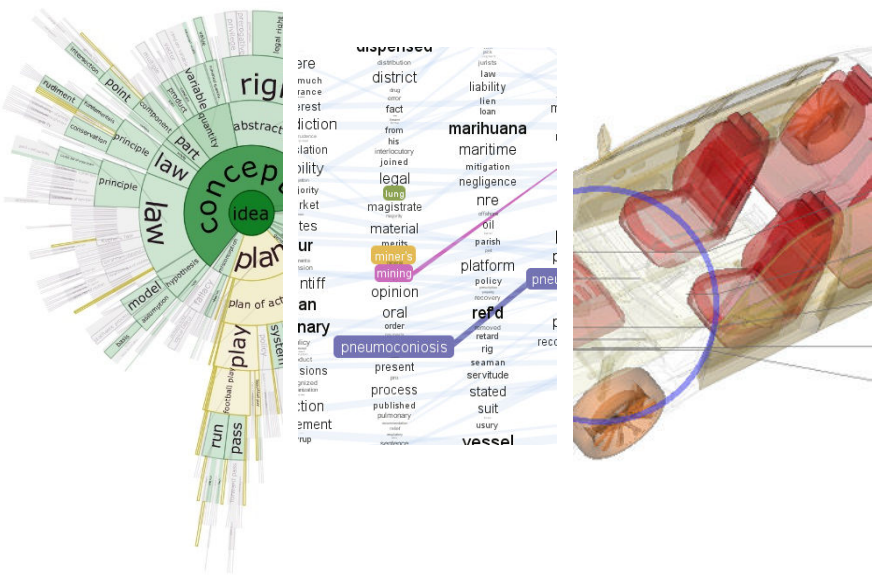
Semantics and Sentiment in Visual Text Analytics

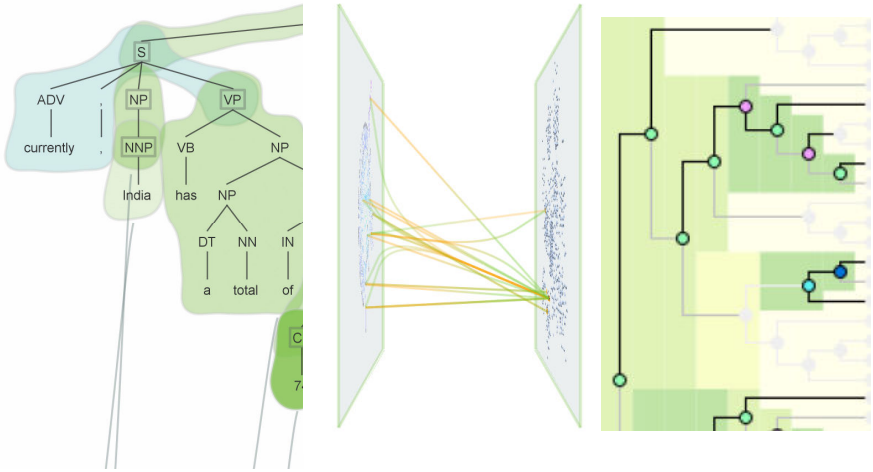
Christopher Collins
Assistant Professor, Canada Research Chair
University of Ontario Institute of Technology

vialab
visualization for information analysis

UOIT
CHALLENGE INNOVATE CONNECT

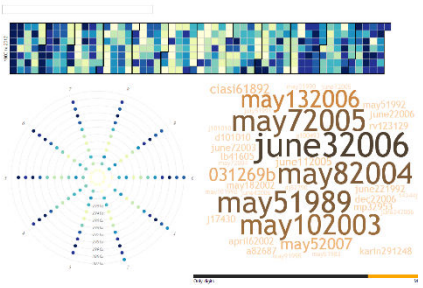


Linguistic Information Visualization

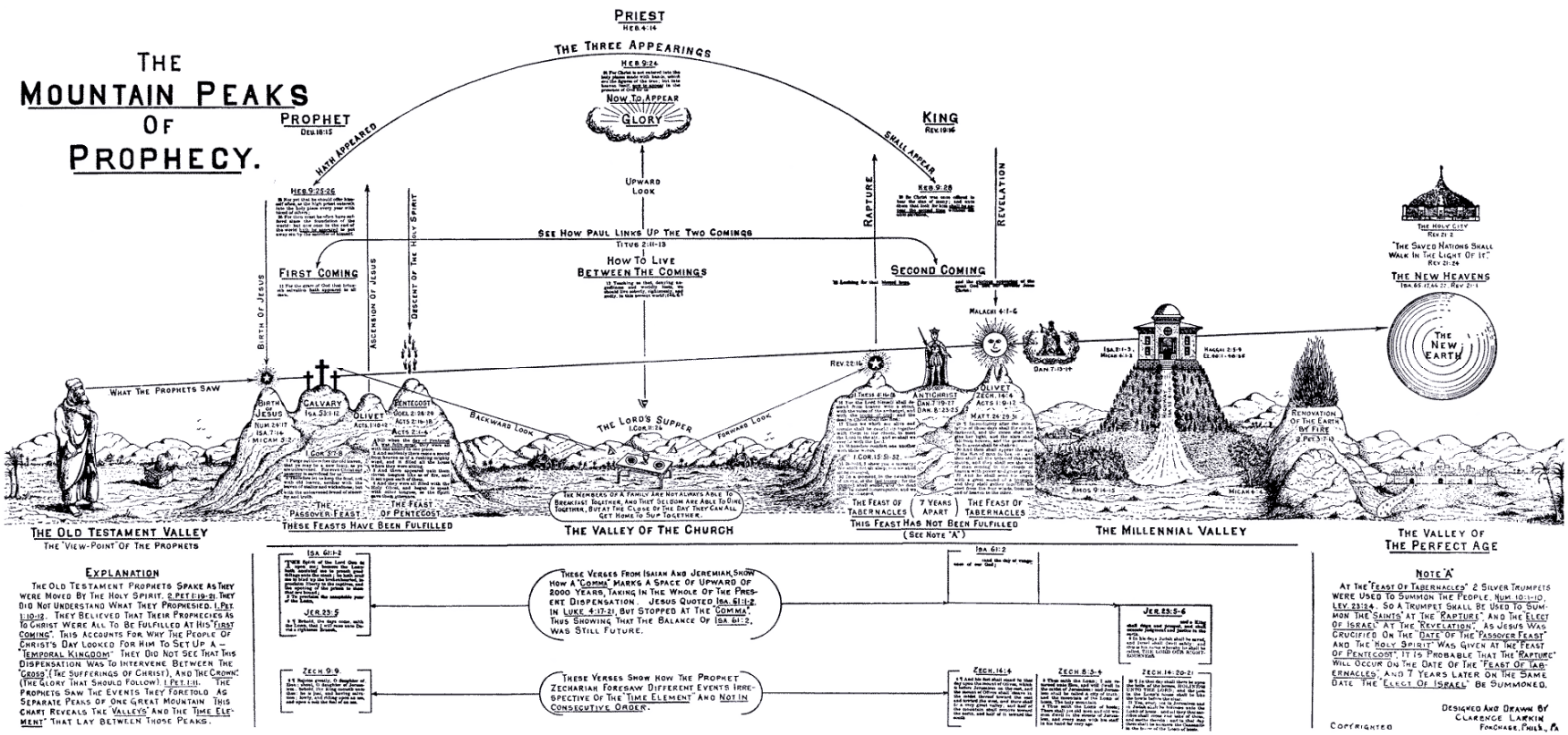


Visualization Technique
and Interaction Design

NUIs for visualization:
tables, walls, gestures



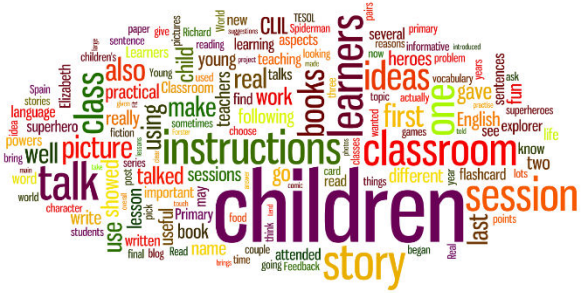
Applied visualization:
software, security,
humanities



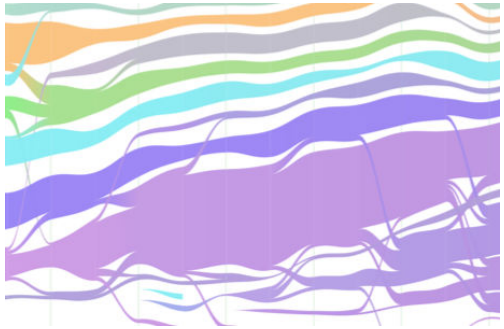
Mountain Peaks of Prophecy (Larkin, 1918)

Visual Text Analytics

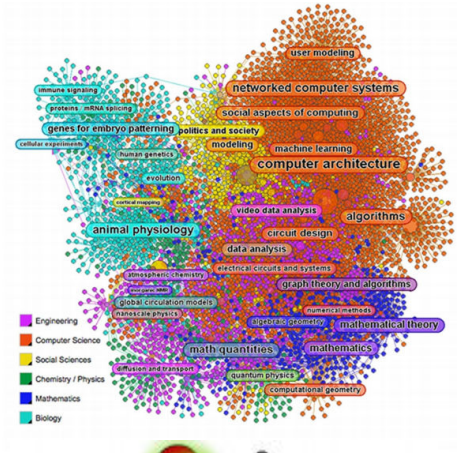
- Visual techniques for words, documents, sets of documents to support rapid summarization, trend analysis, exploration, search, comparative analysis, …
- Application areas include market analysis, legal studies, e-discovery, readability, literary studies, personal reflection, information retrieval and exploration, intelligence analysis

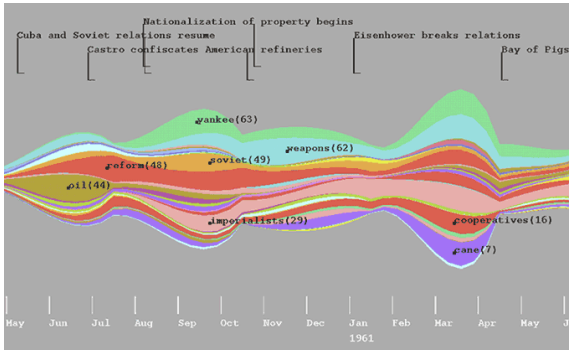


Word Clouds

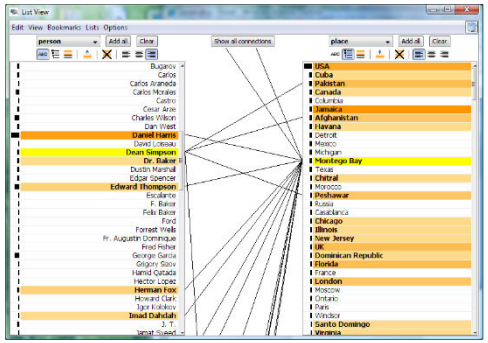


TextFlow

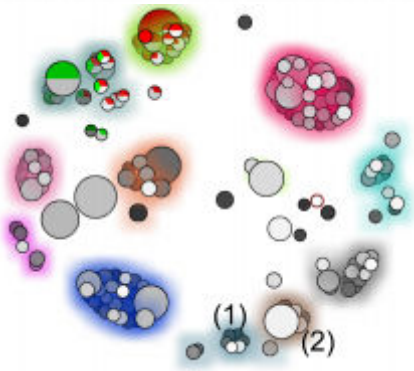




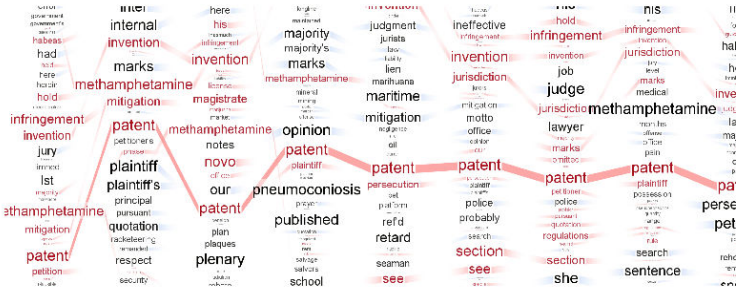
Theme River



Jigsaw

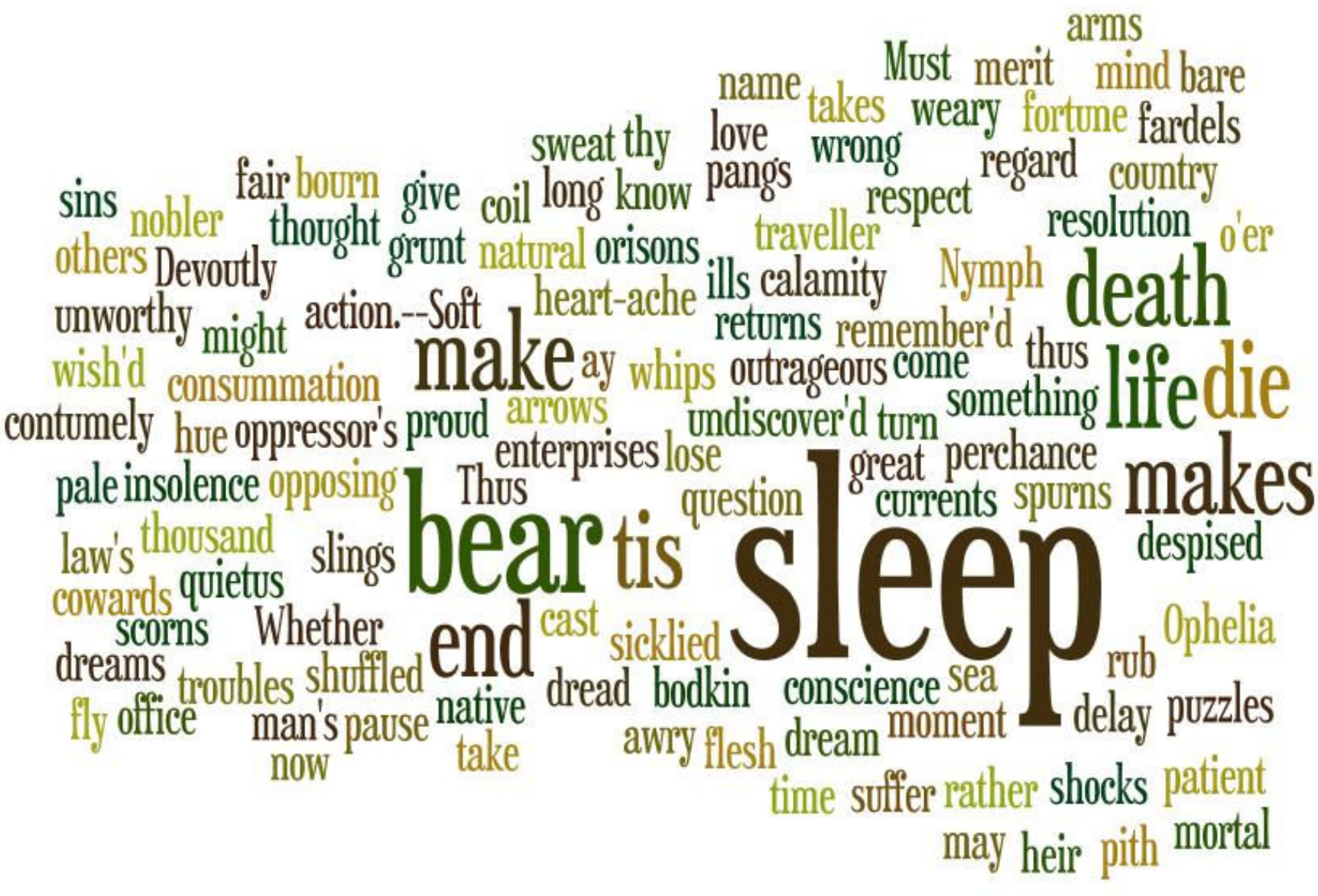


Topic Models

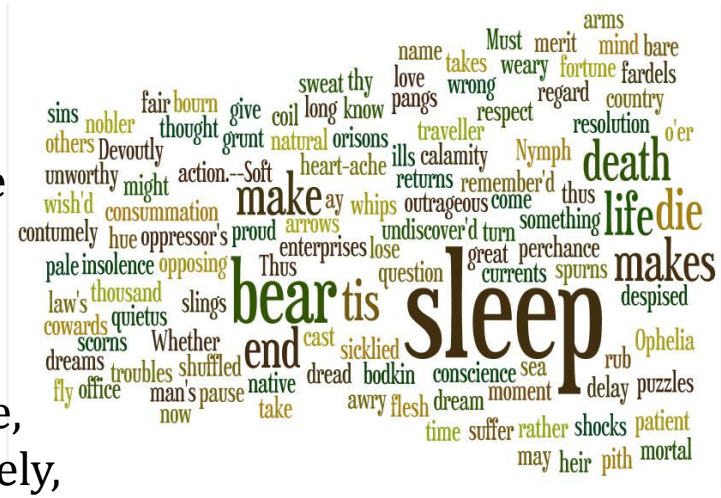


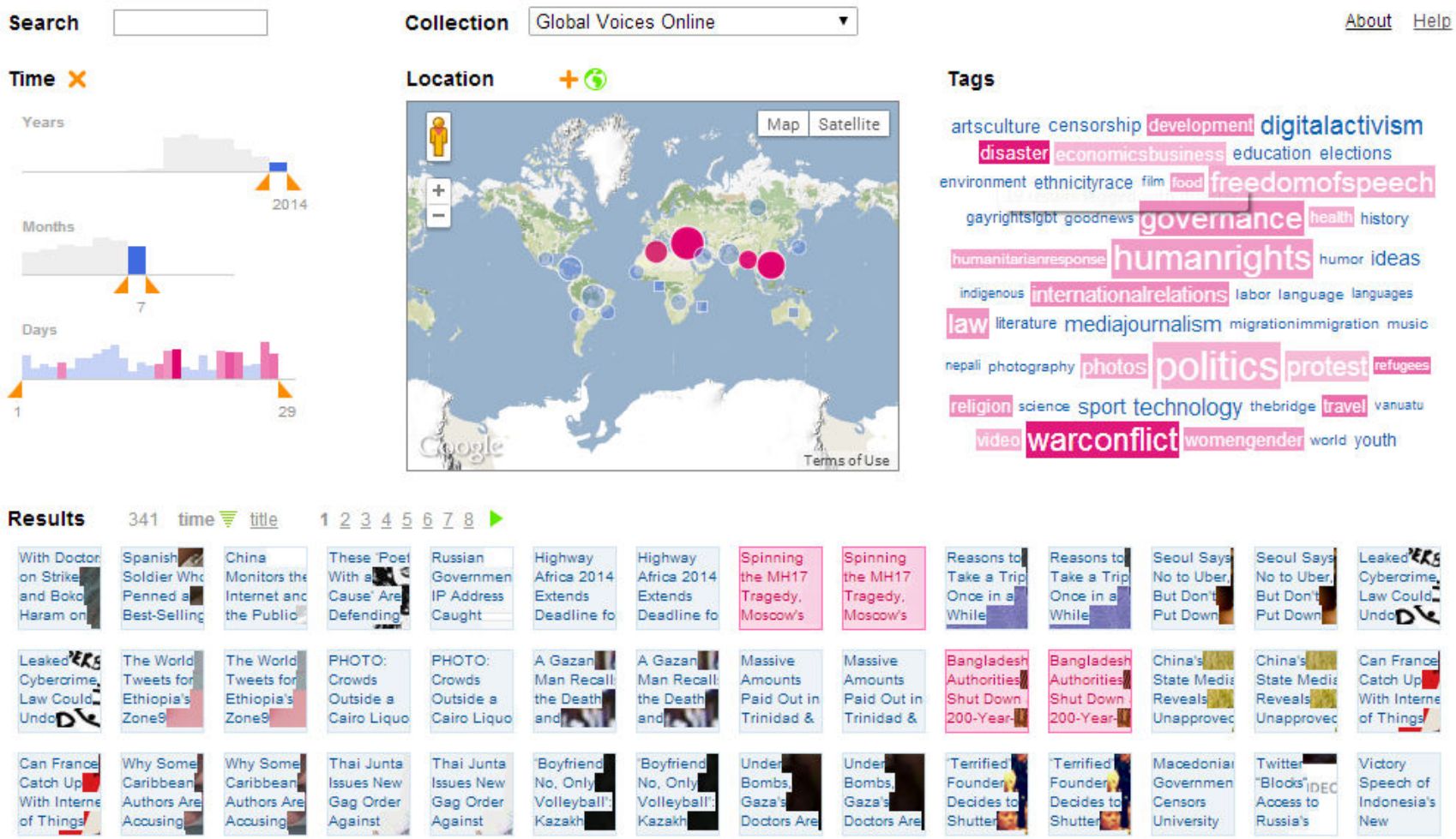
Parallel Tag Clouds

Themescape


arms
Must
merit
mind
bare
name
takes
weary
fortune
fardels
sweat
thy
love
wrong
regard
country
fair
bourn
give
coil
long
know
pangs
respect
sins
nobler
thought
grunt
natural
orisons
traveller
resolution
o'er
others
Devoutly
ills
calamity
Nymph
death
unworthy
might
action.--Soft
heart-ache
returns
remember'd
make
ay
whips
outrageous
come
thus
life
die
wish'd
consummation
arrows
undiscover'd
turn
something
contumely
hue
oppressor's
proud
enterprises
lose
great
perchance
makes
pale
insolence
opposing
Thus
question
currents
spurns
bear
tis
sleep
despised
law's
thousand
slings
quietus
cowards
scorns
Whether
end
cast
sicklied
Ophelia
dreams
troubles
shuffled
dread
bodkin
conscience
sea
rub
fly
office
man's
pause
native
delay
puzzles
now
take
awry
flesh
dream
moment
time
suffer
rather
shocks
patient
may
heir
pith
mortal

To be, or not to be: that is the question:
Whether 'tis nobler in the mind to suffer
The slings and arrows of outrageous fortune,
Or to take arms against a sea of troubles,
And by opposing end them? To die: to sleep;
No more; and by a sleep to say we end
The heart-ache and the thousand natural shocks
That flesh is heir to, 'tis a consummation
Devoutly to be wish'd. To die, to sleep;
To sleep: perchance to dream: ay, there's the rub;
For in that sleep of death what dreams may come
When we have shuffled off this mortal coil,
Must give us pause: there's the respect
That makes calamity of so long life;
For who would bear the whips and scorns of time,
The oppressor's wrong, the proud man's contumely,
The pangs of despised love, the law's delay,
The insolence of office and the spurns
That patient merit of the unworthy takes,
When he himself might his quietus make
With a bare bodkin? who would fardels bear,
…





Marian Dörk et al. VisGets: Coordinated Visualizations for Web-based Information Exploration and Discovery. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1205-1212, November-December, 2008.

10 Topics from Psych Review

'memory recognition retrieval recall items item list'

'representations representation bias single scale known relation'

'information processing action levels automatic components controlled'

'theory theories predictions account explain formation esteem'

'visual attention brain mechanism selection color attentional'

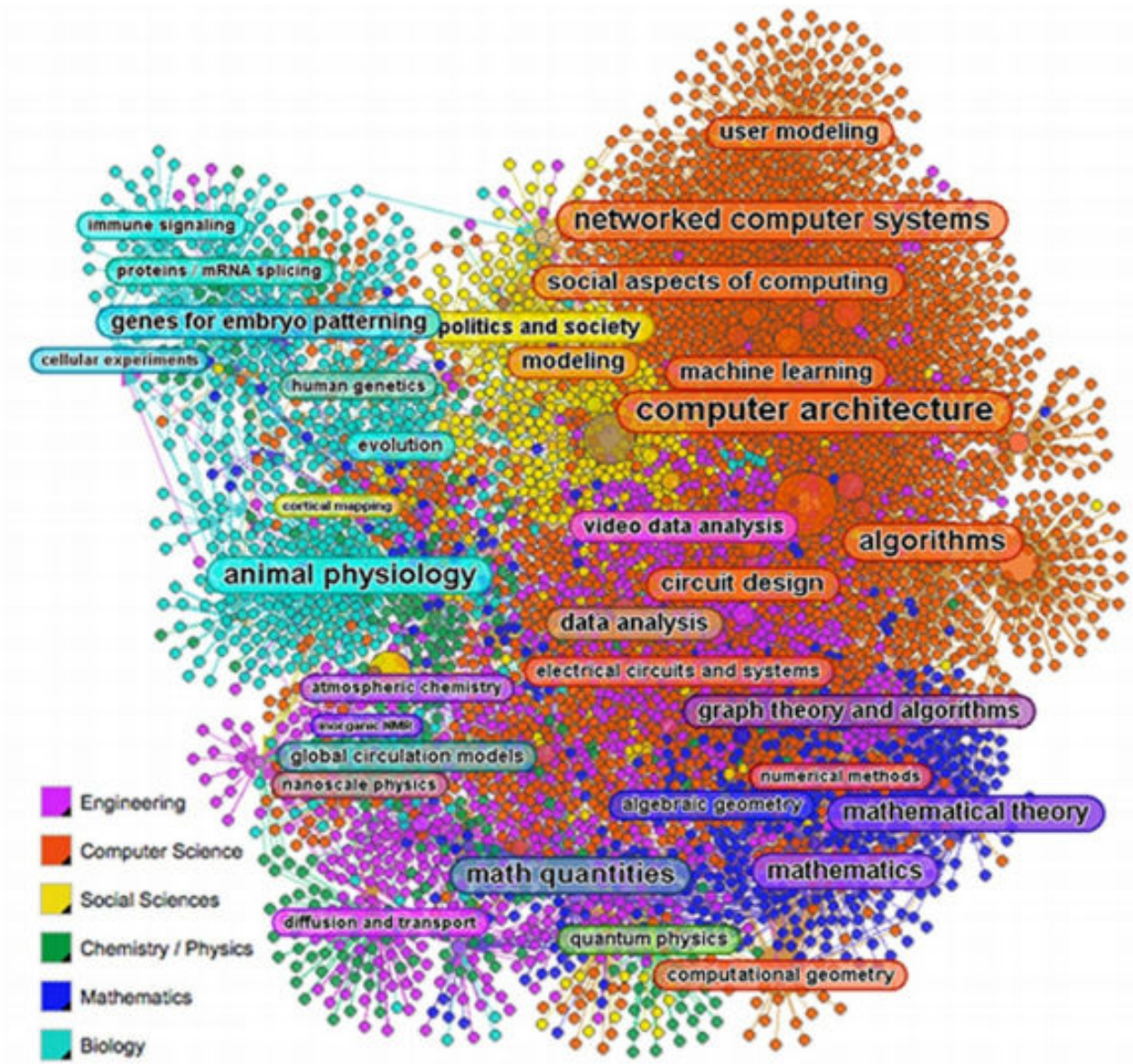
'speech system target motor masking neural relative'

'network semantic ability test lexical predictions normal'

'states related emotional positive primary arousal motivation'

'control conditioning responses conditions avoidance procedures suggested'

'stimulus effects shown concepts trial free generalization'



TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling.
In ACM Transactions on Intelligent Systems and Technology, TIST 2012.

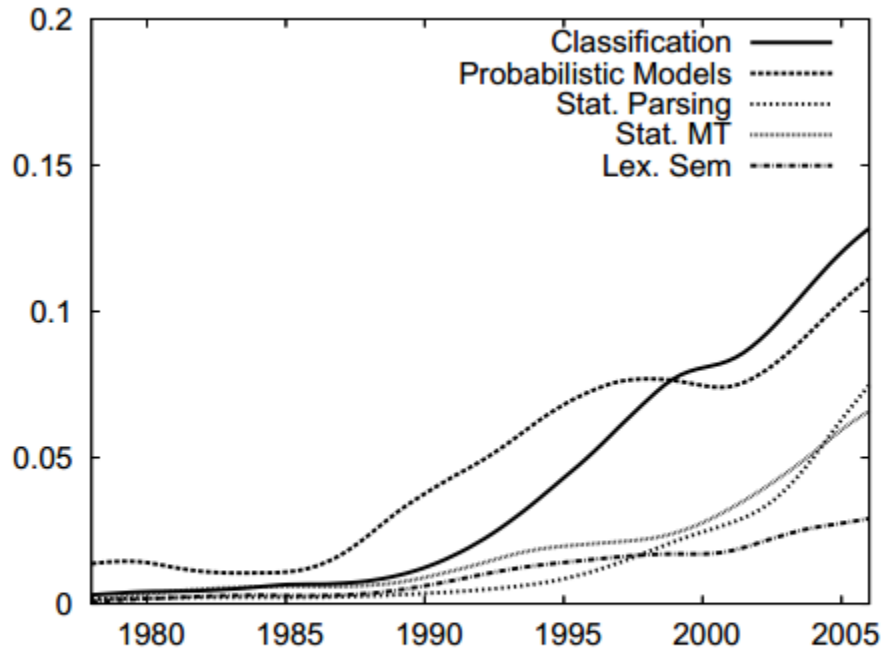


Figure 1: Topics in the ACL Anthology that show a strong recent increase in strength.

Anaphora Resolution	resolution anaphora pronoun discourse antecedent p
Automata	string state set finite context rule algorithm strings l
Biomedical	medical protein gene biomedical wkh abstracts med
Call Routing	call caller routing calls destination vietnamese route
Categorial Grammar	proof formula graph logic calculus axioms axiom th
Centering*	centering cb discourse cf utterance center utterances
Classical MT	japanese method case sentence analysis english dict
Classification/Tagging	features data corpus set feature table word tag al test
Comp. Phonology	vowel phonological syllable phoneme stress phoneti
Comp. Semantics*	semantic logical semantics john sentence interpretat
Dialogue Systems	user dialogue system speech information task spoke
Discourse Relations	discourse text structure relations rhetorical relation
Discourse Segment.	segment segmentation segments chain chains bound
Events/Temporal	event temporal time events tense state aspect referen
French Function	de le des les en une est du par pour
Generation	generation text system language information knowl
Genre Detection	genre stylistic style genres fiction humor register biber authorship registers
Info. Extraction	system text information muc extraction template names patterns pattern domain
Information Retrieval	document documents query retrieval question information answer term text web
Lexical Semantics	semantic relations domain noun corpus relation nouns lexical ontology patterns
MUC Terrorism	slot incident tgt target id hum phys type fills perp
Metaphor	metaphor literal metonymy metaphors metaphorical essay metonymic essays qualia analogy
Morphology	word morphological lexicon form dictionary analysis morphology lexical stem arabic
Named Entities*	entity named entities ne names ner recognition ace nes mentions mention
Paraphrase/RTE	paraphrases paraphrase entailment paraphrasing textual para rte pascal entailed dagan

David Hall et al. 2008. Studying the history of ideas using topic models. In Proc. *EMNLP*, pp. 363-371.

Linguistic Methods

- Word Counting (yesterday)
- Word Scoring (yesterday)
- Stemming
- Stop Word Removal (yesterday – revisit)
- Part of Speech Tagging
- Parsing
- Word Sense Disambiguation
- Named Entity Recognition
- Semantic Categorization
- Sentiment Analysis
- Topic Modeling (yesterday)

NLTK: Natural Language Toolkit

- NLTK.org
- Python

NLTK 3.0 documentation

NEXT | MODULES | INDEX

Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, and an active discussion forum.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called "a wonderful tool for teaching, and working in, computational linguistics using Python," and "an amazing library to play with natural language."

Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The book is being updated for Python 3 and NLTK 3. (The original Python 2 version is still available at http://nltk.org/book_1ed.)

TABLE OF CONTENTS

NLTK News
Installing NLTK
Installing NLTK Data
Contribute to NLTK
FAQ
Wiki
API
HOWTO

SEARCH

Go

Enter search terms or a module, class or function name.

Stemming

- Reduce words to their ‘stems’ by removing endings (morphology)
  - running -> run
  - runs -> run

- A good way to increase signal and reduce fracturing of the corpus if there aren’t many words.
- Note: Keep the original words somewhere! Also keep the case if you choose to lowercase the word; you never know when you’ll need this data

Stop Word Removal

- Common words such as “and”, “the”, “I” are removed from view to highlight content words
- Domain specific stop words, e.g. in legal domain:
  - Court, attorney, honour, plaintiff, etc.
- Caution! These words have been shown to be useful for stylistic analysis! When working with text corpora, KEEP EVERYTHING.

Part of Speech Tagging

- Assign grammatical roles to words
- Conventional tagsets and representation:
  - The/AT grand/JJ jury/NN commented/VBD on/IN a/AT number/NN of/IN …
- Many words are ambiguous: fly, chair, run, store, table, and more!
  - Fly/NN
  - Fly/VB

Fly/NN


Fly/VB

Term / Concept Ambiguity

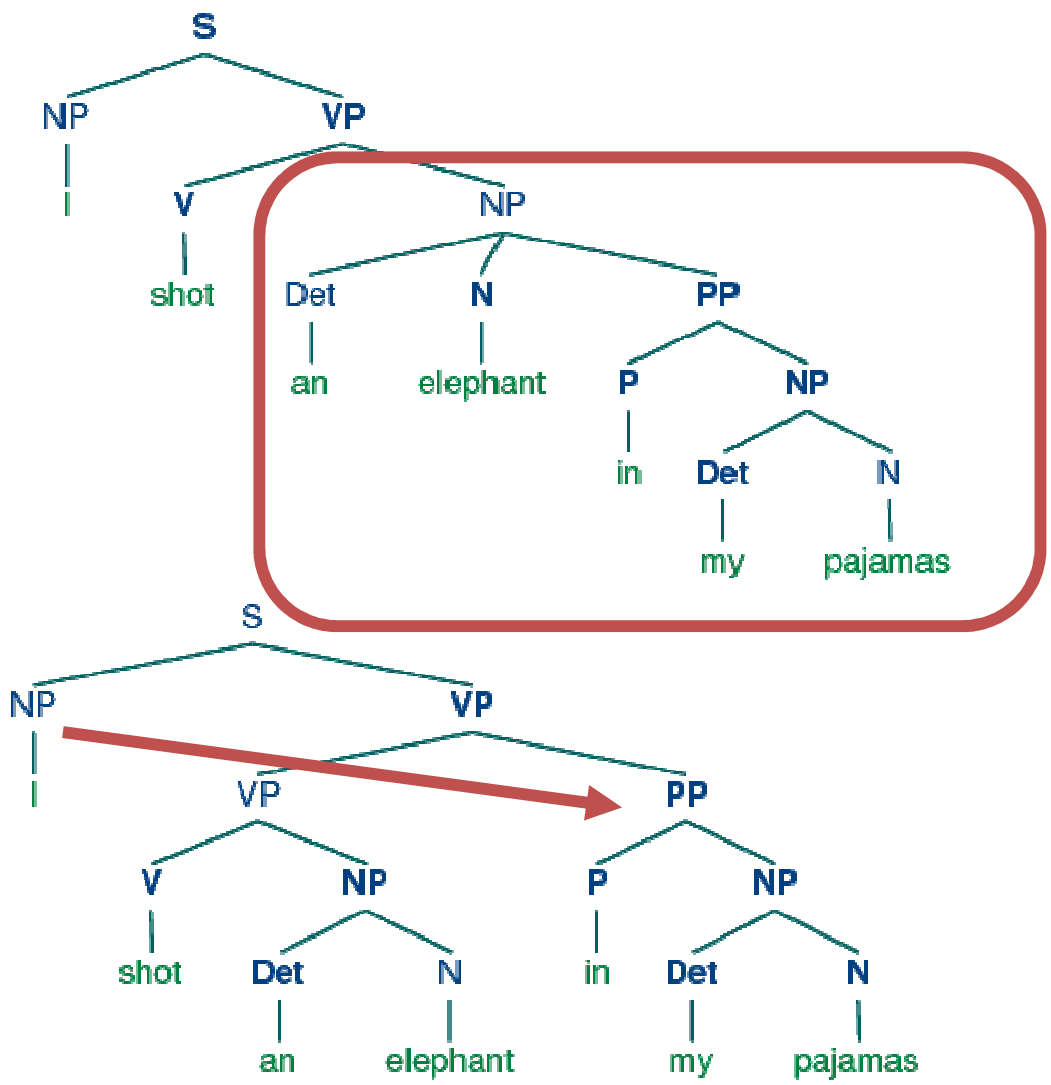
- Most meaning comes from our minds and common understanding.

- "How much is that doggy in the window?"
  - how much: social system of barter and trade (not the size of the dog)
  - "doggy" implies childlike, plaintive, probably cannot do the purchasing on their own
  - "in the window" implies behind a store window, not really inside a window, requires notion of window shopping

(Hearst, 2006)

Parsing

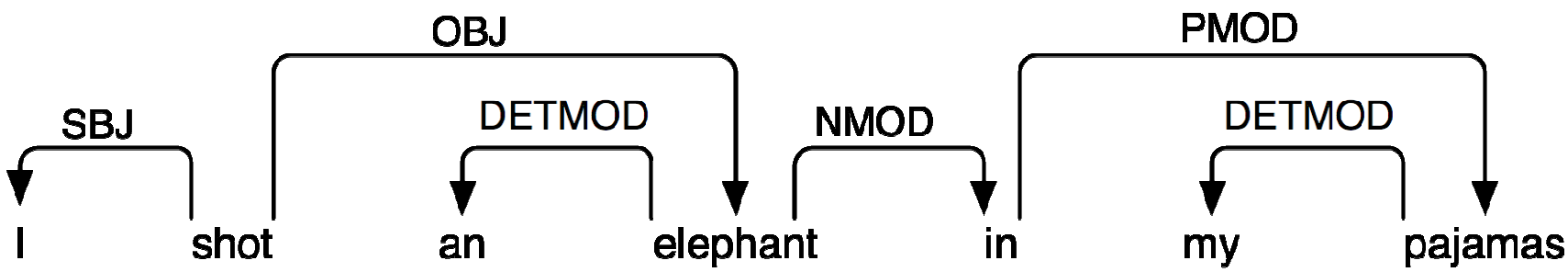
- Determining language structure
- Can reveal word-word relationships
- Useful for processing negation



https://nltk.googlecode.com/svn/trunk/doc/book/ch08.html

Dependency Parsing

- Labelled directed graph
- Arcs represent relationships from heads to dependents



Head of sentence

https://nltk.googlecode.com/svn/trunk/doc/book/ch08.html

Word Sense Disambiguation

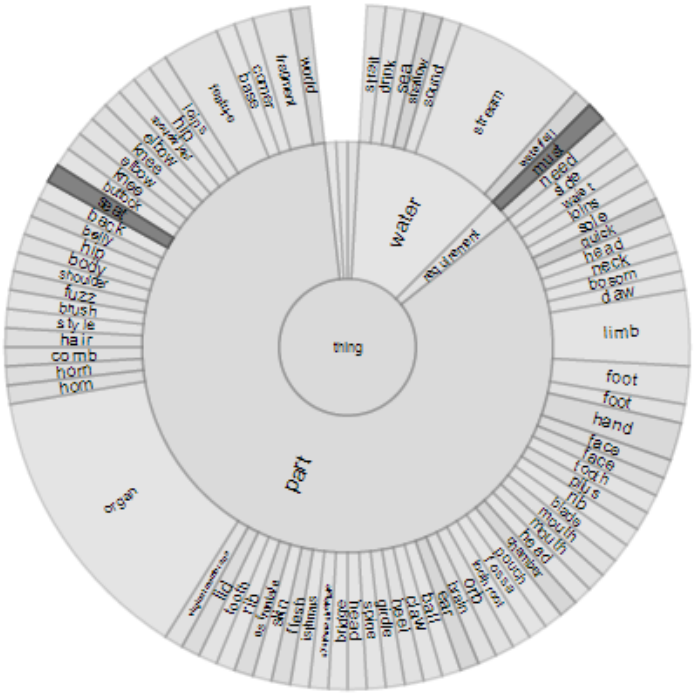
- Susan, the meeting chair, chaired the meeting well from the big chair in the front of the room.
  - Leader of a meeting
  - Action of leading a meeting
  - An object to sit upon

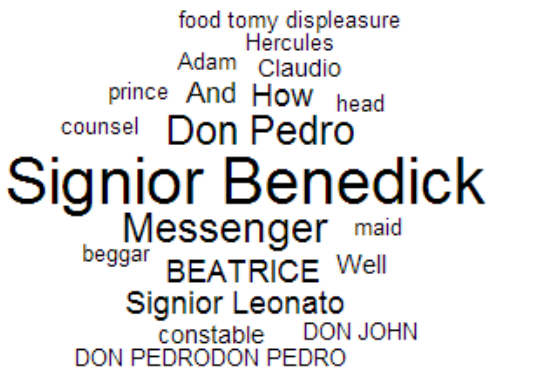
Word Sense Disambiguation

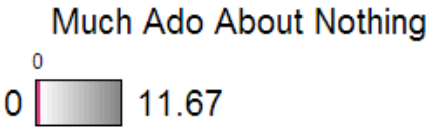
- This is VERY difficult for a computer.
- Contexts are often the same and meanings can be quite fine-grained:
  - bank the financial institution, bank the building in which the financial institution is housed
- Annual contest: SENSEVAL
- My method: assume the most common sense

Named Entity Recognition

- What are the people, places in the text?
- Use NLTK – it's very good at this.







http://vialab.science.uoit.ca/docuburst

Semantic Categorization

- Placing a word into an ontology or sense thesaurus based on *meaning.*

- Common resources include:
  - WordNet
  - Roget's Thesaurus

WordNet

- A large lexical database, or “digital dictionary”
- Covers most English nouns, verbs, adjectives, adverbs
- Organizes *synsets* by *meaning*
- Words are related to one another through many different relationship types:
- X is a kind of Y, X has part Y, an X Ys, X is Y/has property Y

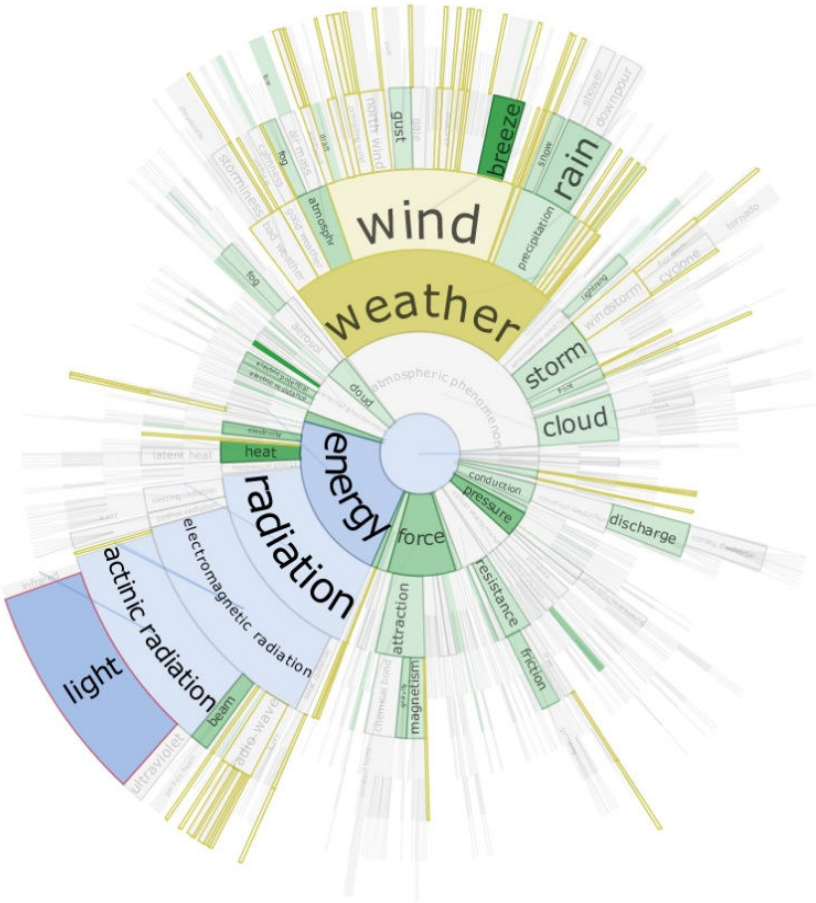
Hyponymy

- The “IS-A” relation for nouns

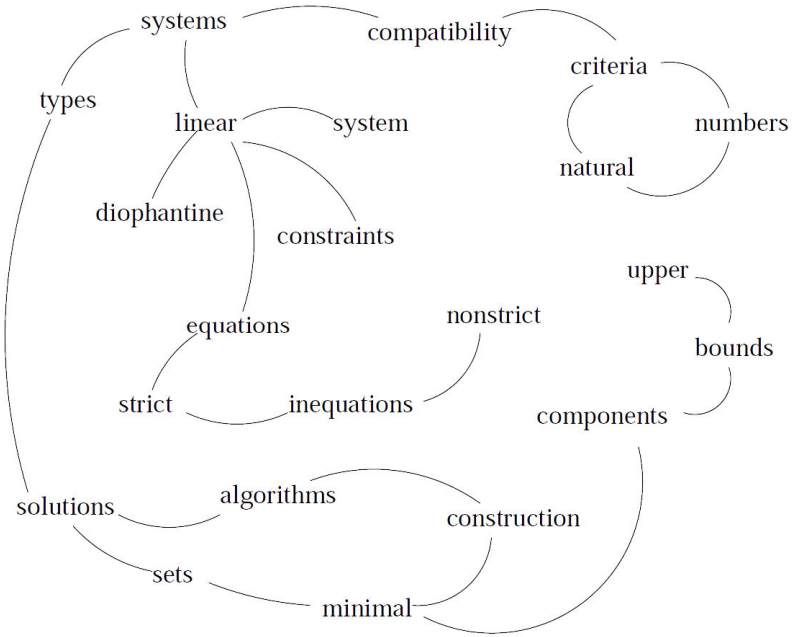
```
                {vehicle}
                  /   \
{car, automobile}     {bicycle, bike}
      /             \             \
  {convertible}  {SUV}  {mountain bike}
```

SEMANTIC VISUALIZATIONS

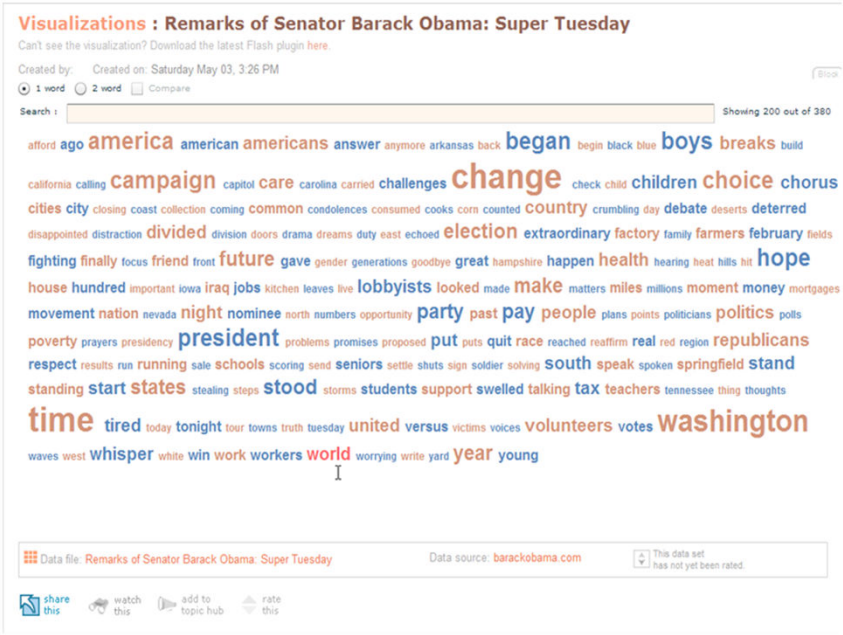
DocuBurst



Collins, C.; Carpendale, S.; Penn, G. DocuBurst: Visualizing Document Content using Language Structure. Proceedings of Eurographics/IEEE VGTC Symposium on Visualization, June, 2009.



Mihalcea and Tarau, 2004

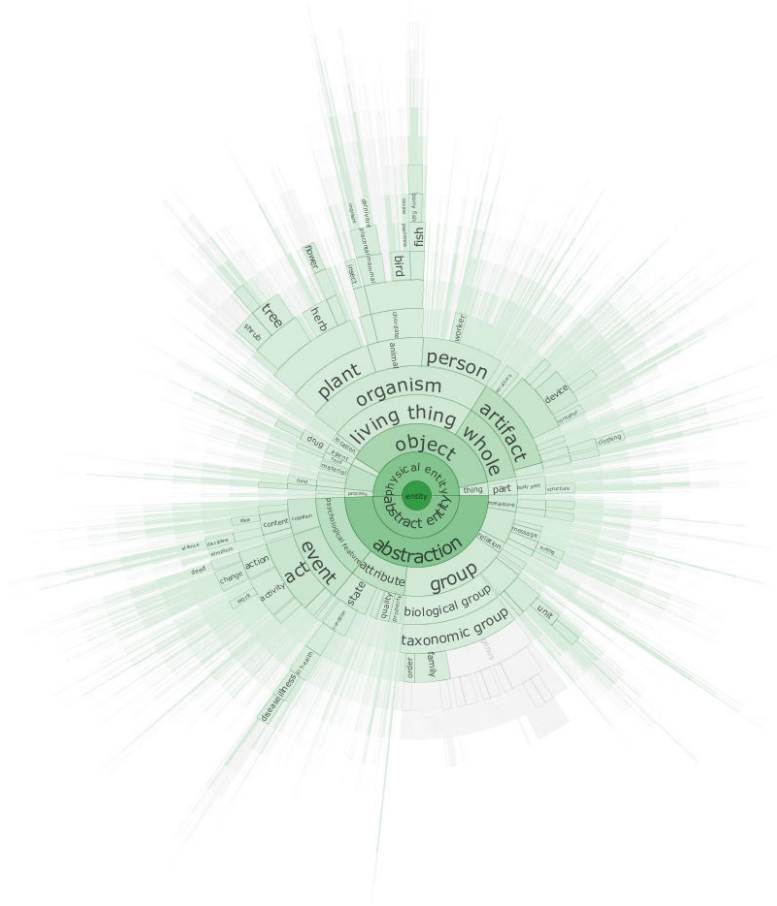


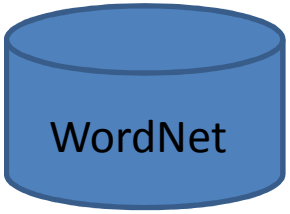
Wattenberg et al., 2008

DocuBurst

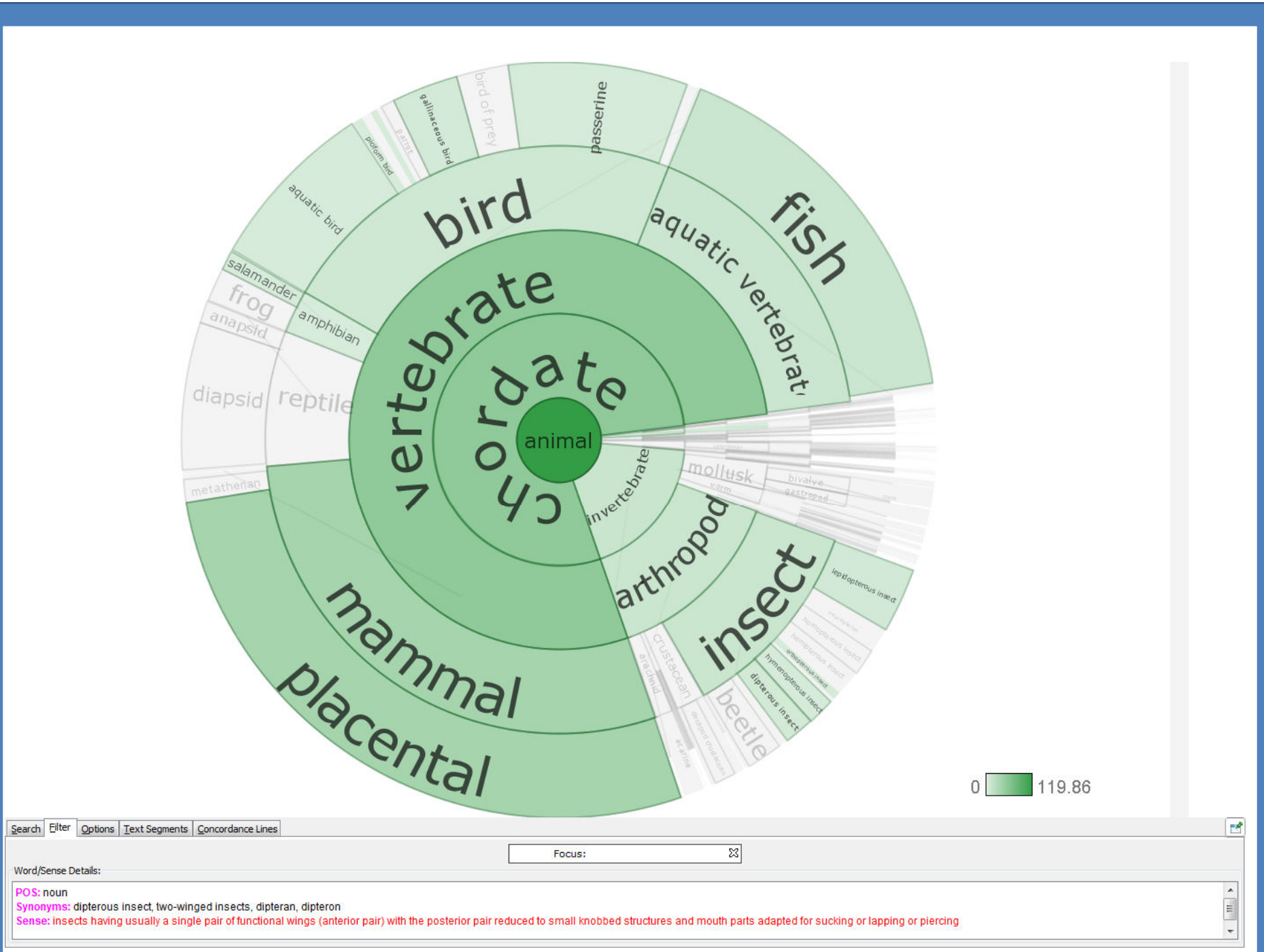
games→game
taken→take

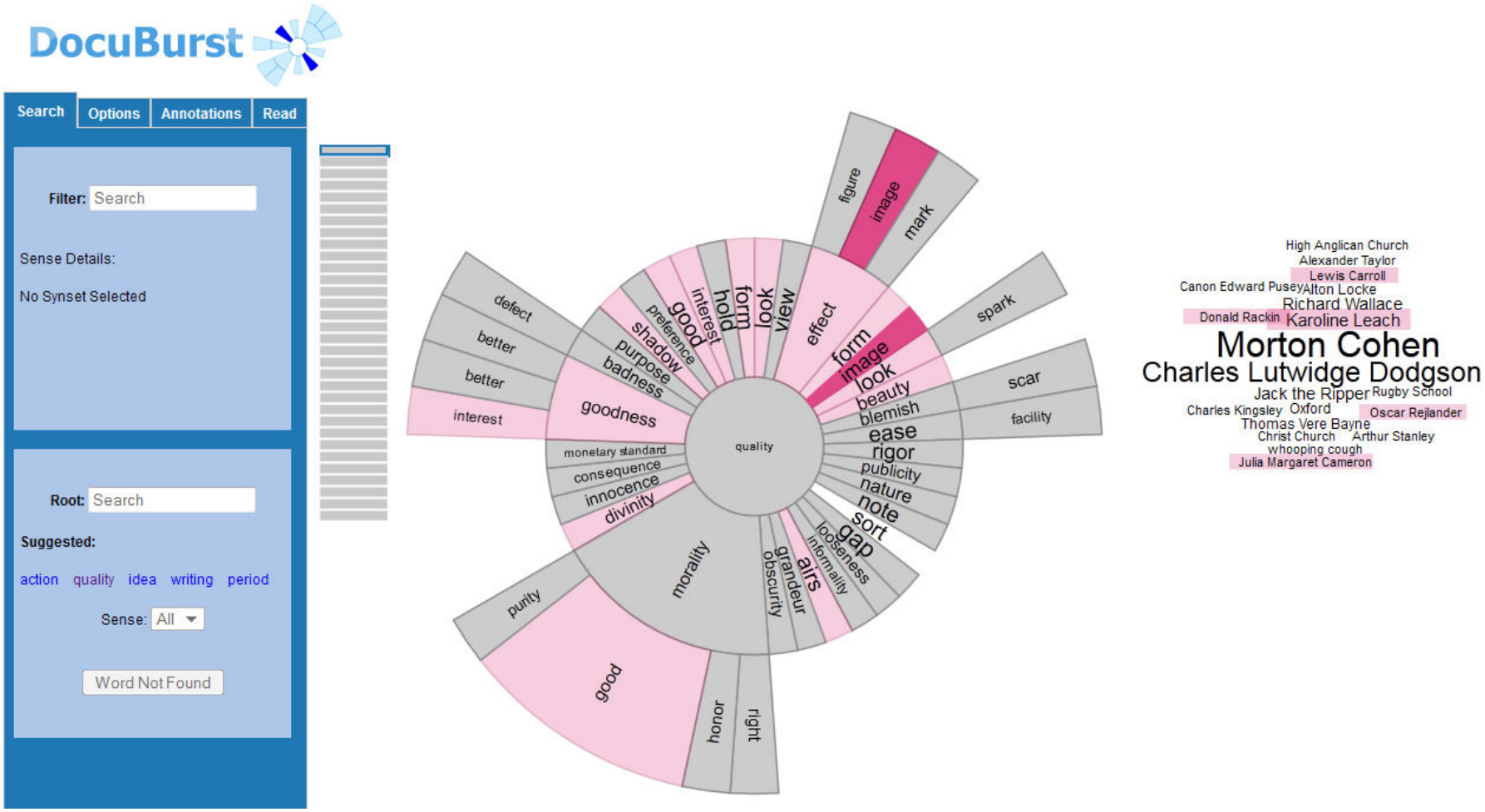
absolute,noun,10
chair,noun,2
moment,noun,11
game,noun,30
reality,noun,3
take,verb,13
represent,verb,17
...





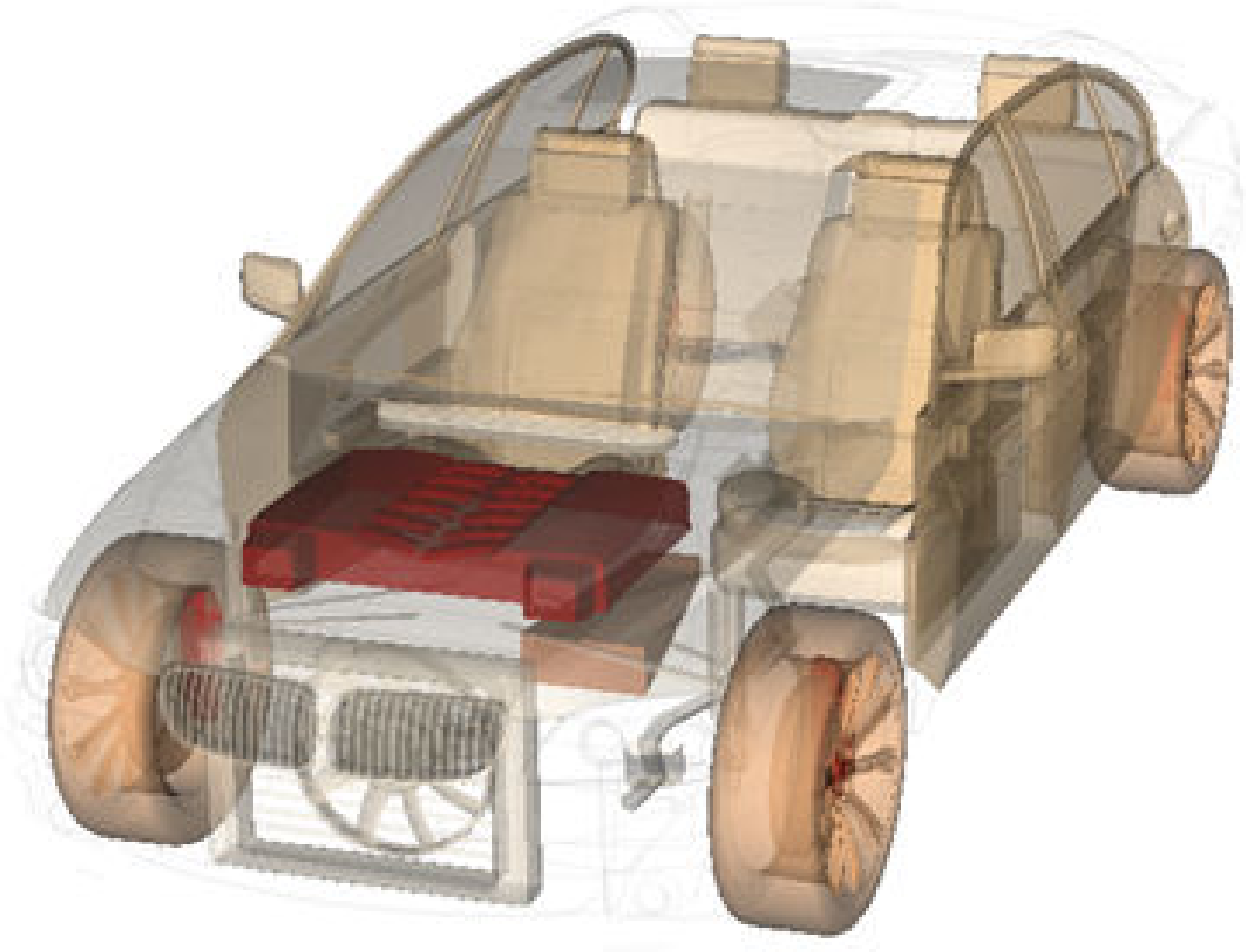
game IS activity
chair IS furniture


animal
chordate
vertebrate
bird
passerine
aquatic vertebrate
fish
gallinaceous bird
bird of prey
pigeon bird
aquatic bird
salamander
frog
anapsid
amphibian
diapsid
reptile
metatherian
mammal
placental
invertebrate
mollusk
arthropod
insect
lepidopterous insect
hymenopterous insect
dipterous insect
beetle
crustacean
arachnid
0
119.86
Search
Filter
Options
Text Segments
Concordance Lines
Focus:
Word/Sense Details:
POS: noun
Synonyms: dipterous insect, two-winged insects, dipteran, dipteron
Sense: insects having usually a single pair of functional wings (anterior pair) with the posterior pair reduced to small knobbed structures and mouth parts adapted for sucking or lapping or piercing



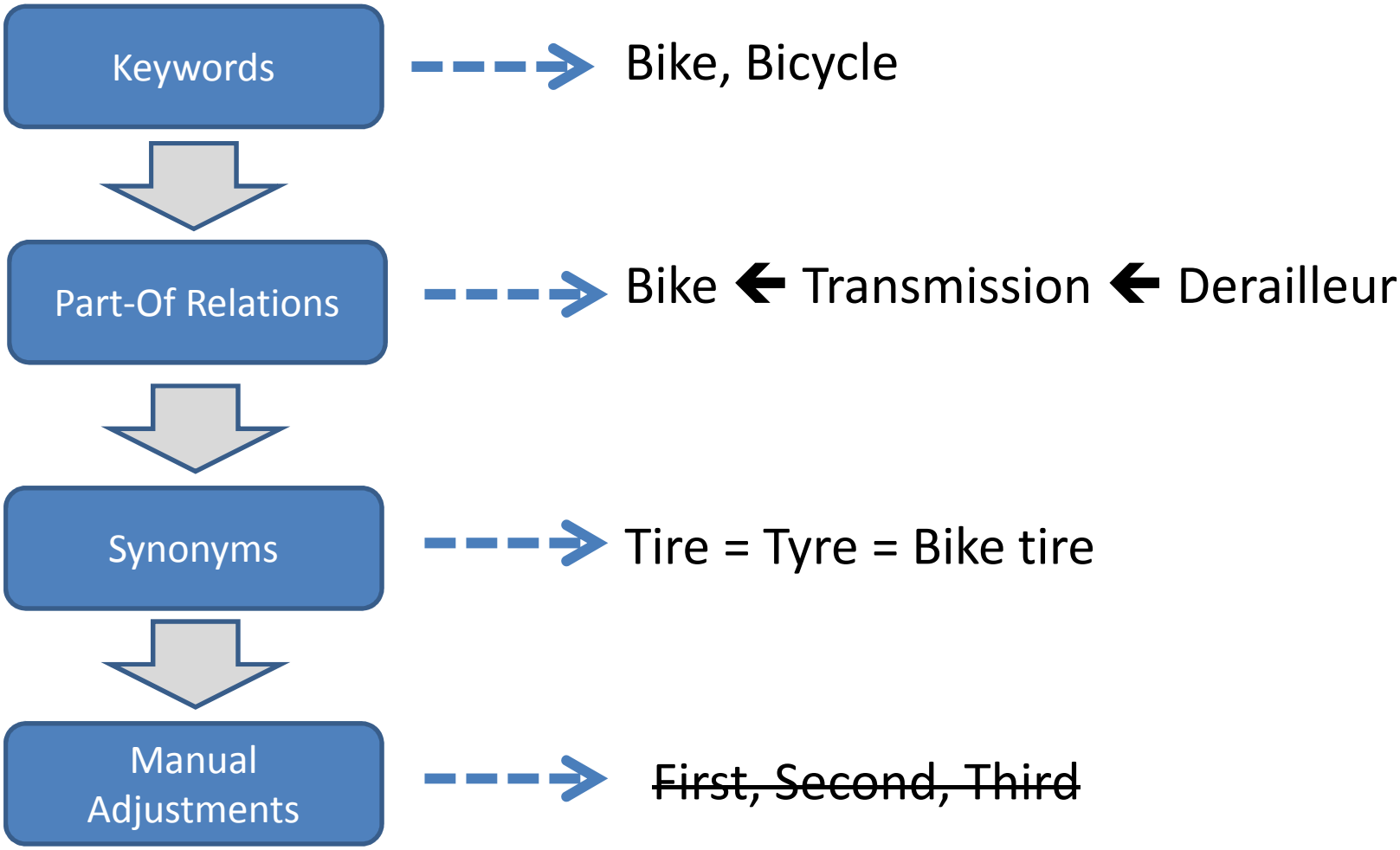
Try it! http://vialab.science.uoit.ca/docuburst

Descriptive Non-Photorealistic Rendering



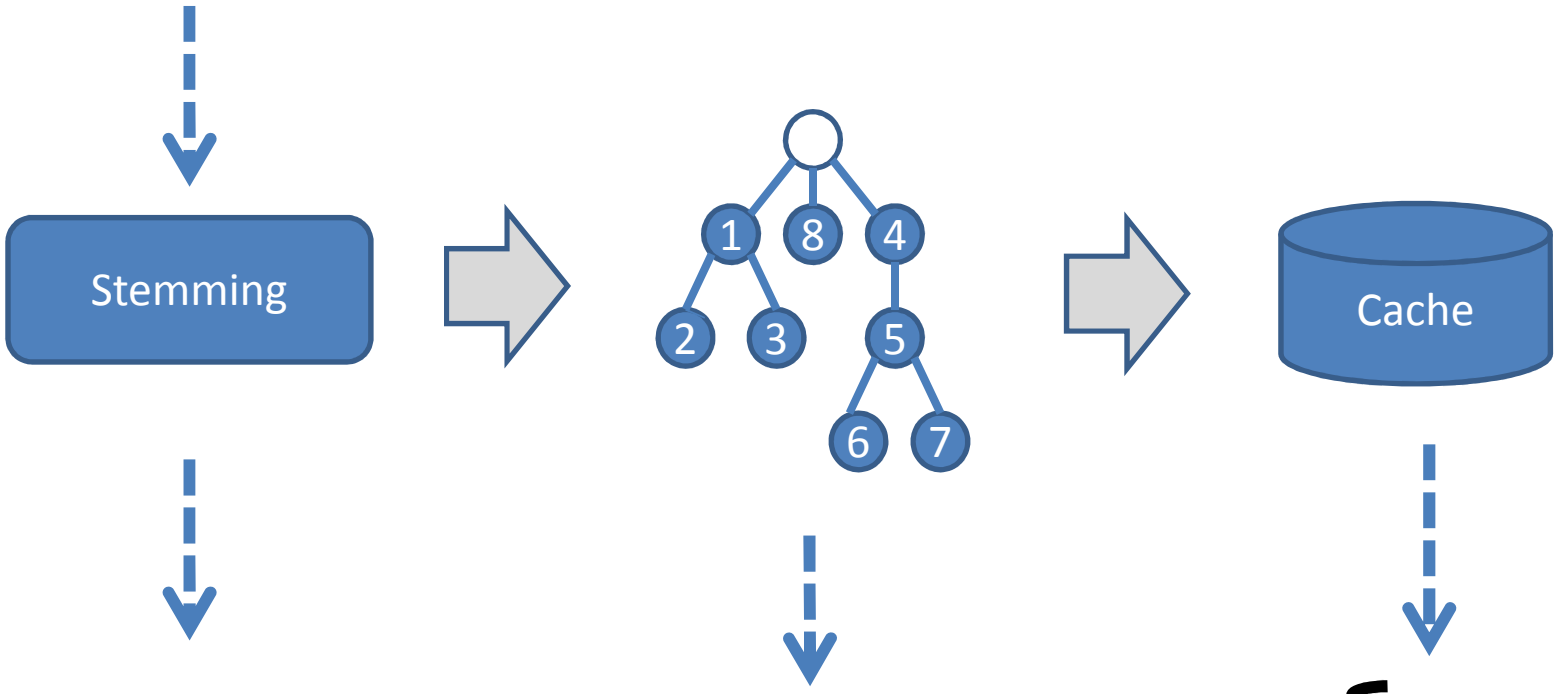
M. Chang and C. Collins, "Exploring Entities in Text with Descriptive Non-photorealistic Rendering," in*Proc. of the 2013 IEEE Pacific Visualization Symposium (PACIFICVIS '13)*, 2013.

Ontology Generation



Entity Extraction

"**Brakes** failed going at 35 mph."

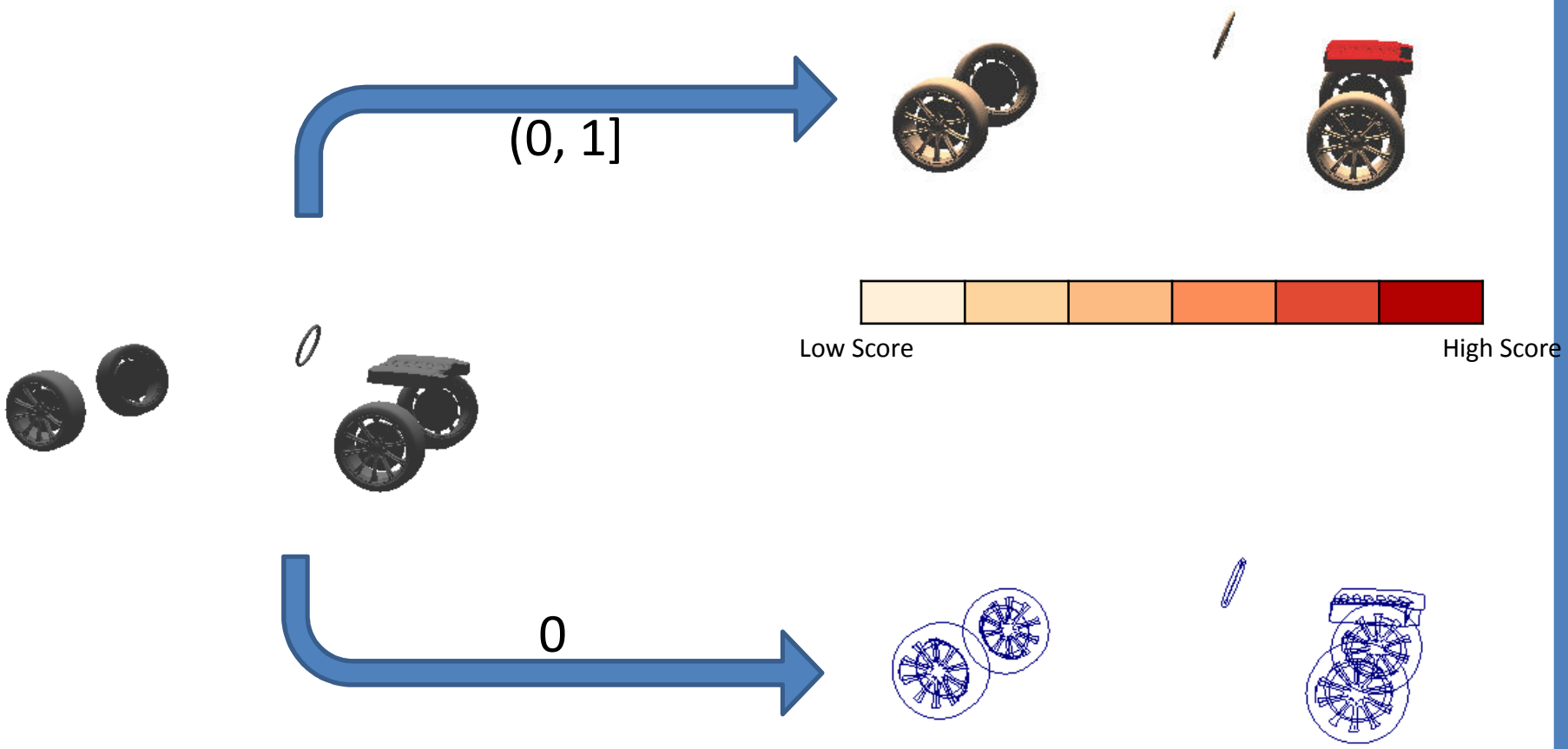


"brak"

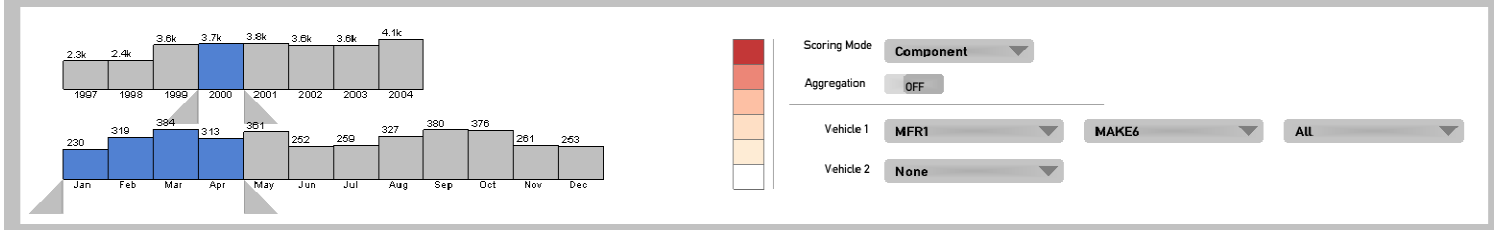
6: {brak, brak system…}

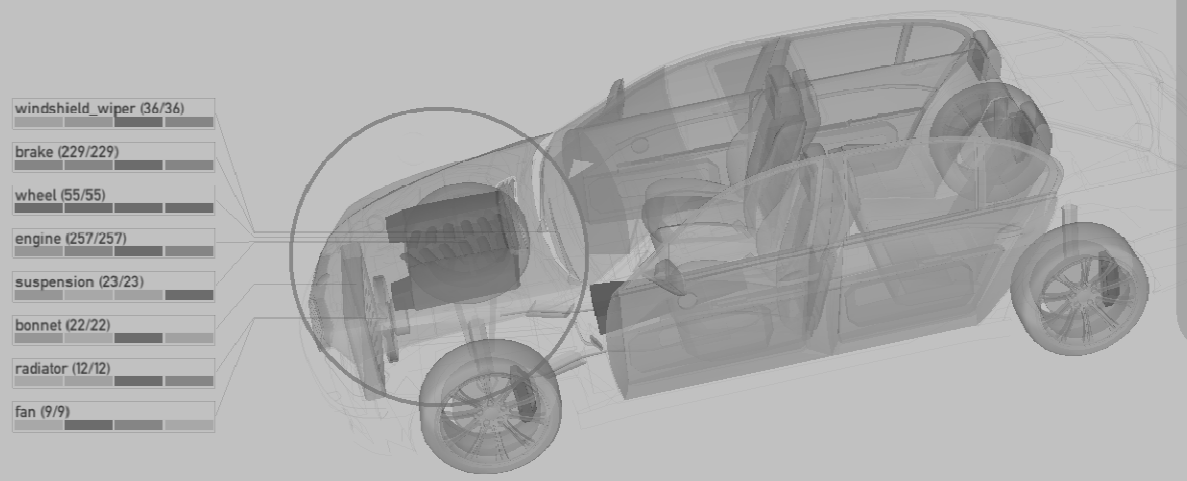
Document: 123
Entity: 6
Location: 0

Visual Representation



Main Interface

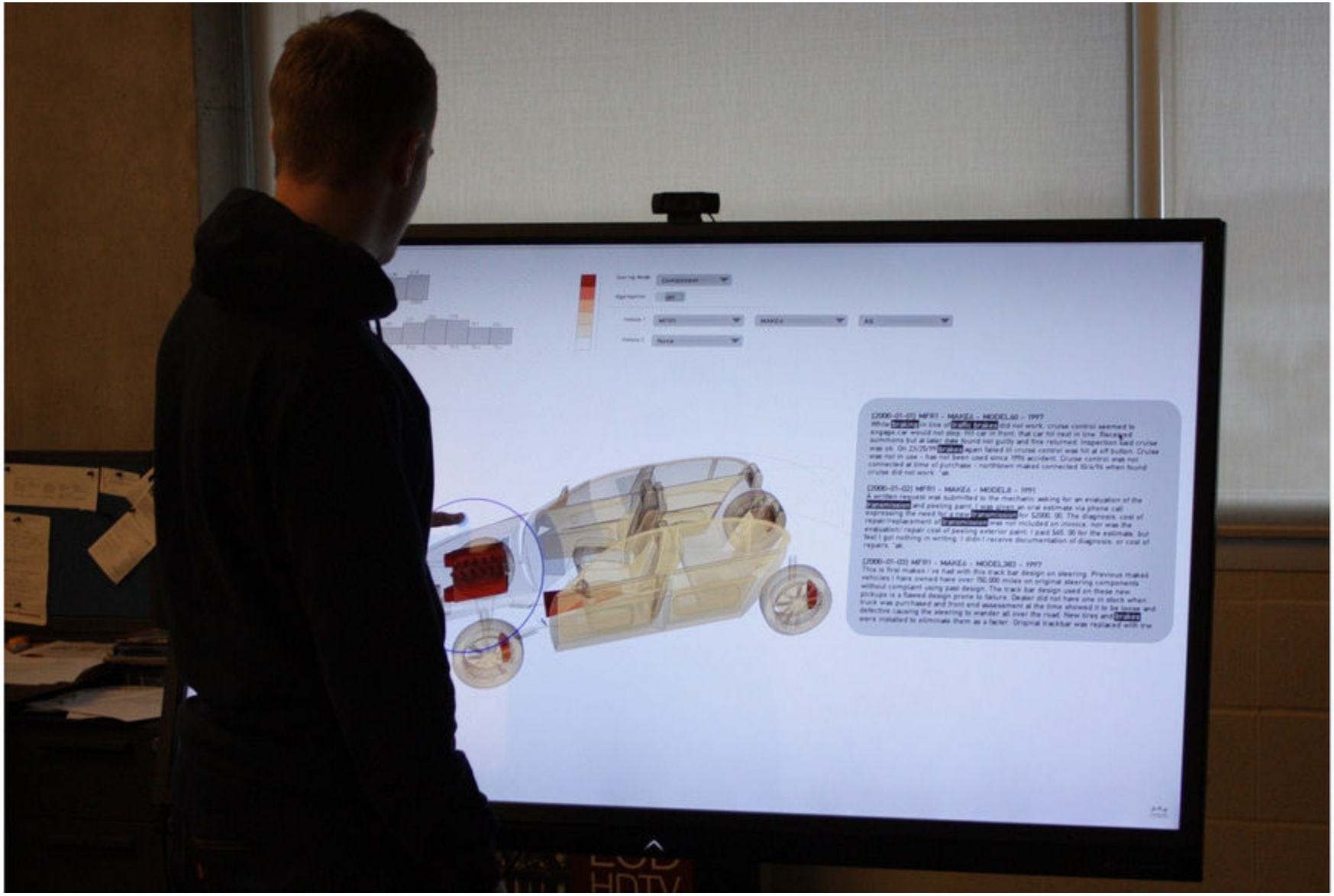




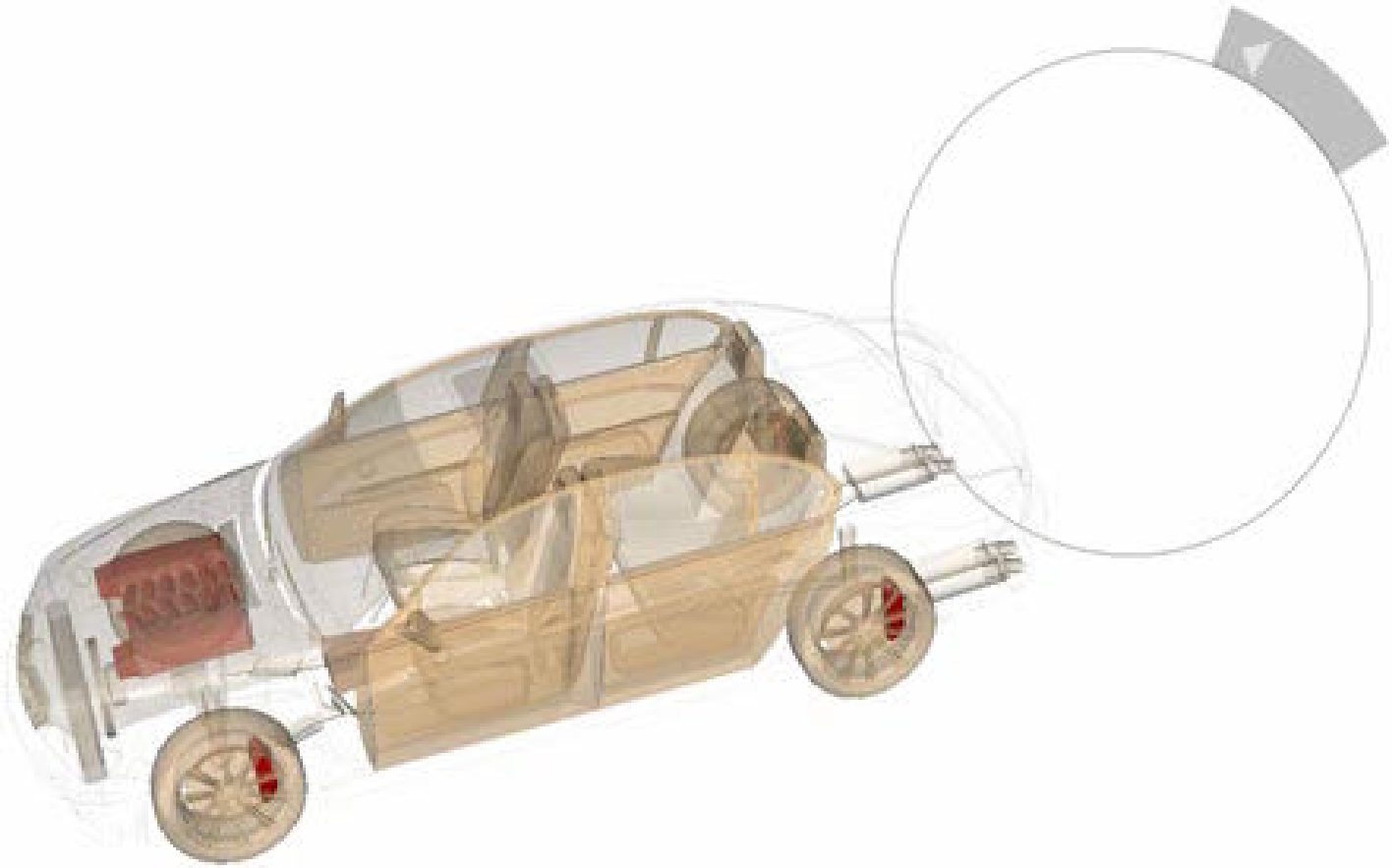
[2000-01-01] MFR1 - MAKE6 - MODEL60 - 1997
While braking in line of traffic brakes did not work, cruise control seemed to engage,car would not stop. Hit car in front, that car hit next in line. Received summons but at later date found not guilty and fine returned. Inspection said cruise was ok. On 23/25/99 brakes again failed til cruise control was hit at off button. Cruise was not in use - has not been used since 1996 accident. Cruise control was not connected at time of purchase - northtown make6 connected 10/4/96 when found cruise did not work. *ak.

[2000-01-02] MFR1 - MAKE6 - MODEL8 - 1991
A written request was submitted to the mechanic asking for an evaluation of the transmission and peeling paint, I was given an oral estimate via phone call expressing the need for a new transmission for $2000. 00. The diagnosis, cost of repair/replacement of transmission was not included on invoice, nor was the evaluation/ repair cost of peeling exterior paint. I paid $65. 00 for the estimate, but feel I got nothing in writing. I didn't receive documentation of diagnosis, or cost of repairs. *ak.

[2000-01-03] MFR1 - MAKE6 - MODEL383 - 1997
This is first make6 i've had with this track bar design on steering. Previous make6 vehicles I have owned have over 150,000 miles on original steering components without complaint using past design. The track bar design used on these new pickups is a flawed design prone to failure. Dealer did not have one in stock when truck was purchased and front end assessment at the time showed it to be loose and defective causing the steering to wander all over the road. New tires and brakes were installed to eliminate them as a facter. Original trackbar was replaced with trw



Exploration with Lens



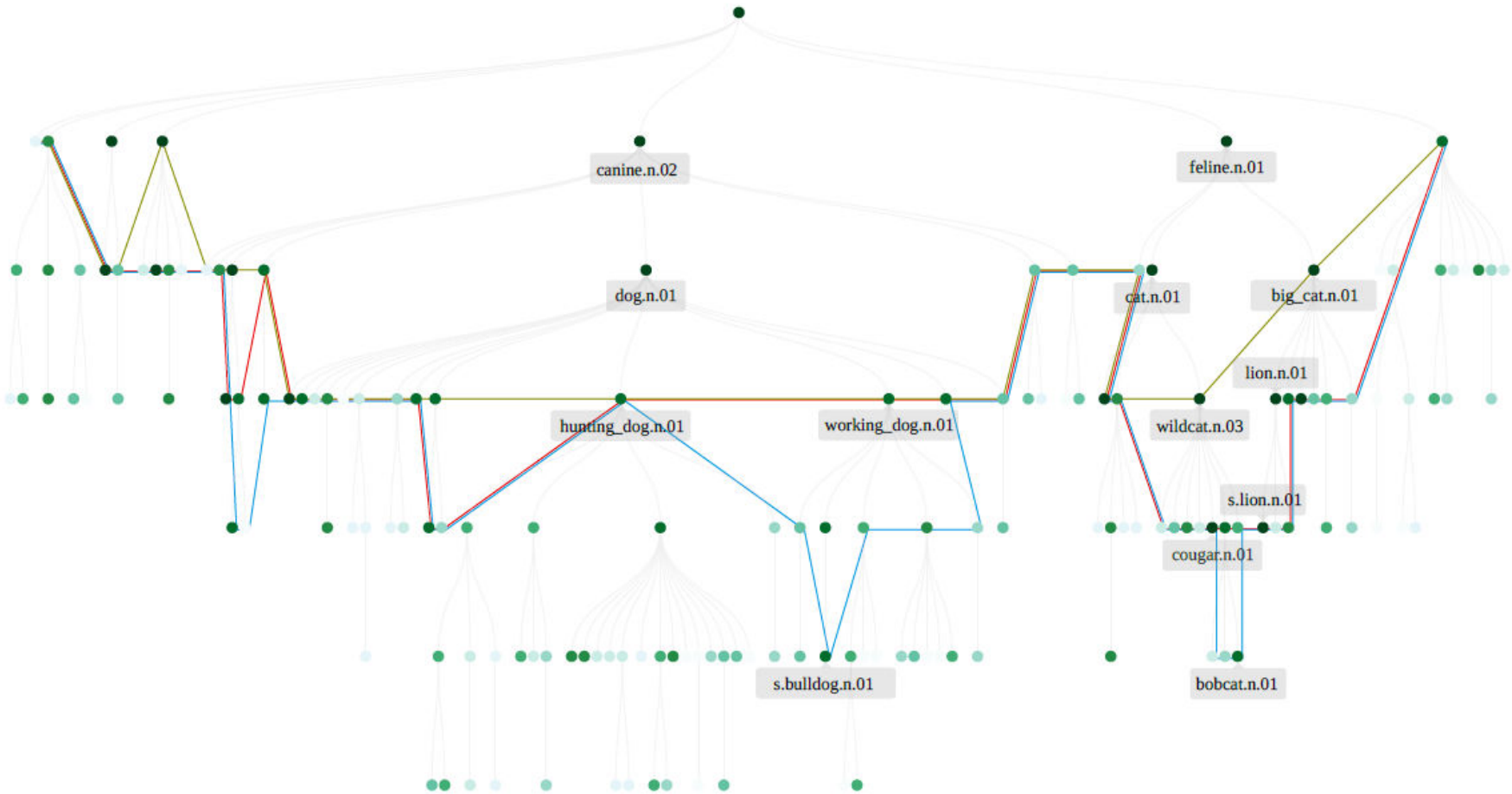
Semantic Password Analysis

- What types of words do people use in their passwords?
- Do the patterns of word use represent security vulnerabilities

R. Veras, C. Collins, and J. Thorpe, "On Semantic Patterns of Passwords and their Security Impact," *In Proceeding of the Network and Distributed System Security Symposium (NDSS'14)*, 2014.

- Extract words from passwords
- Categorize them
- Parse the results to find structure

Password	Segment	Semantic tag
hope87	hope	wish.v.01
hope87	87	number
serenity	serenity	trait.n.01
bishop5	bishop	status.n.01
bishop5	5	number
goblue0507	go	s.travel.v.01
goblue0507	blue	
goblue0507	507	number
looted	looted	take.v.21
drift21	drift	force.n.02
drift21	21	number
candysinger	candy	s.candy.n.01
candysinger	singer	musician.n.01
671soldier	671	number
671soldier	soldier	worker.n.01
bravo100	bravo	murderer.n.01
bravo100	100	number
egobrain	ego	pride.n.01
egobrain	brain	structure.n.04
pitcher9	pitcher	athlete.n.01
pitcher9	9	number
puppies	puppies	puppy.n.01
church	church	religion.n.02
‘ale‘8	‘	special
‘ale‘8	ale	alcohol.n.01
‘ale‘8	‘8	num+special


canine.n.02
feline.n.01
dog.n.01
cat.n.01
big_cat.n.01
lion.n.01
hunting_dog.n.01
working_dog.n.01
wildcat.n.03
s.lion.n.01
cougar.n.01
s.bulldog.n.01
bobcat.n.01

Results

- Place names, male names very popular
- “Cute” animals more common
- Emotional verbs like “love” are common
- Profanity is very common

WordsEye.com

SENTIMENT VISUALIZATION

Sentiment Analysis

- Business intelligence:
  - Do people like my product/restaurant/movie/hotel?
  - Why or why not?
- Forensics and medicine:
  - State of mind analysis based on social media
- Emotional profiling / psycholinguistics
  - Understanding users -> individualization
  - Targeted advertising

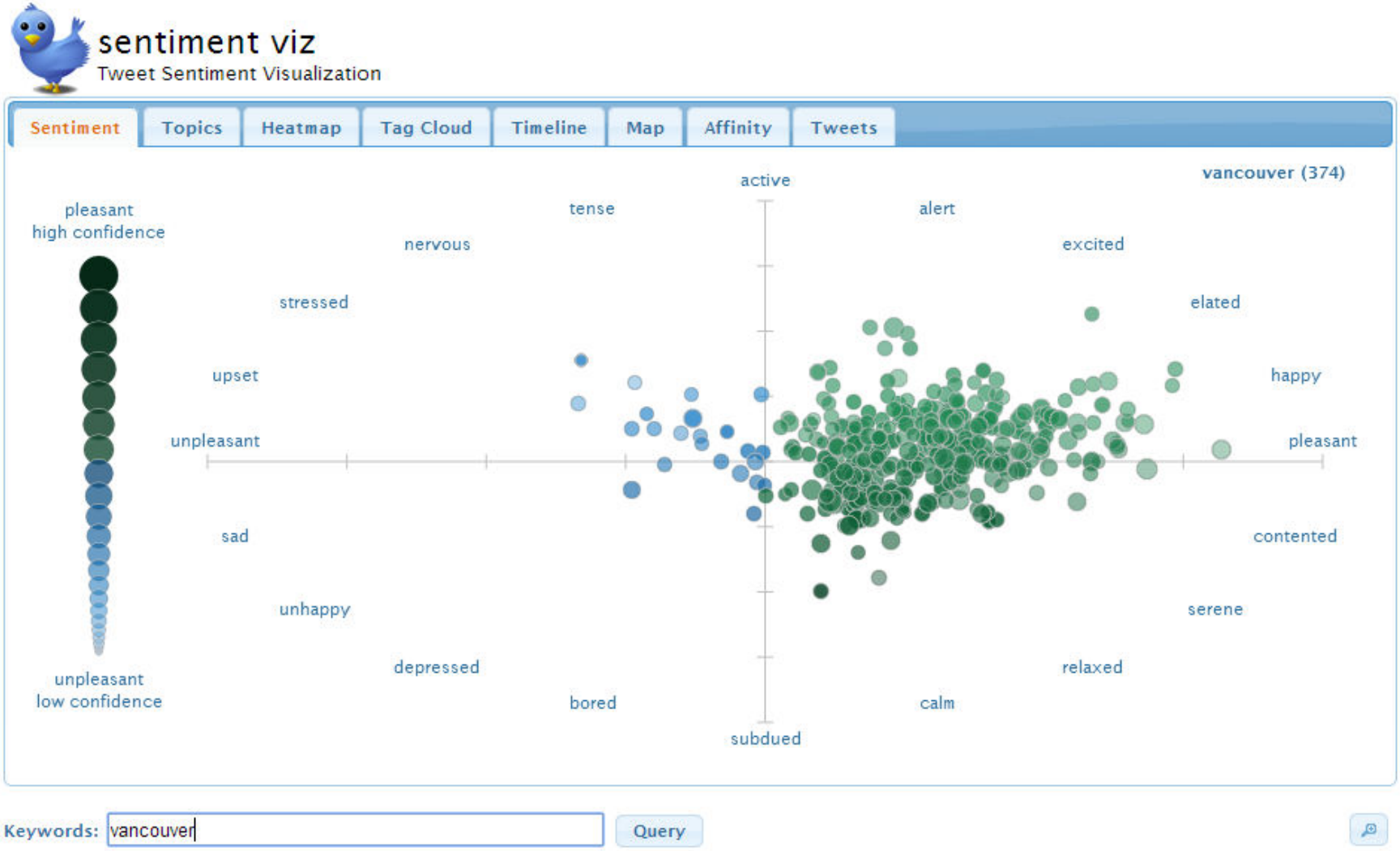
Sentiment Analysis

- Language Processing:
  - Stemming
  - POS Tagging
  - Dependency Parsing
  - Named Entity Detection
- Granularity:
  - Positive/negative
  - 8+ emotions
  - Word, sentence, paragraph, or document level

Resources and Datasets

- NRC Word-Emotion Lexicon:
  - Saif Mohammad, 2013
    http://www.saifmohammad.com/WebPages/ResearchInterests.html
- LIWC:
  - James Pennybaker et al., 2007:
    http://www.liwc.net/
- Opinion Mining Dataset:
  - Bing Liu, 2004—current
    http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

Twitter Sentiment Viz



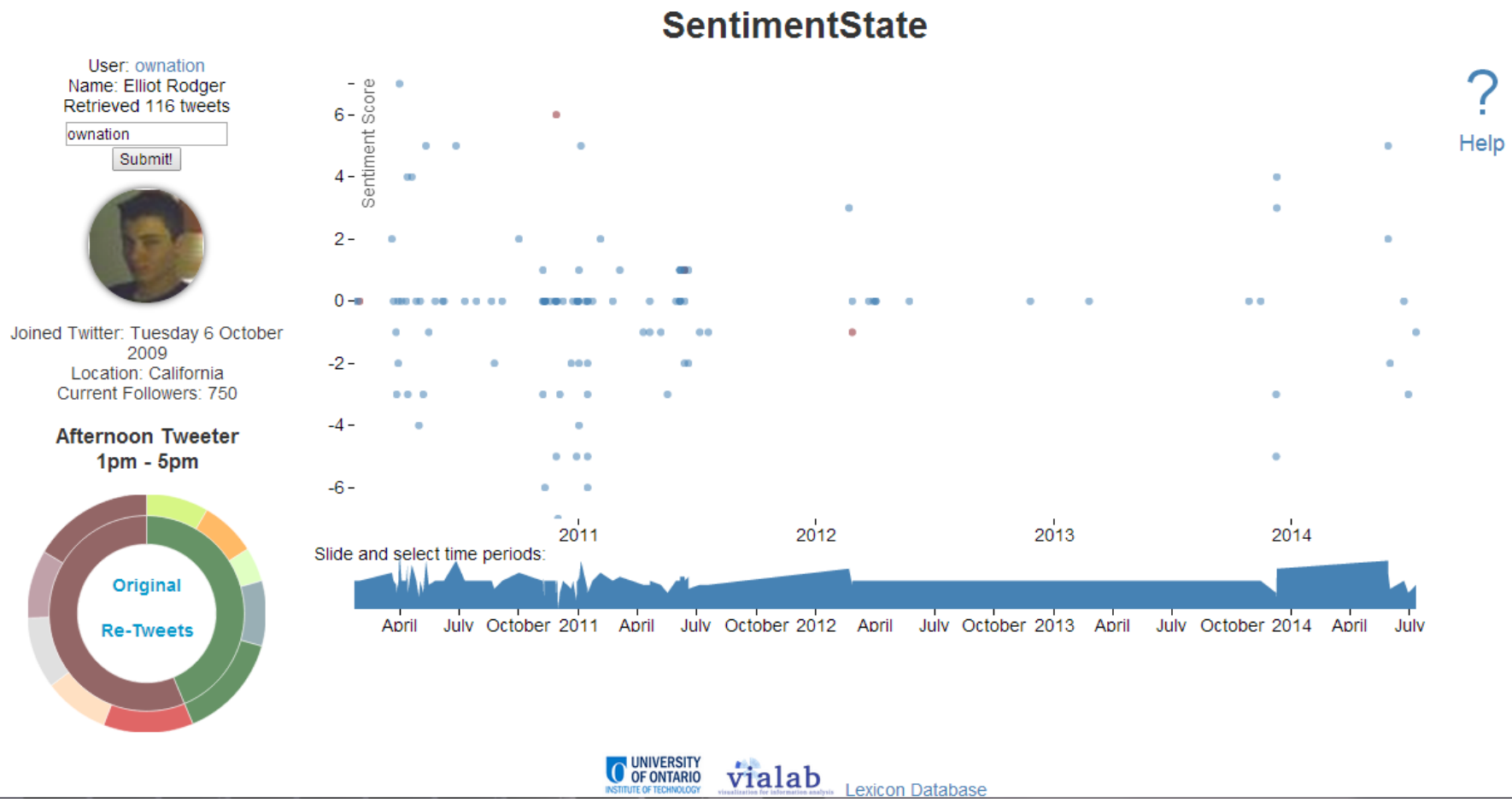
Healey and Ramaswamy, 2013. http://www.csc.ncsu.edu/faculty/healey/tweet_viz/

SentimentState

- Tweets over time, categorized using an emotion lexicon
- Examine Tweets in context, filter based on time and emotions

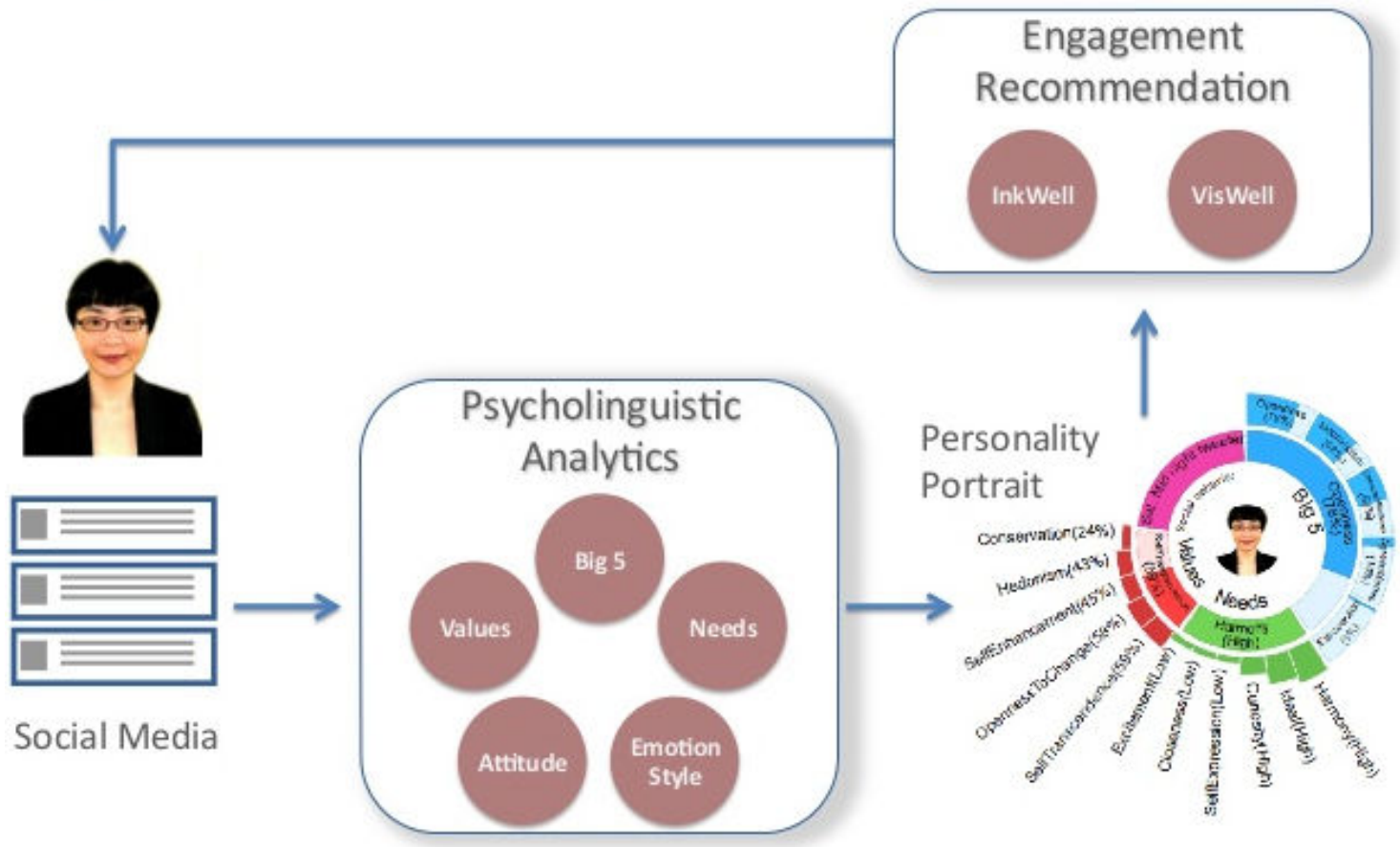
Scantlebury and Collins, 2014. http://vialab.science.uoit.ca/sentimentstate

SentimentState

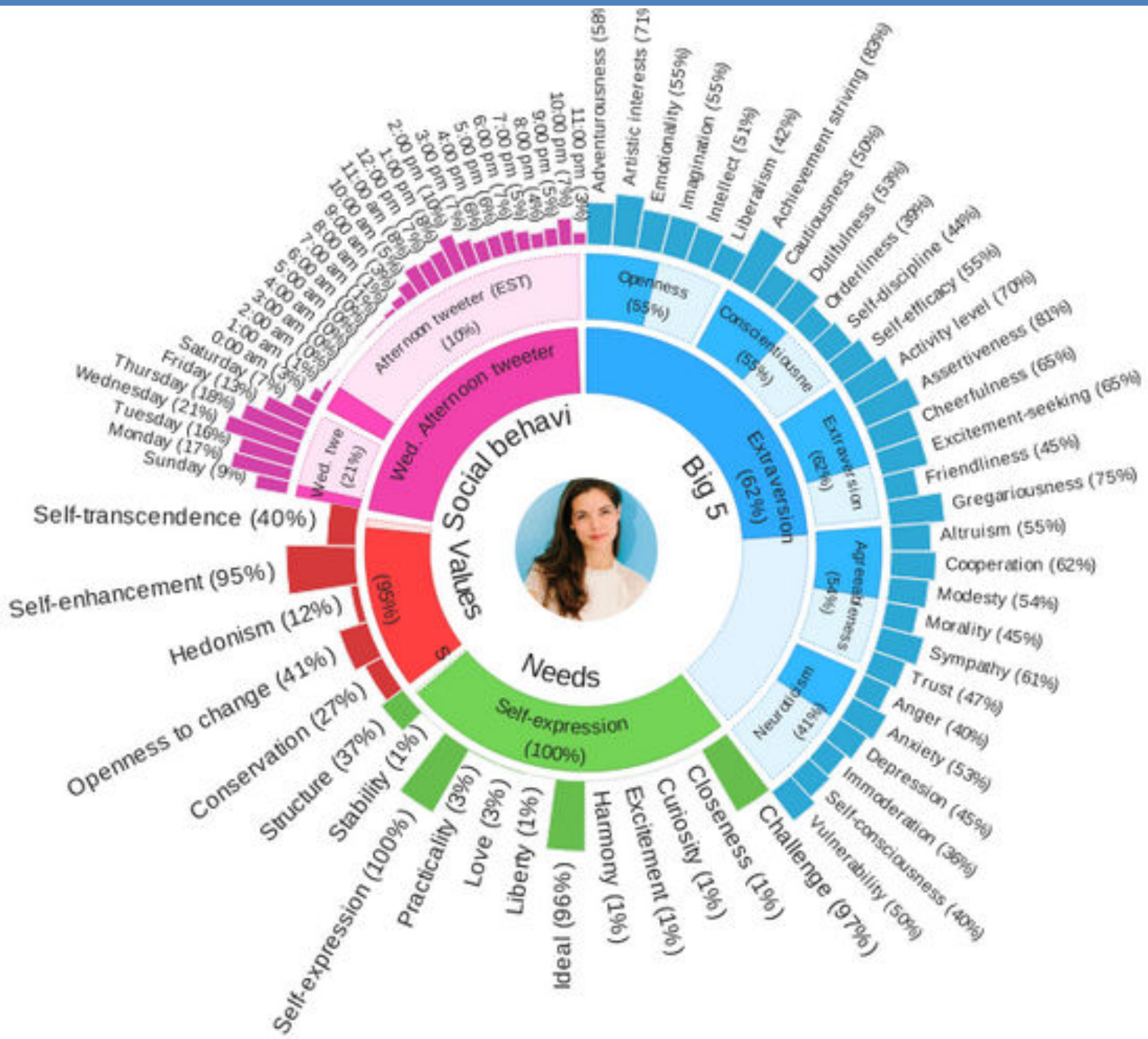


Scantlebury and Collins, 2014. http://vialab.science.uoit.ca/sentimentstate

IBM System U



Michelle Zhou, System U: Computational Discovery of Personality Traits from Social Media for Individualized Experience, 2014.



Christopher Collins, CANVAS 2014

This movie was actually neither that funny, nor super witty.

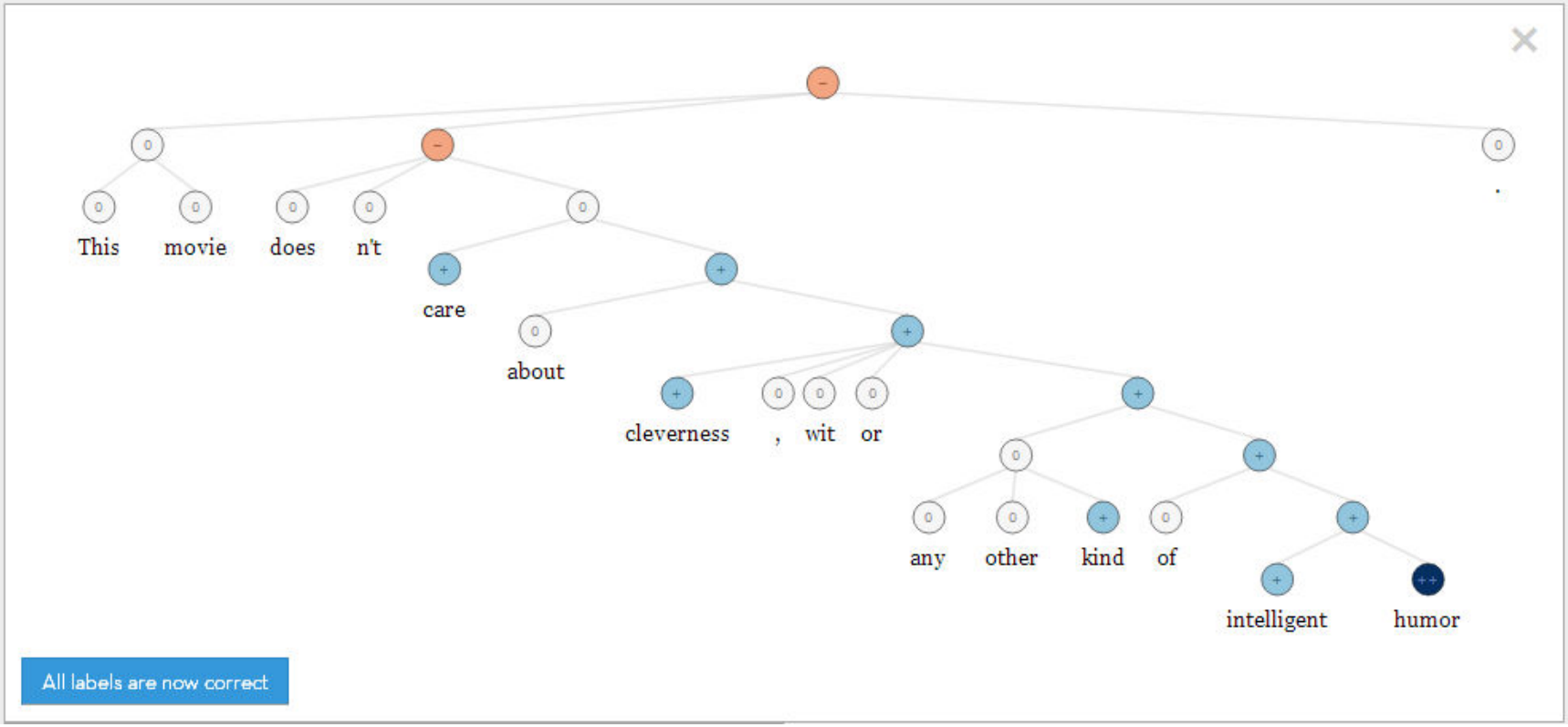
Stanford Sentiment Parser

- Recursive neural network built on top of grammatical structures
- Trained on Stanford Sentiment Treebank
  - Parse trees labelled with sentiment scores
  - Crowed-sourced and editable

Socher et al. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. Conference on Empirical Methods in Natural Language Processing (EMNLP 2013).

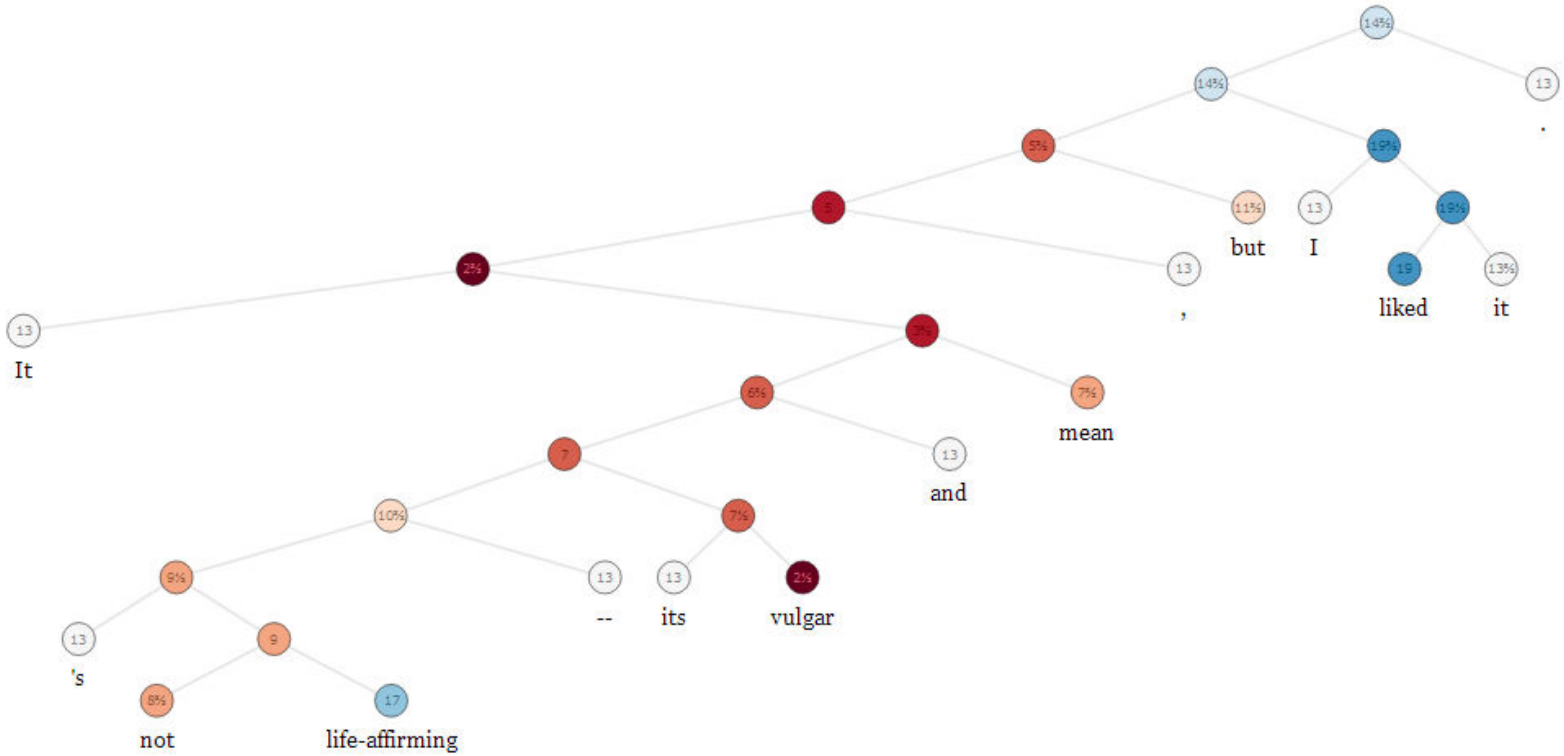
Sentiment Trees

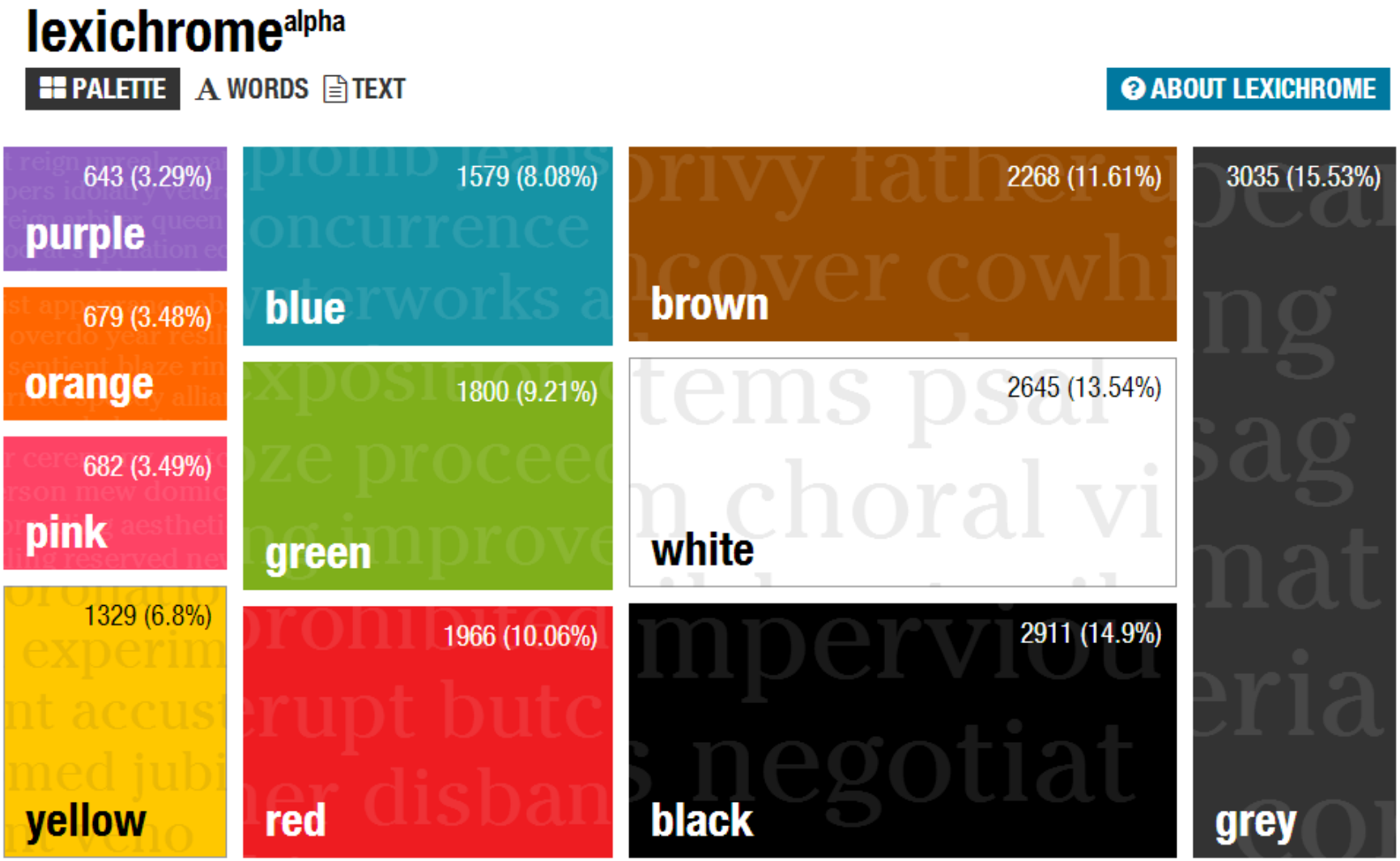
You can double-click on each tree figure to see its expanded version with greater details. There are 5 classes of sentiment classification: very negative, negative, neutral, positive, and very positive.



Parsing is Needed!

- Stanford Sentiment Treebank:
  - http://nlp.stanford.edu/sentiment/treebank.html





Kim and Collins, 2014. http://lexichrome.com

< all words associated with yellow

PALETTE | WORDS | TEXT

ABOUT LEXICHROME

RELEVANCE (DESC) | ALPHABETICAL

cowardly 10 out of 10	nugget 7 out of 7	sun 7 out of 7	sunny 9 out of 10
saffron 8 out of 9	treasure 7 out of 8	lion 6 out of 7	mustard 6 out of 7
radiant 6 out of 7	bee 11 out of 13	butter 11 out of 13	insecure 6 out of 8
sandy 6 out of 8	scatter 6 out of 8	lightning 8 out of 11	beehive 10 out of 14
practically 5 out of 7	radiate 5 out of 7	enlighten 7 out of 10	sunshine 7 out of 10

lexichrome[alpha]

PALETTE A WORDS TEXT

ABOUT LEXICHROME

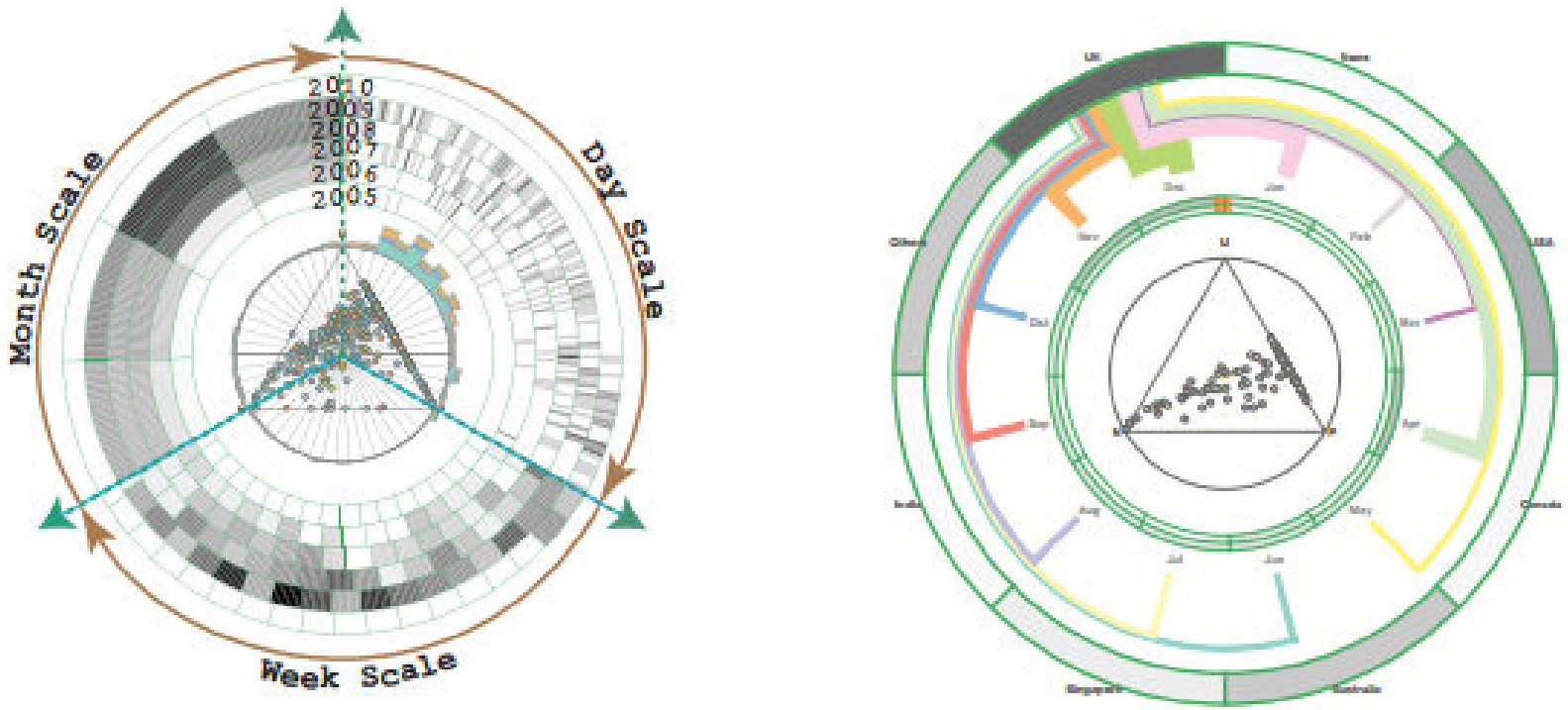
Nameless here for evermore.

And the silken sad uncertain rustling
of each purple curtain
Thrilled me - filled me with fantastic
terrors never felt before;
So that now, to still the beating of my
heart, I stood repeating
`'Tis some visitor entreating entrance
at my chamber door -
Some late visitor entreating entrance
at my chamber door; -
This it is, and nothing more,'

ANALYZE

Once upon a midnight dreary, while
I pondered weak and weary,
Over many a quaint and curious
volume of forgotten lore,
While I nodded, nearly napping,
suddenly there came a tapping,
As of some one gently rapping,
rapping at my chamber door.
`'Tis some visitor,' I muttered,
`tapping at my chamber door -
Only this, and nothing more.'

Opinion Seer



(a) (b)

Yingcai Wu et al. 2010. OpinionSeer: Interactive Visualization of Hotel Customer Feedback. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (November 2010.

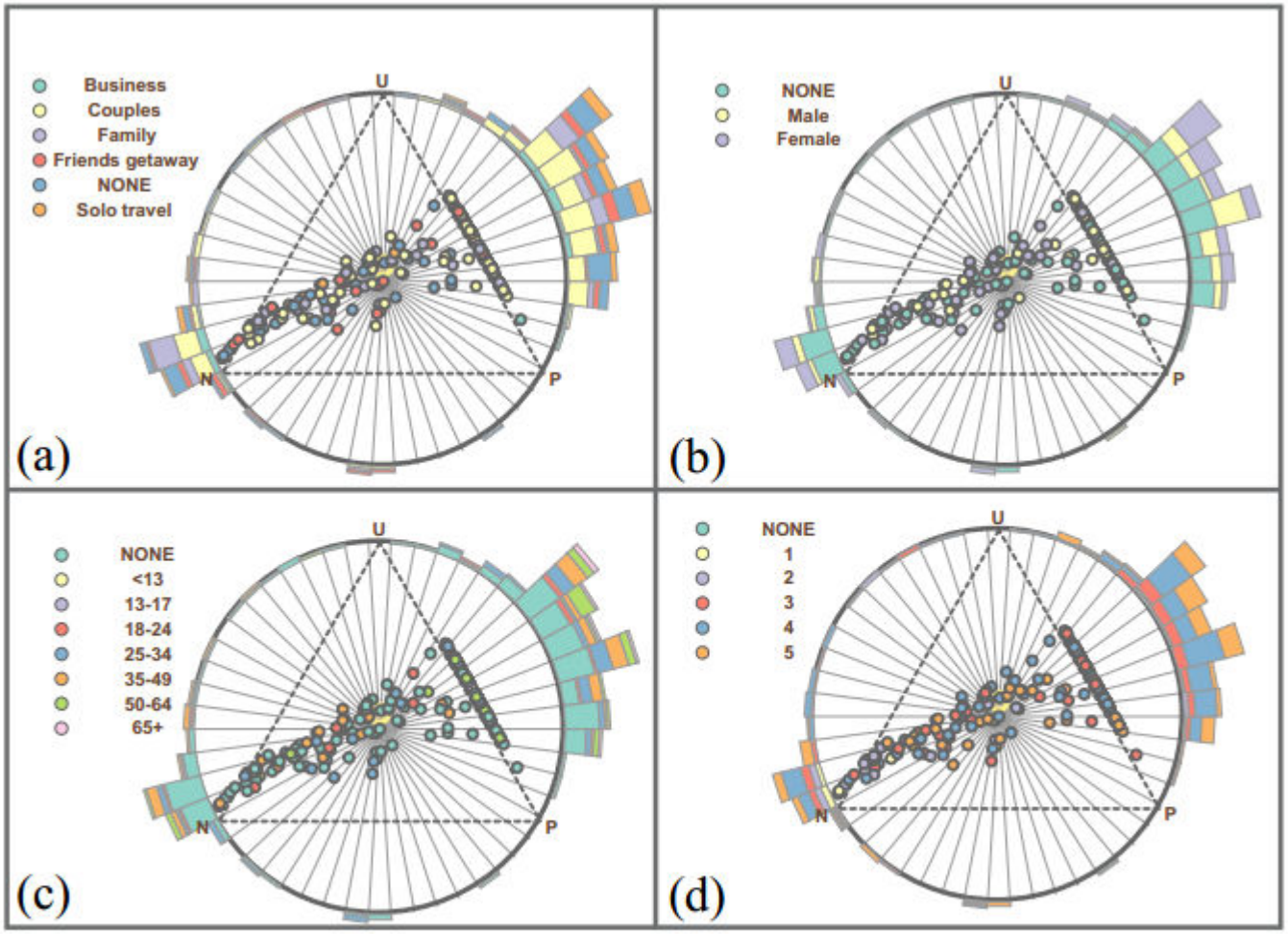


Fig. 8. OpinionSeer results showing how customer opinions are correlated with trip type, gender, age range, and ratings.

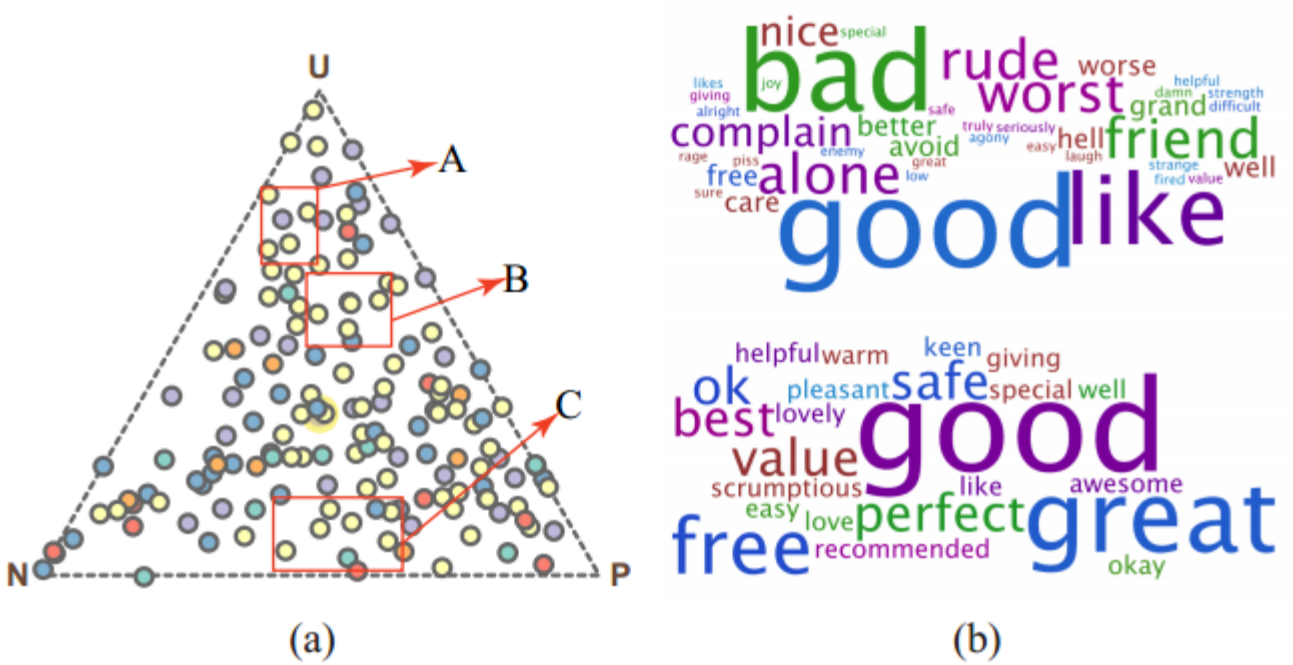


Fig. 7. (a) An opinion triangle where three regions A, B, and C are selected; (b) Top and bottom: two tag clouds of the opinion words associated with Region A and B in (a), respectively.

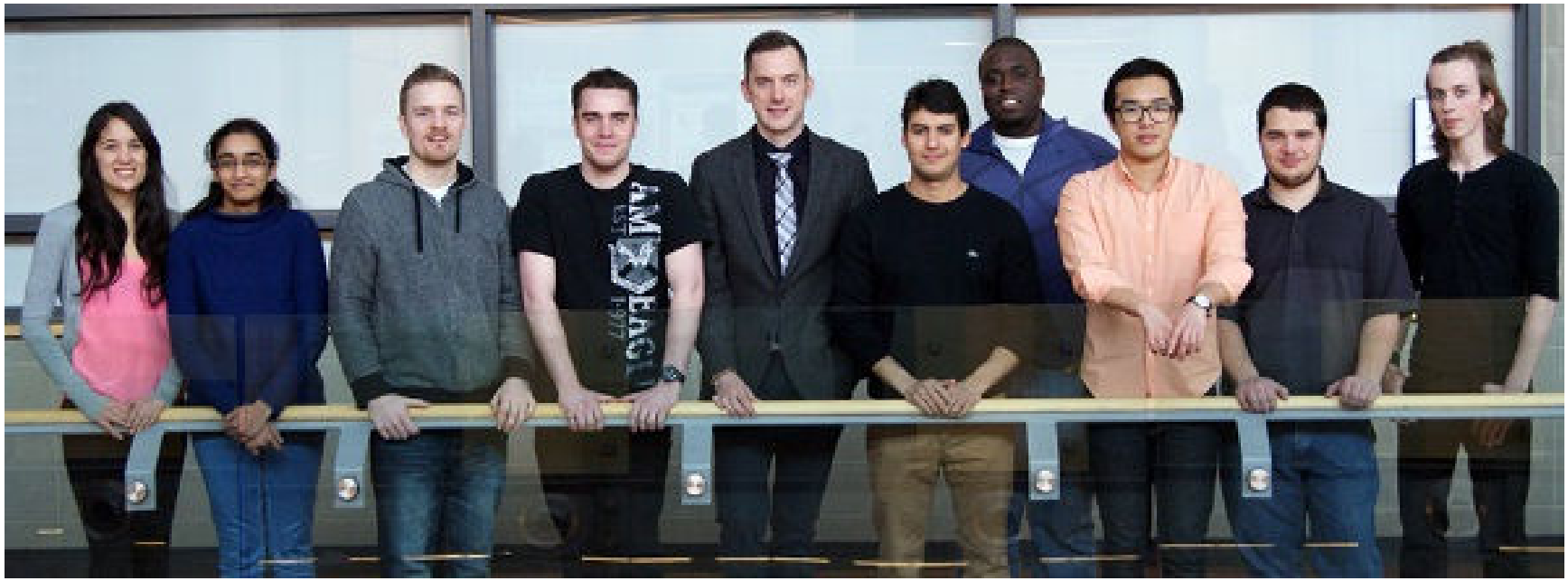
Research Ideas

- Putting it all together: NNPs + sentiment + semantic structure
- Incorporating uncertainty in sentiment analysis
- Suggesting analytic starting points – "Overview first" is just too general to be useful

Visual Text Analytics Best Practice

- Problem-driven, real questions about real data
- Generalizable techniques
- Human-understandable outputs of linguistic processing (issues of trust, transparency, usability)
- Interactive link to original text

- “Clarify, don’t simplify” – Marti Hearst



Undergrad

Zachary Cook
Taurean Scantlebury
Kalev Sikes

Grad

Rafael Veras
Erik Paluka
Brittany Kondo
Hrim Mehta
Stephen McIntyre
Daniel Chang
Chris Kim

UOIT
CHALLENGE INNOVATE CONNECT

Sheelagh Carpendale
Martin Wattenberg
Uta Hinrichs
Mark Hancock
Jeffrey Heer
Julie Thorpe
Dan McFarland
Gerald Penn
Fernanda ViÉgas
Petra Isenberg
Fanny Chevalier
Jian Zhao
Jeremy Bradbury
Marian Dörk

NSERC CRSNG
Federal Economic Development Agency for Southern Ontario
Canada Foundation for Innovation
Fondation canadienne pour l'innovation
Graphisme, animation et nouveaux médias
grand
Graphics, Animation and New Media
SHARCNET™
SurfNet