

European Summer School 2017

Text Mining with Canonical Text Services
Theory Session 6 – Canonical Text Miner



Federal Ministry
of Education
and Research



Topic Models

Abstract thematic structure of a document set

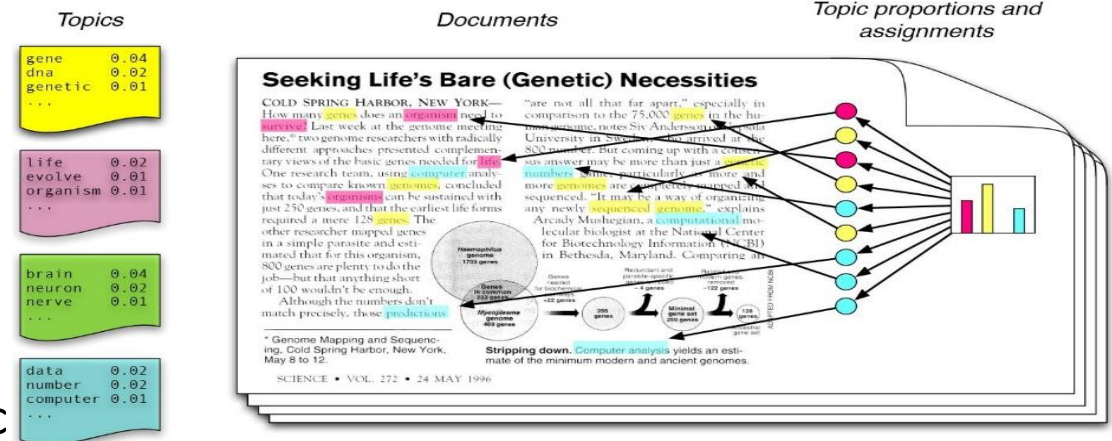
Documents belong to Topic with probabilities
(90% Evolution, 10% Disease)

One (of many) Algorithms:
Latent Dirichlet Allocation

(LDA)

(David M. Blei: Introduction to Probabilistic Topic
<http://www.cs.princeton.edu/~blei/papers/Blei2011.pdf>)

Genetics	Evolution	Disease	Computers
Human	Evolution	Disease	Computer
Genome	Species	Host	Model
DNA	Organism	Bacteria	information



Topic Models

Example Topic:

labor	jury
workers	trial
employees	crime
union	defendant
employer	sentencing
work	judges
job	punishment
bargaining	evidence
unions	sentence
collective	offense

Topic Models

- Manage and explore document sets

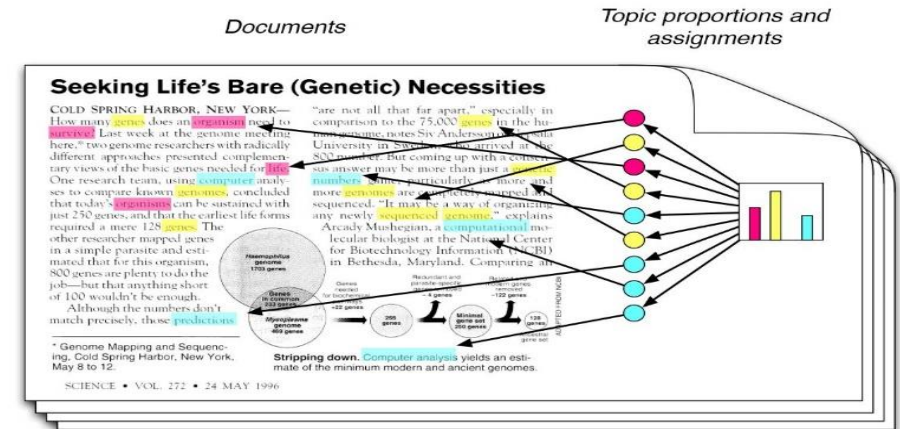
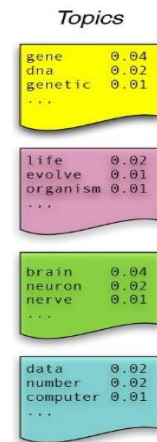
Zoom in and out of Topics

Find further or more specialized topics

Semantic Analysis

user has “topic-fingerprint”

Genetics	Evolution	Disease	Computers
Human	Evolution	Disease	Computer
Genome	Species	Host	Model
DNA	Organism	Bacteria	information



Topic Models

- Data: TED Subtitle Corpus, 51770 documents, 105 languages, 1938 english documents, big variety of topics, CTS access <http://ctstest.informatik.uni-leipzig.de/ted/cts/?request=GetCapabilities>

Tokens for Topic 23

brain, neurons, brains, memory, body, consciousness, autism, human, cells sleep

Texts about topic 23

"Re-engineering the brain", "The linguistic genius of babies", "A light switch for neurons", "The mystery of chronic pain", "The quest to understand consciousness", "A prosthetic eye to treat blindness", "How your brain tells you where you are", "The mysterious workings of the adolescent brain", "A monkey that controls a robot with its thoughts. No", "really.", "How a fly flies", "Your brain is more than a bag of chemicals", "Parkinson's", "depression and the switch that might turn them off", "A mouse. A laser beam. A manipulated memory.", "3 clues to understanding your brain", "The paralyzed rat that walked", "The neuroscience of restorative justice", "A neural portrait of the human mind", "One more reason to get a good nights sleep", "Brain-to-brain communication has arrived. How we did it", "A look inside the brain in real time", "Growing evidence of brain plasticity", "What hallucination reveals about our minds", "The neurons that shaped civilization", "A second opinion on developmental disorders", "I am my connectome"

Topics for text "A monkey that controls a robot with its thoughts. No, really."

15 -> computer, data, machine, information, show, computers, video, simple, using, each

23 -> brain, neurons, brains, memory, body, consciousness, autism, human, cells sleep

Topics generated from TED Subtitles (engl)

0 yeah, hand, yes, thank, four, five, audience, okay, show, number
1 women, men, woman, girls, love, sex, girl, children, young, gay
2 car, cars, fly, miles, power, road, drive, driving, vehicle, flying
3 universe, earth, space, science, planet, theory, stars, mars, sun, physics
4 school, kids, children, students, education, teachers, learning, child, schools, learn
5 robot, film, robots, movie, story, head, him, tail, character, shot
6 might, question, fact, example, should, problem, find, better, any, whether
7 light, water, air, made, its, energy, material, nature, off, inside
8 him, after, went, story, never, didn, came, started, old, thank
9 data, internet, information, media, online, web, phone, social, google, facebook
10 music, play, sound, game, games, video, playing, song, hear, voice
11 language, book, words, books, word, english, read, writing, write, poem
12 human, god, feel, self, believe, compassion, happiness, love, live, experience
13 health, disease, care, hiv, children, countries, virus, malaria, percent, treatment
14 percent, today, countries, per, data, growth, change, billion, population, million
15 computer, data, machine, information, show, computers, video, simple, using, each
16 talk, mean, bit, great, tell, didn, maybe, start, sort, big
17 cancer, cells, disease, body, heart, patient, patients, blood, surgery, medical
18 food, energy, oil, water, waste, eat, carbon, climate, percent, plant
19 dog, him, black, white, man, legs, smell, bear, wine, dogs
20 ocean, water, sea, fish, ice, animals, species, forest, earth, planet
21 money, dollars, business, companies, company, market, percent, value, buy, jobs
22 technology, create, today, system, able, idea, design, build, together, working
23 brain, neurons, brains, memory, body, consciousness, autism, human, cells, sleep
24 war, violence, police, military, prison, security, killed, states, peace, united
25 africa, country, china, power, india, political, countries, government, chinese, democracy
26 nand, nto, nof, nthat, nthe, nin, nis, nfor, nwith, new
27 art, design, made, sort, project, museum, artist, image, images, show
28 species, dna, animals, human, bacteria, humans, evolution, its, genes, genetic
29 city, building, cities, buildings, space, place, public, built, community, york

Topics generated from TED Subtitles (arab)

0 الأطفال, التعليم, المدرسة, طفل, الطلاب, الطفل, المدارس, مدرسة, تعليم, هؤلاء
1 الإنترنت, المعلومات, البيانات, الانترنت, الشبكة, موقع, التكنولوجيا, الكمبيوتر, شبكة, جهاز
2 النووي, الن, الحمض, الت, سنة, الس, الجينات, الجينوم, البكتيريا, الص
3 بأن, الحياة, الأشخاص, لنا, فإن, علينا, أننا, بالنسبة, نكون, مما
4 النساء, الرجال, المرأة, الجنس, الفتيات, امرأة, نساء, الجنسية, الزواج, الحب
5 والتي, مرة, قمنا, ومن, تقوم, القيام, وهي, استخدام, مختلفة, نقوم
6 الصور, صورة, الصورة, القصص, الفيلم, قصة, صور, التصوير, فيلم, الأفلام
7 الموسيقى, موسيقى, الصوت, صوت, الروبوت, الروبوتات, الأصوات, الموسيقية, فيديو, موسيقية
8 الطاقة, المياه, الكربون, الماء, المحيط, النفط, الغابات, البحر, الحيوانات, طاقة
9 عام, المتحدة, أفريقيًا, الهند, الولايات, الصين, الدول, سنة, عدد, دولة
10 اللغة, كلمة, الكلمات, كلمات, لغة, الكتب, الإنجليزية, المرور, الكتابة, الصينية
11 حسنا, حين, إذن, وبالتالي, كذلك, نعم, بذلك, الجمهور, القيام, ثلاثة
12 حسنا, أعتقد, إنها, إنه, أنها, أيضا, أحد, حقًا, آخر, لأنه
13 السرطان, مرض, المرضى, الخلايا, القلب, العلاج, المرض, الصحية, الرعاية, المريض
14 لوحة, متحف, الرسم, amp, quot, التصميم, الفن, تصميم, العمل, نوع
15 الحرب, المتحدة, الولايات, العنف, الحكومة, السجن, القوة, الشرطة, السلطة, السلام
16 أنني, لدي, علي, بدأت, نفسي, حياتي, أستطيع, أني, وأنا, أريد
17 اللعب, الألعاب, اللعبة, لعبة, ألعاب, لعب, كرة, هنالك, اللاعبين, ذلك
18 الطعام, النحل, الغذاء, النباتات, الحشرات, النمل, الخبز, الغذائية, الحيوانات, الزراعة
19 الدماغ, الخلايا, العصبية, المخ, علم, دماغ, الذاكرة, الأعصاب, خلية, العقل
20 الأرض, قدم, المحيط, البحر, تحت, فوق, الفضاء, عبر, الجليد, الماء
21 الله, التعاطف, الدين, الرب, الموت, الحياة, الكتاب, قال, الدينية, والتراحم
22 دولار, المال, شركة, الشركات, العمل, الأعمال, السوق, التجارية, مليون, الحصول
23 المدينة, المدن, مدينة, المبني, بناء, المباني, السيارات, نيويورك, البناء, المكان
24 وكان, يكن, قال, كانوا, الرجل, عام, هؤلاء, منذ, يوم, سنوات
25 أو, إلى, أن, حسن, شكر, التي, على, من, أيضا, n
26 الكون, الفضاء, الأرض, الحياة, النجوم, الفيزياء, نظرية, الشمس, المادة, الكواكب
27 التي, علي, الذي, إلي, الآن, لابد, شيء, هذة, حتي, آخر
28 اذا, انها, الان, انا, اكثر, ايضا, الاشياء, اعتقد, حسنا, انهم
29 لكي, ومن, بصورة, الامر, والتي, وان, انها, وهي, لان, فحسب

Trend Detection

CTS-TM can sort result by publication date

-> Trend Detection applicable for any included tool

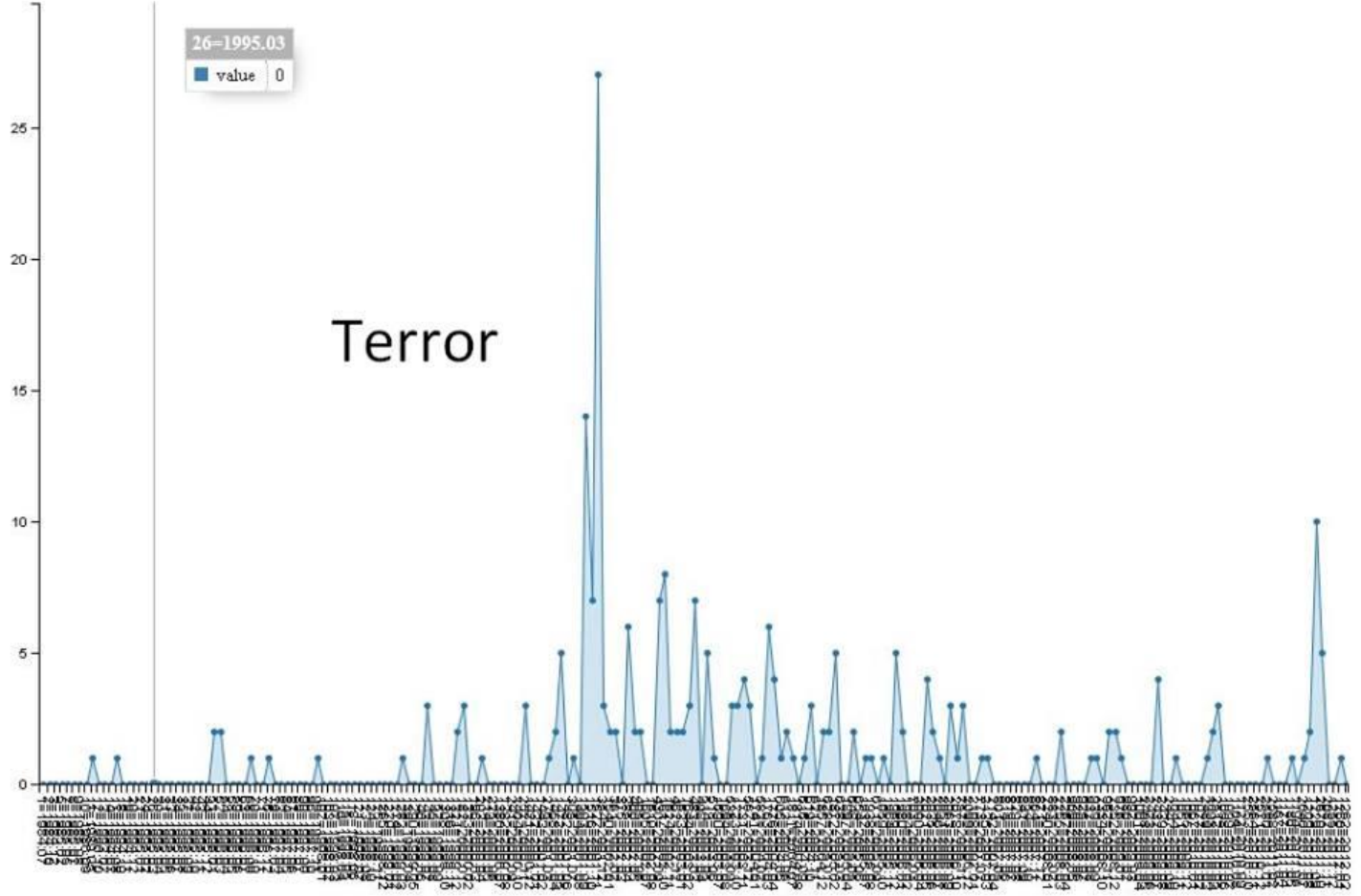
-> Implemented for Token Frequency

```
-<edition urn="urn:cts:pbc:bible.parallel.deu.luther1545letztehand:">  
  <title>Luther 1545 (Letzte Hand). The Bible in German</title>  
  -<license>  
    Source & Publisher: Hans Jürgen Herbst, http://www.lutherbibel.net. Rights: Umsonst habt ihrs empfangen,  
    umsonst gebet es auch. (Matthäus 10:8)  
  </license>  
  -<source>  
    http://www.bibel-online.net/buch/luther\_1545\_letzte\_hand/. http://www.zeno.org/Literatur/M/Luther,+Martin/Luther-Bibel+1545  
  </source>  
  <publicationDate>1545</publicationDate>  
  <language>deu</language>  
  <contentType>xml</contentType>  
</edition>
```



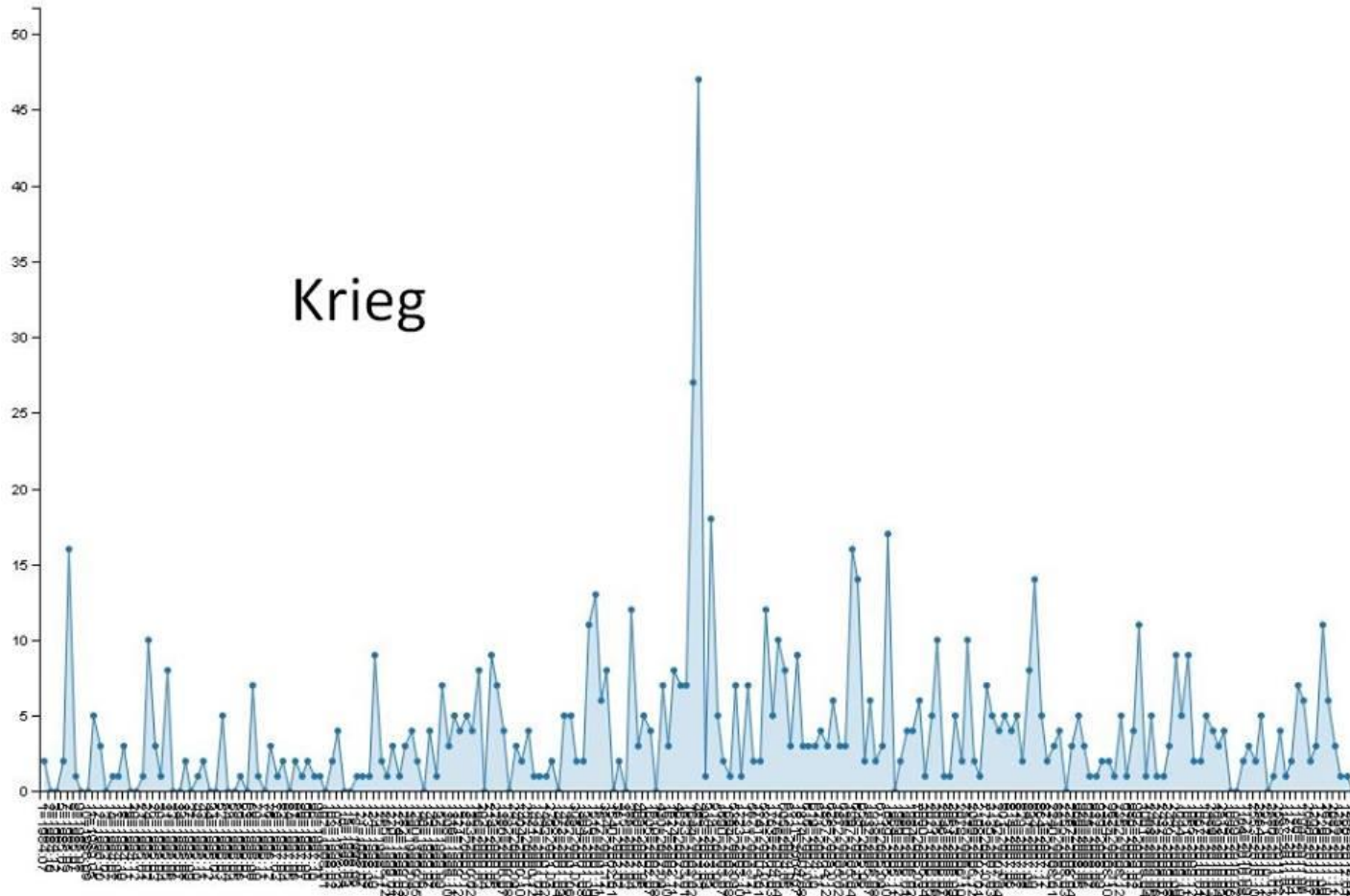
Trend Analysen

Based on German Political Speeches



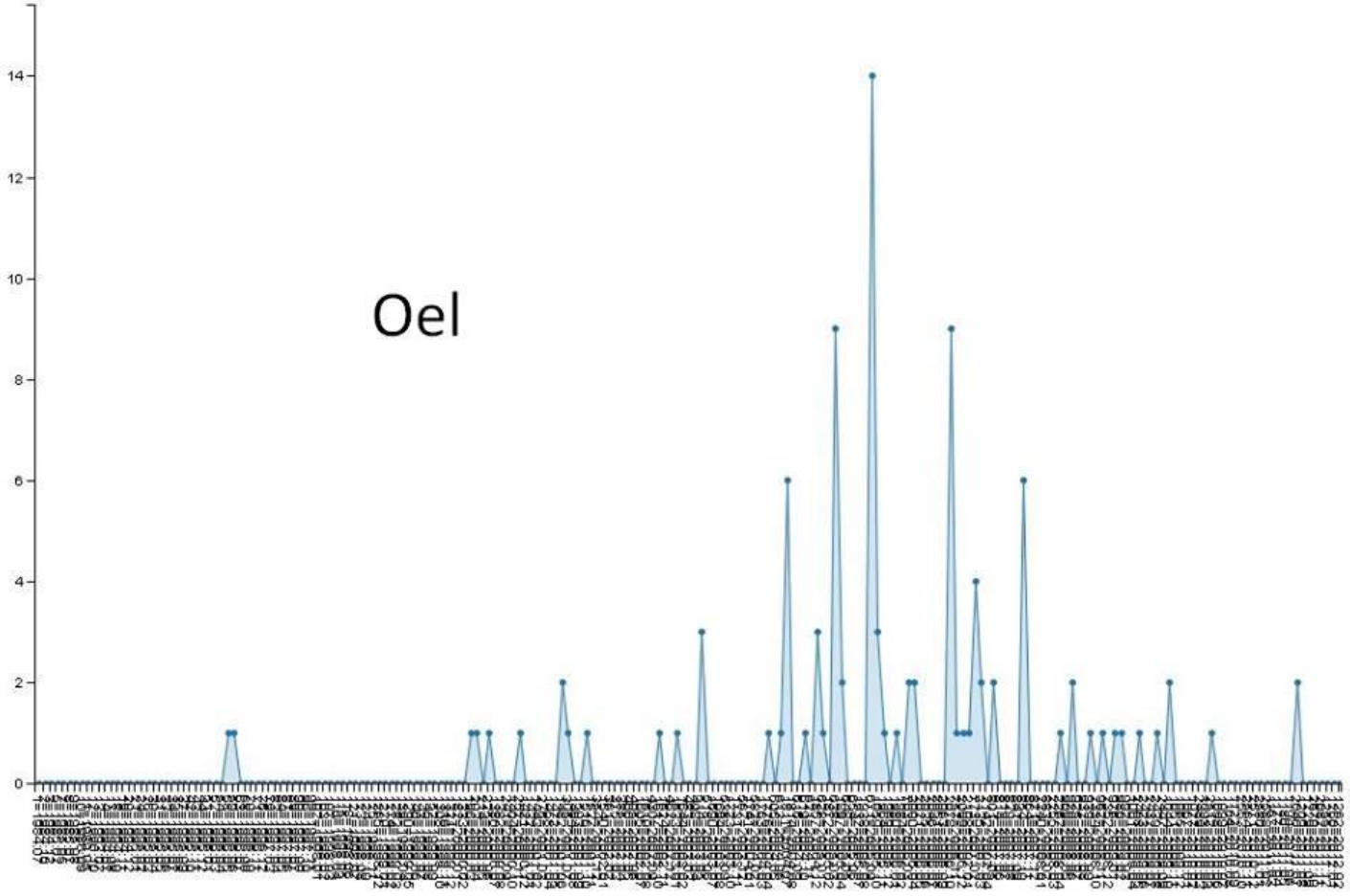
Trend Analysen

Based on German Political Speeches



Trend Analysen

Based on German Political Speeches



Canonical Text Miner

- Broad & comprehensive Text Mining Framework
- Already working:
 - Statistics, Term Document Matrices, Neighbour Cooccurrence, Zipf Ranking, Stopwordlists per Pruning and Zipf, 3 Methods for Volltextsuche, Topic Models with Mallet, Basic Text Reuse analysis
- CTS as standardised Access point
 - Repeating an experiment only requires the configuration file

Canonical Text Miner

Results available via webservice

URNs as Filter

Data sets and results can be connected across projects

URN == Unique key

Persistently citable functions calls

Results connected to Canonical Text Services

Text passages can be requested

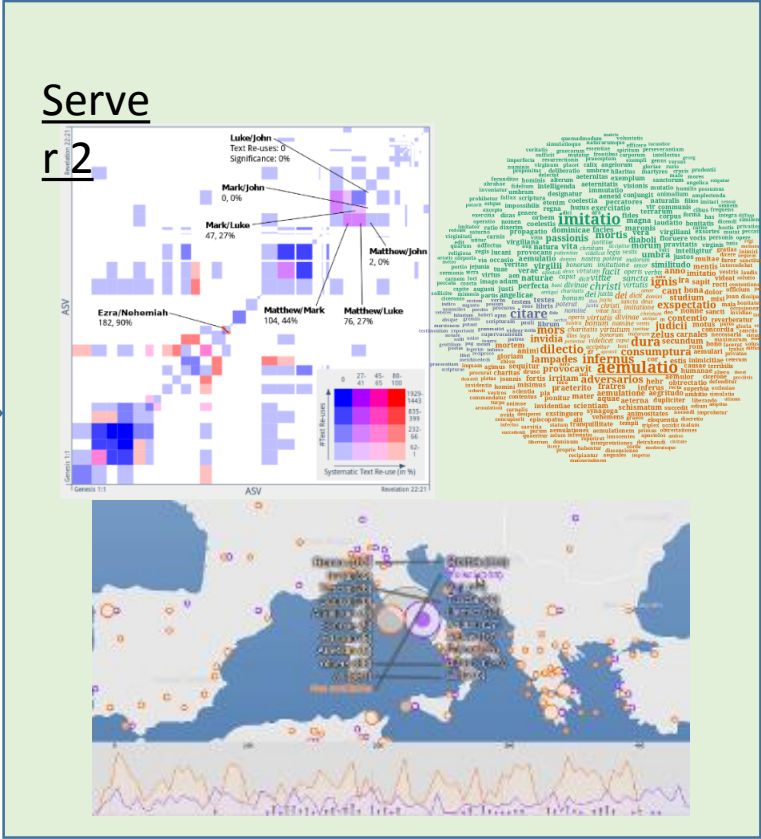
Basic Idea

Server 1

urn:cts:demo:[work]:1.1.1
 urn:cts:demo:[work]:1.2.1

```


<passage>
  <div1 n="1" type="song">
    <div2 n="1" type="strophe">
      <div3 n="1" type="line">
      </div3>
    </div2>
  <div2 n="2" type="strophe">
    <div3 n="1" type="line">
    </div3>
  </div2>
</div1>
</passage>
  
```



(Example Visualizations from work of [Stefan Jaenicke](#))

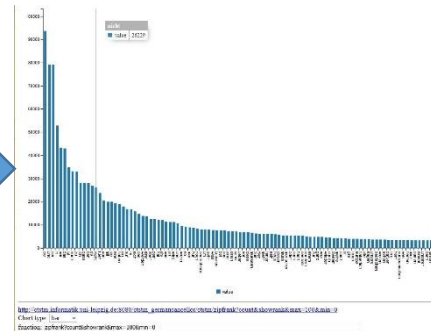
Canonical Text Miner

3 layered webservice architecture

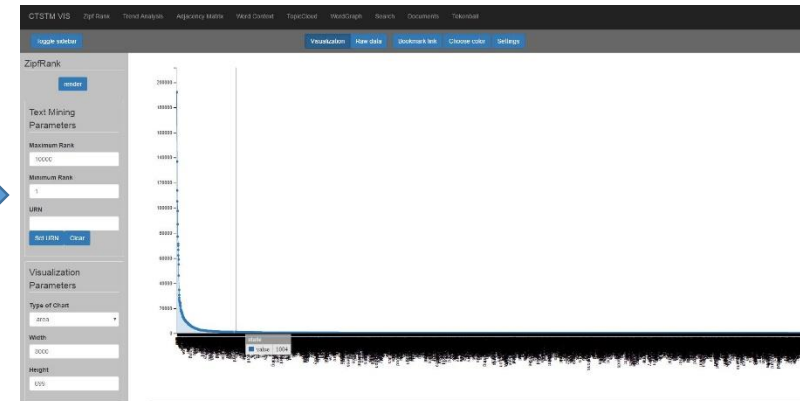


1	die	93703
2	und	79139
3	der	78999
4	in	52893
5	wir	43250
6	das	42996
7	ist	34754
8	auch	33023
9	zu	32870
10	ich	28174
11	dass	28017
12	den	27955
13	es	26908
14	nicht	26229
15	fuer	23742
16	von	20522
17	mit	20013

RESTful Data Requests



RESTful Visualisation Requests



Graphical User Interface

Layer 1 Raw Data

- Raw Data as webservice



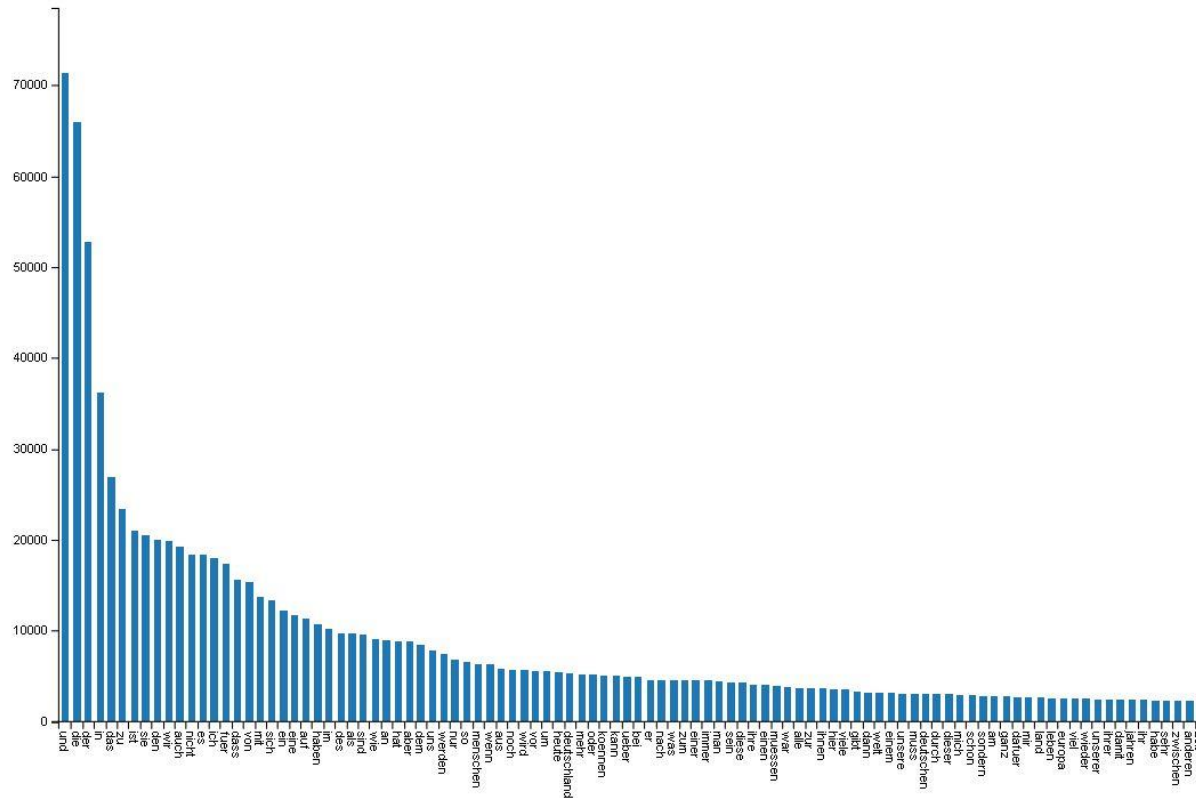
The screenshot shows a web browser window with the URL `ctstm.informatik.uni-leipzig.de:8080/ctstm_germanpresident/ctstm/zipfrank?count&showrank&max=100&min=0`. The browser displays a list of words and their corresponding frequencies, ordered from highest to lowest. The list is as follows:

1	und	71299
2	die	65911
3	der	52718
4	in	36139
5	das	26908
6	zu	23339
7	ist	21051
8	sie	20461
9	den	20048
10	wir	19859
11	auch	19300
12	nicht	18412
13	es	18304
14	ich	18010
15	fuer	17364
16	dass	15584
17	von	15373
18	mit	13758
19	sich	13384
20	ein	12158
21	eine	11734
22	auf	11355
23	haben	10664
24	im	10204
25

Layer 2 Visualisations

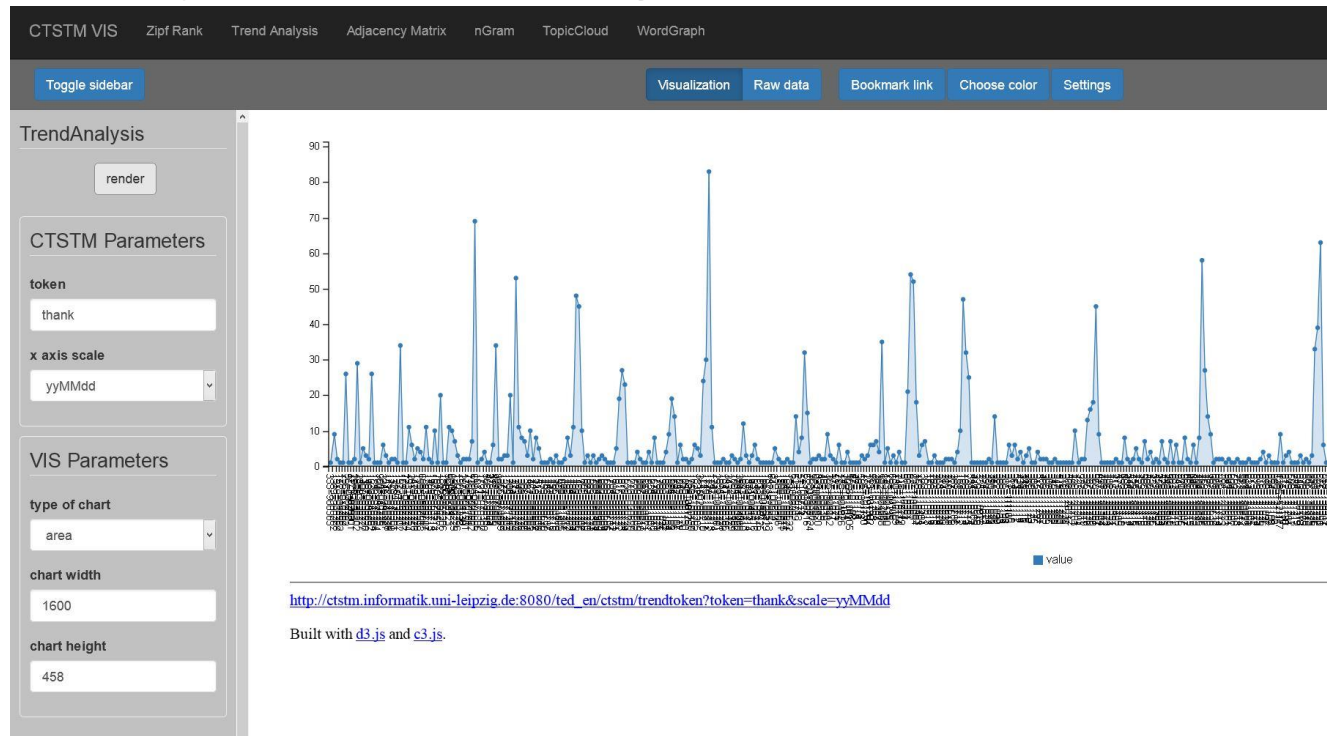
- Generic Data Visualisations as webservice

ctstm.informatik.uni-leipzig.de:8080/ctstm_germanpresident/vis/chart/index.html?function="zipfrank?count&showrank&max=100&min=0"&chartType=bar&width=1000&minValue=3&li



Layer 3 Graphical User Interface

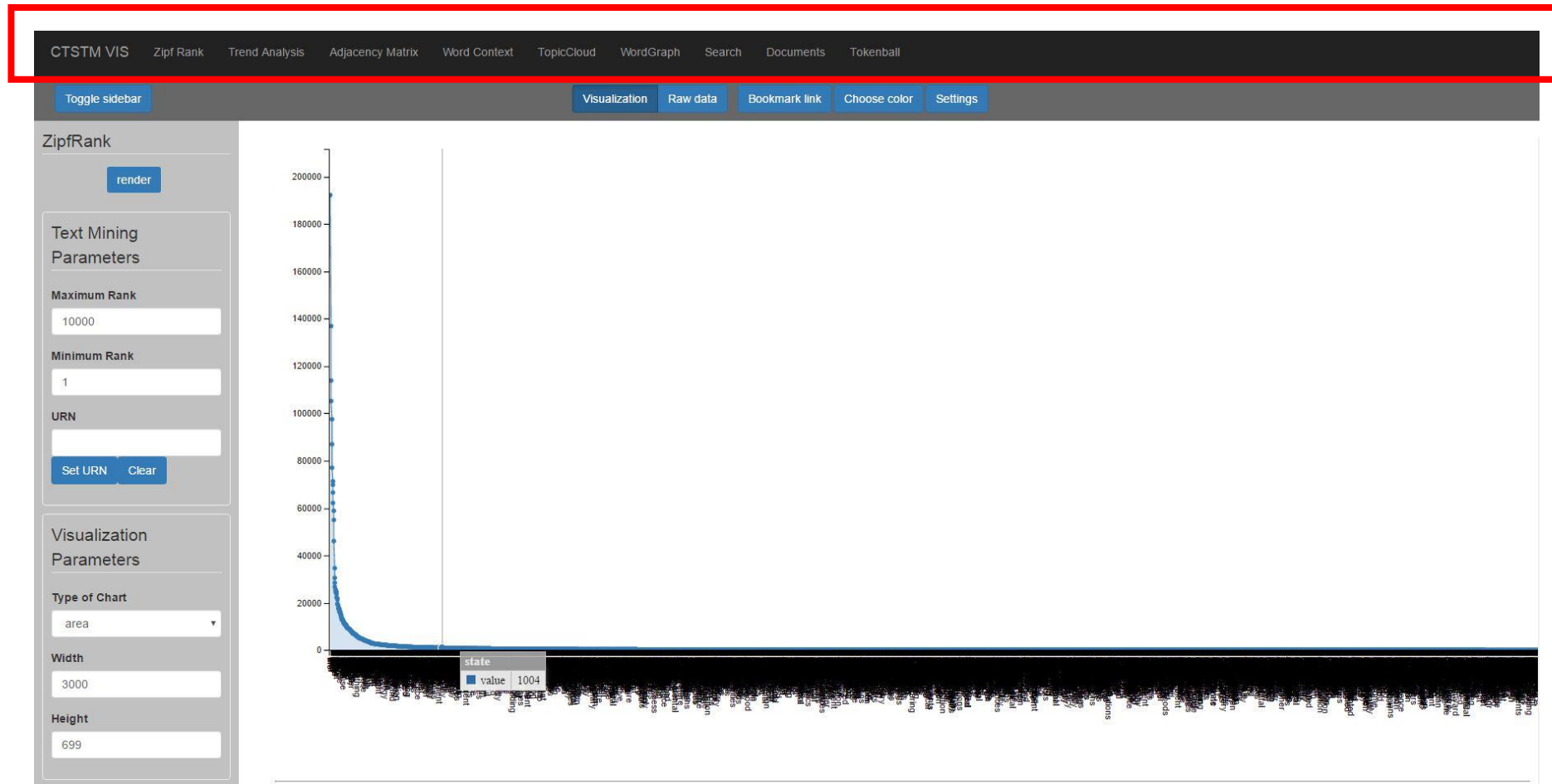
- Open Text Mining Tool as webservice



http://ctstm.informatik.uni-leipzig.de:8080/ted_en/vis/

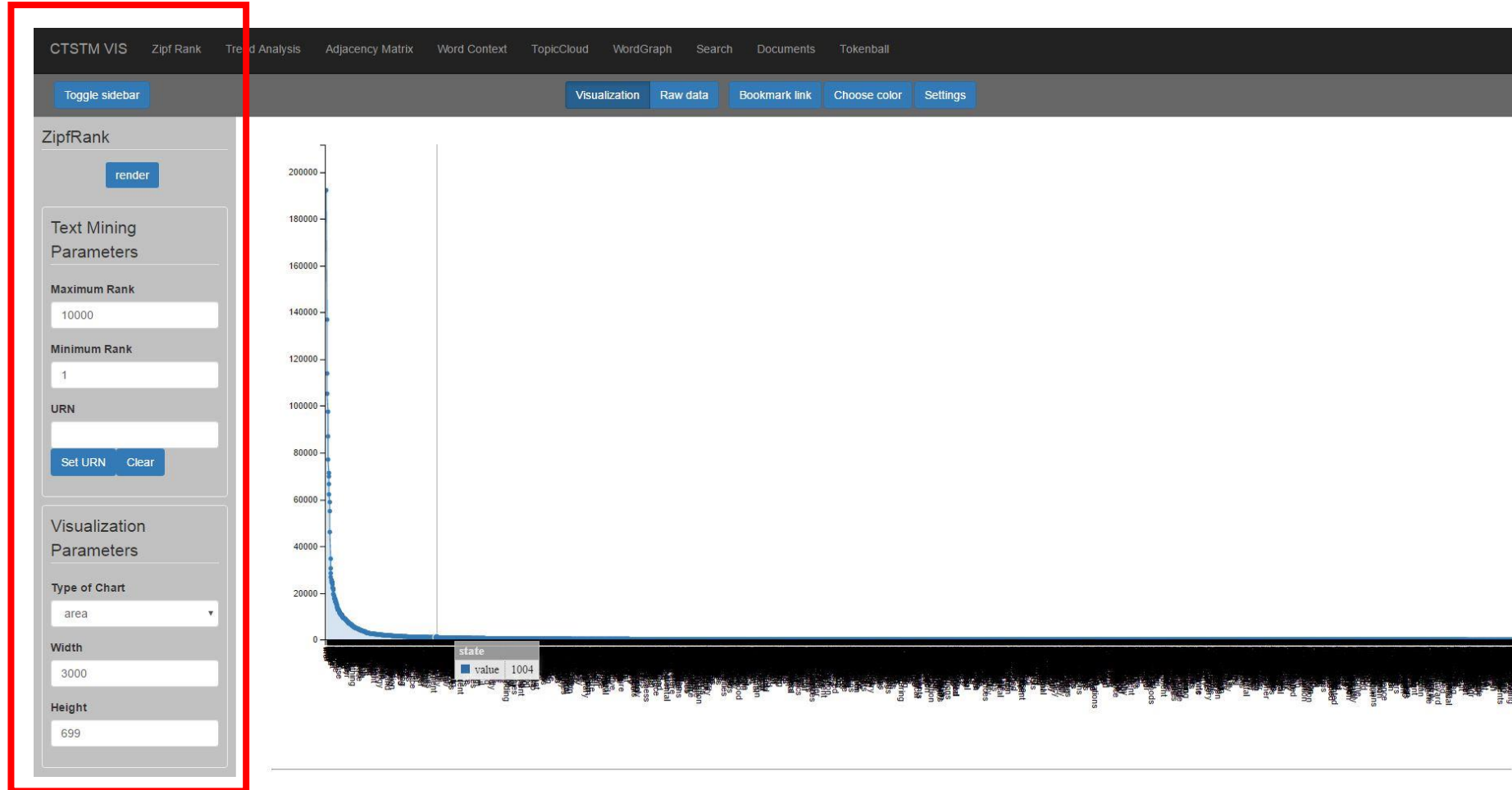
Interface

Toolbox

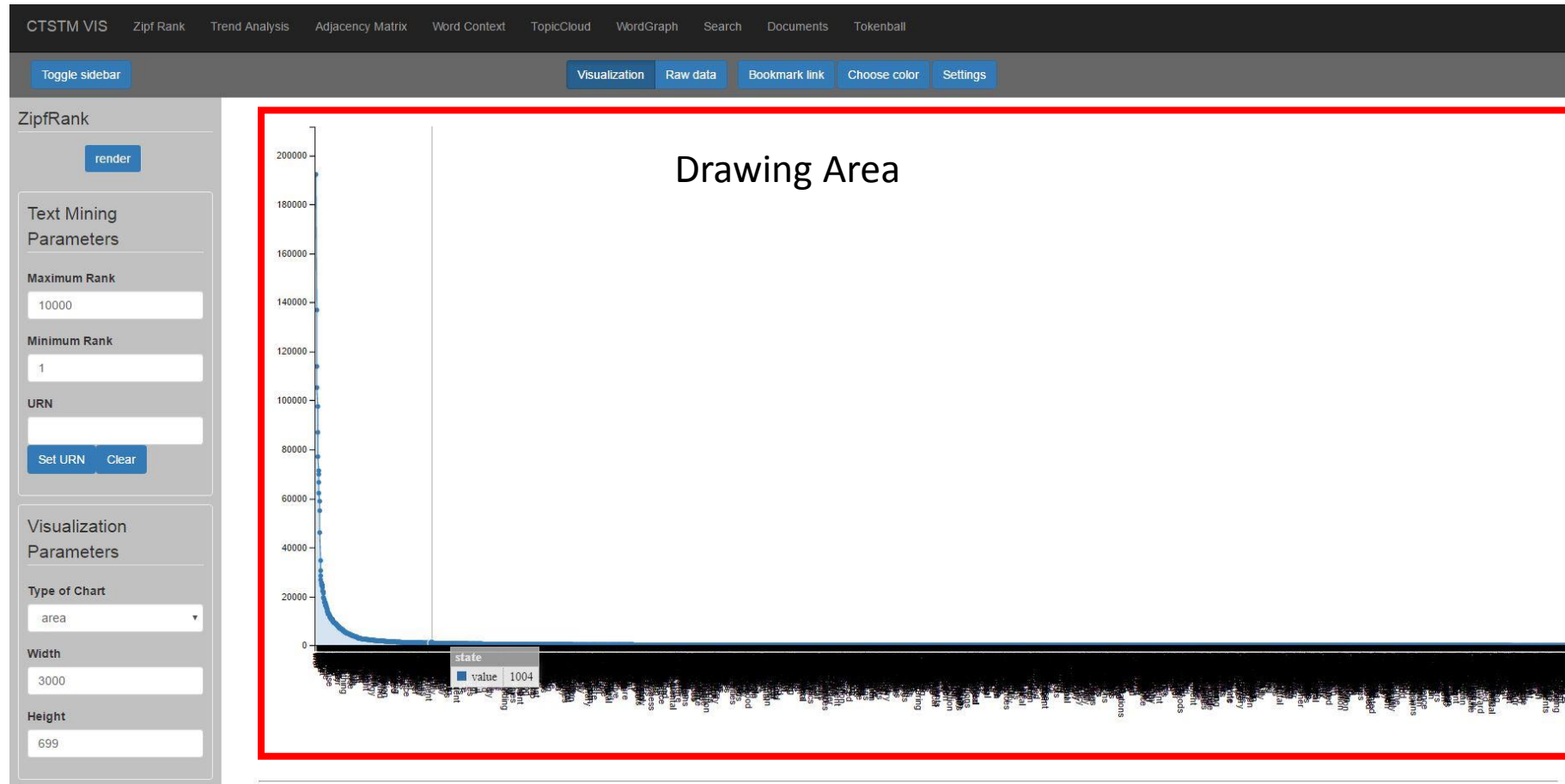


Interface

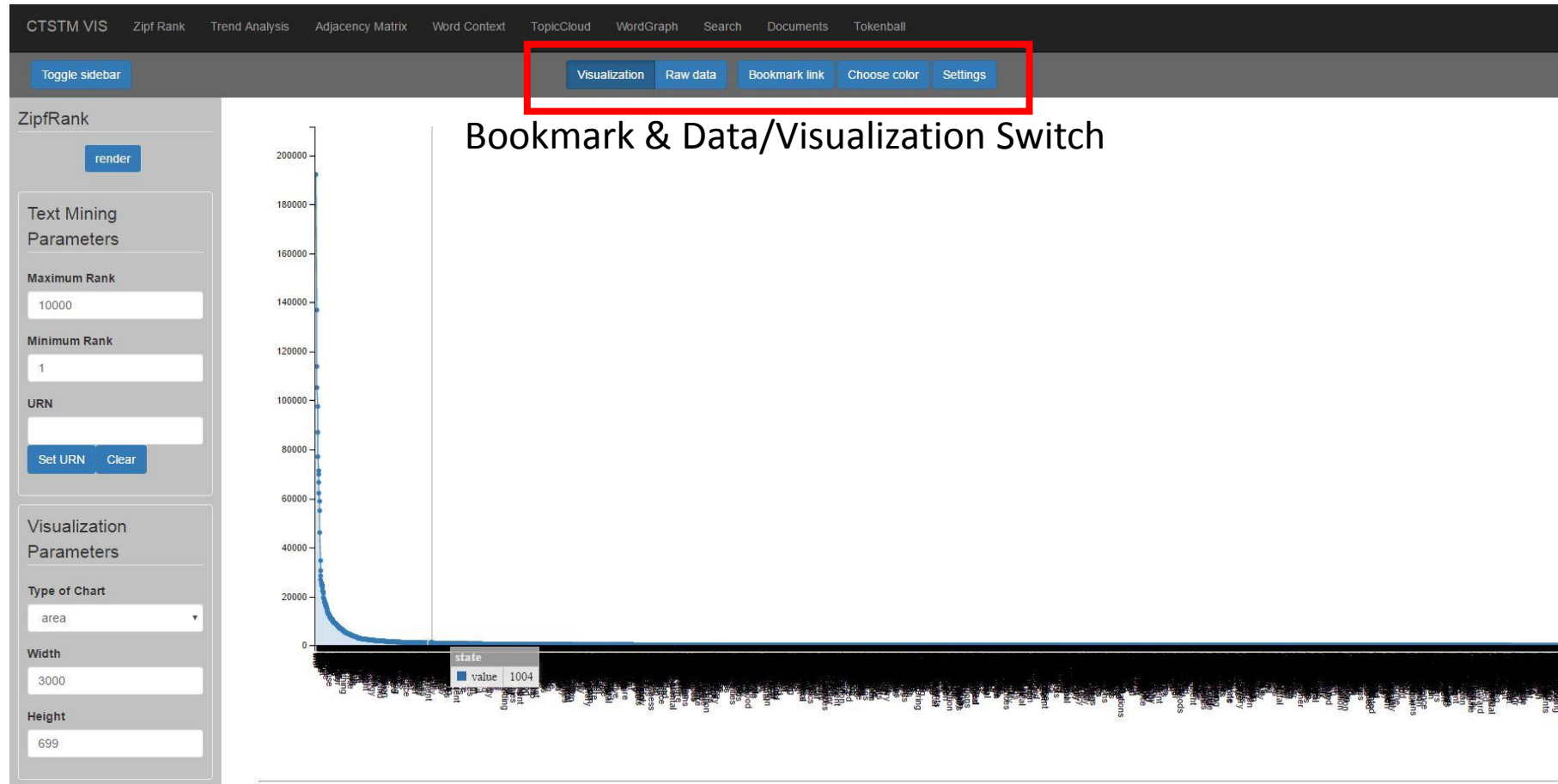
Tool Parameters



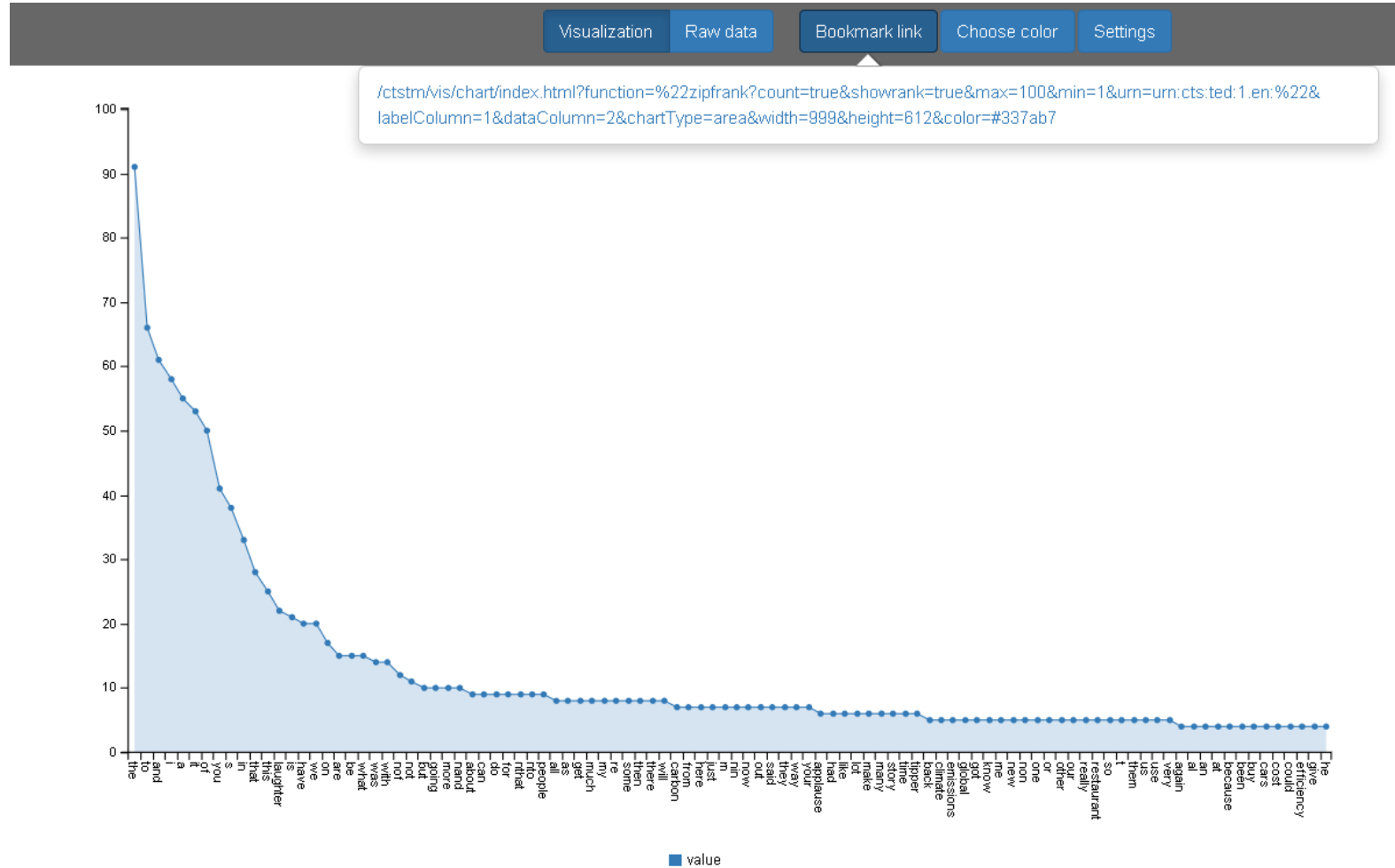
Interface



Interface



Bookmark & Data/Visualization Switch

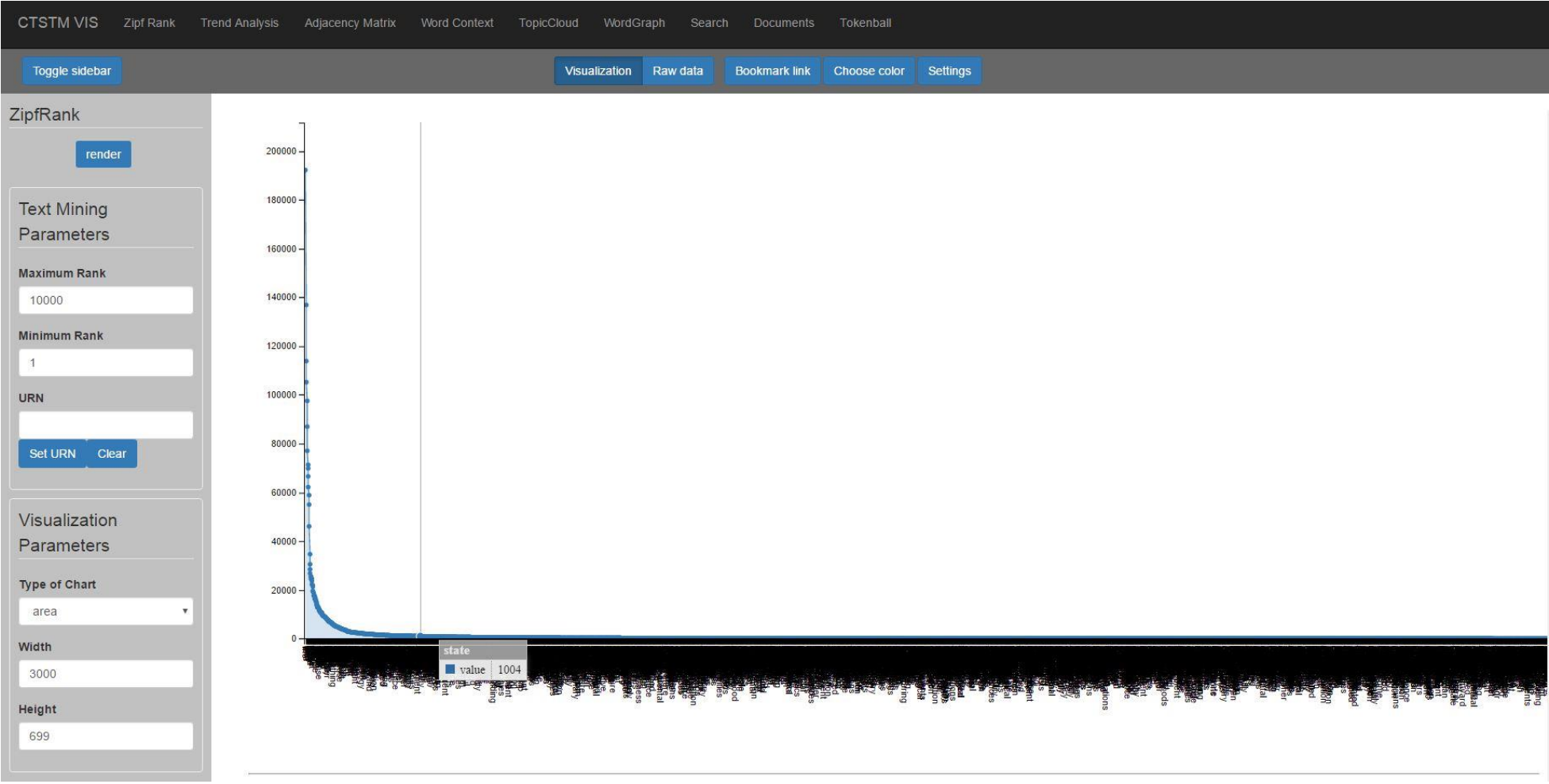


Bookmark & Data/Visualization Switch

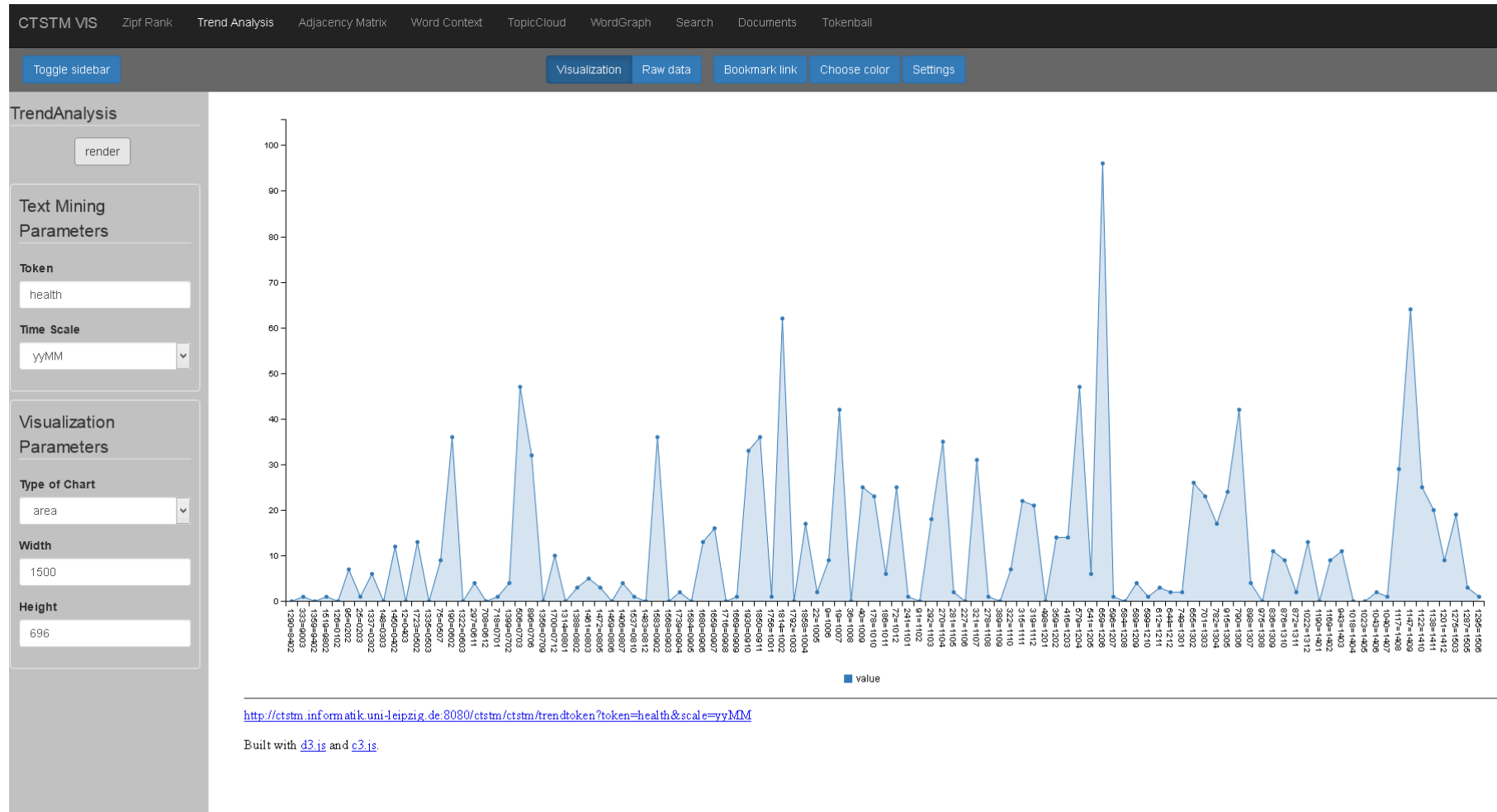
The screenshot shows a web application interface with a dark grey navigation bar at the top containing five buttons: "Visualization", "Raw data", "Bookmark link", "Choose color", and "Settings". The "Bookmark link" button is highlighted with a white mouse cursor. Below the navigation bar, a white tooltip box displays the URL: <http://ctstn.informatik.uni-leipzig.de:8080/ctstn/ctstn/zipfrank?count=true&showrank=true&max=100&min=1&urn=urn:cts:ted:1.en>. Below the tooltip, a word frequency list is displayed in a monospaced font. The list consists of 50 rows, each with a line number, a word, and a frequency value. The word "laughter" is highlighted in blue, and the number "22" is positioned to its right.

1	the	91	
2	to	66	
3	and	61	
4	i	58	
5	a	55	
6	it	53	
7	of	50	
8	you	41	
9	s	38	
10	in	33	
11	that	28	
12	this	25	
13	laughter		22
14	is	21	
15	have	20	
16	we	20	
17	on	17	
18	are	15	
19	be	15	
20	what	15	
21	was	14	
22	with	14	
23	no	12	
24	not	11	
25	but	10	
26	going	10	
27	more	10	
28	and	10	
29	about	9	
30	can	9	
31	do	9	
32	for	9	
33	that	9	
34	to	9	
35	people	9	
36	all	8	
37	as	8	
38	get	8	
39	much	8	
40	my	8	
41	re	8	
42	some	8	
43	then	8	
44	there	8	
45	will	8	
46	carbon	7	
47	from	7	
48	here	7	
49	just	7	
50	m	7	

Tools – Zipf Ranking / Token Frequency



Tools – Trend Detection



Tools – Adjacency Matrix

CTSTM VIS Zipf Rank Trend Analysis Adjacency Matrix Word Context TopicCloud WordGraph Search Documents Tokenball

Toggle sidebar Visualization Raw data Bookmark link Choose color Settings

Matrix

render

Text Mining Parameters

Left Neighbours

URN

Set URN Clear

Visualization Parameters

Word List (Comma Separated)

the, and, to, of, a, that, i, in, it, you

Number of Neighbours per Token

100

Zoom

3.0

Align Input Tokens Vertically

	the	and	to	of	a	that	i	in	it	you
of	Dark	Light	Light	Light	Light	Light	Light	Light	Light	Light
in	Dark	Light	Light	Light	Light	Light	Light	Light	Light	Light
and	Light	Dark	Light	Light	Light	Light	Light	Light	Light	Light
to	Light	Light	Dark	Light	Light	Light	Light	Light	Light	Light
on	Light	Light	Light	Light	Light	Light	Light	Light	Light	Light
at	Light	Light	Light	Light	Light	Light	Light	Light	Light	Light
is	Light	Light	Light	Light	Light	Light	Light	Light	Light	Light
for	Light	Light	Light	Light	Light	Light	Light	Light	Light	Light
that	Light	Light	Light	Light	Light	Dark	Light	Light	Light	Light
with	Light	Light	Light	Light	Light	Light	Light	Light	Light	Light
from	Light	Light	Light	Light	Light	Light	Light	Light	Light	Light
all	Light	Light	Light	Light	Light	Light	Light	Light	Light	Light
s	Light	Light	Light	Light	Light	Light	Light	Light	Light	Light
about	Light	Light	Light	Light	Light	Light	Light	Light	Light	Light
by	Light	Light	Light	Light	Light	Light	Light	Light	Light	Light
into	Light	Light	Light	Light	Light	Light	Light	Light	Light	Light
so	Light	Light	Light	Light	Light	Light	Light	Light	Light	Light

Tools – Word Context

CTSTM VIS Zipf Rank Trend Analysis Adjacency Matrix Word Context TopicCloud WordGraph Search Documents Tokenball

Toggle sidebar Visualization Raw data Bookmark link Choose color Settings

Word Context

render

Parameters

Number of Results
100

nGram Function
docbyngram

N (as in nGram)
5

Token/nGram
health

Token/nGram to Highlight
health

		words	occurences	document
on	the	lung health of	5	urn:cts:ted:1320.en:
		lung health of asthmatic patients	5	urn:cts:ted:1320.en:
		improve health where it begins	4	urn:cts:ted:2076.en:
		improve the health of women	4	urn:cts:ted:1910.en:
in	the	public health sector	3	urn:cts:ted:2098.en:
		women s health to chance	3	urn:cts:ted:1910.en:
to	improve	the health of	3	urn:cts:ted:1910.en:
leave	women	s health to	3	urn:cts:ted:1910.en:
away	from	a health clinic	2	urn:cts:ted:977.en:
		the health of its predators	2	urn:cts:ted:790.en:
scale	of	basic health services	2	urn:cts:ted:62.en:
		got the health benefits of	2	urn:cts:ted:38.en:
		qualified health workers nlike nurses	2	urn:cts:ted:2167.en:
		less qualified health workers nlike	2	urn:cts:ted:2167.en:
		the health media collaboratory and	2	urn:cts:ted:2112.en:
impact	on	our health than	2	urn:cts:ted:2076.en:
		can improve health where it	2	urn:cts:ted:2076.en:
		s health to chance and	2	urn:cts:ted:1910.en:
		monitoring the health of these	2	urn:cts:ted:1872.en:
for	monitoring	the health of	2	urn:cts:ted:1872.en:
		to health and human services	2	urn:cts:ted:1688.en:
of	men	s health and	2	urn:cts:ted:1606.en:
about	men	s health i	2	urn:cts:ted:1606.en:
		range of health care interventions	2	urn:cts:ted:1557.en:
need	to	deliver health care	2	urn:cts:ted:1557.en:
impact	of	a health condition	2	urn:cts:ted:1557.en:
		global health point of view	2	urn:cts:ted:1557.en:
from	a	global health point	2	urn:cts:ted:1557.en:
a	range	of health care	2	urn:cts:ted:1557.en:

Tools – Topic Cloud

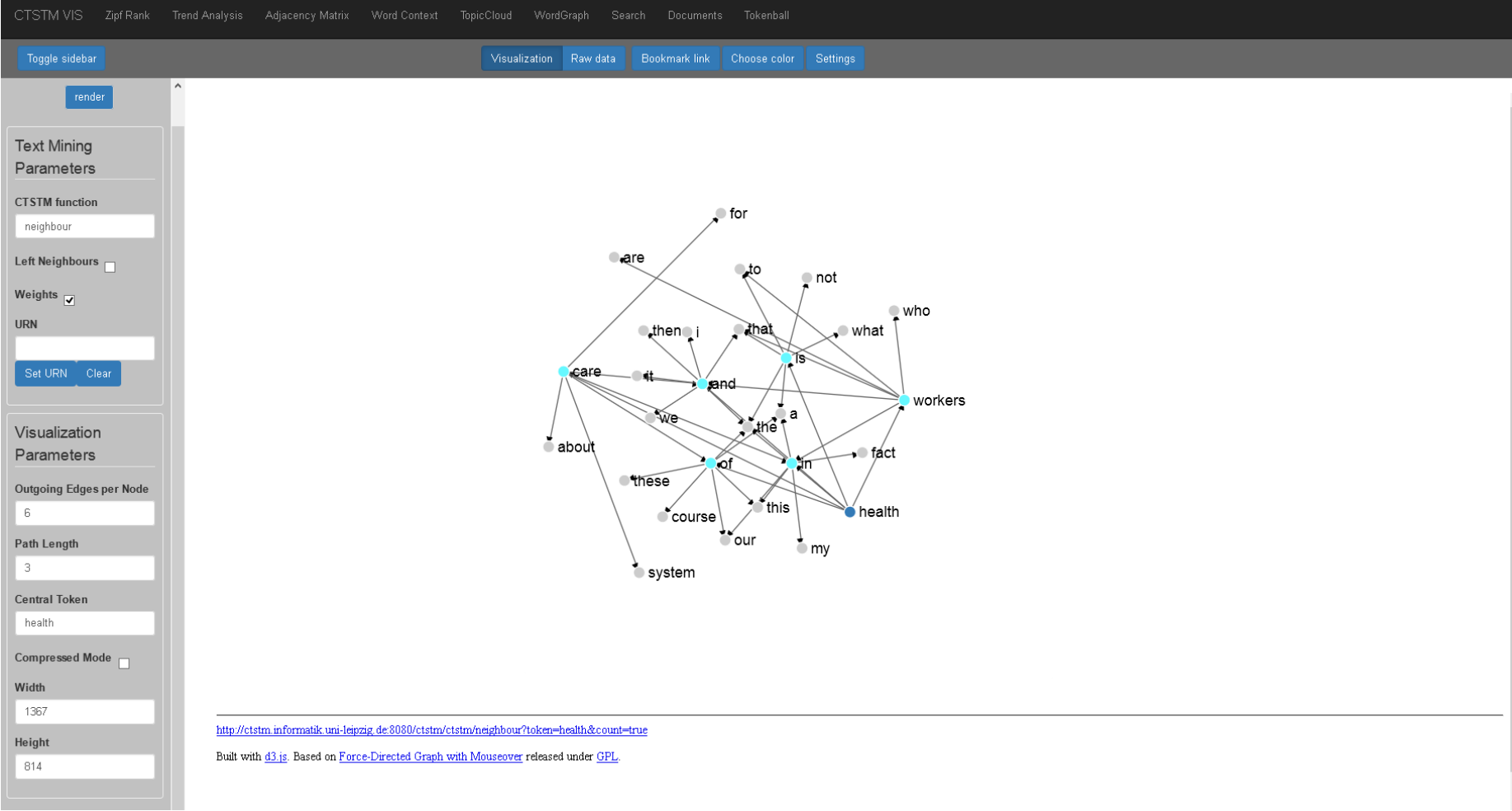
The screenshot displays the CTSTM V4 Topic Cloud tool interface. The main area contains a grid of 12 topic clouds, each representing a different topic. The words in the clouds are sized and colored based on their frequency and importance. The topics shown are: data, cities, water, space, art, machine, ocean, water, sea, land, video, game, india, country, nornis, war, economy, market, percent, company, animals, eat, bees, plants, ideas, god, human, science, language, family, man, him, saw, patient, patients, cancer, and hiv.

The interface includes several control panels on the left side:

- TopicCloud:** A search bar and a "render" button.
- Text Mining Parameters:** A checkbox for "Use token weights" and a "Tokens per Big Cloud" input field set to 150.
- Visualization Parameters:** Sliders for "Probability" (0.05), "Base" (30), "Width of Big Clouds" (500), and "Height of Big Clouds" (300). A checkbox for "Load Document Names" is also present.

At the bottom of the interface, there is a URL and a "Print with" option.

Tools – Word Graph



CTSTM VIS Zipf Rank Trend Analysis Adjacency Matrix Word Context TopicCloud WordGraph Search Documents Tokenball

Toggle sidebar Visualization Raw data Bookmark link Choose color Settings

render

Text Mining Parameters

CTSTM function
neighbour

Left Neighbours

Weights

URN

Set URN Clear

Visualization Parameters

Outgoing Edges per Node
6

Path Length
3

Central Token
health

Compressed Mode

Width
1367

Height
814

<http://ctstm.informatik.uni-leipzig.de/8080/ctstm/ctstm/neighbour?token=health&count=true>

Built with [d3.js](#). Based on [Force-Directed Graph with Mouseover](#) released under [GPL](#).

Tools – Search

CTSTM VIS Zipf Rank Trend Analysis Adjacency Matrix Word Context TopicCloud WordGraph Search Documents Tokenball

Toggle sidebar Visualization Raw data Bookmark link Choose color Settings

Search

render

Text Mining Parameters

Search Function

fulltextsearch
 searchcandidates
 tokensearch

Search String

water

URN

Set URN Clear

Similarity

0.05

Search Method

Document Count Pruning

Term Count Pruning

Lucene Fulltext Index

MySQL Fulltext Index

Token Length Index

Visualization Parameters

CTS link	TagCloud link
urn:cts:ted:1093.en:	urn:cts:ted:1093.en:
urn:cts:ted:1360.en:	urn:cts:ted:1360.en:
urn:cts:ted:390.en:	urn:cts:ted:390.en:
urn:cts:ted:1697.en:	urn:cts:ted:1697.en:
urn:cts:ted:1522.en:	urn:cts:ted:1522.en:
urn:cts:ted:1707.en:	urn:cts:ted:1707.en:
urn:cts:ted:1433.en:	urn:cts:ted:1433.en:
urn:cts:ted:648.en:	urn:cts:ted:648.en:
urn:cts:ted:1794.en:	urn:cts:ted:1794.en:
urn:cts:ted:2281.en:	urn:cts:ted:2281.en:
urn:cts:ted:285.en:	urn:cts:ted:285.en:
urn:cts:ted:27.en:	urn:cts:ted:27.en:
urn:cts:ted:471.en:	urn:cts:ted:471.en:
urn:cts:ted:811.en:	urn:cts:ted:811.en:
urn:cts:ted:978.en:	urn:cts:ted:978.en:
urn:cts:ted:1545.en:	urn:cts:ted:1545.en:
urn:cts:ted:141.en:	urn:cts:ted:141.en:
urn:cts:ted:206.en:	urn:cts:ted:206.en:
urn:cts:ted:1987.en:	urn:cts:ted:1987.en:
urn:cts:ted:1561.en:	urn:cts:ted:1561.en:
urn:cts:ted:1336.en:	urn:cts:ted:1336.en:
urn:cts:ted:1054.en:	urn:cts:ted:1054.en:
urn:cts:ted:1611.en:	urn:cts:ted:1611.en:
urn:cts:ted:2199.en:	urn:cts:ted:2199.en:
urn:cts:ted:664.en:	urn:cts:ted:664.en:
urn:cts:ted:1417.en:	urn:cts:ted:1417.en:

Tools – Token Ball

CTSTM VIS Zipf Rank Trend Analysis Adjacency Matrix Word Context TopicCloud WordGraph Search Documents Tokenball

Toggle sidebar Visualization Raw data Bookmark link Choose color Settings

Tokenball

render

Text Mining Parameters

Maximum Token Rank: 600

Minimum Token Rank: 400

URN:


Visualization Parameters

Shape of Cloud: sphere

Weight Mode: both colour size

Width: 1367

Height: 814



<http://ctstm.informatik.uni-leipzig.de/8080/ctstm/ctstm/zipfrank?max=600&min=400&count=true>

Built with [d3.js](#) and [TagCanvas](#).

Contact

Jochen Tiepmar

E-Mail: jtiepmar@informatik.uni-leipzig.de

Scalable Data Solutions (ScaDS) Leipzig

Universität Leipzig

Ritterstraße 9-13

04109 Leipzig

