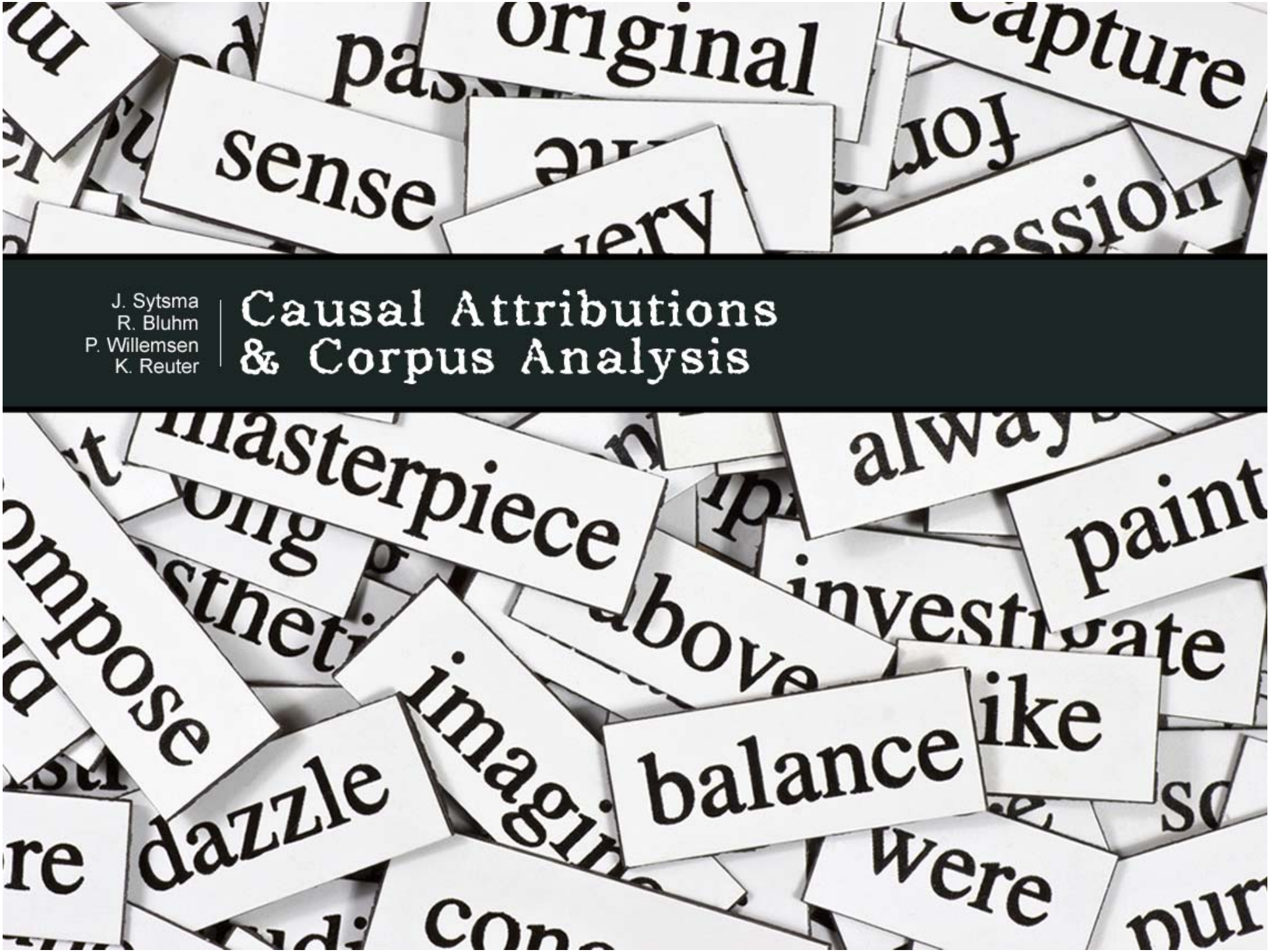


J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

Two goals:

(a) Introduce methods of corpus analysis

(b) Apply these methods to a recent debate concerning ordinary causal attributions

J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

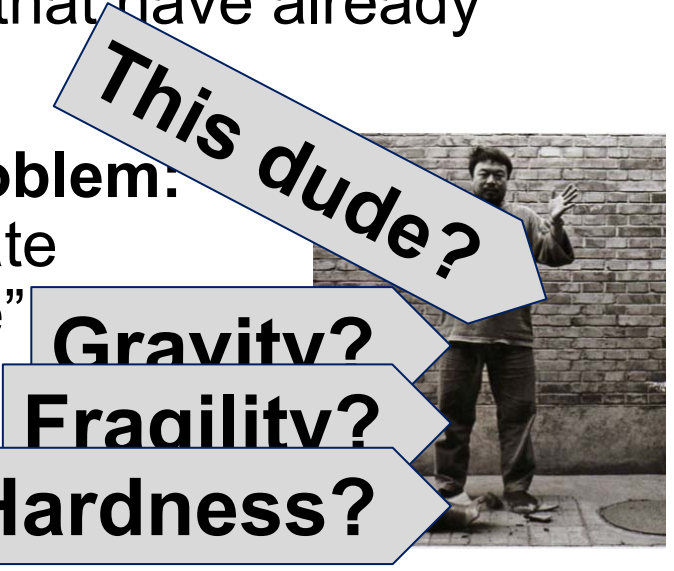
Causal Attributions & Corpus Analysis

Two varieties of causation…

Structural Causation:
Patterns of statistical association…

Actual Causation:
Accounting for things that have already happened…

Causal Selection Problem:
Which factors to elevate to the status of "cause" of the actual effect?



J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

Two goals:

(a) Introduce methods of corpus analysis

(b) Apply these methods to a recent debate concerning ordinary causal attributions

This dude caused this vase to break.

X caused Y.

J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

Two goals:

(a) Introduce methods of corpus analysis

(b) Apply these methods to a recent debate concerning ordinary causal attributions

J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

Background: Norms (especially injunctive norms) seem to matter for ordinary causal attributions…

- Hilton and Slugoski (1986)
- Alicke (1992)
- Knobe and Fraser (2008)
- Hitchcock and Knobe (2009)
- Sytsma, Livengood, and Rose (2012)
- Reuter et al. (2014)
- Kominsky et al. (2015)
- Livengood, Sytsma, and Rose (2016)
- And many others…

J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

Typical effect is a direct comparison in a conjunction case:

- Two "agents" perform the same action, bringing about an outcome
- **Conjunction:** jointly necessary
- One agent violates an injunctive norm, the other does not
- **Comparison:** causal ratings higher for agent who violates the norm

J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

Example:
Knobe & Fraser (2008), Pen Case

The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take the pens, but faculty members are supposed to buy their own.

The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist has repeatedly emailed them reminders that only administrative assistants are allowed to take the pens.

On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's desk. Both take pens. Later that day, the receptionist needs to take an important message… but she has a problem. There are no pens left on her desk.

Agree with statement (-3=not at all; 0=somewhat; 3=fully):

Professor Smith caused the problem. **Mean=2.2**

The administrative assistant caused the problem. **Mean=-1.2**

J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

Example:
Livengood et al. (2016) based on Knobe (2006)

Lauren and Jane both work for a company that uses a mainframe that can be accessed from terminals on different floors of its building. The mainframe has recently become unstable, so that if more than one person is logged in at the same time, the system crashes. Therefore, the company has instituted a temporary policy restricting the use of terminals so that two terminals are not used at the same time until the mainframe is repaired. The policy prohibits logging into the mainframe from the terminal on any floor except the ground floor.

One day, Lauren logged into the mainframe on the authorized terminal on the ground floor at the exact same time that Jane logged into the mainframe on the unauthorized terminal on the second floor. Lauren and Jane were both unaware that the other was logging in. Sure enough, the system crashed.

Agree with statement (1=strongly disagree; 4=neutral; 7=strongly agree):

Jane caused the system to crash. **Mean=5.21**

Lauren caused the system to crash. **Mean=2.42**

J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

Striking finding from recent work on causal attribution:

Causal ratings correspond with whether or not an injunctive norm was violated.

Four types of explanation:

- pragmatic view
- counterfactual view
- bias view
- responsibility view

Pragmatic View

e.g., Samland and Waldmann (2016)

Skeptical position: features of probes lead participants to interpret the question in terms of an alternative concept, such as accountability or responsibility.

Counterfactual View

e.g., Hitchcock and Knobe (2009)

Causal judgments identify suitable intervention points; injunctive norms relevant to suitability.

Basic idea: people think about how an outcome could have been prevented, focusing on changing those aspects of the situation that are *abnormal*….

Responsibility View

e.g., Sytsma, Livengood, and Rose (2012)

Relevant causal attributions are normative.

Used to indicate more than that an entity brought about or produced an outcome; gives a normative evaluation similar to saying the entity is responsible for or accountable for the outcome.

J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

Evidence from a variety of different types of questionnaire studies has been put forward for these different views.

Including evidence for my view…

But today I'll focus on another source of evidence that has been brought to bear on these debates: corpus analysis

J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

Two goals:

(a) Introduce methods of corpus analysis

(b) Apply these methods to a recent debate concerning ordinary causal attributions

J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

(1) Can corpus analysis provide *independent support* for the thesis that ordinary causal attributions are sensitive to normative information?

(2) Does the evidence coming from corpus analysis support the contention that outcome valence matters for ordinary causal attributions?

(3) Are ordinary causal attributions similar to responsibility attributions?

(4) Are causal attributions of philosophers different from causal attributions we find in corpora of more ordinary language?

J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

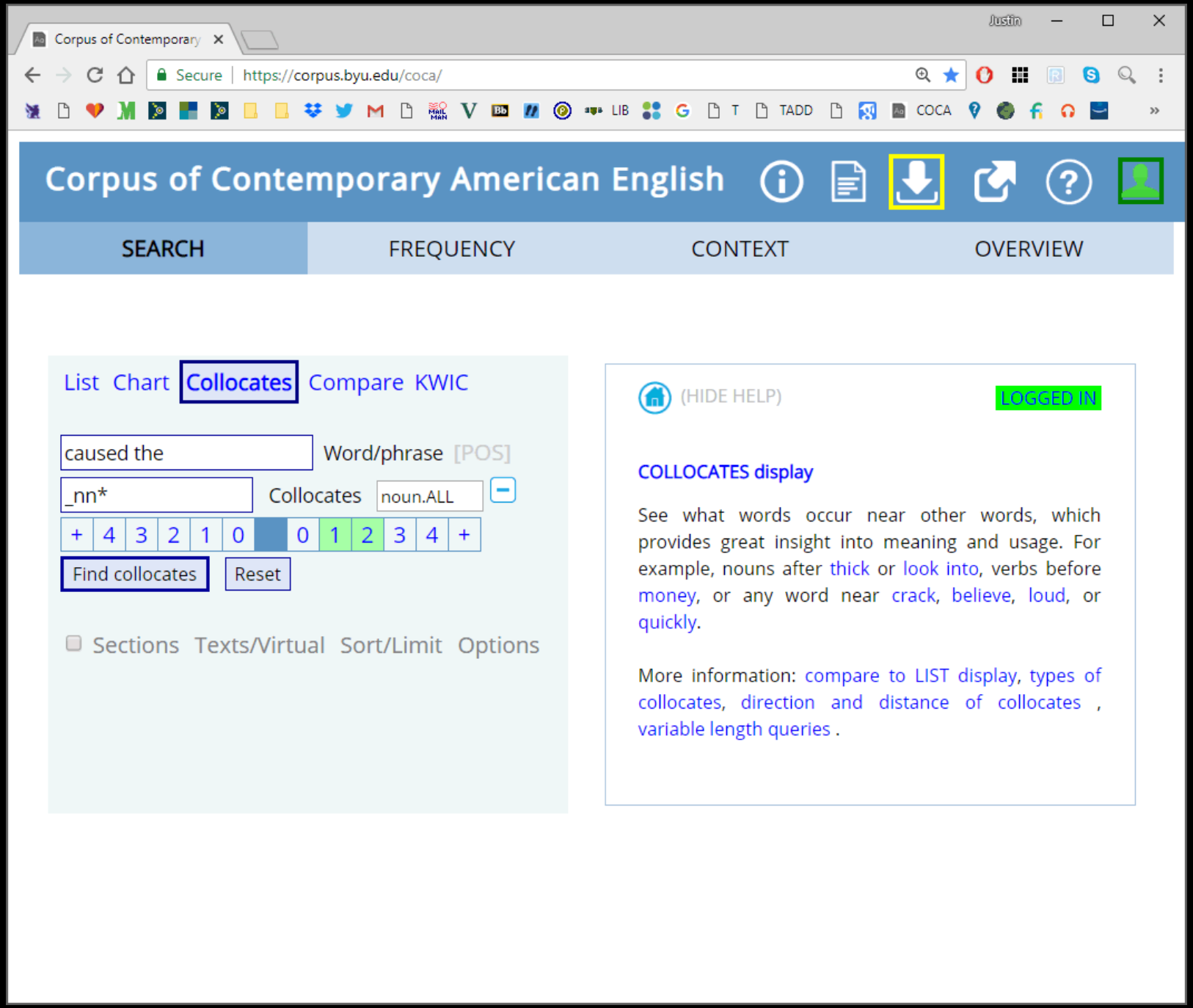
Causal Attributions & Corpus Analysis

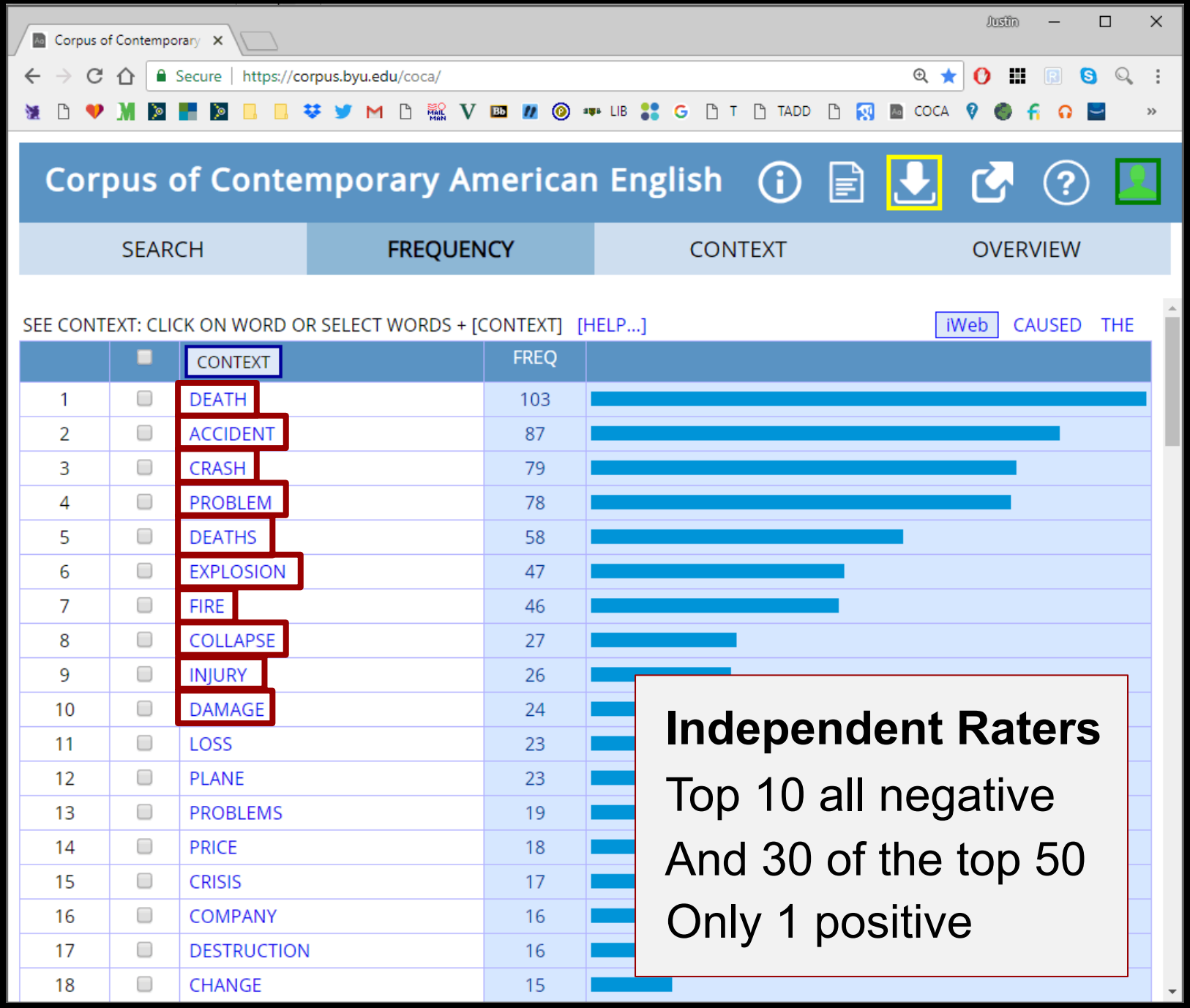
Start with some methods that are non-mathematical…

Use search features available online for COCA: **https://corpus.byu.edu/coca/**

Does the context for ordinary causal attributions tend to be valenced?
(i.e., do we tend to attribute causation more often for bad things than good?)

➔ **Search for nouns occurring after "caused the" in the corpus**


Corpus of Contemporary
Justin
Secure | https://corpus.byu.edu/coca/
Corpus of Contemporary American English
SEARCH
FREQUENCY
CONTEXT
OVERVIEW
List Chart Collocates Compare KWIC
caused the
Word/phrase [POS]
_nn*
Collocates
noun.ALL
+ 4 3 2 1 0 0 1 2 3 4 +
Find collocates
Reset
Sections Texts/Virtual Sort/Limit Options
(HIDE HELP)
LOGGED IN
COLLOCATES display
See what words occur near other words, which provides great insight into meaning and usage. For example, nouns after thick or look into, verbs before money, or any word near crack, believe, loud, or quickly.
More information: compare to LIST display, types of collocates, direction and distance of collocates , variable length queries .

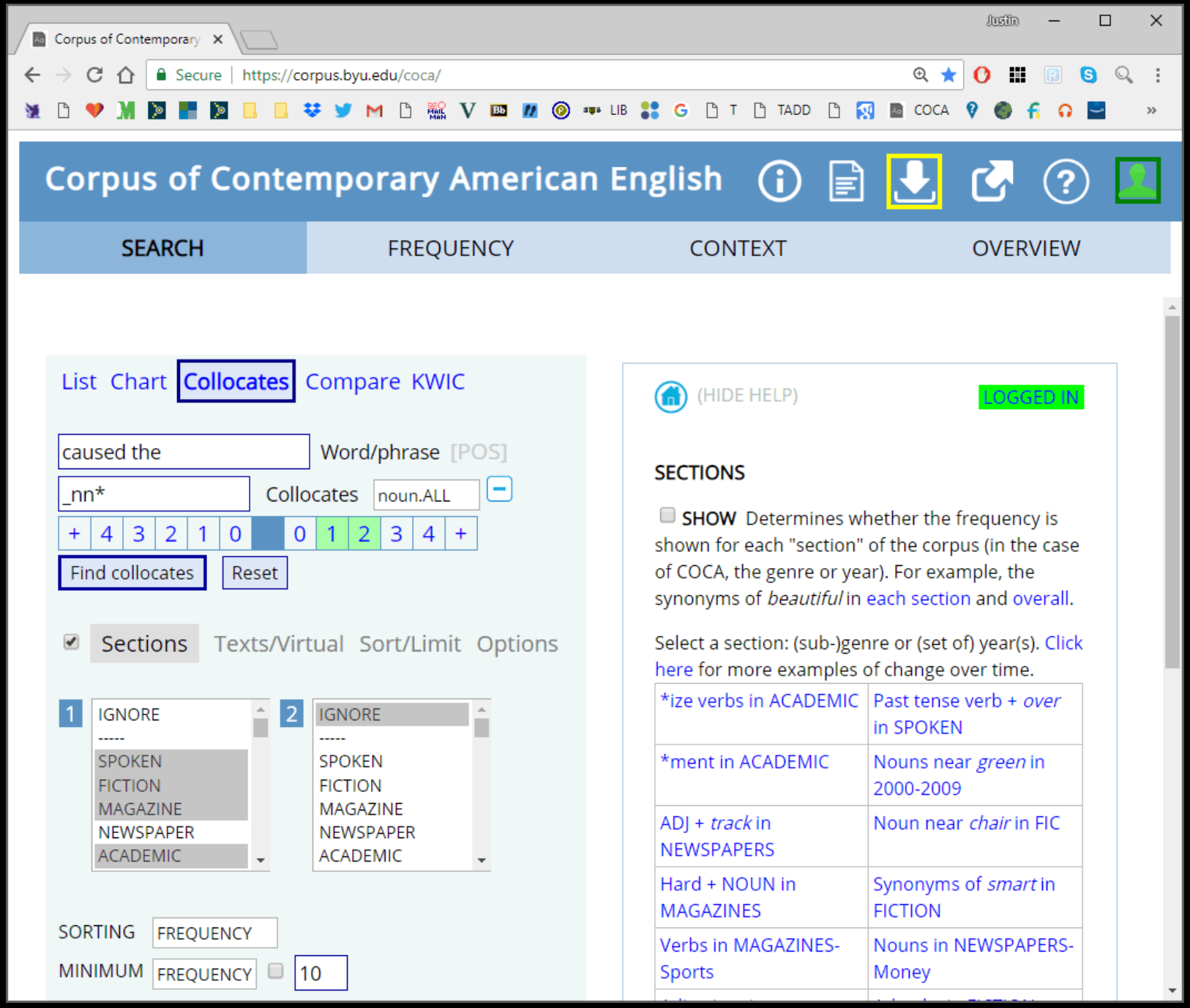

Corpus of Contemporary
Secure https://corpus.byu.edu/coca/
Corpus of Contemporary American English
SEARCH
FREQUENCY
CONTEXT
OVERVIEW
SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT] [HELP...]
iWeb CAUSED THE
CONTEXT
FREQ
1 DEATH 103
2 ACCIDENT 87
3 CRASH 79
4 PROBLEM 78
5 DEATHS 58
6 EXPLOSION 47
7 FIRE 46
8 COLLAPSE 27
9 INJURY 26
10 DAMAGE 24
11 LOSS 23
12 PLANE 23
13 PROBLEMS 19
14 PRICE 18
15 CRISIS 17
16 COMPANY 16
17 DESTRUCTION 16
18 CHANGE 15
Independent Raters
Top 10 all negative
And 30 of the top 50
Only 1 positive

J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

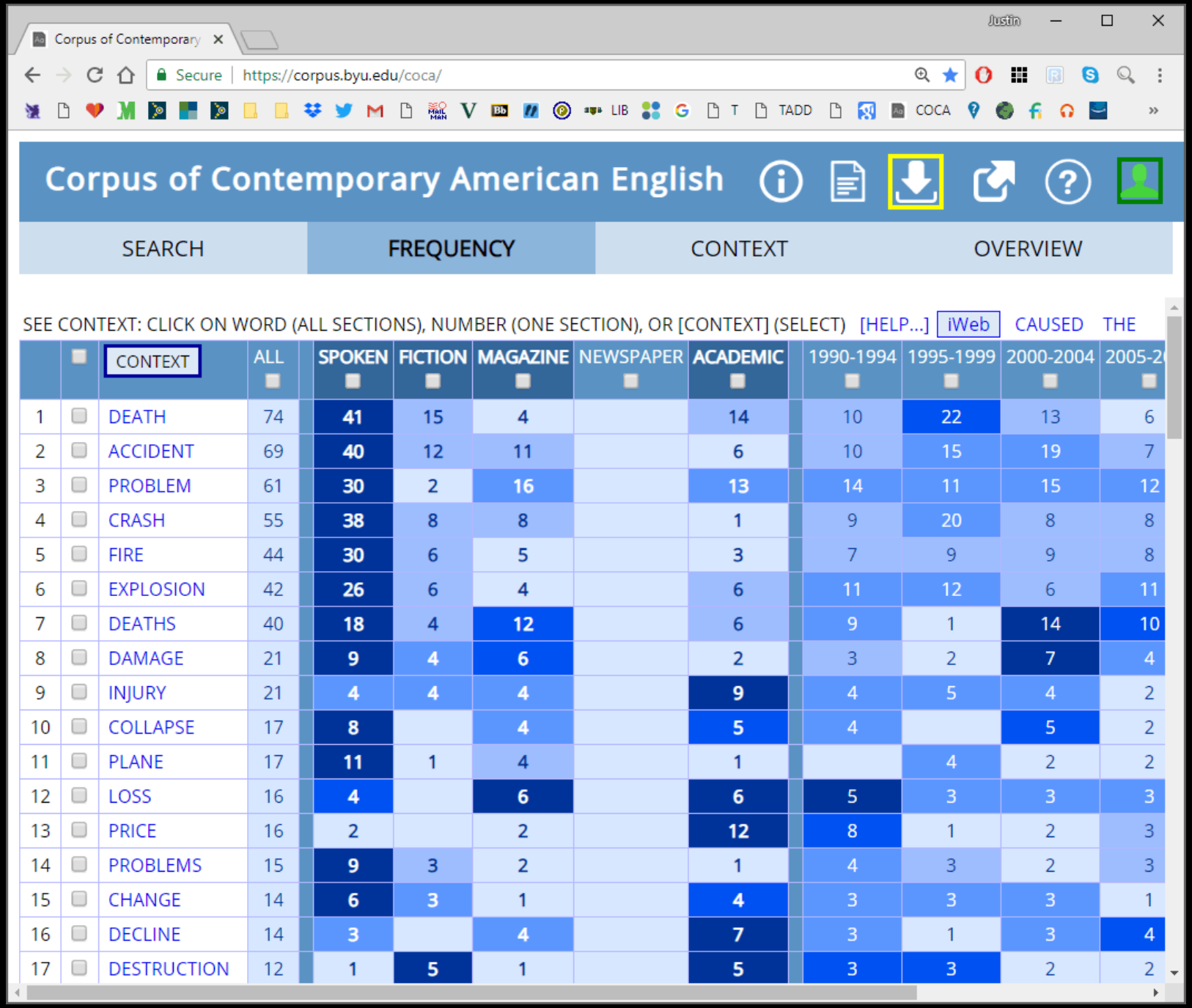
Objection: Newspapers are one source used in COCA. And they are notorious for focusing on negative events.

Reply: Limit the search to other sources.


Corpus of Contemporary
Justin
Secure | https://corpus.byu.edu/coca/
Corpus of Contemporary American English
SEARCH
FREQUENCY
CONTEXT
OVERVIEW
List Chart Collocates Compare KWIC
caused the
Word/phrase [POS]
_nn*
Collocates
noun.ALL
+ 4 3 2 1 0 0 1 2 3 4 +
Find collocates
Reset
Sections
Texts/Virtual
Sort/Limit
Options
1
IGNORE

SPOKEN
FICTION
MAGAZINE
NEWSPAPER
ACADEMIC
2
IGNORE

SPOKEN
FICTION
MAGAZINE
NEWSPAPER
ACADEMIC
SORTING
FREQUENCY
MINIMUM
FREQUENCY
10
(HIDE HELP)
LOGGED IN
SECTIONS
SHOW Determines whether the frequency is shown for each "section" of the corpus (in the case of COCA, the genre or year). For example, the synonyms of beautiful in each section and overall.
Select a section: (sub-)genre or (set of) year(s). Click here for more examples of change over time.
*ize verbs in ACADEMIC
Past tense verb + over in SPOKEN
*ment in ACADEMIC
Nouns near green in 2000-2009
ADJ + track in NEWSPAPERS
Noun near chair in FIC
Hard + NOUN in MAGAZINES
Synonyms of smart in FICTION
Verbs in MAGAZINES-Sports
Nouns in NEWSPAPERS-Money



		CONTEXT	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC	1990-1994	1995-1999	2000-2004	2005-2
1		DEATH	74	41	15	4		14	10	22	13	6
2		ACCIDENT	69	40	12	11		6	10	15	19	7
3		PROBLEM	61	30	2	16		13	14	11	15	12
4		CRASH	55	38	8	8		1	9	20	8	8
5		FIRE	44	30	6	5		3	7	9	9	8
6		EXPLOSION	42	26	6	4		6	11	12	6	11
7		DEATHS	40	18	4	12		6	9	1	14	10
8		DAMAGE	21	9	4	6		2	3	2	7	4
9		INJURY	21	4	4	4		9	4	5	4	2
10		COLLAPSE	17	8		4		5	4		5	2
11		PLANE	17	11	1	4		1		4	2	2
12		LOSS	16	4		6		6	5	3	3	3
13		PRICE	16	2		2		12	8	1	2	3
14		PROBLEMS	15	9	3	2		1	4	3	2	3
15		CHANGE	14	6	3	1		4	3	3	3	1
16		DECLINE	14	3		4		7	3	1	3	4
17		DESTRUCTION	12	1	5	1		5	3	3	2	2

J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

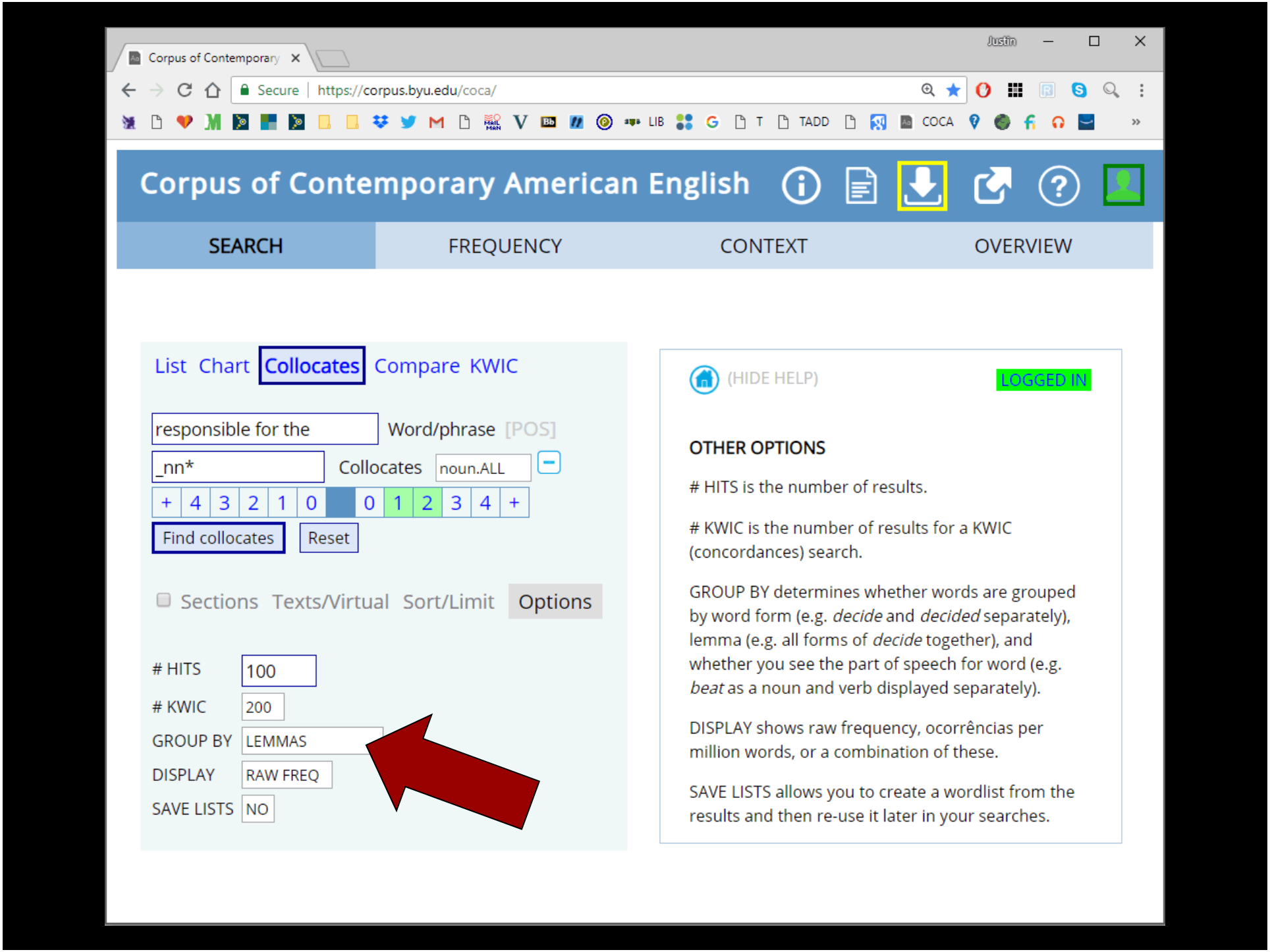
Objection: Negativity is part of the human condition.

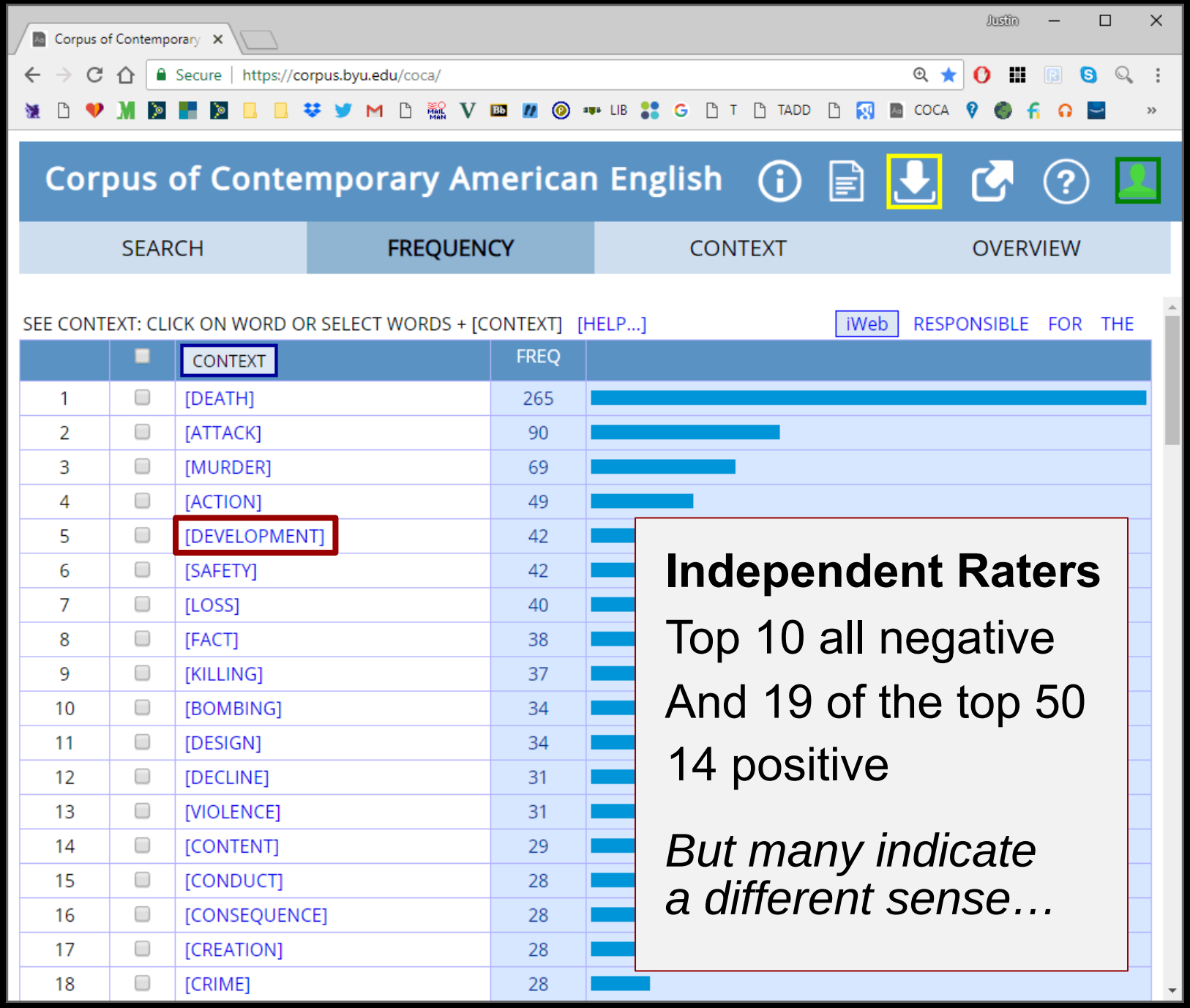
Reply: Check other terms listed as synonyms of the verb “cause.”

phrase	number of negative terms (out of 20)	negative terms
caused the	16	death, accident, crash, problem, collapse, injury, damage, loss, crisis, destruction, decline, extinction, harm, demise, explosion, fire
created the	3	problem, need, illusion
generated the	3	waste, war, killing
induced the	4	coma, panic, defendant, opposition
led to the	7	death, arrest, collapse, demise, loss, firing, end
made the	2	mistake, cut
precipitated the	13	crisis, war, attack, decline, conflict, collapse, downfall, fight, invasion, violence, demise, end, split
produced the	1	plutonium
provoked the	9	anger, fight, violence, murder, resignation, rebellion, strike, crisis, evacuation

Table 1: Most frequent negatively connotated nouns after phrases synonymous to 'caused the'

Are responsibility attributions similar?


Corpus of Contemporary
Justin
Secure | https://corpus.byu.edu/coca/
Corpus of Contemporary American English
SEARCH
FREQUENCY
CONTEXT
OVERVIEW
List Chart Collocates Compare KWIC
responsible for the
Word/phrase [POS]
_nn*
Collocates
noun.ALL
+ 4 3 2 1 0 0 1 2 3 4 +
Find collocates
Reset
Sections Texts/Virtual Sort/Limit Options
HITS 100
KWIC 200
GROUP BY LEMMAS
DISPLAY RAW FREQ
SAVE LISTS NO
(HIDE HELP)
LOGGED IN
OTHER OPTIONS
HITS is the number of results.
KWIC is the number of results for a KWIC (concordances) search.
GROUP BY determines whether words are grouped by word form (e.g. decide and decided separately), lemma (e.g. all forms of decide together), and whether you see the part of speech for word (e.g. beat as a noun and verb displayed separately).
DISPLAY shows raw frequency, ocorrências per million words, or a combination of these.
SAVE LISTS allows you to create a wordlist from the results and then re-use it later in your searches.


Corpus of Contemporary
Secure https://corpus.byu.edu/coca/
Corpus of Contemporary American English
SEARCH
FREQUENCY
CONTEXT
OVERVIEW
SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT] [HELP...]
iWeb RESPONSIBLE FOR THE
CONTEXT
FREQ
1 [DEATH] 265
2 [ATTACK] 90
3 [MURDER] 69
4 [ACTION] 49
5 [DEVELOPMENT] 42
6 [SAFETY] 42
7 [LOSS] 40
8 [FACT] 38
9 [KILLING] 37
10 [BOMBING] 34
11 [DESIGN] 34
12 [DECLINE] 31
13 [VIOLENCE] 31
14 [CONTENT] 29
15 [CONDUCT] 28
16 [CONSEQUENCE] 28
17 [CREATION] 28
18 [CRIME] 28
Independent Raters
Top 10 all negative
And 19 of the top 50
14 positive
But many indicate
a different sense…

Causal Attributions & Corpus Analysis

Also more mathematical methods…

"you shall know a word by the company it keeps" (Firth 1957, 11)

Distributional Semantic Models (DSMs)

Represent terms as geometric vectors in high-dimensional "semantic" space.

Cosine of vectors as measure of semantic relatedness.

Causal Attributions & Corpus Analysis

Latent Semantic Analysis (LSA):
"bag-of-words" approach

Window-based methods (HAL)

Context-predicting methods (word2vec)

Details can get daunting.

But some relatively easy tools you can use to get started…

LSA @ CU Boulder

Not secure | lsa.colorado.edu

Latent Semantic Analysis @ CU Boulder

Main Menu

Information	Affiliations
Applications	
Demos	Mail to...

Click on **Main Menu** Items to reveal sub-menus in this frame.

IMPORTANT NOTICE

It is **essential** that you understand the LSA modeling methods before using the applications on this website. Selecting incorrect semantic spaces, number of dimensions, or types of comparisons will result in **flawed** analyses.

PLEASE consult the Information provided on this website BEFORE attempting analyses.

Applications

Near Neighbors	Matrix Comparison	Sentence Comparison	One-To-Many Comparison	Pairwise Comparison
Info	Info	Info	Info	Info

Demonstrations

Educational Text Selection

Info

New! How to Use Web Site from Handbook of LSA Amazon

Executive Summary	1st Time User Help File	LSA News Updated:06/10/15	Download LSA Publications	Mail to Webmaster LSA-NLP.support@colorado.edu

June 2015: We have moved the site to a new, faster server and upgraded the operating system. We have noticed that the site is no longer working in the Safari web browser, which is the default on Mac OS X. If you are on a Mac, we recommend using Google Chrome or Firefox to access the site until this is fixed. If the website doesn't work for you, please submit bug reports against the "lsa.colorado.edu" product on Bugzilla. In the bug report please include your exact operating system version, exact browser version, and the exact set of steps you performed on the website, including the text you used.

LSA @ CU Boulder

Not secure | lsa.colorado.edu

Latent Semantic Analysis @ CU Boulder

Main Menu

Information	Affiliations
Applications	
Demos	Mail to...

Click on **Main Menu** Items to reveal sub-menus in this frame.

IMPORTANT NOTICE

It is **essential** that you understand the LSA modeling methods before using the applications on this website. Selecting incorrect semantic spaces, number of dimensions, or types of comparisons will result in **flawed** analyses.

PLEASE consult the Information provided on this website BEFORE attempting analyses.

One-To-Many Comparison

This interface allows you to compare the similarity of multiple texts within a particular LSA space. One designated text is compared to all other texts.

To compute the similarity of a particular text to many other texts, enter the main text in the first edit field and each of the others in the second box below. Use a blank line to separate each text in the second box. Then press the **'Submit Texts'** button. The system will compute a similarity score between -1 and 1 between the main text and the other submitted texts.

Select a topic space: General_Reading_up_to_1st_year_college (300 factors)

Select the comparison type: term to term

Number of factors to use: (Leave blank for maximum factors available.)

Show vector lengths:

Main text (to be compared to each of the others):

```
cause
```

Texts to compare (separate different texts with a blank line):

```
responsible

blame

fault

praise
```

Submit Texts | Reset Form

IMPORTANT NOTICE

It is **essential** that you understand the LSA modeling methods before using the applications on this website. Selecting incorrect semantic spaces, number of dimensions, or types of comparisons will result in **flawed** analyses.

PLEASE consult the Information provided on this website BEFORE attempting analyses.

One-to-Many Comparison Results

The submitted texts' similarity matrix (in **term to term** space):

Texts	caused
responsible	0.33
blame	0.28
fault	0.30
praise	0.05

Causal Attributions & Corpus Analysis

What do these numbers mean?!

"dog" and "wolf" = 0.30
"dog" and "animal" = 0.15
"dog" and "sandwich" = 0.08

BETTER: a systematic set of comparisons, such as MEN benchmark…

Bruni, Tran, and Baroni (2013), "Multimodal Distributional Semantics," *Journal of Artificial Intelligence Research*, 48.

J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

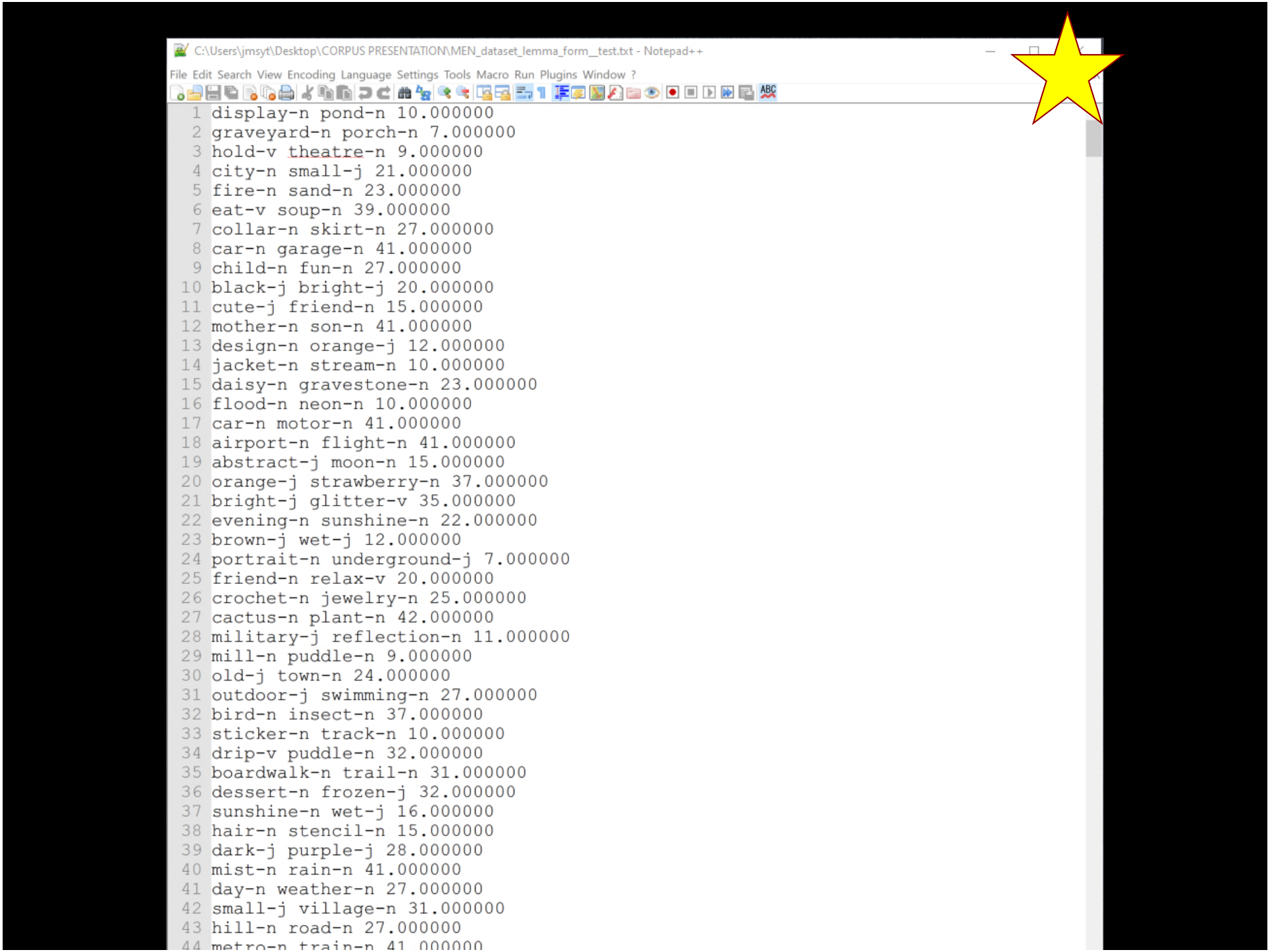
~~http://clic.cimec.unitn.it/~elia.bruni/MEN~~

https://github.com/vecto-ai/word-benchmarks

1,000 pairs of words for evaluation

Semantic relatedness of each pair assessed using Mturk workers

0-50 scale


C:\Users\jmsyt\Desktop\CORPUS PRESENTATION\MEN_dataset_lemma_form__test.txt - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
1 display-n pond-n 10.000000
2 graveyard-n porch-n 7.000000
3 hold-v theatre-n 9.000000
4 city-n small-j 21.000000
5 fire-n sand-n 23.000000
6 eat-v soup-n 39.000000
7 collar-n skirt-n 27.000000
8 car-n garage-n 41.000000
9 child-n fun-n 27.000000
10 black-j bright-j 20.000000
11 cute-j friend-n 15.000000
12 mother-n son-n 41.000000
13 design-n orange-j 12.000000
14 jacket-n stream-n 10.000000
15 daisy-n gravestone-n 23.000000
16 flood-n neon-n 10.000000
17 car-n motor-n 41.000000
18 airport-n flight-n 41.000000
19 abstract-j moon-n 15.000000
20 orange-j strawberry-n 37.000000
21 bright-j glitter-v 35.000000
22 evening-n sunshine-n 22.000000
23 brown-j wet-j 12.000000
24 portrait-n underground-j 7.000000
25 friend-n relax-v 20.000000
26 crochet-n jewelry-n 25.000000
27 cactus-n plant-n 42.000000
28 military-j reflection-n 11.000000
29 mill-n puddle-n 9.000000
30 old-j town-n 24.000000
31 outdoor-j swimming-n 27.000000
32 bird-n insect-n 37.000000
33 sticker-n track-n 10.000000
34 drip-v puddle-n 32.000000
35 boardwalk-n trail-n 31.000000
36 dessert-n frozen-j 32.000000
37 sunshine-n wet-j 16.000000
38 hair-n stencil-n 15.000000
39 dark-j purple-j 28.000000
40 mist-n rain-n 41.000000
41 day-n weather-n 27.000000
42 small-j village-n 31.000000
43 hill-n road-n 27.000000
44 metro-n train-n 41.000000

LSA @ CU Boulder

Not secure | lsa.colorado.edu

Reading list

Latent Semantic Analysis @ CU Boulder

Main Menu

Information	Affiliations
Applications	
Demos	Mail to...

Click on **Main Menu** Items to reveal sub-menus in this frame.

IMPORTANT NOTICE

It is **essential** that you understand the LSA modeling methods before using the applications on this website. Selecting incorrect semantic spaces, number of dimensions, or types of comparisons will result in **flawed** analyses.

PLEASE consult the Information provided on this website BEFORE attempting analyses.

Pairwise Comparison

This interface allows you to compare the similarity of multiple texts within a particular LSA space. Each pair of texts is compared to one another.

To compute the similarity of any number of text segment pairs, enter them into the edit field below. Use a blank line to separate each text you enter. The first and second texts will be compared to one another, the third and fourth will be compared to one another, and so on. Then press the **'Submit Texts'** button. The system will compute a similarity score between -1 and 1 between each pair of texts.

Select a topic space: General_Reading_up_to_1st_year_college (300 factors)

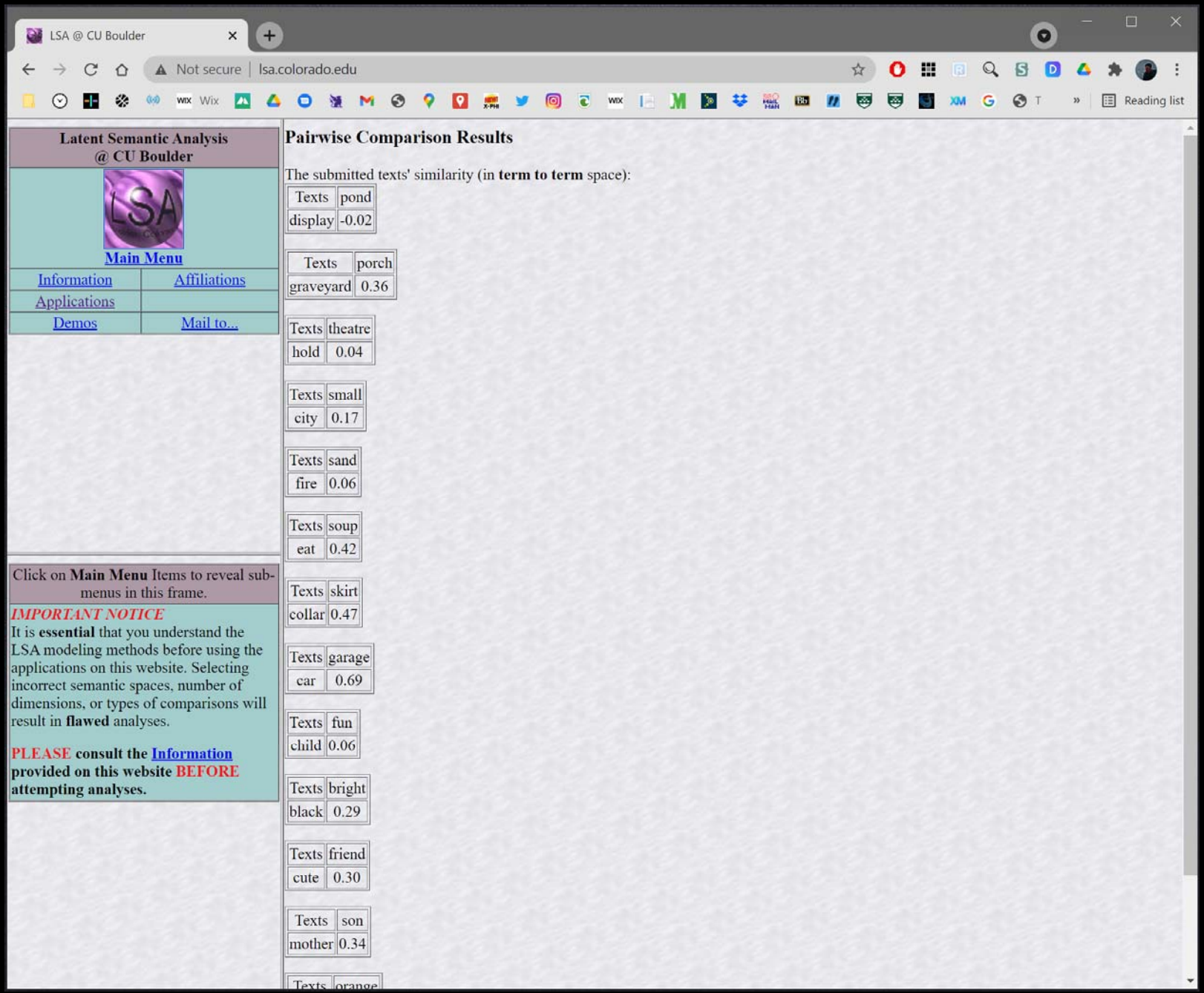
Select the comparison type: term to term

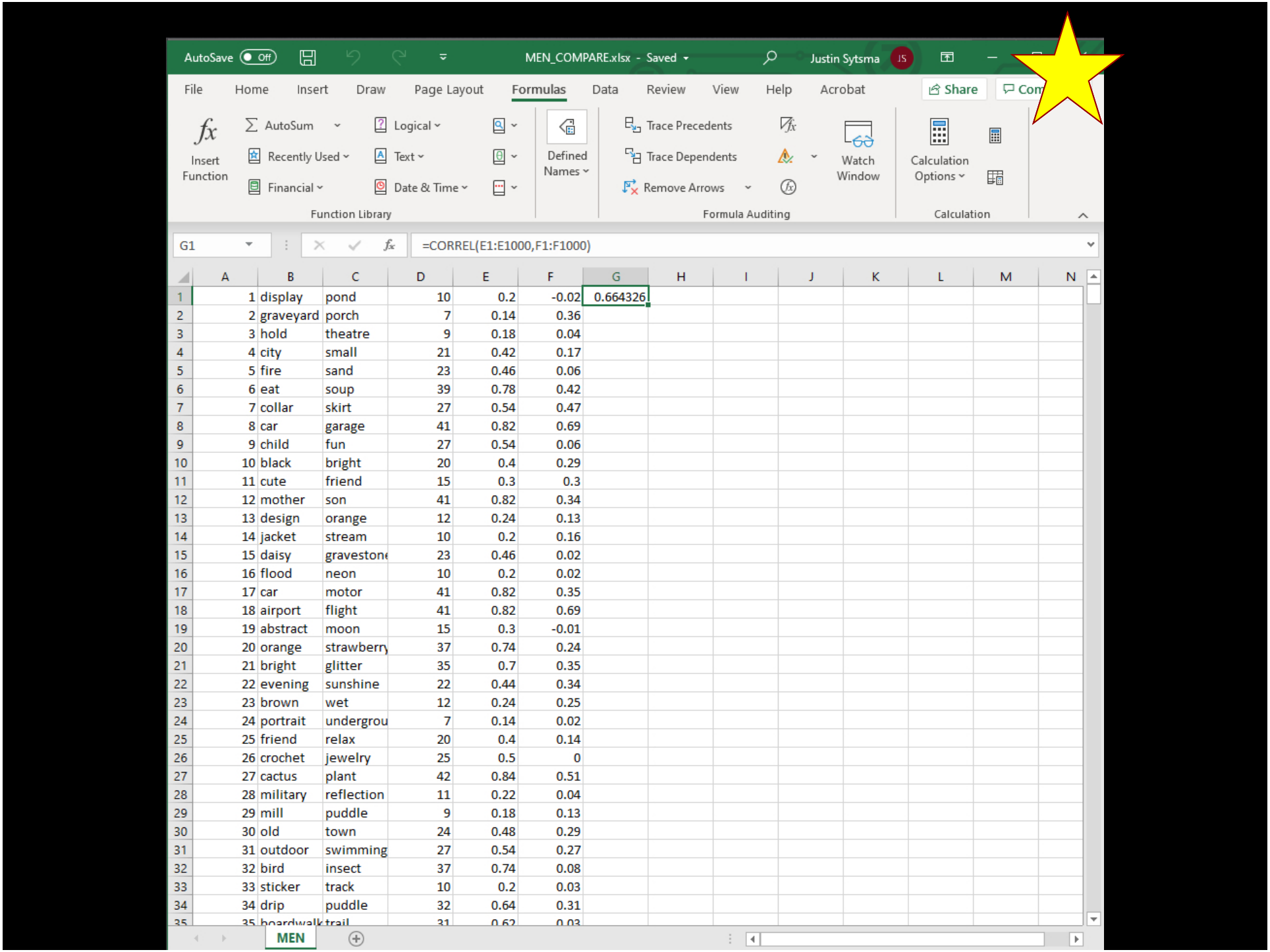
Number of factors to use: (Leave blank for maximum factors available.)

Texts to compare (separate different texts with a blank line):

```
display

pond

graveyard

porch

hold

theatre
```

Submit Texts | Reset Form


LSA @ CU Boulder
Not secure | lsa.colorado.edu
Reading list
Latent Semantic Analysis
@ CU Boulder
Main Menu
Information
Affiliations
Applications
Demos
Mail to...
Click on Main Menu Items to reveal sub-menus in this frame.
IMPORTANT NOTICE
It is essential that you understand the LSA modeling methods before using the applications on this website. Selecting incorrect semantic spaces, number of dimensions, or types of comparisons will result in flawed analyses.
PLEASE consult the Information provided on this website BEFORE attempting analyses.
Pairwise Comparison Results
The submitted texts' similarity (in term to term space):
Texts pond
display -0.02
Texts porch
graveyard 0.36
Texts theatre
hold 0.04
Texts small
city 0.17
Texts sand
fire 0.06
Texts soup
eat 0.42
Texts skirt
collar 0.47
Texts garage
car 0.69
Texts fun
child 0.06
Texts bright
black 0.29
Texts friend
cute 0.30
Texts son
mother 0.34
Texts orange



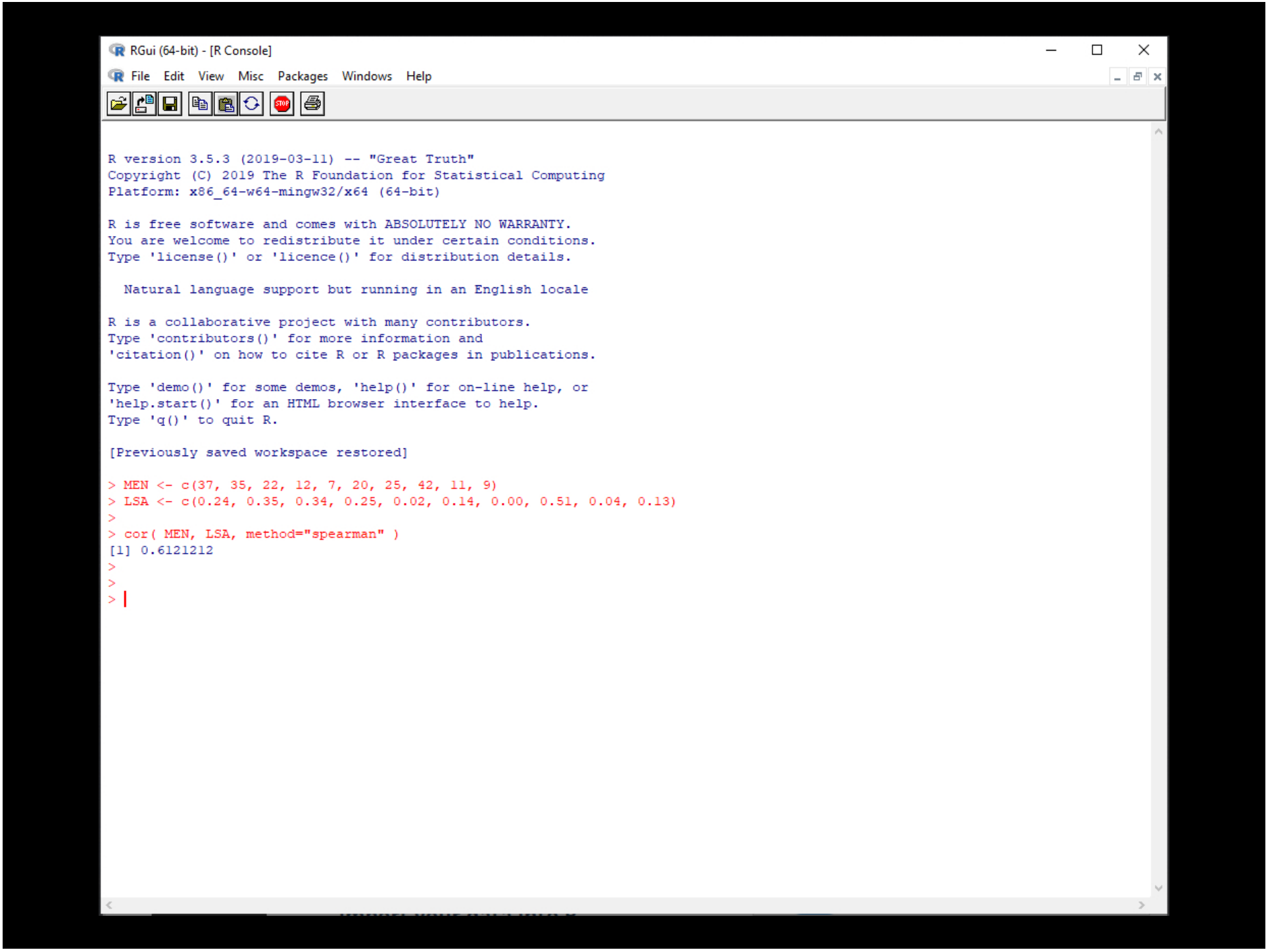
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	1	display	pond	10	0.2	-0.02	0.664326							
2	2	graveyard	porch	7	0.14	0.36								
3	3	hold	theatre	9	0.18	0.04								
4	4	city	small	21	0.42	0.17								
5	5	fire	sand	23	0.46	0.06								
6	6	eat	soup	39	0.78	0.42								
7	7	collar	skirt	27	0.54	0.47								
8	8	car	garage	41	0.82	0.69								
9	9	child	fun	27	0.54	0.06								
10	10	black	bright	20	0.4	0.29								
11	11	cute	friend	15	0.3	0.3								
12	12	mother	son	41	0.82	0.34								
13	13	design	orange	12	0.24	0.13								
14	14	jacket	stream	10	0.2	0.16								
15	15	daisy	gravestone	23	0.46	0.02								
16	16	flood	neon	10	0.2	0.02								
17	17	car	motor	41	0.82	0.35								
18	18	airport	flight	41	0.82	0.69								
19	19	abstract	moon	15	0.3	-0.01								
20	20	orange	strawberry	37	0.74	0.24								
21	21	bright	glitter	35	0.7	0.35								
22	22	evening	sunshine	22	0.44	0.34								
23	23	brown	wet	12	0.24	0.25								
24	24	portrait	undergrou	7	0.14	0.02								
25	25	friend	relax	20	0.4	0.14								
26	26	crochet	jewelry	25	0.5	0								
27	27	cactus	plant	42	0.84	0.51								
28	28	military	reflection	11	0.22	0.04								
29	29	mill	puddle	9	0.18	0.13								
30	30	old	town	24	0.48	0.29								
31	31	outdoor	swimming	27	0.54	0.27								
32	32	bird	insect	37	0.74	0.08								
33	33	sticker	track	10	0.2	0.03								
34	34	drip	puddle	32	0.64	0.31								
35	35	boardwalk	trail	31	0.62	0.03								

J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

Let's find **Spearman's Rho** for items **20-29** ("MEN_dataset_lemma_form__test.txt") for "General_Reading_up_to_1st_year_college" at **lsa.colorado.edu**

19	19	abstract	moon	15	-0.01
20	20	orange	strawberry	37	0.24
21	21	bright	glitter	35	0.35
22	22	evening	sunshine	22	0.34
23	23	brown	wet	12	0.25
24	24	portrait	undergrou	7	0.02
25	25	friend	relax	20	0.14
26	26	crochet	jewelry	25	0
27	27	cactus	plant	42	0.51
28	28	military	reflection	11	0.04
29	29	mill	puddle	9	0.13
30	30	old	town	24	0.29


RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help
R version 3.5.3 (2019-03-11) -- "Great Truth"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Previously saved workspace restored]
> MEN <- c(37, 35, 22, 12, 7, 20, 25, 42, 11, 9)
> LSA <- c(0.24, 0.35, 0.34, 0.25, 0.02, 0.14, 0.00, 0.51, 0.04, 0.13)
>
> cor(MEN, LSA, method="spearman")
[1] 0.6121212
>
>
>

Causal Attributions & Corpus Analysis

Overall, Spearman's rho = **0.67**

Another option is to look at "nearest Neighbors" for the terms of interest:

The terms closest to a term in the semantic space (i.e., terms with highest cosine values with the target term)

LSA @ CU Boulder

Not secure | lsa.colorado.edu

Latent Semantic Analysis @ CU Boulder

Main Menu

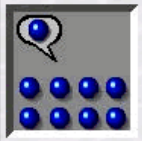
Information	Affiliations
Applications	
Demos	Mail to...

Click on **Main Menu** Items to reveal sub-menus in this frame.

IMPORTANT NOTICE

It is **essential** that you understand the LSA modeling methods before using the applications on this website. Selecting incorrect semantic spaces, number of dimensions, or types of comparisons will result in **flawed** analyses.

PLEASE consult the Information provided on this website BEFORE attempting analyses.



Near Neighbors

This interface allows you to select a set of **n** near neighbor terms based on a submitted term or piece of text **(pseudodoc)**. The terms returned are those in the LSA space which are nearest the submitted term or pseudodoc.

At the end of the return page is a text list of the return items to cut and paste into other applications if you like.

To try the system, enter a term or piece of text in the input area below. Then press the **'Submit Text'** button.

Select a topic space: General_Reading_up_to_1st_year_college (300 factors)

Number of terms to return: 50

Number of factors to use: (Leave blank for maximum factors available.)

Remove terms from return list that appear in corpus with frequency less than (<=): 0

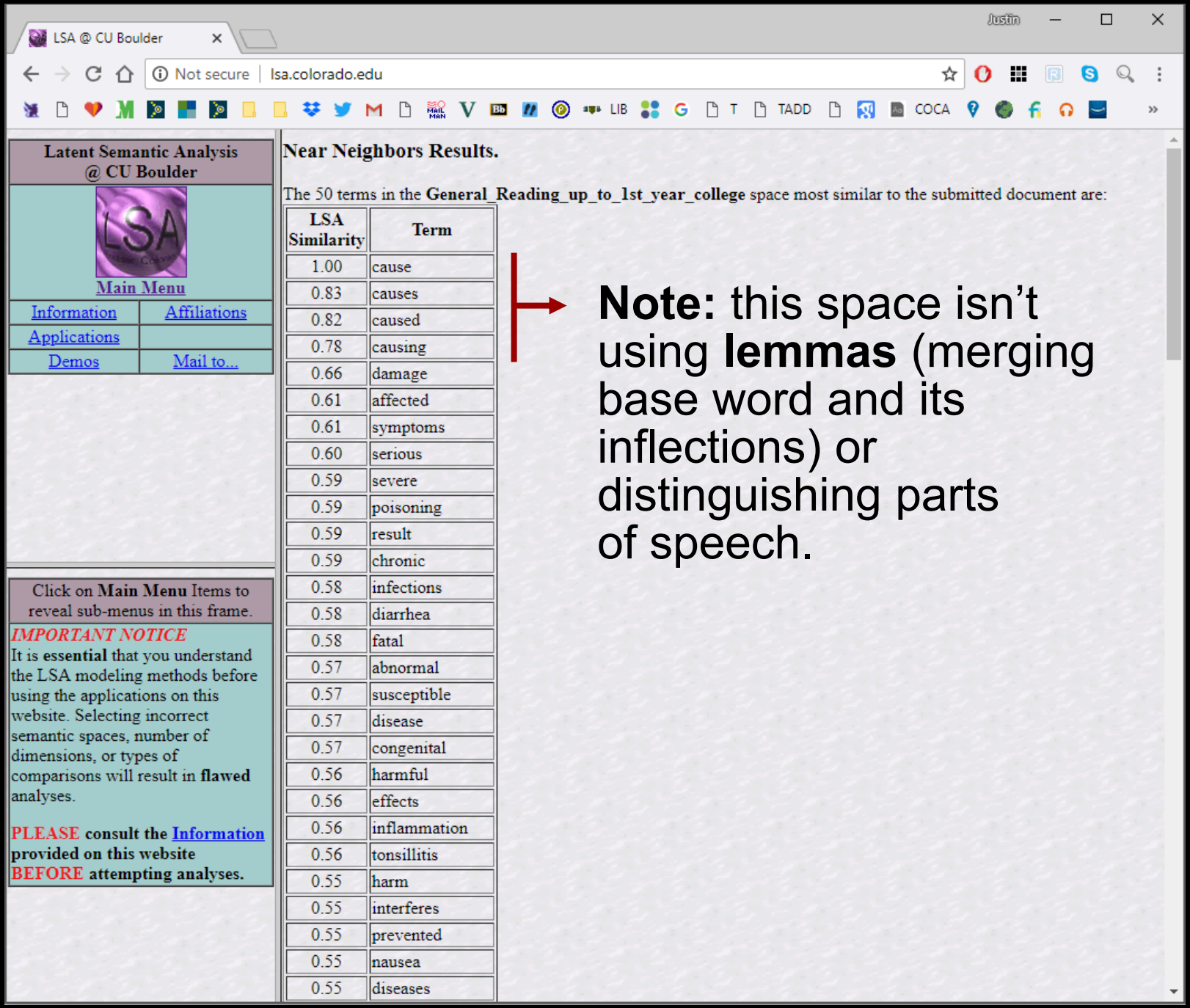
Select the type of input text: term

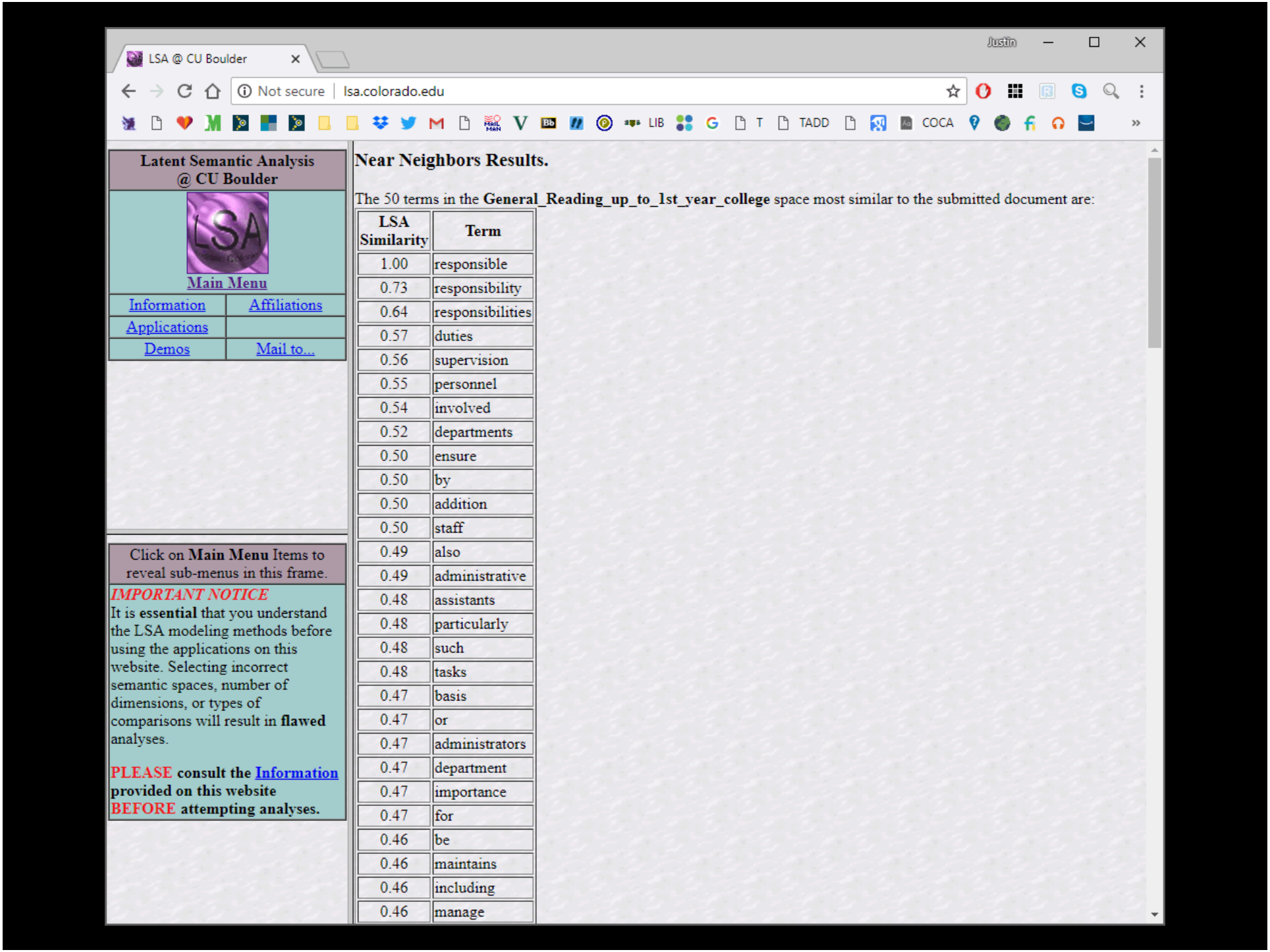
Note: By selecting *term* no weighting is used. Selecting *pseudodoc* uses log entropy weighting.

Text to submit:

cause

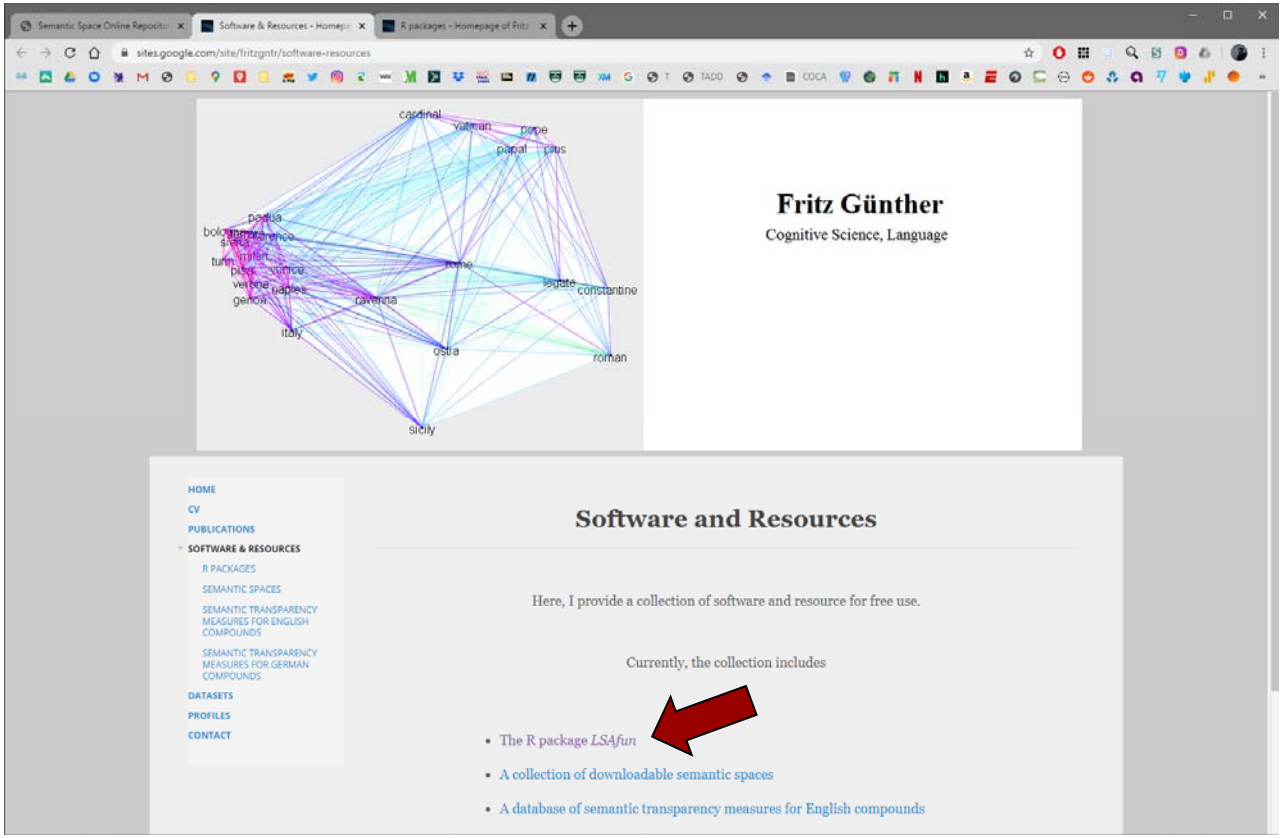
Submit Text | Reset to Defaults


LSA @ CU Boulder
Justin
Not secure
lsa.colorado.edu
Latent Semantic Analysis @ CU Boulder
Main Menu
Information
Affiliations
Applications
Demos
Mail to...
Click on Main Menu Items to reveal sub-menus in this frame.
IMPORTANT NOTICE
It is essential that you understand the LSA modeling methods before using the applications on this website. Selecting incorrect semantic spaces, number of dimensions, or types of comparisons will result in flawed analyses.
PLEASE consult the Information provided on this website BEFORE attempting analyses.
Near Neighbors Results.
The 50 terms in the General_Reading_up_to_1st_year_college space most similar to the submitted document are:
LSA Similarity
Term
1.00 cause
0.83 causes
0.82 caused
0.78 causing
0.66 damage
0.61 affected
0.61 symptoms
0.60 serious
0.59 severe
0.59 poisoning
0.59 result
0.59 chronic
0.58 infections
0.58 diarrhea
0.58 fatal
0.57 abnormal
0.57 susceptible
0.57 disease
0.57 congenital
0.56 harmful
0.56 effects
0.56 inflammation
0.56 tonsillitis
0.55 harm
0.55 interferes
0.55 prevented
0.55 nausea
0.55 diseases
Note: this space isn't using lemmas (merging base word and its inflections) or distinguishing parts of speech.

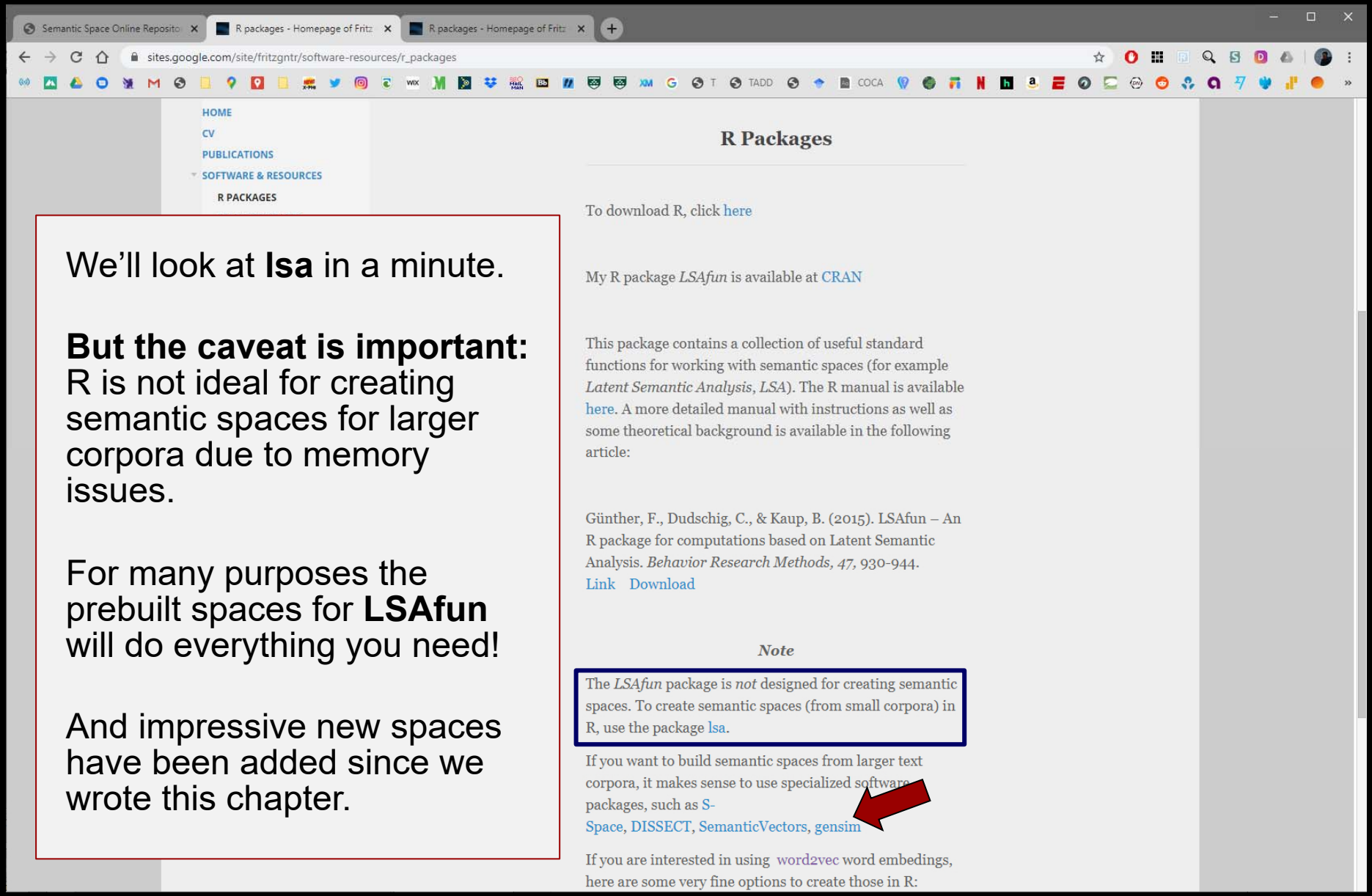

LSA @ CU Boulder
Justin
Not secure
lsa.colorado.edu
LIB
T
TADD
COCA
Latent Semantic Analysis @ CU Boulder
Main Menu
Information
Affiliations
Applications
Demos
Mail to...
Click on Main Menu Items to reveal sub-menus in this frame.
IMPORTANT NOTICE
It is essential that you understand the LSA modeling methods before using the applications on this website. Selecting incorrect semantic spaces, number of dimensions, or types of comparisons will result in flawed analyses.
PLEASE consult the Information provided on this website BEFORE attempting analyses.
Near Neighbors Results.
The 50 terms in the General_Reading_up_to_1st_year_college space most similar to the submitted document are:
LSA Similarity
Term
1.00 responsible
0.73 responsibility
0.64 responsibilities
0.57 duties
0.56 supervision
0.55 personnel
0.54 involved
0.52 departments
0.50 ensure
0.50 by
0.50 addition
0.50 staff
0.49 also
0.49 administrative
0.48 assistants
0.48 particularly
0.48 such
0.48 tasks
0.47 basis
0.47 or
0.47 administrators
0.47 department
0.47 importance
0.47 for
0.46 be
0.46 maintains
0.46 including
0.46 manage

Causal Attributions & Corpus Analysis

Can also load prebuilt spaces into R



https://sites.google.com/site/fritzgntr/home

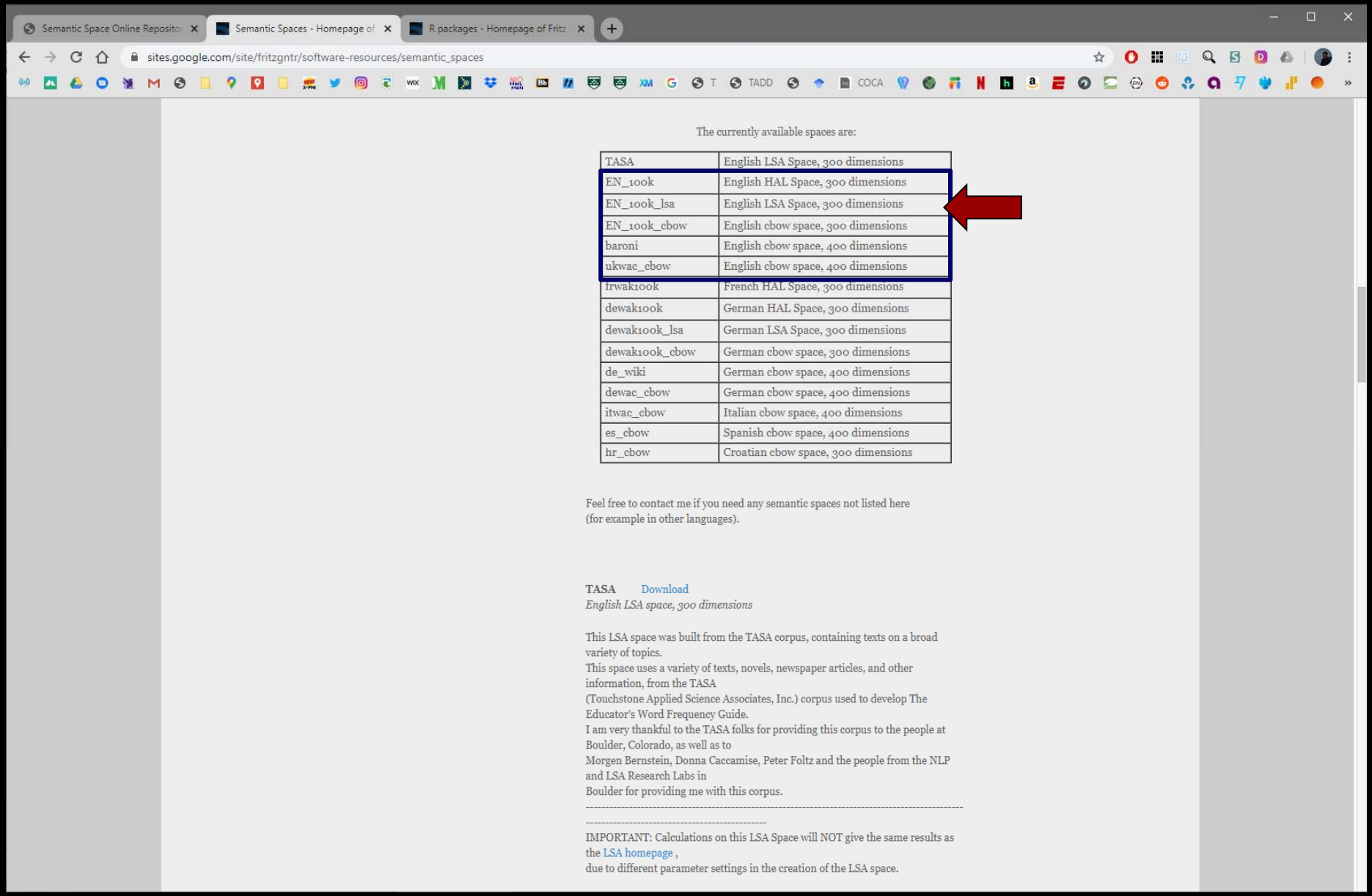


We'll look at **lsa** in a minute.

But the caveat is important: R is not ideal for creating semantic spaces for larger corpora due to memory issues.

For many purposes the prebuilt spaces for **LSAfun** will do everything you need!

And impressive new spaces have been added since we wrote this chapter.



The currently available spaces are:

TASA	English LSA Space, 300 dimensions
EN_100k	English HAL Space, 300 dimensions
EN_100k_lsa	English LSA Space, 300 dimensions
EN_100k_cbow	English cbow space, 300 dimensions
baroni	English cbow space, 400 dimensions
ukwac_cbow	English cbow space, 400 dimensions
frwak100k	French HAL Space, 300 dimensions
dewak100k	German HAL Space, 300 dimensions
dewak100k_lsa	German LSA Space, 300 dimensions
dewak100k_cbow	German cbow space, 300 dimensions
de_wiki	German cbow space, 400 dimensions
dewac_cbow	German cbow space, 400 dimensions
itwac_cbow	Italian cbow space, 400 dimensions
es_cbow	Spanish cbow space, 400 dimensions
hr_cbow	Croatian cbow space, 300 dimensions

Feel free to contact me if you need any semantic spaces not listed here
(for example in other languages).

TASA Download
English LSA space, 300 dimensions

This LSA space was built from the TASA corpus, containing texts on a broad variety of topics.
This space uses a variety of texts, novels, newspaper articles, and other information, from the TASA
(Touchstone Applied Science Associates, Inc.) corpus used to develop The Educator's Word Frequency Guide.
I am very thankful to the TASA folks for providing this corpus to the people at Boulder, Colorado, as well as to
Morgen Bernstein, Donna Caccamise, Peter Foltz and the people from the NLP and LSA Research Labs in
Boulder for providing me with this corpus.

IMPORTANT: Calculations on this LSA Space will NOT give the same results as the LSA homepage ,
due to different parameter settings in the creation of the LSA space.

EN_100k Download
English HAL space, 300 dimensions

Created from a 2 Billion word corpus, which was created by concatenating the British National Corpus (BNC), the ukWaC corpus and a 2009 Wikipedia dump (see here and here).
This space was built using a HAL-like moving window model, with a window size of 5 (2 to the left, 2 to the right), with the 100k most frequent words in the corpus as row words as well as content (column) words for the co-occurrence matrix. A Positive Pointwise Mutual Information weighting scheme was applied, as well as a Singular Value Decomposition to reduce the space from 100k to 300 dimensions.

This space therefore contains vectors for 100,000 different words.

EN_100k_lsa Download
English LSA space, 300 dimensions

Created from a 2 Billion word corpus, which was created by concatenating the British National Corpus (BNC), the ukWaC corpus and a 2009 Wikipedia dump. This corpus is divided into 5,386,653 individual documents (see here and here).
This space was created from a term-document matrix with the 100k most frequent words in the corpus as rows and the 5.4 million documents the corpus consists of as columns (as in LSA). Other than in LSA, a A Positive Pointwise Mutual Information weighting scheme was applied instead of the standard log-entropy weighting (this should however not have a large influence on the results). As in standard LSA, an SVD was applied to reduce the space from the ~5.4 million dimensions to 300 dimensions.

This space therefore contains vectors for 100,000 different words.

EN_100k_cbow Download
English cbow space, 300 dimensions

Created from a 2 Billion word corpus, which was created by concatenating the British National Corpus (BNC), the ukWaC corpus and a 2009 Wikipedia dump (see here and here).
This semantic space was created using the cbow algorithm as implemented in the word2vec model (Mikolov et al., 2013), using only the 100k most frequent words in the corpus as target and context words. The model parameters were as follows: A context window size of 5 words (i.e., 2 to the left, 2 to the right), and 300-dimensional vectors (negative sampling with k = 10, subsampling with t = 1e−5), corresponding to the second-best word2vec model examined by Baroni et al. (2014).

This space contains vectors for 100,000 different words.

baroni Download (large file, 700 MB)
English cbow space, 400 dimensions

The semantic space shown to produce the best empirical results by Baroni et al. (2014). This semantic space is the "best predict vectors" space available here, converted to .rda format.
Created from a 2 Billion word corpus, which was created by concatenating the British National Corpus (BNC), the ukWaC corpus and a 2009 Wikipedia dump (see here and here). This semantic space was created using the cbow algorithm as implemented in the word2vec model (Mikolov et al., 2013), using the parameter set shown to produce the best results by Baroni et al. (2014): A context window size of 11 words (5 to the left, 5 to the right), and 400-dimensional vectors (negative sampling with k = 10, subsampling with t = 1e−5).

ukwac_cbow Download (large file, 500 MB)
English cbow space, 400 dimensions

Created from the 2 billion word ukWaC corpus (see here). This semantic space was created using the cbow algorithm as implemented in the word2vec model (Mikolov et al., 2013), using the following parameter set: A context window size of 5 words, and 400-dimensional vectors (negative sampling with k = 10, subsampling with t = 1e−5).

This space contains vectors for all words that appear at least 50 times in the ukWaC corpus (211,358 different words).

C:\Users\jmsyt\Desktop\CoLiPhi 2021\Causal Attributions and Corpus Analysis\LSAfun.txt - Notepad++

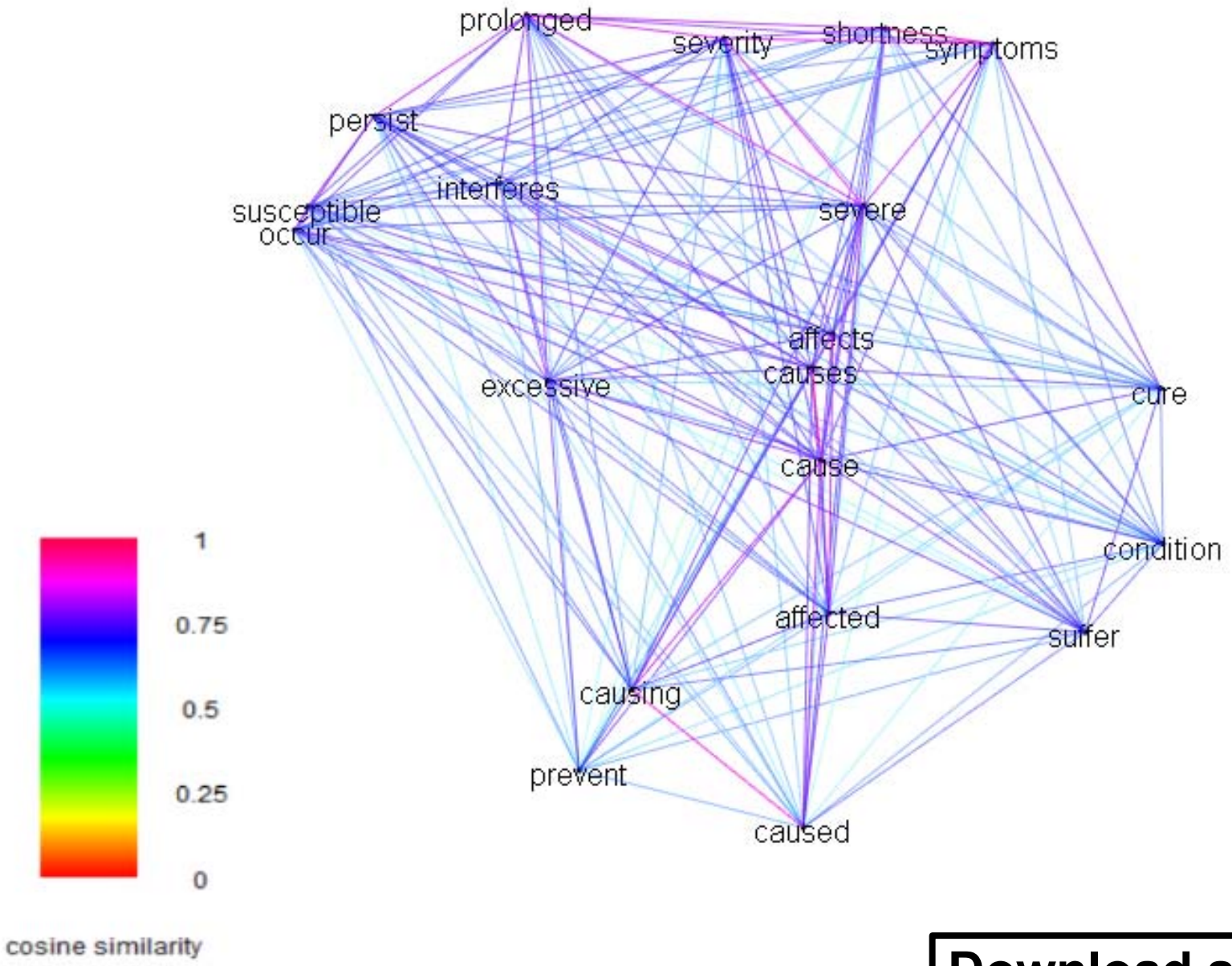
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?

```
# install.packages("LSAfun", dependencies=TRUE)
library(LSAfun)

load("C:/Users/jmsyt/Desktop/Causal Attributions and Corpus Analysis/LSAfun/EN_100k_lsa.rda")
#load("C:/Users/jmsyt/Desktop/Causal Attributions and Corpus Analysis/LSAfun/EN_100k.rda")
#load("C:/Users/jmsyt/Desktop/Causal Attributions and Corpus Analysis/LSAfun/EN_100k_cbow.rda")
#load("C:/Users/jmsyt/Desktop/Causal Attributions and Corpus Analysis/LSAfun/baroni.rda")
#load("C:/Users/jmsyt/Desktop/Causal Attributions and Corpus Analysis/LSAfun/ukwac_cbow.rda")


Cosine("responsible", "cause", tvectors=EN_100k_lsa)
Cosine("blame", "cause", tvectors=EN_100k_lsa)

multicos("cause", "responsible blame fault praise", tvectors=EN_100k_lsa)
multicos("caused", "responsible blame fault praise", tvectors=EN_100k_lsa)

neighbors("cause",20,tvectors=EN_100k_lsa,breakdown=FALSE)
neighbors("caused",20,tvectors=EN_100k_lsa,breakdown=FALSE)
neighbors("responsible",20,tvectors=EN_100k_lsa,breakdown=FALSE)

plot_neighbors(x="cause",n=20,connect.lines="all",start.lines=T,
               method="PCA",dims=3,axes=F,box=F,cex=1,alpha="shade",
               col="rainbow",tvectors=EN_100k_lsa,breakdown=FALSE)

plot_neighbors(x="cause",n=20,method="PCA",dims=2,tvectors=EN_100k_lsa)

plot_neighbors(x="responsible",n=20,method="PCA",dims=2,tvectors=EN_100k_lsa)

#-----------
# BENCHMARK
#-----------

MEN <- c(10,7,9,21,23,39,27,41,27,20,15,41,12,10,23,10,41,41,15,37,35,22,12,7,20,25,42,11,9,24,2

LSA_EN_100k_lsa <- c(
    Cosine("display", "pond", tvectors=EN_100k_lsa),
    Cosine("graveyard", "porch", tvectors=EN_100k_lsa),
    Cosine("hold", "theatre", tvectors=EN_100k_lsa),
    Cosine("city", "small", tvectors=EN_100k_lsa),
    Cosine("fire", "sand", tvectors=EN_100k_lsa),
    Cosine("eat", "soup", tvectors=EN_100k_lsa),
```

Normal text file length : 55,438 lines : 1,041 Ln : 8 Col : 70 Sel : 0 | 0 Windows (CR LF) UTF-8 IN

```
>
> multicos("cause", "responsible blame fault praise", tvectors=EN_100k_lsa)
      responsible     blame     fault    praise
cause   0.3233889 0.4573165 0.4963336 0.2338599
>
> multicos("caused", "responsible blame fault praise", tvectors=EN_100k_lsa)
       responsible     blame     fault    praise
caused   0.3434053 0.4664445 0.5663609 0.1838482
>
```

```
> neighbors("cause",20,tvectors=EN_100k_lsa,breakdown=FALSE)
     cause      causes     causing      caused     affects   excessive      suffer
 1.0000000   0.9231290   0.8368305   0.8205370   0.8042862   0.7848970   0.7848847
    severe    affected       occur     prevent susceptible   condition    severity
 0.7804020   0.7788130   0.7618391   0.7560800   0.7522059   0.7313918   0.7313401
 shortness        cure    symptoms   interferes     persist   prolonged
 0.7297577   0.7293867   0.7276848   0.7256855   0.7245498   0.7236668
```



Download space & run in R to rotate!

J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

Benchmark?

*C:\Users\jmsyt\Desktop\CORPUS PRESENTATION\LSAfun.txt - Notepad++

File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?

```
#-----------
# BENCHMARK
#-----------

MEN <- c(10,7,9,21,23,39,27,41,27,20,15,41,12,10,23,10,41,41,15,37,35,22,12,7,20,25

LSA__EN_100k_lsa <- c(
    Cosine("display", "pond", tvectors=EN_100k_lsa),
    Cosine("graveyard", "porch", tvectors=EN_100k_lsa),
    Cosine("hold", "theatre", tvectors=EN_100k_lsa),
    Cosine("city", "small", tvectors=EN_100k_lsa),
    Cosine("fire", "sand", tvectors=EN_100k_lsa),
    Cosine("eat", "soup", tvectors=EN_100k_lsa),
    Cosine("collar", "skirt", tvectors=EN_100k_lsa),
    Cosine("car", "garage", tvectors=EN_100k_lsa),
    Cosine("child", "fun", tvectors=EN_100k_lsa),
    Cosine("black", "bright", tvectors=EN_100k_lsa),
    Cosine("cute", "friend", tvectors=EN_100k_lsa),
    Cosine("mother", "son", tvectors=EN_100k_lsa),
    Cosine("design", "orange", tvectors=EN_100k_lsa),
    Cosine("jacket", "stream", tvectors=EN_100k_lsa),
    Cosine("daisy", "gravestone", tvectors=EN_100k_lsa),
    Cosine("flood", "neon", tvectors=EN_100k_lsa),
    Cosine("car", "motor", tvectors=EN_100k_lsa),
    Cosine("airport", "flight", tvectors=EN_100k_lsa),
```

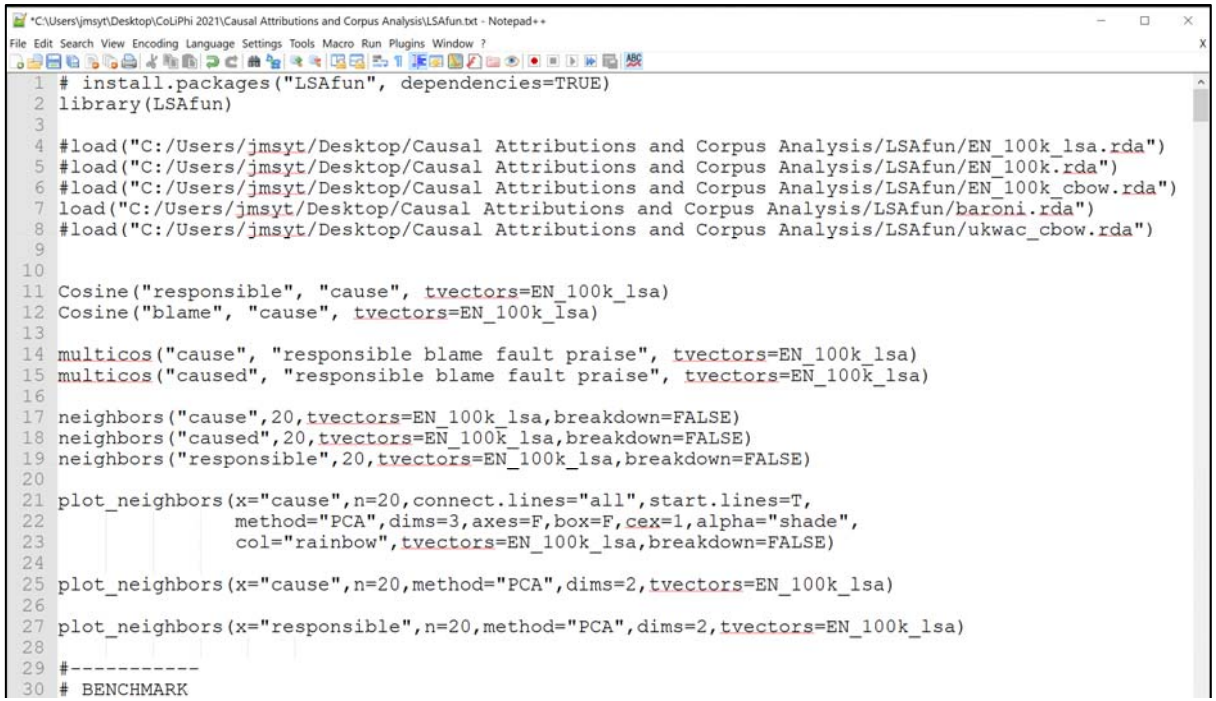
Normal text file length : 55,235 lines : 1,041 Ln : 27 Col : 1 Sel : 77 | 1 Windows (CR LF) UTF-8 IN

```
> cor( MEN, LSA__EN_100k_lsa, method="spearman" )
[1] 0.6744176
>
```

J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

Let's download other semantic spaces from https://sites.google.com/site/fritzgntr/home and modify "LSAfun.txt" to see how the perform on the MEN benchmark.




```
# install.packages("LSAfun", dependencies=TRUE)
library(LSAfun)

#load("C:/Users/jmsyt/Desktop/Causal Attributions and Corpus Analysis/LSAfun/EN_100k_lsa.rda")
#load("C:/Users/jmsyt/Desktop/Causal Attributions and Corpus Analysis/LSAfun/EN_100k.rda")
#load("C:/Users/jmsyt/Desktop/Causal Attributions and Corpus Analysis/LSAfun/EN_100k_cbow.rda")
load("C:/Users/jmsyt/Desktop/Causal Attributions and Corpus Analysis/LSAfun/baroni.rda")
#load("C:/Users/jmsyt/Desktop/Causal Attributions and Corpus Analysis/LSAfun/ukwac_cbow.rda")


Cosine("responsible", "cause", tvectors=EN_100k_lsa)
Cosine("blame", "cause", tvectors=EN_100k_lsa)

multicos("cause", "responsible blame fault praise", tvectors=EN_100k_lsa)
multicos("caused", "responsible blame fault praise", tvectors=EN_100k_lsa)

neighbors("cause",20,tvectors=EN_100k_lsa,breakdown=FALSE)
neighbors("caused",20,tvectors=EN_100k_lsa,breakdown=FALSE)
neighbors("responsible",20,tvectors=EN_100k_lsa,breakdown=FALSE)

plot_neighbors(x="cause",n=20,connect.lines="all",start.lines=T,
               method="PCA",dims=3,axes=F,box=F,cex=1,alpha="shade",
               col="rainbow",tvectors=EN_100k_lsa,breakdown=FALSE)

plot_neighbors(x="cause",n=20,method="PCA",dims=2,tvectors=EN_100k_lsa)

plot_neighbors(x="responsible",n=20,method="PCA",dims=2,tvectors=EN_100k_lsa)

#-----------
# BENCHMARK
```

EN_100k Download
English HAL space, 300 dimensions

Created from a 2 Billion word corpus, which was created by concatenating the British National Corpus (BNC), the ukWaC corpus and a 2009 Wikipedia dump (see here and here).
This space was built using a HAL-like moving window model, with a window size of 5 (2 to the left, 2 to the right), with the 100k most frequent words in the corpus as row words as well as content (column) words for the co-occurrence matrix. A Positive Pointwise Mutual Information weighting scheme was applied, as well as a Singular Value Decomposition to reduce the space from 100k to 300 dimensions.

This space therefore contains vectors for 100,000 different words.

```
> cor( MEN, LSA__EN_100k, method="spearman" )
[1] 0.7096647
>
```

EN_100k_lsa Download
English LSA space, 300 dimensions

Created from a 2 Billion word corpus, which was created by concatenating the British National Corpus (BNC), the ukWaC corpus and a 2009 Wikipedia dump. This corpus is divided into 5,386,653 individual documents (see here and here).
This space was created from a term-document matrix with the 100k most frequent words in the corpus as rows and the 5.4 million documents the corpus consists of as columns (as in LSA). Other than in LSA, a A Positive Pointwise Mutual Information weighting scheme was applied instead of the standard log-entropy weighting (this should however not have a large influence on the results). As in standard LSA, an SVD was applied to reduce the space from the ~5.4 million dimensions to 300 dimensions.

This space therefore contains vectors for 100,000 different words.

```
> cor( MEN, LSA__EN_100k_lsa, method="spearman" )
[1] 0.6744176
>
```

EN_100k_cbow Download
English cbow space, 300 dimensions

Created from a 2 Billion word corpus, which was created by concatenating the British National Corpus (BNC), the ukWaC corpus and a 2009 Wikipedia dump (see here and here).
This semantic space was created using the cbow algorithm as implemented in the word2vec model (Mikolov et al., 2013), using only the 100k most frequent words in the corpus as target and context words. The model parameters were as follows: A context window size of 5 words (i.e., 2 to the left, 2 to the right), and 300-dimensional vectors (negative sampling with k = 10, subsampling with t = 1e−5), corresponding to the second-best word2vec model examined by Baroni et al. (2014).

This space contains vectors for 100,000 different words.

```
> cor( MEN, LSA__EN_100k_cbow, method="spearman" )
[1] 0.7485425
>
```

baroni Download (large file, 700 MB)
English cbow space, 400 dimensions

The semantic space shown to produce the best empirical results by Baroni et al. (2014). This semantic space is the "best predict vectors" space available here, converted to .rda format.
Created from a 2 Billion word corpus, which was created by concatenating the British National Corpus (BNC), the ukWaC corpus and a 2009 Wikipedia dump (see here and here). This semantic space was created using the cbow algorithm as implemented in the word2vec model (Mikolov et al., 2013), using the parameter set shown to produce the best results by Baroni et al. (2014): A context window size of 11 words (5 to the left, 5 to the right), and 400-dimensional vectors (negative sampling with k = 10, subsampling with t = 1e−5).

```
> cor( MEN, LSA__baroni, method="spearman" )
[1] 0.7979672
>
```

ukwac_cbow Download (large file, 500 MB)
English cbow space, 400 dimensions

Created from the 2 billion word ukWaC corpus (see here). This semantic space was created using the cbow algorithm as implemented in the word2vec model (Mikolov et al., 2013), using the following parameter set: A context window size of 5 words, and 400-dimensional vectors (negative sampling with k = 10, subsampling with t = 1e−5).

This space contains vectors for all words that appear at least 50 times in the ukWaC corpus (211,358 different words).

```
> cor( MEN, LSA__ukwac, method="spearman" )
[1] 0.756142
>
```

J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

Limited to the spaces that have been built.

Wanted to lemmatize and look at two multi-word expressions ("caused the," "responsible for the").

And we wanted to compare to philosophical use.

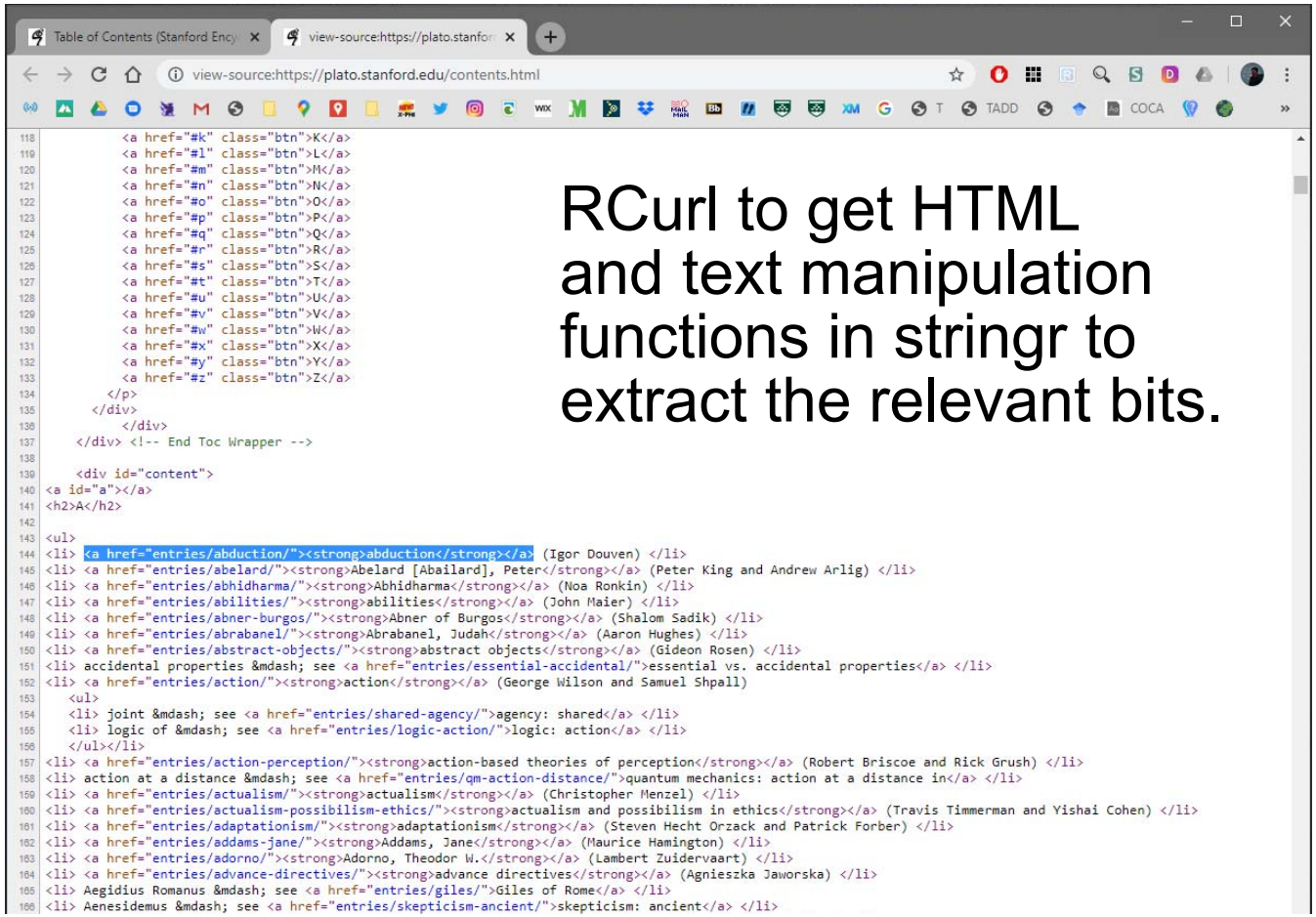
Can also build corpora and semantic spaces in R (although notable limitations for large corpora).

We'll create a toy "philosophy" corpus, then build LSA and word2vec spaces for it in R.

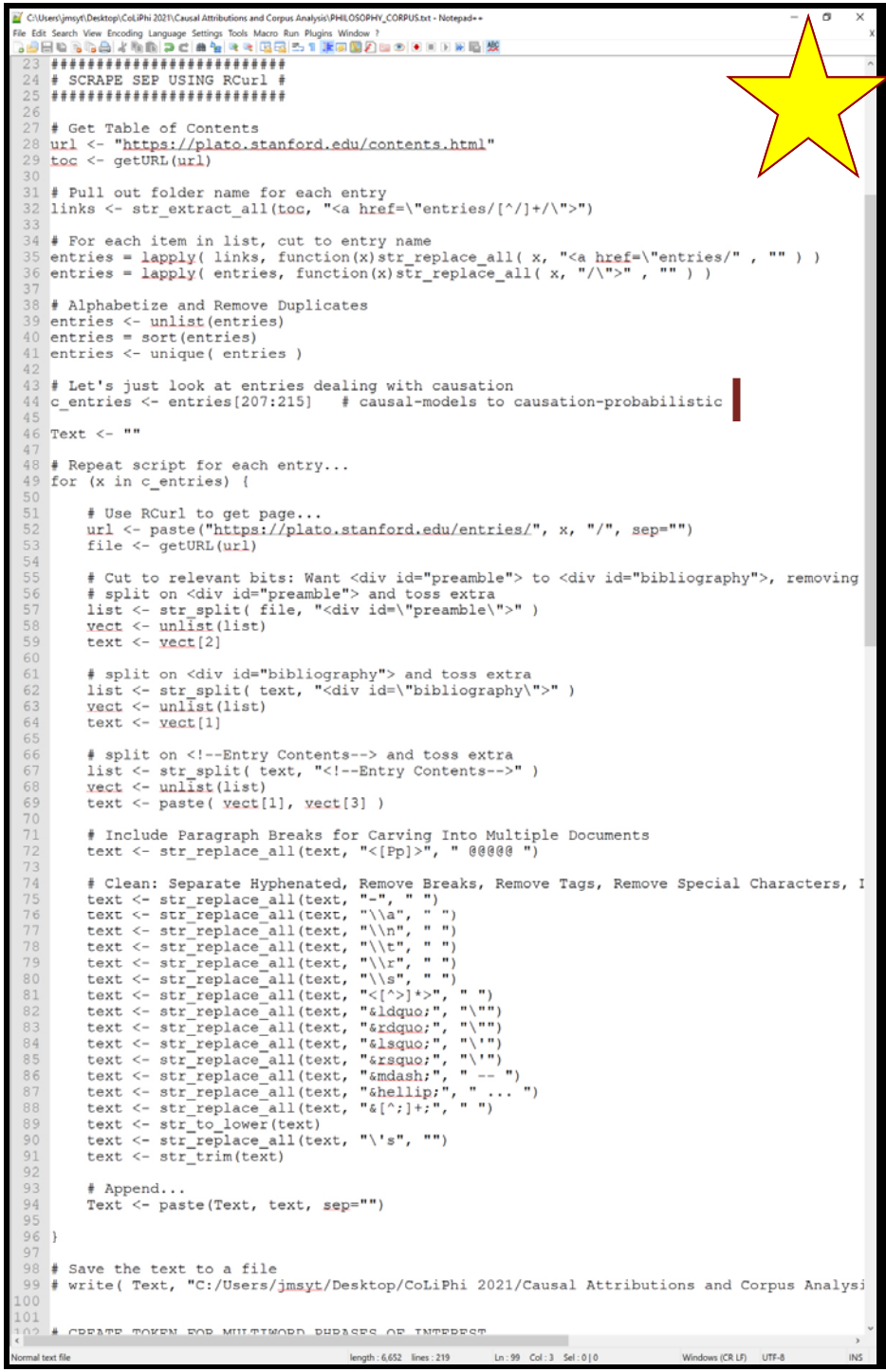
J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

We'll begin by scraping text from the SEP:

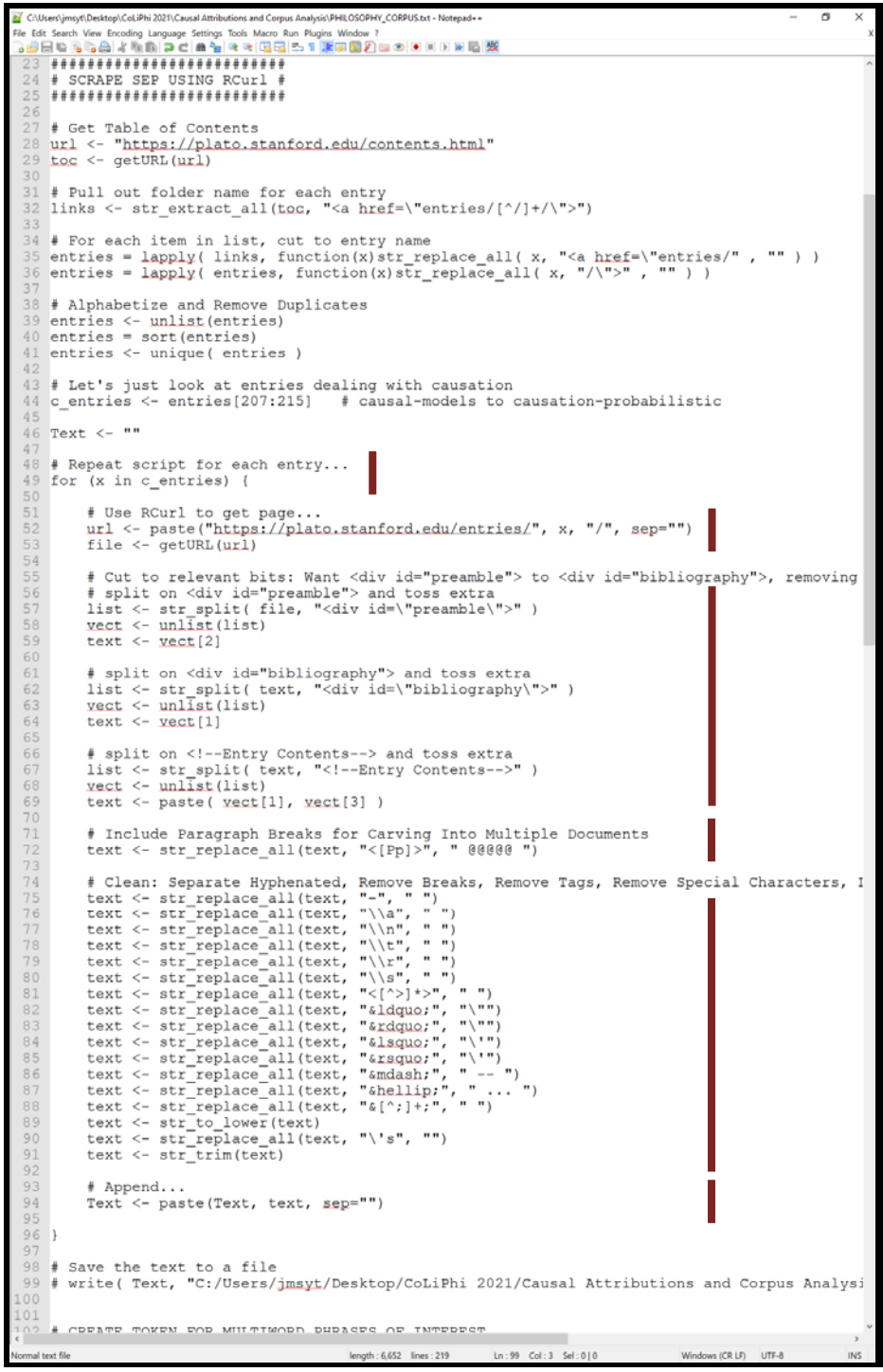


RCurl to get HTML and text manipulation functions in stringr to extract the relevant bits.

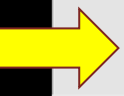


```
C:\Users\jmsyt\Desktop\CoLiPhi 2021\Causal Attributions and Corpus Analysis\PHILOSOPHY_CORPUS.txt - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
23 ##########################
24 # SCRAPE SEP USING RCurl #
25 ##########################
26
27 # Get Table of Contents
28 url <- "https://plato.stanford.edu/contents.html"
29 toc <- getURL(url)
30
31 # Pull out folder name for each entry
32 links <- str_extract_all(toc, "<a href=\"entries/[^/]+/\">")
33
34 # For each item in list, cut to entry name
35 entries = lapply( links, function(x)str_replace_all( x, "<a href=\"entries/" , "" ) )
36 entries = lapply( entries, function(x)str_replace_all( x, "/\">" , "" ) )
37
38 # Alphabetize and Remove Duplicates
39 entries <- unlist(entries)
40 entries = sort(entries)
41 entries <- unique( entries )
42
43 # Let's just look at entries dealing with causation
44 c_entries <- entries[207:215]   # causal-models to causation-probabilistic
45
46 Text <- ""
47
48 # Repeat script for each entry...
49 for (x in c_entries) {
50
51     # Use RCurl to get page...
52     url <- paste("https://plato.stanford.edu/entries/", x, "/", sep="")
53     file <- getURL(url)
54
55     # Cut to relevant bits: Want <div id="preamble"> to <div id="bibliography">, removing
56     # split on <div id="preamble"> and toss extra
57     list <- str_split( file, "<div id=\"preamble\">" )
58     vect <- unlist(list)
59     text <- vect[2]
60
61     # split on <div id="bibliography"> and toss extra
62     list <- str_split( text, "<div id=\"bibliography\">" )
63     vect <- unlist(list)
64     text <- vect[1]
65
66     # split on <!--Entry Contents--> and toss extra
67     list <- str_split( text, "<!--Entry Contents-->" )
68     vect <- unlist(list)
69     text <- paste( vect[1], vect[3] )
70
71     # Include Paragraph Breaks for Carving Into Multiple Documents
72     text <- str_replace_all(text, "<[Pp]>", " @@@@ ")
73
74     # Clean: Separate Hyphenated, Remove Breaks, Remove Tags, Remove Special Characters, I
75     text <- str_replace_all(text, "-", " ")
76     text <- str_replace_all(text, "\\a", " ")
77     text <- str_replace_all(text, "\\n", " ")
78     text <- str_replace_all(text, "\\t", " ")
79     text <- str_replace_all(text, "\\r", " ")
80     text <- str_replace_all(text, "\\s", " ")
81     text <- str_replace_all(text, "<[^>]*>", " ")
82     text <- str_replace_all(text, "&ldquo;", "\"")
83     text <- str_replace_all(text, "&rdquo;", "\"")
84     text <- str_replace_all(text, "&lsquo;", "\'")
85     text <- str_replace_all(text, "&rsquo;", "\'")
86     text <- str_replace_all(text, "&mdash;", " -- ")
87     text <- str_replace_all(text, "&hellip;", " ... ")
88     text <- str_replace_all(text, "&[^;]+;", " ")
89     text <- str_to_lower(text)
90     text <- str_replace_all(text, "\'s", "")
91     text <- str_trim(text)
92
93     # Append...
94     Text <- paste(Text, text, sep="")
95
96 }
97
98 # Save the text to a file
99 # write( Text, "C:/Users/jmsyt/Desktop/CoLiPhi 2021/Causal Attributions and Corpus Analysi
100
101
102 # CREATE TOKEN FOR MULTIWORD PHRASES OF INTEREST
Normal text file   length : 6,652   lines : 219   Ln : 99   Col : 3   Sel : 0 | 0   Windows (CR LF)   UTF-8   INS
```

```
[187] "buddhism-tiantai"                "buridan"
[189] "burke"                           "burley"
[191] "butler-moral"                    "byzantine-philosophy"
[193] "callicles-thrasymachus"          "cambridge-platonists"
[195] "campanella"                      "camus"
[197] "cancer"                          "capability-approach"
[199] "cardano"                         "carnap"
[201] "carneades"                       "cassirer"
[203] "categories"                      "category-mistakes"
[205] "category-theory"                 "catharine-macaulay"
[207] "causal-models"                   "causation-backwards"
[209] "causation-counterfactual"        "causation-law"
[211] "causation-mani"                  "causation-medieval"
[213] "causation-metaphysics"           "causation-physics"
[215] "causation-probabilistic"         "cell-biology"
[217] "cellular-automata"               "certainty"
[219] "ceteris-paribus"                 "chance-randomness"
[221] "change"                          "chaos"
[223] "chemistry"                       "childhood"
[225] "children"                        "chimeras"
[227] "chinese-change"                  "chinese-epistemology"
[229] "chinese-legalism"                "chinese-logic-language"
[231] "chinese-metaphysics"             "chinese-phil-medicine"
[233] "chinese-phil-science"            "chinese-room"
[235] "chinese-social-political"        "chinese-translate-interpret"
[237] "chisholm"                        "christiantheology-philosophy"
[239] "church-turing"                   "citizenship"
[241] "civic-education"                 "civil-disobedience"
[243] "civil-rights"                    "clarke"
[245] "climate-science"                 "clinical-research"
[247] "cloning"                         "closure-epistemic"
[249] "cockburn"                        "coercion"
[251] "cognition-animal"                "cognitive-disability"
[253] "cognitive-science"               "cohen"
[255] "collective-intentionality"       "collective-responsibility"
[257] "collingwood"                     "collingwood-aesthetics"
[259] "collins"                         "colonialism"
[261] "color"                           "common-good"
[263] "common-knowledge"                "communitarianism"
[265] "comparphil-chiwes"               "compatibilism"
[267] "compositionality"                "computability"
[269] "computation-physicalsystems"     "computational-complexity"
[271] "computational-linguistics"       "computational-mind"
[273] "computational-philosophy"        "computer-science"
[275] "computing-history"               "computing-responsibility"
[277] "comte"                           "concept-emotion-india"
[279] "concept-evil"                    "concepts"
[281] "concepts-god"                    "conceptual-art"
[283] "condemnation"                    "condillac"
[285] "conditionals"                    "confirmation"
[287] "confucius"                       "connectionism"
[289] "connectives-logic"               "conscience"
[291] "conscience-medieval"             "consciousness"
[293] "consciousness-17th"              "consciousness-animal"
[295] "consciousness-higher"            "consciousness-intentionality"
[297] "consciousness-neuroscience"      "consciousness-representational"
[299] "consciousness-temporal"          "consciousness-unity"
[301] "consequence-medieval"            "consequentialism"
[303] "consequentialism-rule"           "conservation-biology"
[305] "conservatism"                    "constitutionalism"
[307] "constructive-empiricism"         "constructivism-metaethics"
[309] "content-causal"                  "content-externalism"
[311] "content-narrow"                  "content-nonconceptual"
```



```
C:\Users\jmsyt\Desktop\CoLiPhi 2021\Causal Attributions and Corpus Analysis\PHILOSOPHY_CORPUS.txt - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
23 ##########################
24 # SCRAPE SEP USING RCurl #
25 ##########################
26
27 # Get Table of Contents
28 url <- "https://plato.stanford.edu/contents.html"
29 toc <- getURL(url)
30
31 # Pull out folder name for each entry
32 links <- str_extract_all(toc, "<a href=\"entries/[^/]+/\">")
33
34 # For each item in list, cut to entry name
35 entries = lapply( links, function(x)str_replace_all( x, "<a href=\"entries/" , "" ) )
36 entries = lapply( entries, function(x)str_replace_all( x, "/\">" , "" ) )
37
38 # Alphabetize and Remove Duplicates
39 entries <- unlist(entries)
40 entries = sort(entries)
41 entries <- unique( entries )
42
43 # Let's just look at entries dealing with causation
44 c_entries <- entries[207:215]   # causal-models to causation-probabilistic
45
46 Text <- ""
47
48 # Repeat script for each entry...
49 for (x in c_entries) {
50
51     # Use RCurl to get page...
52     url <- paste("https://plato.stanford.edu/entries/", x, "/", sep="")
53     file <- getURL(url)
54
55     # Cut to relevant bits: Want <div id="preamble"> to <div id="bibliography">, removing
56     # split on <div id="preamble"> and toss extra
57     list <- str_split( file, "<div id=\"preamble\">" )
58     vect <- unlist(list)
59     text <- vect[2]
60
61     # split on <div id="bibliography"> and toss extra
62     list <- str_split( text, "<div id=\"bibliography\">" )
63     vect <- unlist(list)
64     text <- vect[1]
65
66     # split on <!--Entry Contents--> and toss extra
67     list <- str_split( text, "<!--Entry Contents-->" )
68     vect <- unlist(list)
69     text <- paste( vect[1], vect[3] )
70
71     # Include Paragraph Breaks for Carving Into Multiple Documents
72     text <- str_replace_all(text, "<[Pp]>", " @@@@@ ")
73
74     # Clean: Separate Hyphenated, Remove Breaks, Remove Tags, Remove Special Characters, I
75     text <- str_replace_all(text, "-", " ")
76     text <- str_replace_all(text, "\\a", " ")
77     text <- str_replace_all(text, "\\n", " ")
78     text <- str_replace_all(text, "\\t", " ")
79     text <- str_replace_all(text, "\\r", " ")
80     text <- str_replace_all(text, "\\s", " ")
81     text <- str_replace_all(text, "<[^>]*>", " ")
82     text <- str_replace_all(text, "&ldquo;", "\"")
83     text <- str_replace_all(text, "&rdquo;", "\"")
84     text <- str_replace_all(text, "&lsquo;", "'")
85     text <- str_replace_all(text, "&rsquo;", "'")
86     text <- str_replace_all(text, "&mdash;", " -- ")
87     text <- str_replace_all(text, "&hellip;", " ... ")
88     text <- str_replace_all(text, "&[^;]+;", " ")
89     text <- str_to_lower(text)
90     text <- str_replace_all(text, "\'s", "")
91     text <- str_trim(text)
92
93     # Append...
94     Text <- paste(Text, text, sep="")
95
96 }
97
98 # Save the text to a file
99 # write( Text, "C:/Users/jmsyt/Desktop/CoLiPhi 2021/Causal Attributions and Corpus Analysi
100
101
102 # CREATE TOKEN FOR MULTIWORD PHRASES OF INTEREST
Normal text file   length : 6,652   lines : 219   Ln : 99   Col : 3   Sel : 0 | 0   Windows (CR LF)   UTF-8   INS
```



```
C:\Users\jmsyt\Desktop\CoLiPhi 2021\Causal Attributions and Corpus Analysis\SEP_TEXT.txt - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
1 @@@@@  causal models are mathematical models r
within an individual system or population. the
causal relationships from statistical data. th
the epistemology of causation, and about the r
probability. they have also been applied to to
such as the logic of counterfactuals, decision
actual causation.                    1. introdu
is an interdisciplinary field that has its ori
of the 1920s, especially in the work of the am
sewall wright (1921). important contributions
econometrics, epidemiology, philosophy, statis
given the importance of causation to many area
growing philosophical interest in the use of m
major works -- spirtes, glymour, and scheines
2009 -- have been particularly influential.
predictions about the behavior of a system. in
entails the truth value, or the probability, o
system; it predicts the effects of interventio
probabilistic dependence or independence of va
causal models also facilitate the inverse of t
observed probabilistic correlations among vari
experimental interventions, we can determine w
with these observations. the discussion will f
in "in principle". for example, we will consid
infer the correct causal structure of a system
the probability distribution over the variable
very real problem of inferring the true probab
in addition, the entry will discuss the applic
logic of counterfactuals, the analysis of caus
basic tools    @@@@@  this section introduces
used in causal modeling, as well as terminolog
2.1 variables, logic, and language    @@@@@
blocks of causal models. they will be represen
letters. a variable is a function that can tak
of a variable can represent the occurrence or
range of incompatible events, a property of an
individuals, or a quantitative value. for inst
situation in which suzy throws a stone and a w
s  and  w  such that:       \(s = 1\) represen
0\) represents her not throwing    \(w = 1\) r
\(w = 0\) represents the window remaining inta
Normal text file   length : 849,427   lines : 2   Ln : 1   Col
```

Here we can mark any multiword phrases we want to analyze as a single unit.

Divide into documents (our bags-of-words); here we'll do paragraphs

```
 [1] "causal models are mathematical models representing c
 [2] "causal modeling is an interdisciplinary field that h
 [3] "a causal model makes predictions about the behavior
 [4] "this section introduces some of the basic formal too
 [5] "variables  are the basic building blocks of causal m
 [6] "if we are modeling the influence of education on inc
 [7] "the set of possible values of a variable is the  ran
 [8] "a  world  is a complete specification of a causal mo
 [9] "if  x  is a variable in a causal model, and  x  is a
[10] "we will use basic notation from set theory. sets wil
[11] "if \\(\\bs = \\{x_1 , \\ldots ,x_n\\}\\) is a set of
[12] "in   section 4 ,  we will consider causal models tha
[13] "some standard properties of probability are the foll
[14] "some further definitions:"
[15] "the  conditional probability of a given b,  written
[16] "we will ignore problems that might arise when \\(\\p
[17] "as a convenient shorthand, a probabilistic statement
[18] "as shorthand for    \\[ \\begin{aligned} \\forall x_
[19] "where the domain of quantification for each variable
[20] "we will not presuppose any particular interpretation
[21] "if \\(\\bv\\) is the set of variables included in a
[22] "a  path  in a directed graph is a non repeating sequ
[23] "the relationships in the graph are often described u
[24] "an arrow from  y  to  z  in a dag represents that  y
[25] "a second type of graph that we will consider is an
[26] "we can be a bit more precise, we only need to repres
```

```
# CREATE TOKEN FOR MULTIWORD PHRASES OF INTEREST
Text <- str_replace_all(Text, "caused the", "caused_the")
Text <- str_replace_all(Text, "responsible for the", "responsible_for_the")

Documents <- str_split(Text, "@@@@@")
Documents <- unlist(Documents)
Documents <- Documents[2:length(Documents)]
Documents <- str_trim(Documents)



###################################
# Put into Corpus using tm        #
# Preprocess                      #
# Lemmatize using textstem        #
###################################

Corpus <- Corpus( VectorSource(Documents) )

# Preprocess: Remove Stopwords, Numbers, Punctuation...

Corpus <- tm_map(Corpus, removeWords, stopwords("english") )
# stopwords("english")
Corpus <- tm_map(Corpus, removeNumbers )
Corpus <- tm_map(Corpus, removePunctuation, preserve_intra_word_contractions=TRUE )
# Fix tokens after punctuation was removed
Corpus$content <- str_replace_all(Corpus$content, "causedthe", "caused_the")
Corpus$content <- str_replace_all(Corpus$content, "responsibleforthe", "responsible_for_th

# Preprocess: Lemmatize
Corpus <- tm_map(Corpus, content_transformer(lemmatize_words))
Corpus <- tm_map(Corpus, stripWhitespace )

# Preprocess: Remove "non-words"
TDM    <- TermDocumentMatrix(Corpus)
TOKENS <- findFreqTerms(TDM, 1)
# Check tokens against dictionary, excluding "caused_the" and "responsible_for_the"
REMOVE <- setdiff(TOKENS, GradyAugmented)
REMOVE <- REMOVE[ REMOVE != "responsible_for_the" ]
REMOVE <- REMOVE[ REMOVE != "caused_the" ]
# Remove from Corpus
Corpus <- tm_map(Corpus, content_transformer(removeWords), REMOVE)

# Corpus$content



#####
# Create Vector Space
######

DTM <- DocumentTermMatrix(Corpus)
inspect(DTM)

# DTM <- removeSparseTerms(DTM, 0.9955)
# inspect(DTM)

DTM_M <- as.matrix(DTM)
sum(DTM_M[,"cause"])
sum(DTM_M[,"responsible"])
sum(DTM_M[,"caused_the"])
sum(DTM_M[,"responsible_for_the"])


# Weight
########
DTM_W <- weightTfIdf(DTM, normalize=FALSE)


# CREATE SPACE
###############
SPACE = lsa(DTM_W)
MAT <- SPACE$dk


# ANALYZE
#########

multicos("cause", "responsible_blame", tvectors=MAT)
```

Normal text file length : 6,729 lines : 223 Ln : 157 Col : 3 Sel : 0 | 0 Windows (CR LF) UTF-8 INS

Build a corpus from the documents using tm…

Do some further cleaning

Lemmatize (PoS tag, etc.)

Remove non-words (?)

```
#####
# Create Vector Space
######

DTM <- DocumentTermMatrix(Corpus)
inspect(DTM)

# DTM <- removeSparseTerms(DTM, 0.9955)
# inspect(DTM)

DTM_M <- as.matrix(DTM)
sum(DTM_M[,"cause"])
sum(DTM_M[,"responsible"])
sum(DTM_M[,"caused_the"])
sum(DTM_M[,"responsible_for_the"])


# Weight
########
DTM_W <- weightTfIdf(DTM, normalize=FALSE)


# CREATE SPACE
##############
SPACE = lsa(DTM_W)
MAT <- SPACE$dk


# ANALYZE
#########

multicos("cause", "responsible blame", tvectors=MAT)
multicos("blame", "responsible", tvectors=MAT)

neighbors("cause", 50, tvectors=MAT, breakdown=FALSE)
neighbors("responsible", 50, tvectors=MAT, breakdown=FALSE)
neighbors("blame", 50, tvectors=MAT, breakdown=FALSE)

multicos("caused_the", "responsible_for_the", tvectors=MAT)
neighbors("caused_the", 50, tvectors=MAT, breakdown=FALSE)
neighbors("responsible_for_the", 50, tvectors=MAT, breakdown=FALSE)


# Most Frequent Terms:
sorted <- sort( colSums(DTM_M), decreasing=TRUE )
d <- data.frame(word = names(sorted),freq=sorted)
head(d, 100)


# Wordcloud
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
          max.words=300, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))



# rword2vec
###########
#
# write( Corpus$content, "C:/Users/jmsyt/Desktop/CORPUS PRESENTATION/word2vec.txt" )
#
# Model=word2vec(train_file = "C:/Users/jmsyt/Desktop/CORPUS PRESENTATION/word2vec.txt",ou
#
# distance(file_name="w2v.bin", search_word="cause",num = 20)
# distance(file_name="w2v.bin", search_word="responsible",num = 20)
#
#####
```

```
<<DocumentTermMatrix (documents: 1074, terms: 5618)>>
Non-/sparse entries: 41054/5992678
Sparsity           : 99%
Maximal term length: 19
Weighting          : term frequency (tf)
Sample             :
     Terms
Docs  can causal causation cause causes one probability time variables will
 271    0      0         0     0      0   1           0    0         0    3
 293    4      0         0     1      0   2           0    1         0    5
 422    0      1         0     7      1   3           0    0         0    0
 508    0      7         9    12      1   0           0    3         1    0
 535    0      0         1     4      2   0           0    0         0    1
 543    0      4         0     0      2   2           0    0         0    3
 550    0      5         0     5      3   2           0    0         0    0
 556    2      8         6     2      1   1           0    0         0    0
 559    2      5         2     9      1   1           0    0         3    1
 565    0      6         3     1      2   1           0    0         1    1
```

1074 documents comprised of 5618 unique terms

1074x5618= 6033732= 41054+5992678

5992678 cells are 0
41054 cells are !0

Can reduce the sparsity (removing infrequent terms), which can help with memory issues….

```
#####
# Create Vector Space
######

DTM <- DocumentTermMatrix(Corpus)
inspect(DTM)

# DTM <- removeSparseTerms(DTM, 0.9955)
# inspect(DTM)

DTM_M <- as.matrix(DTM)
sum(DTM_M[,"cause"])
sum(DTM_M[,"responsible"])
sum(DTM_M[,"caused_the"])
sum(DTM_M[,"responsible_for_the"])


# Weight
########
DTM_W <- weightTfIdf(DTM, normalize=FALSE)


# CREATE SPACE
##############
SPACE = lsa(DTM_W)
MAT <- SPACE$dk


# ANALYZE
#########

multicos("cause", "responsible blame", tvectors=MAT)
multicos("blame", "responsible", tvectors=MAT)

neighbors("cause", 50, tvectors=MAT, breakdown=FALSE)
neighbors("responsible", 50, tvectors=MAT, breakdown=FALSE)
neighbors("blame", 50, tvectors=MAT, breakdown=FALSE)

multicos("caused_the", "responsible_for_the", tvectors=MAT)
neighbors("caused_the", 50, tvectors=MAT, breakdown=FALSE)
neighbors("responsible_for_the", 50, tvectors=MAT, breakdown=FALSE)


# Most Frequent Terms:
sorted <- sort( colSums(DTM_M), decreasing=TRUE )
d <- data.frame(word = names(sorted),freq=sorted)
head(d, 100)


# Wordcloud
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
          max.words=300, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))



# rword2vec
###########
#
# write( Corpus$content, "C:/Users/jmsyt/Desktop/CORPUS PRESENTATION/word2vec.txt" )
#
# Model=word2vec(train_file = "C:/Users/jmsyt/Desktop/CORPUS PRESENTATION/word2vec.txt",ou
#
# distance(file_name="w2v.bin", search_word="cause",num = 20)
# distance(file_name="w2v.bin", search_word="responsible",num = 20)
#
#####
```

```
<<DocumentTermMatrix (documents: 1074, terms: 1750)>>
Non-/sparse entries: 34328/1845172
Sparsity           : 98%
Maximal term length: 19
Weighting          : term frequency (tf)
Sample             :
     Terms
Docs  can causal causation cause causes one probability time variables will
  293   4      0         0     1      0   2           0    1         0    5
  422   0      1         0     7      1   3           0    0         0    0
  508   0      7         9    12      1   0           0    3         1    0
  535   0      0         1     4      2   0           0    0         0    1
  543   0      4         0     0      2   2           0    0         0    3
  550   0      5         0     5      3   2           0    0         0    0
  556   2      8         6     2      1   1           0    0         0    0
  559   2      5         2     9      1   1           0    0         3    1
  563   3      7         0     1      0   0           0    0         0    1
  565   0      6         3     1      2   1           0    0         1    1
```

But make sure you didn't remove the terms you're interested in!

```
> DTM_M <- as.matrix(DTM)
> sum(DTM_M[,"cause"])
[1] 614
> sum(DTM_M[,"responsible"])
[1] 9
> sum(DTM_M[,"caused_the"])
[1] 35
> sum(DTM_M[,"responsible_for_the"])
[1] 5
>
```

*C:\Users\jmsyt\Desktop\CoLiPhi 2021\Causal Attributions and Corpus Analysis\PHILOSOPHY_CORPUS.txt - Notepad++

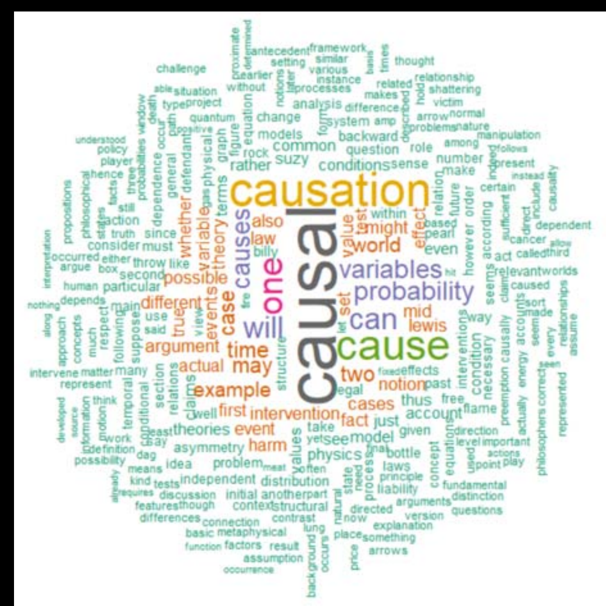
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?

```
#####
# Create Vector Space
######

DTM <- DocumentTermMatrix(Corpus)
inspect(DTM)

# DTM <- removeSparseTerms(DTM, 0.9955)
# inspect(DTM)

DTM_M <- as.matrix(DTM)
sum(DTM_M[,"cause"])
sum(DTM_M[,"responsible"])
sum(DTM_M[,"caused_the"])
sum(DTM_M[,"responsible_for_the"])


# Weight
########
DTM_W <- weightTfIdf(DTM, normalize=FALSE)


# CREATE SPACE
##############
SPACE = lsa(DTM_W)
MAT <- SPACE$dk


# ANALYZE
#########

multicos("cause", "responsible blame", tvectors=MAT)
multicos("blame", "responsible", tvectors=MAT)

neighbors("cause", 50, tvectors=MAT, breakdown=FALSE)
neighbors("responsible", 50, tvectors=MAT, breakdown=FALSE)
neighbors("blame", 50, tvectors=MAT, breakdown=FALSE)

multicos("caused_the", "responsible_for_the", tvectors=MAT)
neighbors("caused_the", 50, tvectors=MAT, breakdown=FALSE)
neighbors("responsible_for_the", 50, tvectors=MAT, breakdown=FALSE)


# Most Frequent Terms:
sorted <- sort( colSums(DTM_M), decreasing=TRUE )
d <- data.frame(word = names(sorted),freq=sorted)
head(d, 100)


# Wordcloud
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
          max.words=300, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))




# rword2vec
###########
#
# write( Corpus$content, "C:/Users/jmsyt/Desktop/CORPUS PRESENTATION/word2vec.txt" )
#
# Model=word2vec(train_file = "C:/Users/jmsyt/Desktop/CORPUS PRESENTATION/word2vec.txt",ou
#
# distance(file_name="w2v.bin", search_word="cause",num = 20)
# distance(file_name="w2v.bin", search_word="responsible",num = 20)
#
#####
```

Normal text file | length : 6,737 lines : 227 | Ln : 226 Col : 1 Sel : 0 | 0 | Windows (CR LF) | UTF-8 | INS

Can explore the space the same way we did before!

```
> multicos("caused_the", "responsible_for_the", tvectors=
            responsible_for_the
caused_the          0.005850152
>
```



```
#####
# Create Vector Space
######

DTM <- DocumentTermMatrix(Corpus)
inspect(DTM)

# DTM <- removeSparseTerms(DTM, 0.9955)
# inspect(DTM)

DTM_M <- as.matrix(DTM)
sum(DTM_M[,"cause"])
sum(DTM_M[,"responsible"])
sum(DTM_M[,"caused_the"])
sum(DTM_M[,"responsible_for_the"])


# Weight
########
DTM_W <- weightTfIdf(DTM, normalize=FALSE)


# CREATE SPACE
##############
SPACE = lsa(DTM_W)
MAT <- SPACE$dk


# ANALYZE
#########

multicos("cause", "responsible blame", tvectors=MAT)
multicos("blame", "responsible", tvectors=MAT)

neighbors("cause", 50, tvectors=MAT, breakdown=FALSE)
neighbors("responsible", 50, tvectors=MAT, breakdown=FALSE)
neighbors("blame", 50, tvectors=MAT, breakdown=FALSE)

multicos("caused_the", "responsible_for_the", tvectors=MAT)
neighbors("caused_the", 50, tvectors=MAT, breakdown=FALSE)
neighbors("responsible_for_the", 50, tvectors=MAT, breakdown=FALSE)


# Most Frequent Terms:
sorted <- sort( colSums(DTM_M), decreasing=TRUE )
d <- data.frame(word = names(sorted),freq=sorted)
head(d, 100)


# Wordcloud
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
          max.words=300, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))



# rword2vec
###########
#
# write( Corpus$content, "C:/Users/jmsyt/Desktop/CORPUS PRESENTATION/word2vec.txt" )
#
# Model=word2vec(train_file = "C:/Users/jmsyt/Desktop/CORPUS PRESENTATION/word2vec.txt",ou
#
# distance(file_name="w2v.bin", search_word="cause",num = 20)
# distance(file_name="w2v.bin", search_word="responsible",num = 20)
#
#####
```

Could also build a context-predicting space in R, although the implementation is rather limited…

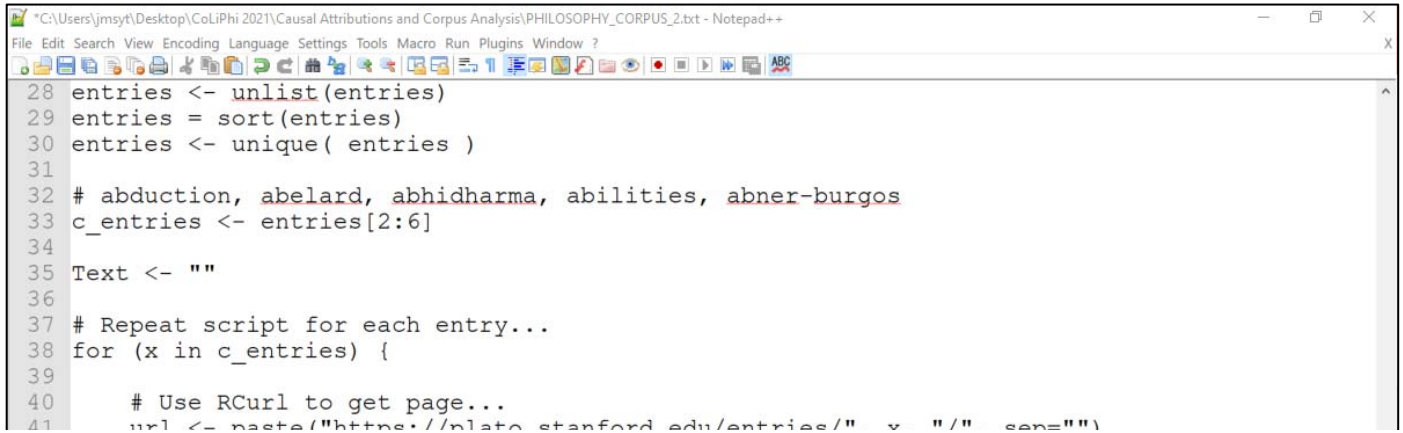
```
> distance(file_name="w2v.bin", search_word="cause",num = 20)
Entered word or sentence: cause

Word: cause  Position in vocabulary: 3
           word              dist
1         joint 0.897010445594788
2        effect  0.87340921163559
3        common 0.808746933937073
4     principle  0.80454808473587
5    maintained  0.80194753408432
6       drawing 0.799688398838043
7      recovery 0.796327173709869
8        direct 0.788753747940063
9  intermediate 0.785856604576111
10         even 0.767904043197632
11      neutral 0.763022720813751
12        raise 0.746761739253998
13      screens 0.741270363330841
14    explained 0.739869356155396
15       though 0.735124230384827
16    dependent 0.733080744743347
17  correlation 0.728201627731323
18  requirement 0.724978089332581
19       select 0.722158789634705
20        thing 0.714402496814728
>
```

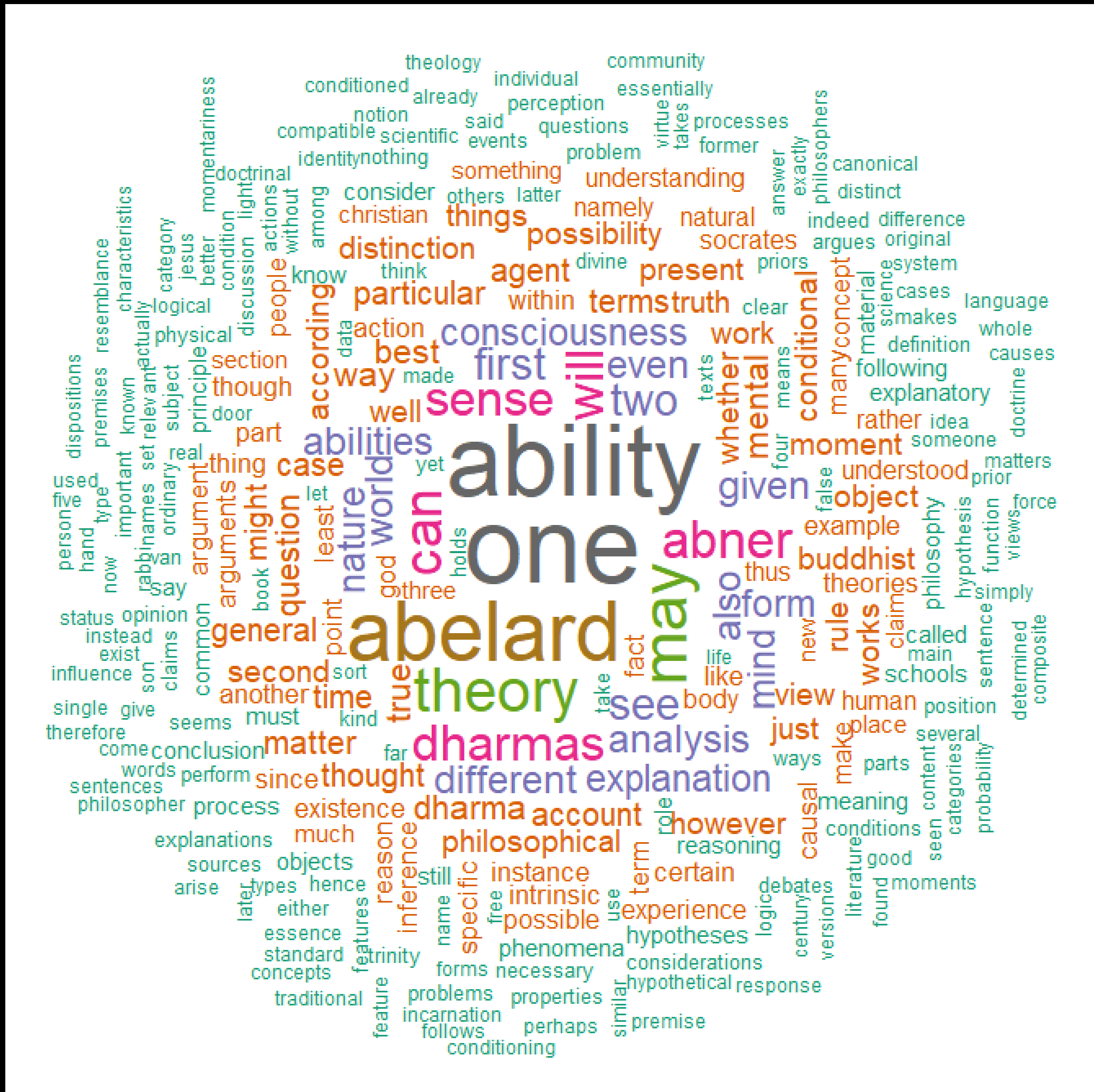
Causal Attributions & Corpus Analysis

Let's modify "PHILOSOPHY_CORPUS.txt" to scrape the entries for:

- abduction
- abelard
- abhidharma
- abilities
- abner-burgos

Then generate a wordcloud for those entries.




ability
one
abelard
theory
may
can
will
sense
abner
dharmas
first
even
two
consciousness
abilities
given
nature
world
also
form
mind
see
analysis
different
explanation
true
dharma
account
philosophical
however
matter
thought
general
second
time
question
might
according
particular
distinction
agent
present
terms
truth
possibility
things
best
way
well
action
case
moment
object
buddhist
rule
works
view
just
conditional
mental
whether
work
understanding
namely
natural
socrates
something
christian
consider
know
think
within
divine
priors
clear
concept
many
rather
understood
example
theories
thus
fact
like
body
human
place
make
causal
certain
experience
possible
intrinsic
instance
specific
reason
inference
existence
much
since
another
must
conclusion
process
objects
hypotheses
phenomena
considerations
hypothetical
response
premise
similar
necessary
properties
perhaps
conditioning
follows
incarnation
problems
forms
trinity
features
feature
traditional
concepts
standard
essence
either
types
arise
sources
explanations
philosopher
sentences
words
perform
come
therefore
single
give
seems
influence
exist
instead
status
opinion
claims
common
argument
arguments
book
least
point
god
three
holds
let
yet
thing
part
though
section
door
principle
subject
real
ordinary
rabbi
names
van
say
important
set
relevant
known
premises
dispositions
person
five
used
type
hand
now
son
actually
physical
logical
resemblance
characteristics
category
jesus
better
condition
discussion
light
actions
without
among
people
data
made
momentariness
doctrinal
identity
nothing
compatible
notion
scientific
conditioned
already
said
events
theology
individual
perception
questions
problem
community
essentially
virtue
takes
processes
former
philosophers
exactly
answer
canonical
distinct
indeed
difference
argues
original
system
science
material
cases
makes
definition
language
whole
causes
following
explanatory
idea
doctrine
someone
matters
prior
force
views
function
hypothesis
philosophy
simply
claim
called
main
schools
sentence
determined
composite
position
several
ways
parts
content
categories
probability
meaning
conditions
good
seen
moments
literature
found
debates
century
versions
logic
use
term
role
reasoning
free
name
still
later
hence
far
kind
sort
take
life
new
false
four
means
texts
others
latter
mean

J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

For creating spaces for larger corpora (even the full text of the SEP), it is better to switch to something like the Gensim toolkit in Python…

COCA (non-academic)

lemmas

word2vec context-predicting model

(Spearman's rho = **0.80**)

	responsible for the	*blame*	*fault*	*praise*
caused the	0.63	0.55	0.42	–0.10
responsible for the		0.61	0.35	0.16

Table 8: Cosine values for term comparisons for the non-academic COCA corpus word2vec space with causal attribution and responsibility attribution tokens.

J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

Philosophy Corpus:
SEP and IEP; lemmatized, word2vec

	responsible for the	*blame*	*fault*	*praise*
caused the	0.55	0.16	0.17	0.03
responsible for the		0.10	0.07	0.00

Table 9: Cosine values for term comparisons for the philosophy corpus word2vec space with causal attribution and responsibility attribution tokens.

J. Sytsma
R. Bluhm
P. Willemsen
K. Reuter

Causal Attributions & Corpus Analysis

Looking across these sets of analyses, find evidence for a positive answer to each of our four questions:

(1) Can corpus analysis provide independent support for the thesis that ordinary causal attributions are sensitive to normative information?

(2) Does the evidence coming from corpus analysis support the contention that outcome valence matters for ordinary causal attributions?

(3) Are ordinary causal attributions similar to responsibility attributions?

(4) Are causal attributions of philosophers different from causal attributions we find in corpora of more ordinary language?