lnews pr0n badsummary ibm yay database books education
encryption hoax oops
gui ooo opensource deadhorse slashvertising retarded java
maybe linux lame php
perl
movies security humor no internet slashvertisement tv yawn entertainment
abouttime jackthompson donotwant court copyright
music aboutdamntime goodluckwiththat good idiots dmca suddenoutbreakofcommonsense mafiaa arstechnica p2p riaa
bsd openbsd finally troll flamebait thecakeisalie
wrong overrated programming
macosx iphone handheld cellphones sweet ipwned apple nokia
communications whocares
drm thinkofthechildren
sco dell unix gnu redhat gplv3 os linuxbusiness gpl vmware caldera ubuntu kde debian
sun
android microsoft nonfree patents software bug
wireless aol
overpriced yes toolittletoolate money duh google
greed japan media business getoffmylawn fairuse badidea
vista standards itsatrap mandriva gnome odf ooxml openoffice novell
defectivebydesign pwned monopoly windows fud
haha donoevil googleisevil eu its fake obvious priorart patentfu
amazon it noshit worms msie spam firefox
bittorrent seriesoftubes sowhat social mozilla statistics networking
comcast yahoo law wikipedia censorship opera tubes australia facebook
SOM 26

5. Tag prediction

- Tag-tag co-occurrence matrix revisited

	Gw-bush	Elections	Cooking	Internet	Web2.0	...
Gw-bush	600	100	1	2	0	
Elections	100	520	0	10	0	
Cooking	1	0	80	0	0	
Internet	2	10	0	400	200	
Web2.0	0	0	0	200	350	
...						

	Gw-bush	Elections	Cooking	Internet	Web2.0	...
Gw-bush	600/600	100/600	1/600	2/600	0	
Elections	100/520	520/520	0	10/520	0	
Cooking	1/80	0	80/80	0	0	
Internet	2/400	10/400	0	400/400	200/400	
Web2.0	0	0	0	200/350	350/350	
...						

Tag-tag-matrix

- Very sparse, feasible to store for whole month as adjacency lists
- Original matrix symmetrical, normalized matrix not
- M (”obama”, ”politics”) = 0.99
- M (”politics”, ”obama”) = 0.02
- Possible to find partial orders between tags (see Mannila, et. al ...)
- General term > specific term

Tags frequently occurring together

- For each tag i find tag j for which M(i, j) maximal (or higher than a threshold)
- If M(j , i) not too high, then j probably more general
- If both are high, terms probably closely related
- Before processing remove diagonal elements
- Note: results can be just as good as the quality of tags

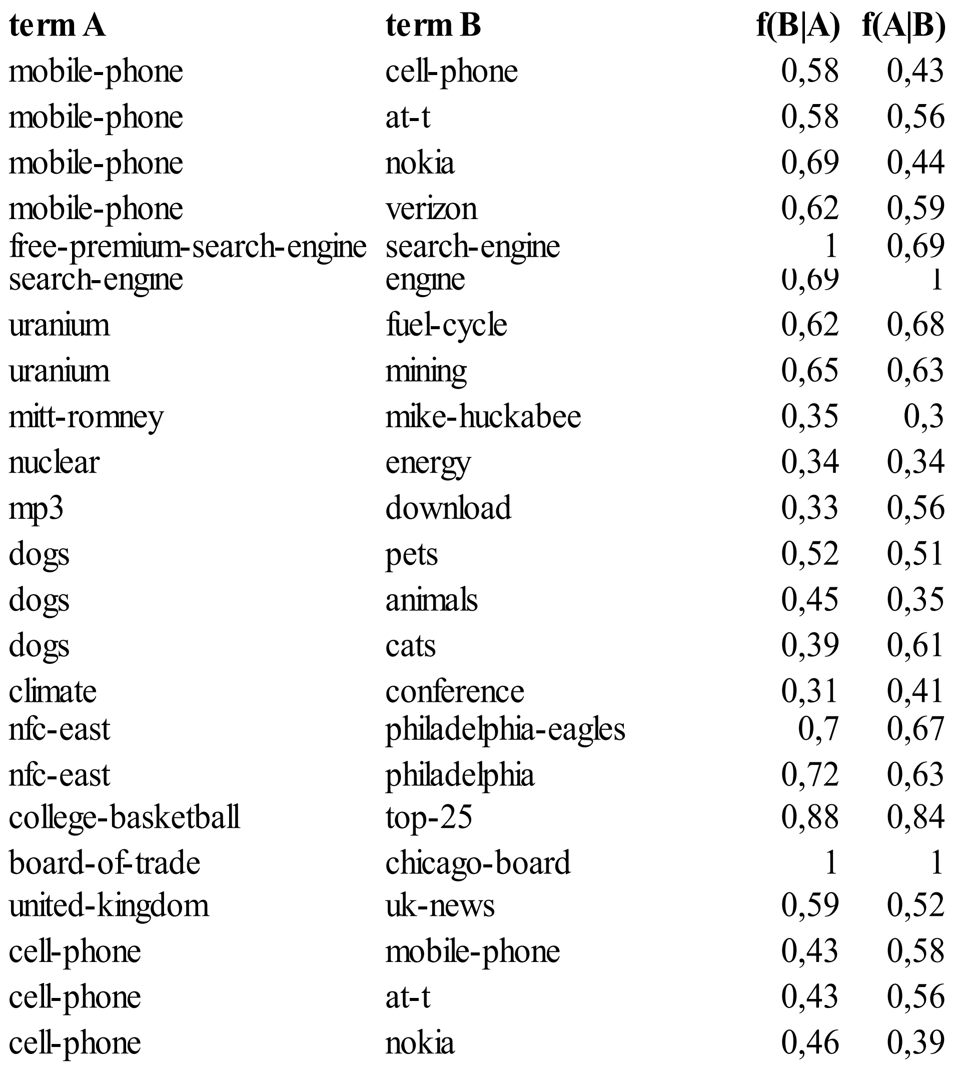
Example 1

- Co-occurrance high both for f(B|A) and f(A|B)
- Best match for each term

term A	**term B**	**f(B|A)**	**f(A|B)**
free-premium-search-engine	search-engine	1	0,69
dogs	pets	0,52	0,51
united-kingdom	uk-news	0,59	0,52
aids	hiv	0,47	0,68
pkk	turkey	1	0,35
ron	paul	1	0,59
chavez	venezuela	0,81	0,39
energy	nuclear	0,34	0,34
au-news	australia	1	0,45
playstation	xbox	0,64	0,47
benazir-bhutto	pakistan	0,84	0,31
on	trial	0,32	0,37
wii	nintendo	0,64	0,68
based	home	1	0,38
bosnia-news	bosnia	1	0,84
pets	animals	0,59	0,47
slain	taylor	0,56	0,86
gay	lesbian	0,43	0,74
maryland	dcmetro	0,66	0,82
dcmetro	maryland	0,82	0,66
plane	crash	0,38	0,32
veterinary	animals	0,57	0,33
animals	pets	0,47	0,59
homosexual	gay	0,77	0,38
climate-change	global-warming	0,73	0,62

Example 2

- Co-occurrance high both for f(B|A) and f(A|B)
- All matches for each term which are higher than specified threshold



term A	term B	f(B|A)	f(A|B)
mobile-phone	cell-phone	0,58	0,43
mobile-phone	at-t	0,58	0,56
mobile-phone	nokia	0,69	0,44
mobile-phone	verizon	0,62	0,59
free-premium-search-engine	search-engine	1	0,69
search-engine	engine	0,69	1
uranium	fuel-cycle	0,62	0,68
uranium	mining	0,65	0,63
mitt-romney	mike-huckabee	0,35	0,3
nuclear	energy	0,34	0,34
mp3	download	0,33	0,56
dogs	pets	0,52	0,51
dogs	animals	0,45	0,35
dogs	cats	0,39	0,61
climate	conference	0,31	0,41
nfc-east	philadelphia-eagles	0,7	0,67
nfc-east	philadelphia	0,72	0,63
college-basketball	top-25	0,88	0,84
board-of-trade	chicago-board	1	1
united-kingdom	uk-news	0,59	0,52
cell-phone	mobile-phone	0,43	0,58
cell-phone	at-t	0,43	0,56
cell-phone	nokia	0,46	0,39

Example 3

- (specific, general)-pairs
- f(B|A) large
- f(A|B) small

term A	term B	f(B|A)	f(A|B)
music	entertainment	0,7	0,11
tech	technology	0,81	0,02
mobile-phone	technology	0,77	0,01
john-edwards	politics	0,97	0,02
dennis-kucinich	politics	0,91	0,01
election-08	politics	0,95	0,01
cheney	politics	0,77	0,01
investment	business	0,78	0,01
company	marketing	0,72	0,22
company	business	0,86	0,01
uranium	nuclear	0,74	0,09
mitt-romney	politics	0,93	0,03
biology	science	0,82	0,03
mp3	music	0,76	0,17
mp3	entertainment	0,76	0,03
bill-clinton	politics	0,92	0,01
britney-spears	entertainment	0,87	0,01
college-basketball	sports	1	0,09
board-of-trade	business	1	0,01
cell-phone	technology	0,71	0,01
movies	entertainment	0,82	0,06
venezuela	world-news	0,8	0,02
bhutto	pakistan	0,75	0,22

Applications of generalization

- Build trees of terms, proceeding from specific to general
- Automatically finding categories
- Predicting new tags

Tag prediction, method #1

- For a set of tags, what are the most plausible next tags
  - Simple method: t_i in tags
  - Predict t_o, such that $\sum_{i=1}^{n} M(t_0, t_i)$ is maximized
- Examples
  - (”john-mccain”, ”mitt-romney”, ”obama”)

    --> ”bill-richardson”, ”mccain”, ”rudy-giuliani”, ”fred-thompson”
  - (”nokia”, ”mobile-phone”)

    --> ”technology”, ”cell-phone”, ”verizon”, ”at-t”
  - (”gordon-brown”, ”george-w-bush”)

    --> ”uk-news”, ”politics”, ”world-news”

Tag prediction, method #2

- Until now we relied on tag co-occurence
  - But ”george-w-bush” and ”g-w-bush” probably don't occur together for the same document, one writer only uses one of these tags in one article.
- They probably still have similar co-occurence patterns with other terms (”politics”, ”iraq-war”, ”united-states”, ...)
- **Idea**: cross-correlation between rows of the tag-tag matrix
  - Matching terms will not necessarily be related

Examples for method #2

- **”finland”**
    - reuters, britain, philippines, somalia, africa, germany, russia, mexico, algeria, lebanon, ireland, brazil, cuba, oil, violence
- **”president-bush”**
    - republicans, constitution, democrats, white-house, budget, congress, george-w-bush, bush, taxes, senate, foreign-policy, election-2008, george-bush
- **”beer”**
    - alcohol, coffee, party, food, children, charity, new-year, relationships, baby, odd, holidays, theft, ireland, medical, santa-claus, christmas

Questions?