ConceptVector: Text Visual Analytics via Interactive Lexicon Building using Word Embedding

Deokgun Park[*], Seungyeon Kim[$], Jurim Lee[&], Jaegul Choo[&]

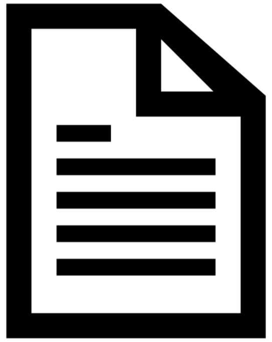
Nicholas Diakopoulos[*], and Niklas Elmqvist[*]

*University of Maryland
USA
{intuinno, nad, elm}@umd.edu

$Google
USA
sylund@gmail.com

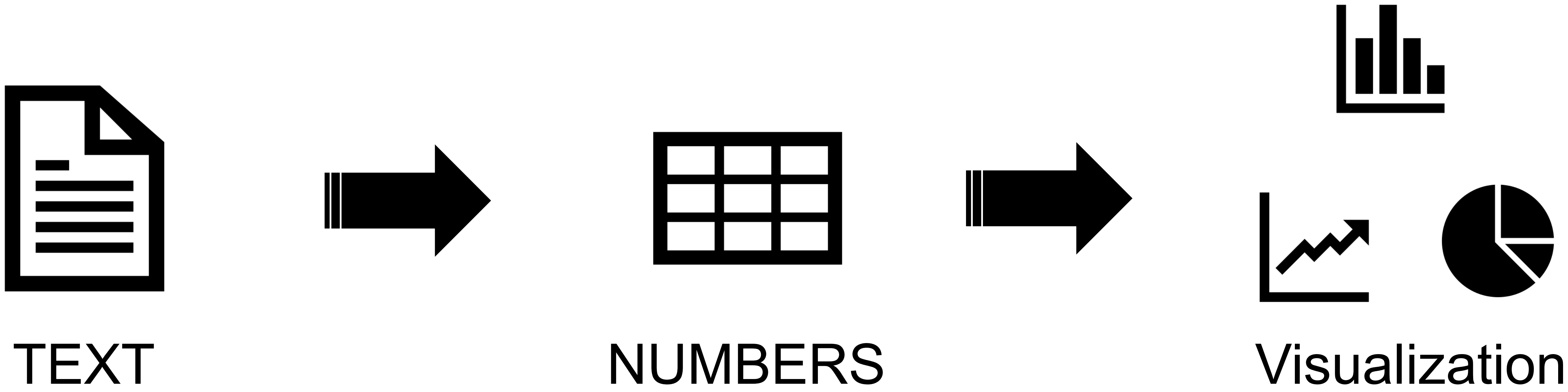
&Korea University
South Korea
{jurim0301, jchoo}@korea.ac.kr

TEXT



TEXT

Visualization


TEXT
NUMBERS
Visualization

Lexicon-based Text Mining

TEXT

NUMBERS

TEXT

Lexicon-based Text Mining

NUMBERS

Lexicon = A set of keywords that is related to specific concept

TEXT

Lexicon-based Text Mining

NUMBERS

Lexicon = A set of keywords that is related to specific concept

Positive = { good, great, happy, … }

Negative = {bad, worst, horrible, ... }

TEXT

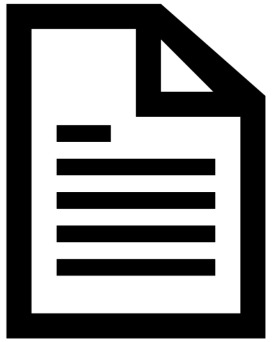
Lexicon-based Text Mining

NUMBERS

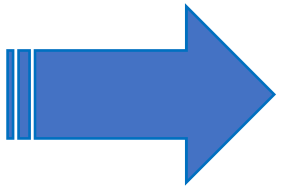
Lexicon = A set of keywords that is related to specific concept

Positive = { good, great, happy, … }

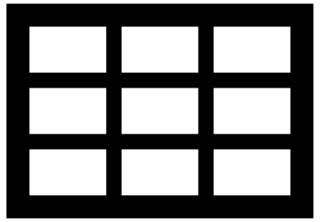
Negative = {bad, worst, horrible, ... }



TEXT



Lexicon-based Text Mining



NUMBERS

Positive = { good, great, happy, … }

Negative = {bad, worst, horrible, ... }

TEXT

Lexicon-based Text Mining

NUMBERS

Positive = { good, great, happy, … }

Negative = {bad, worst, horrible, ... }

D1 *"The movie was **<u>great</u>**. I was **<u>happy</u>**."*

D2 *"The movie opens today."*

D3 *"The movie is **<u>horrible</u>**."*

TEXT

Lexicon-based Text Mining

NUMBERS

Positive = { good, great, happy, … }

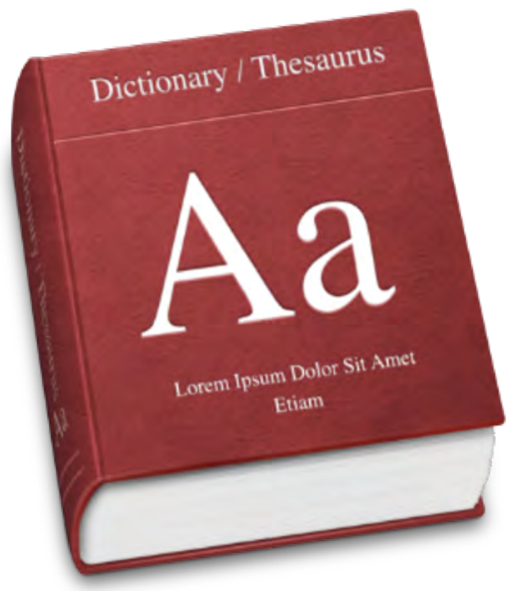
Negative = {bad, worst, horrible, ... }

D1 *“The movie was **great**. I was **happy**.”*

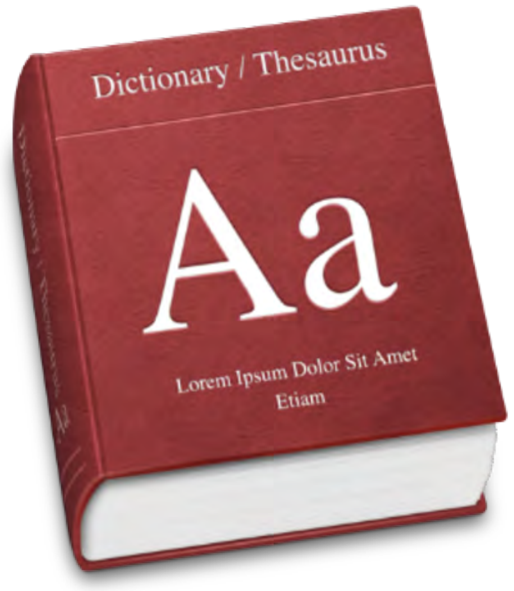
D2 *“The movie opens today.”*

D3 *“The movie is **horrible**.”*

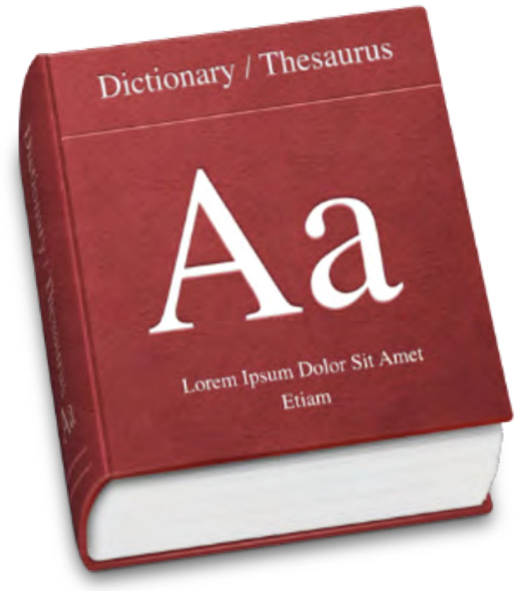
	Positive	Negative
D1	2	0
D2	0	0
D3	0	1


Dictionary / Thesaurus
Aa
Lorem Ipsum Dolor Sit Amet
Etiam

Linguistic Inquiry and Word Count (LIWC)



83 Power				84 Reward	85 Risk	90 FocusPast		
a-list*	forbids	obeyed	under	access*	abstain*	accepted	mapped	we've
above	force*	obeying	underclass*	accrue*	alarm*	added	mastered	weakened
acclaimed	frail*	obeys	underdog*	accumul*	apprehens*	admitted	mated	weighed
administr*	freshm*	officehold*	underling*	achievable	apprenhens*	affected	meant	weirded
age	glori*	officer*	underprivileg*	achieve*	averse	ago	messaged	went
allow*	glory	official*	unimportant	achievi*	aversi*	already	met	wept
amateur*	god	oppose*	unqualified	acquir*	avert*	appeared	might've	were
ambition	god's	oppositi*	unwanted	add	avoid*	approached	mightve	weren't
ambitions	goddess*	order	up	added	bad	arrived	missed	werent
ambitious	good-for-nothing	orders	upper	adding	balk	asked	mocked	weve
apolog*	govern*	outcast	upperclass*	adds	beware	ate	mothered	what'd
apprentic*	greatest	outrank*	useless	advanc*	careful*	attended	moved	whatd
approv*	greatness	over	vanquish*	advantag*	caution*	attracted	must've	where'd
armies	greed*	overpower*	verb	adventur*	cautious*	became	mustve	wished
army	grown-up*	overrul*	veteran*	amass*	cease*	been	named	wobbled
asham*	grownup*	owner*	victim*	approach	concern	began	narrowed	woke
assault*	help	pariah*	victor*	approached	consequen*	begged	neared	woken
assertive	helper*	parliament*	vip	approaches	crises	believed	needed	won
attack*	helpless*	passiv*	vp*	approaching	crisis	born	noticed	wondered
attendant	hierarch*	pathetic	vulnerab*	award*	curb*	bought	obeyed	wore
authorit*	high	pathetically	war	benefit	danger	bounced	obtained	worked
authorize	high-ranking	patriarch*	warfare*	benefits	dangerous	braved	od'ed	worn
battl*	higher	peasant*	warred	best	dangerously	broke	okayed	worsen
beat	highest	peon*	warring	bet	dangers	brought	organized	would've



How to Build a Lexicon?

How to build Lexicon

Manual Automatic

Hand-picking

- Linguistic Inquiry and Word Count (LIWC)
- General Inquirer (GI)

- High Quality, strong signal words
- Hard to build, scale
- Does not adapt to different domain (e.g. Twitter)

Manual ⟺ Automatic

Crowdsourcing

- Affective Norms for English Words (ANEW)
- Hedonometer

- Scales up for single category
- Costly, limited category

Manual

Automatic

How to build Lexicon

- Domain adaptable
- Scales to diverse topic
- Difficult to label each group

Topic Modeling

- Latent Semantic Indexing (LSA)
- Latent Dirichlet Allocation (LDA)

Manual ⟷ Automatic

How to build Lexicon

- Built upon **word embedding**
- Scales to diverse topic
- Easy to interpret

Mixed-initiative

- Empath

Manual ⟷ Automatic

word embedding

word embedding

Frog

0.34	.7	0.67	0.9	...

Car

0.53	0.23	0.35	0.21	...

Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.

Goldberg, Yoav, and Omer Levy. "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method." *arXiv preprint arXiv:1402.3722* (2014).

word embedding

Nearest Neighbor

Frog

1. Frogs
2. Toad
3. Litoria
4. Leptodacitylidae
5. Rana
6. Lizard
7. eleutherodactylus




word embedding

Nearest Neighbor

Frog

1. Frogs
2. Toad
3. Litoria
4. Leptodacitylidae
5. Rana
6. Lizard
7. eleutherodactylus

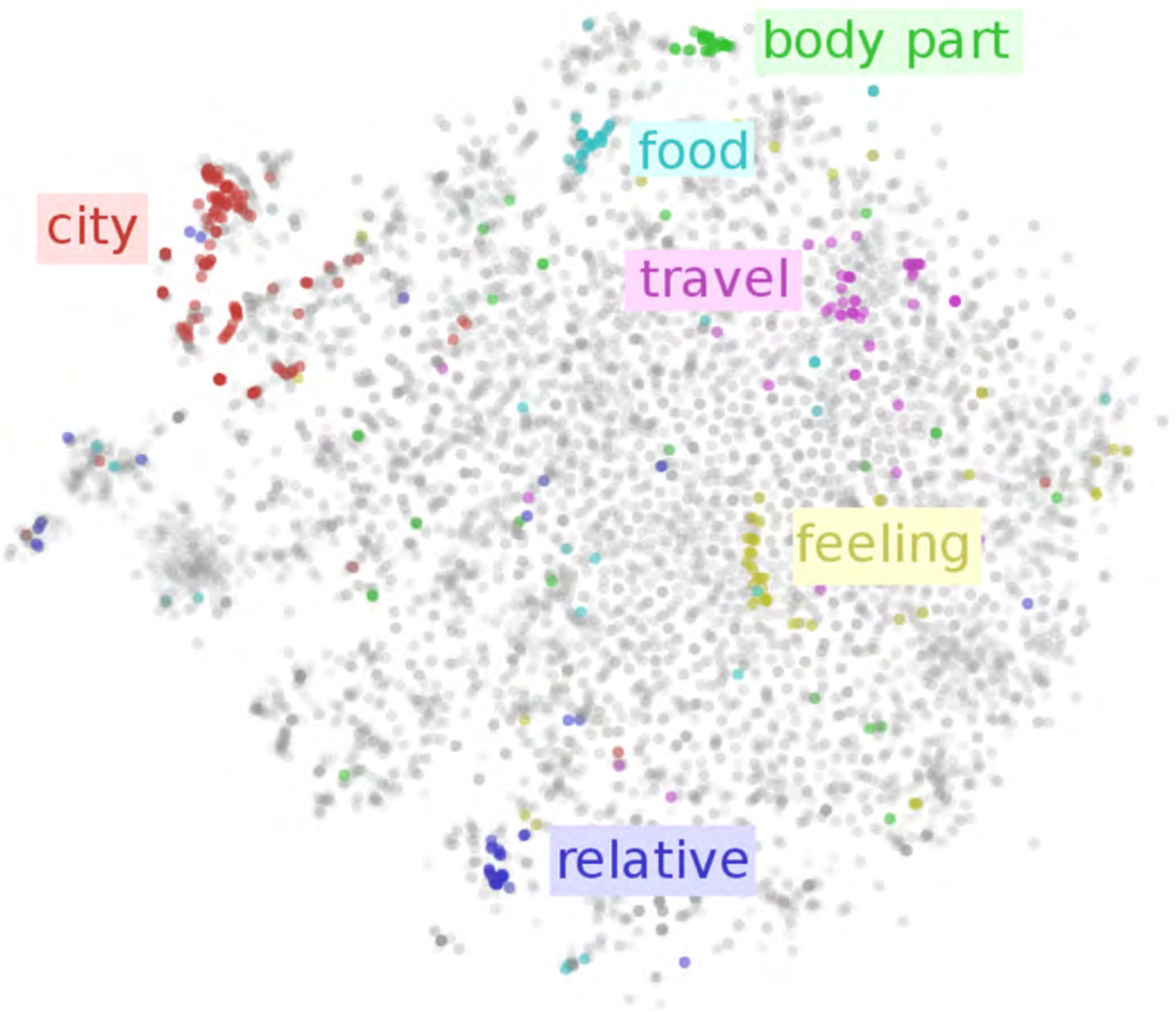
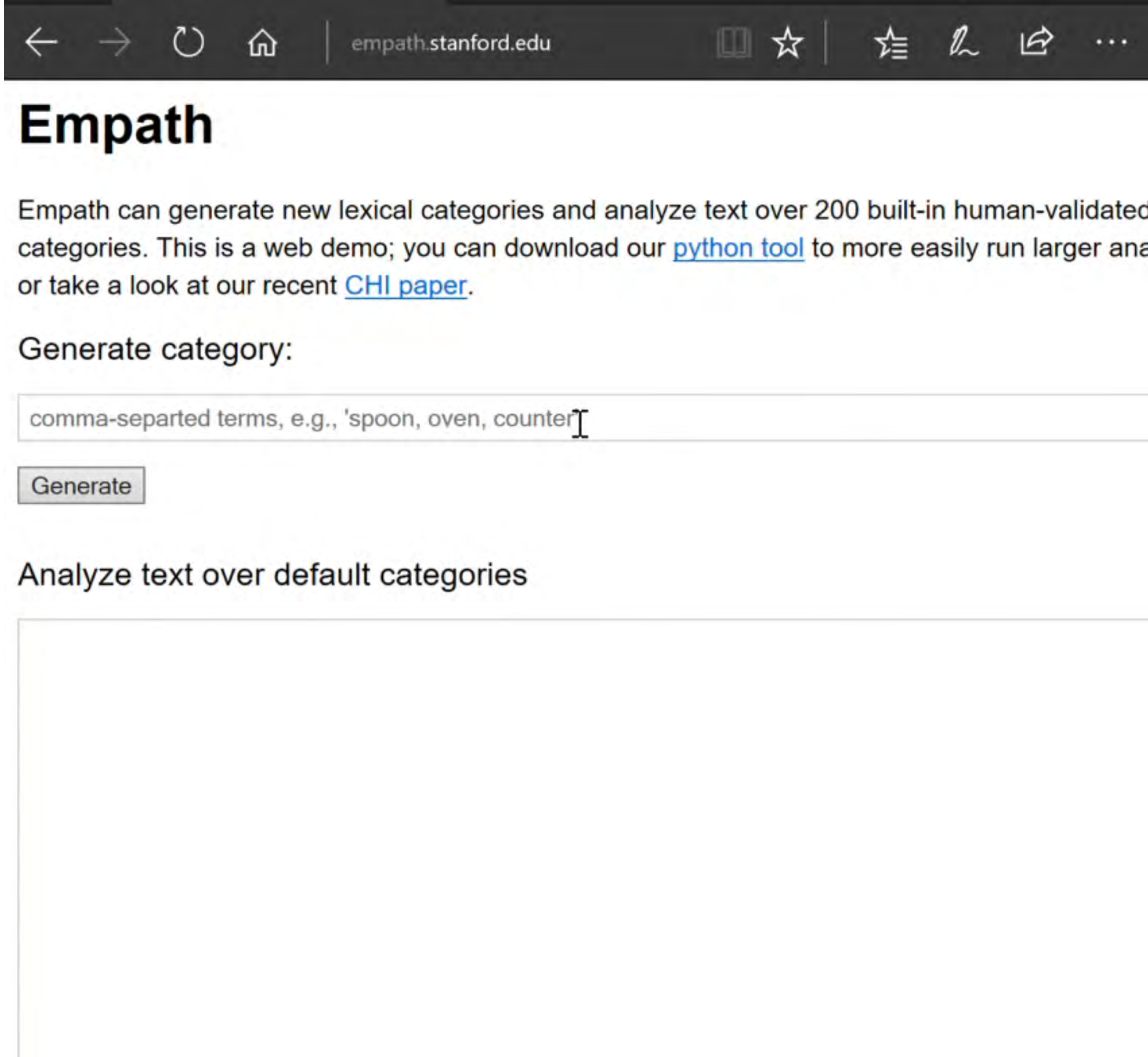
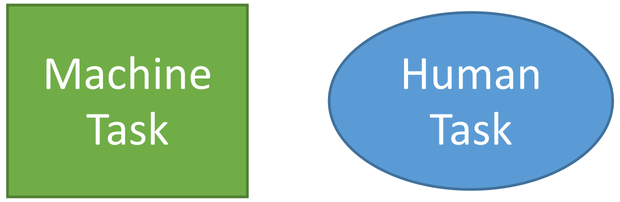


Image from http://cs.stanford.edu/people/karpathy/tsnejs/wordvecs.html

Previous Work



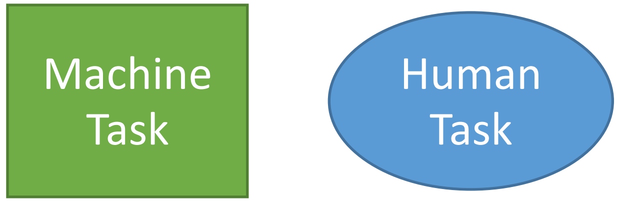
Fast, E., Chen, B., & Bernstein, M. S. (2016, May). Empath: Understanding topic signals in large-scale text. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 4647-4657). ACM.

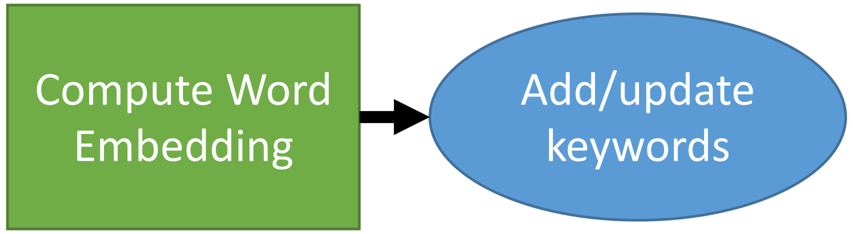

Machine Task
Human Task

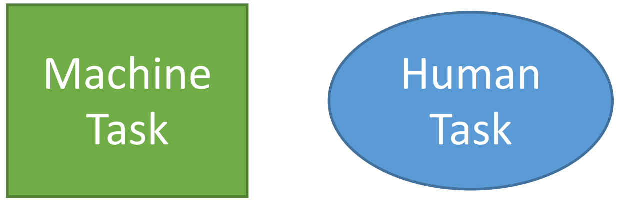
Machine Task

Human Task

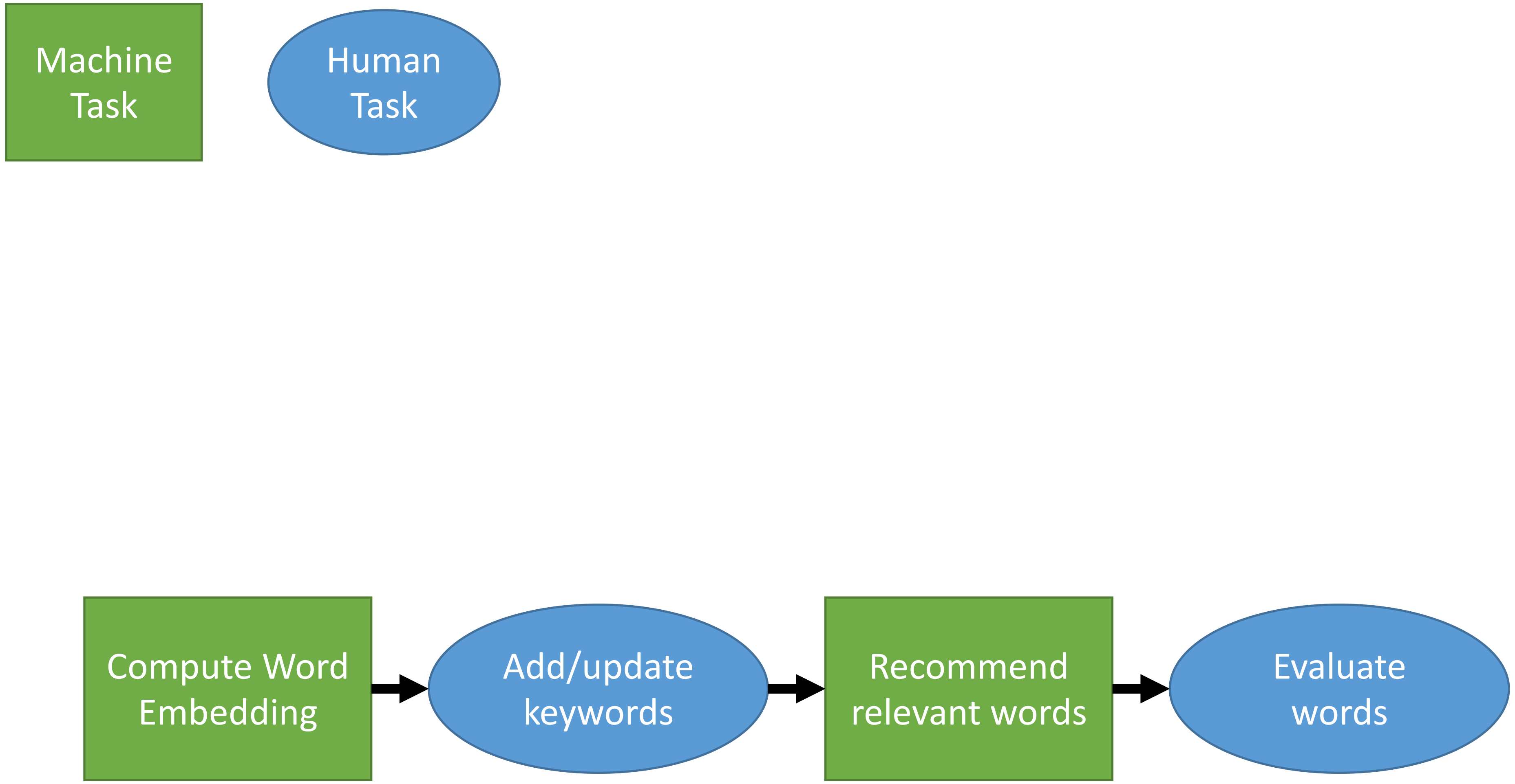
Compute Word Embedding

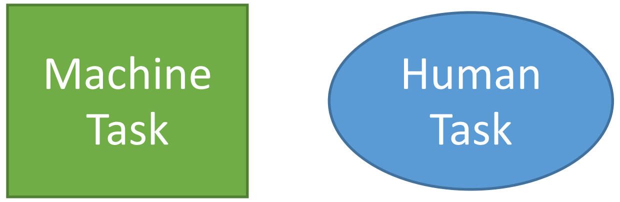

Machine Task
Human Task

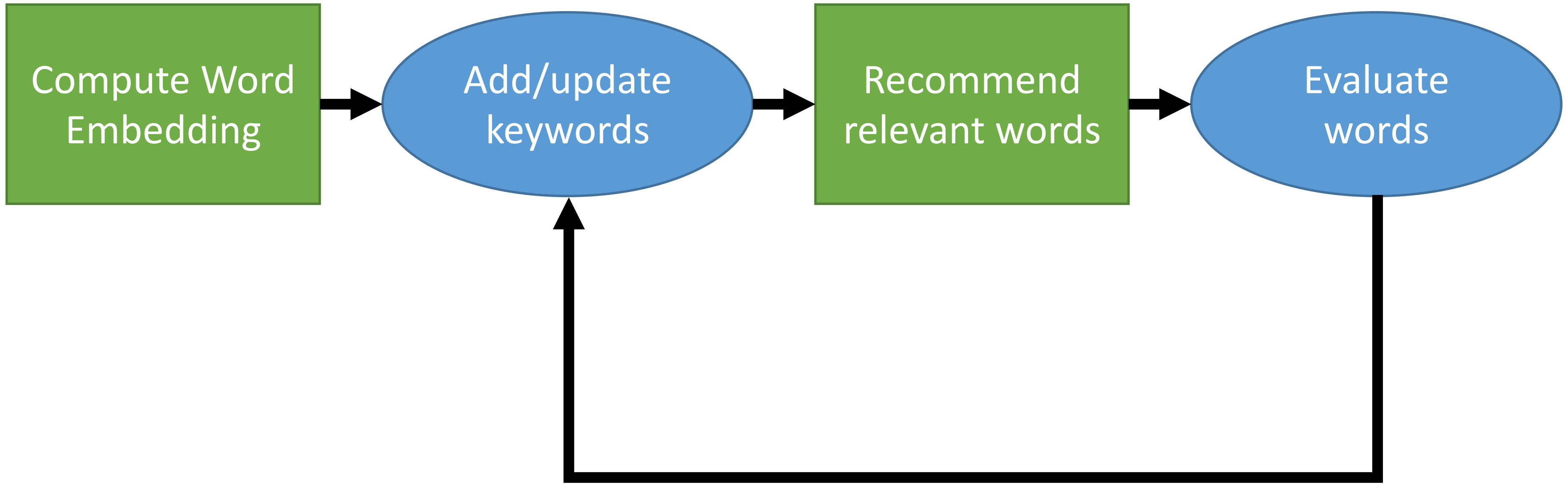

Compute Word Embedding
Add/update keywords


Machine Task
Human Task

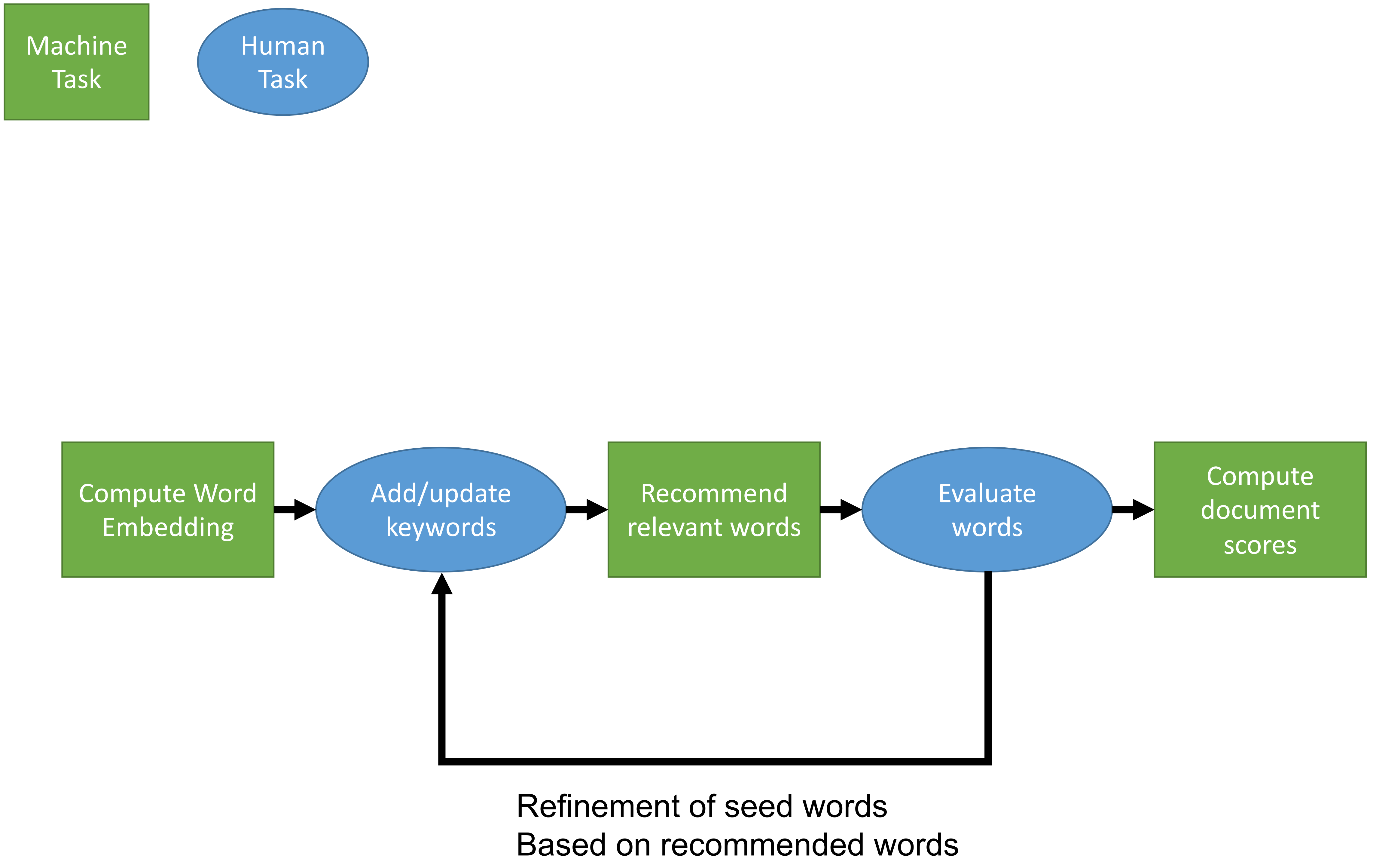

Compute Word Embedding
Add/update keywords
Recommend relevant words

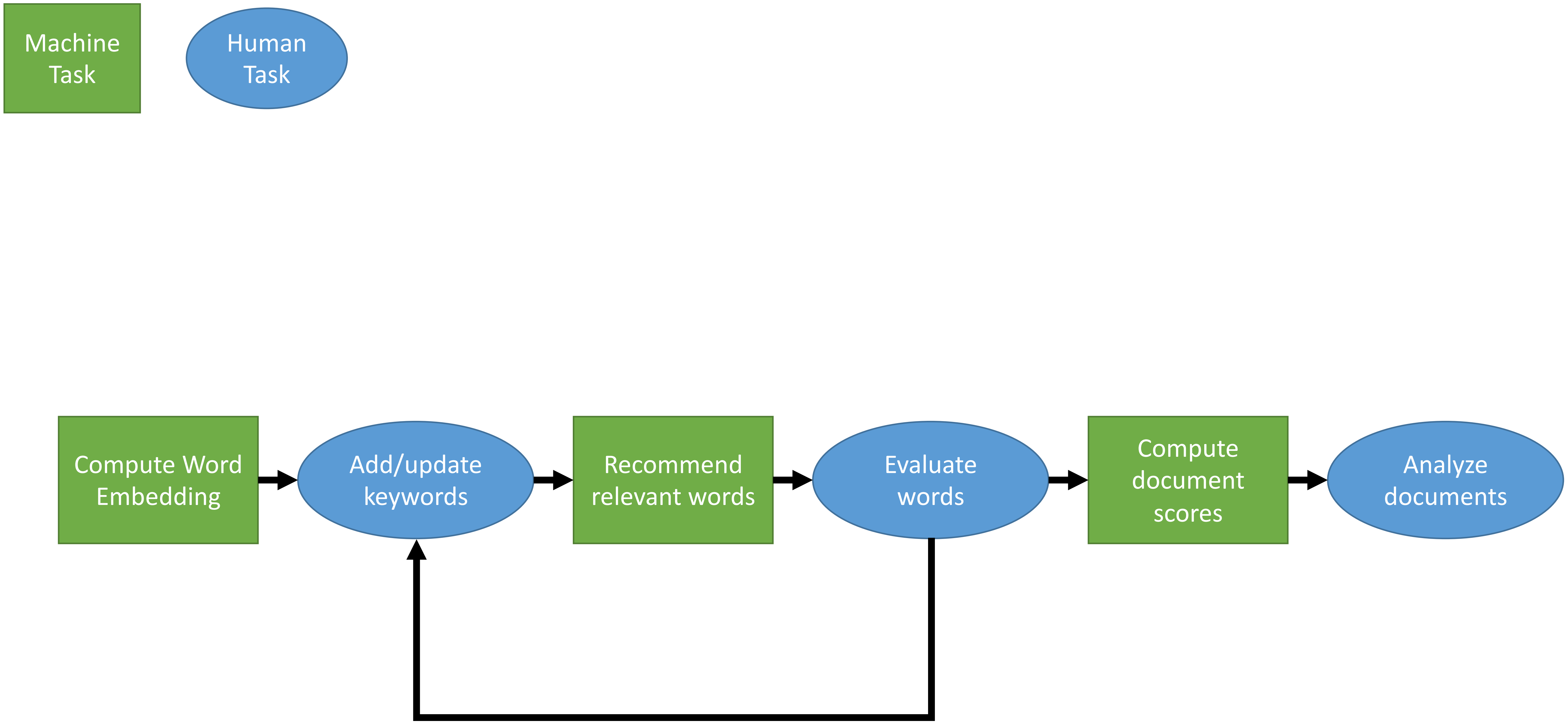

Machine Task
Human Task
Compute Word Embedding
Add/update keywords
Recommend relevant words
Evaluate words



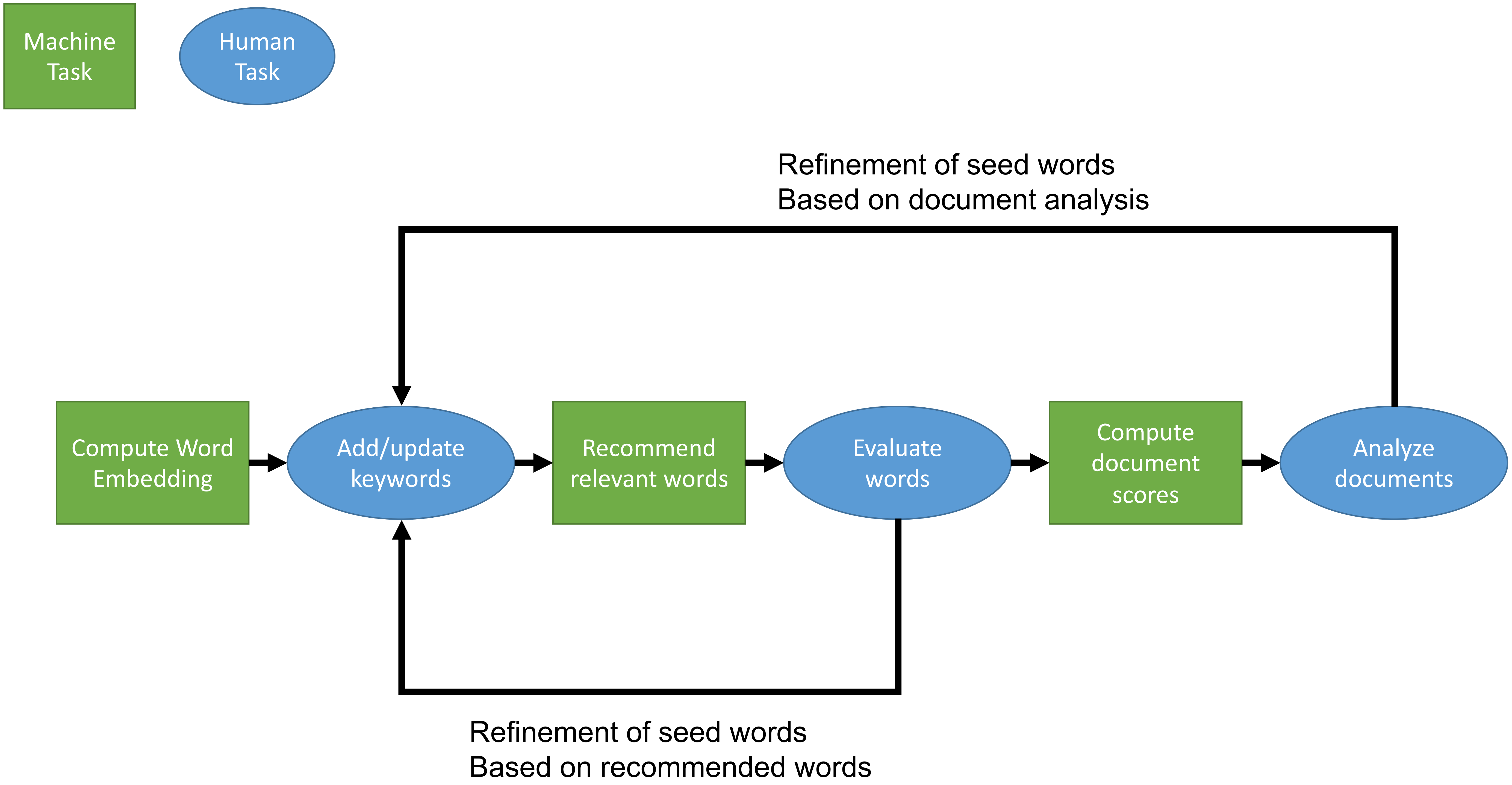


Refinement of seed words
Based on recommended words


Machine Task
Human Task
Compute Word Embedding
Add/update keywords
Recommend relevant words
Evaluate words
Compute document scores
Refinement of seed words
Based on recommended words



Refinement of seed words
Based on recommended words


Machine Task
Human Task
Refinement of seed words
Based on document analysis
Compute Word Embedding
Add/update keywords
Recommend relevant words
Evaluate words
Compute document scores
Analyze documents
Refinement of seed words
Based on recommended words

X 3,000

Tweets from Trump and Hillary during 2016 election

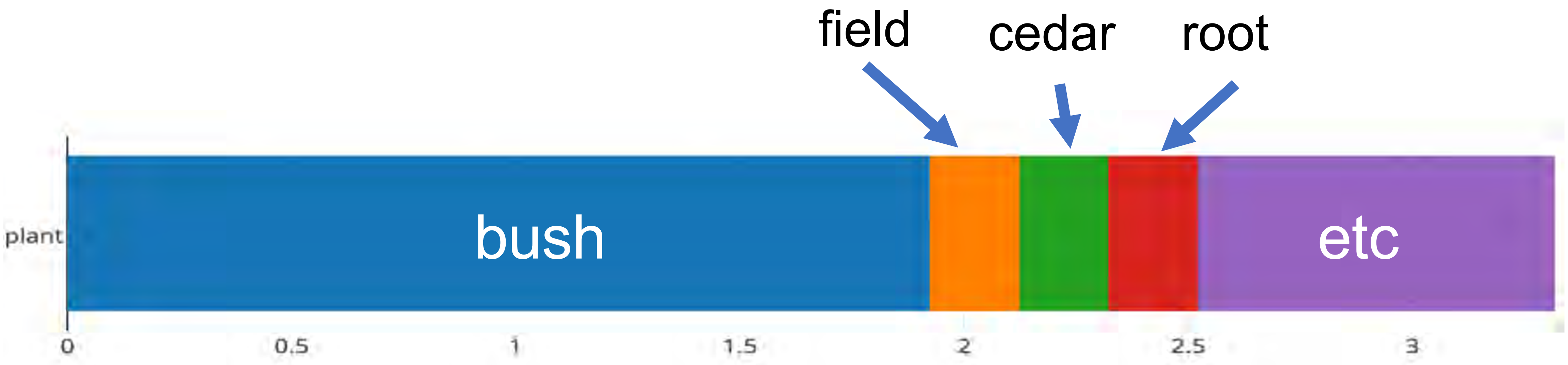
Plant Category

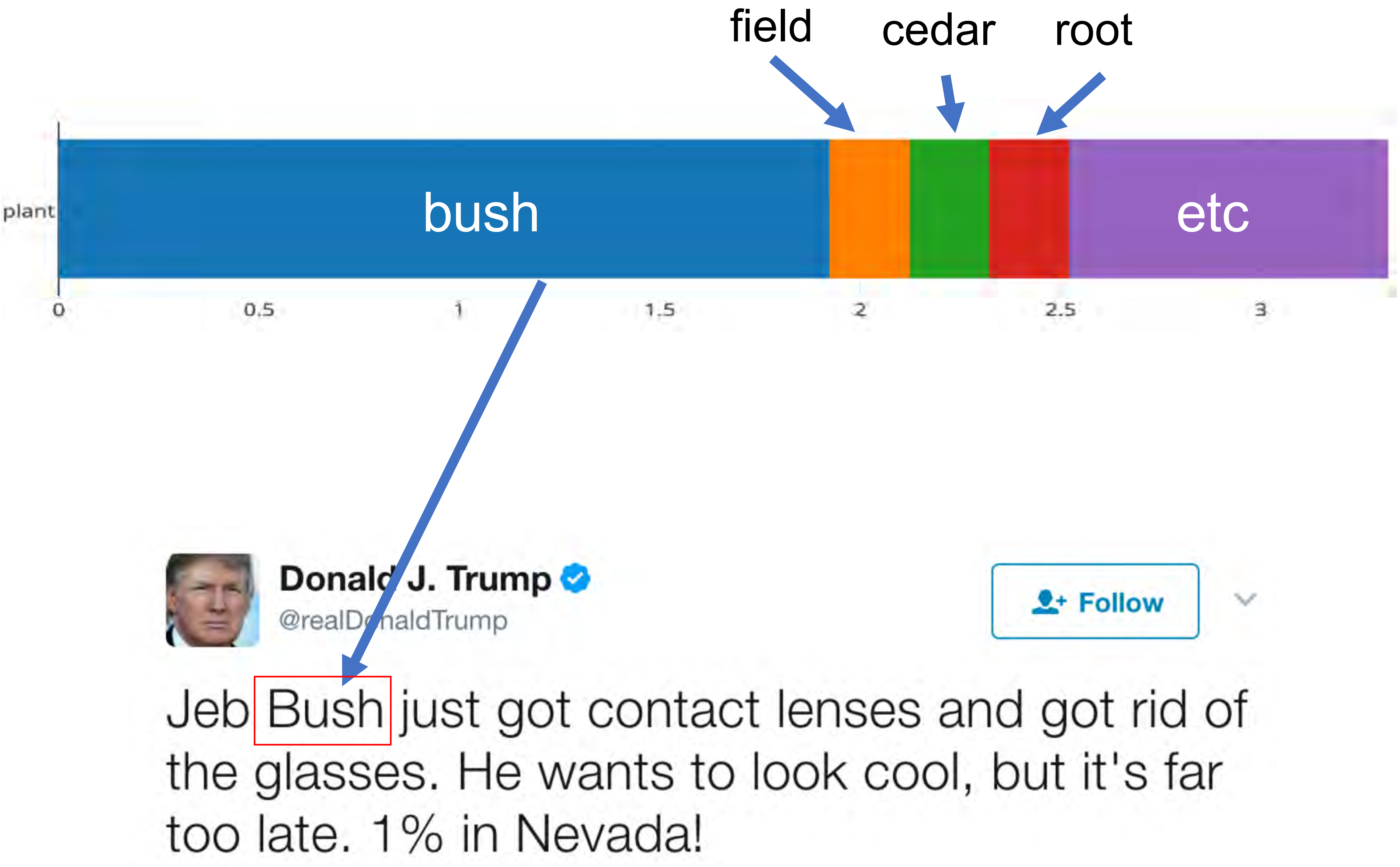
{ pollen, cherry, underbrush, thicket, grape, sunflower, field, willow, rose, fenced, bouquet, flowered … bush … }

Plant Category

{ pollen, cherry, underbrush, thicket, grape, sunflower, field, willow, rose, fenced, bouquet, flowered … bush … }

Trump talked 3.31 times more about Plant than Hillary


field
cedar
root
plant
bush
etc
0
0.5
1
1.5
2
2.5
3


field
cedar
root
plant
bush
etc
0
0.5
1
1.5
2
2.5
3
Donald J. Trump
@realDonaldTrump
Follow
Jeb Bush just got contact lenses and got rid of the glasses. He wants to look cool, but it's far too late. 1% in Nevada!

~~Trump talked 3.31 times more about Plant than Hillary~~

False positive errors

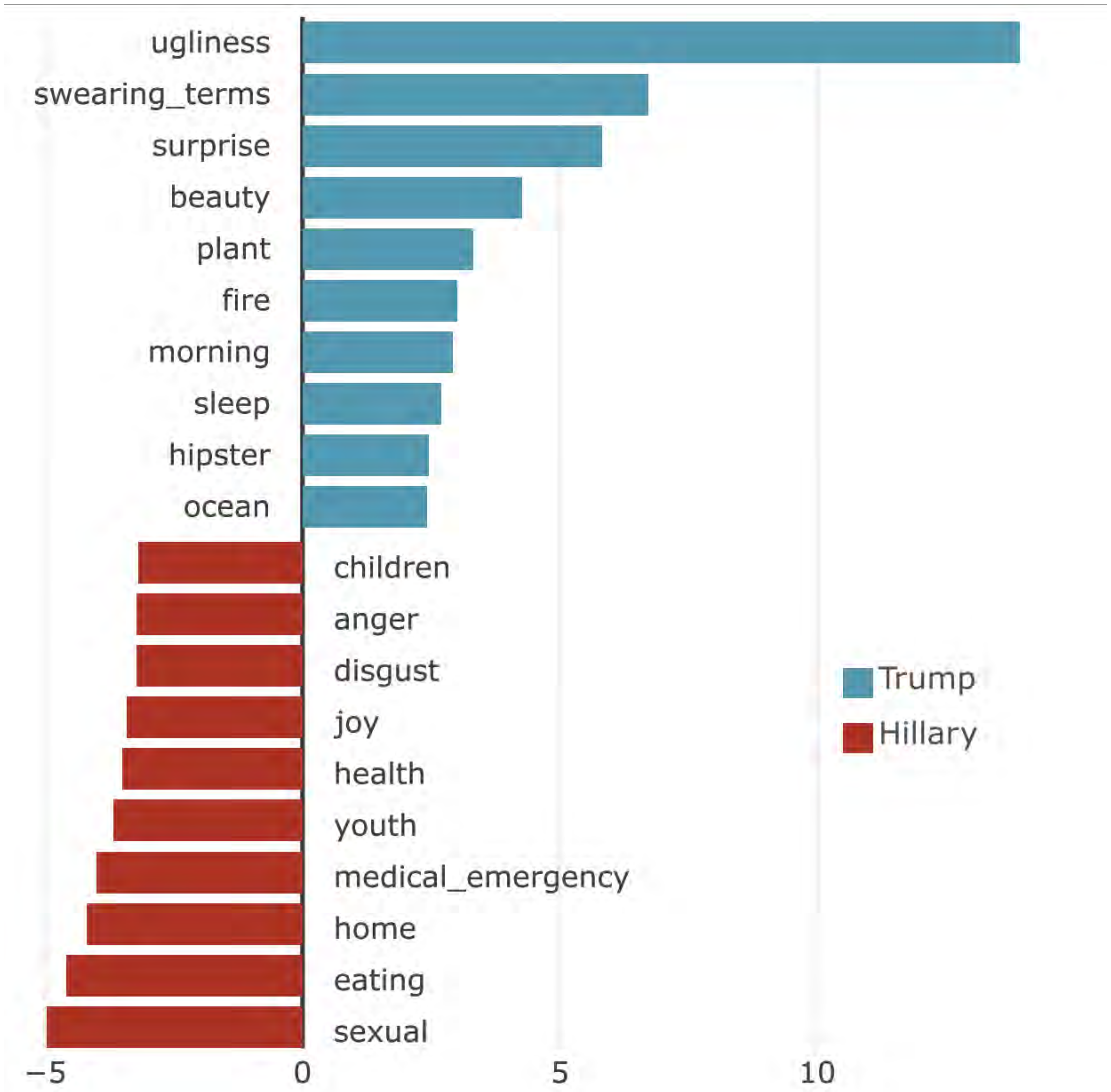
~~Trump talked 3.31 times more about Plant than Hillary~~

False positive errors

Polysemy
bush – n . a low plant with many branches that arise from or near the ground.
Bush – n. Jeff Bush

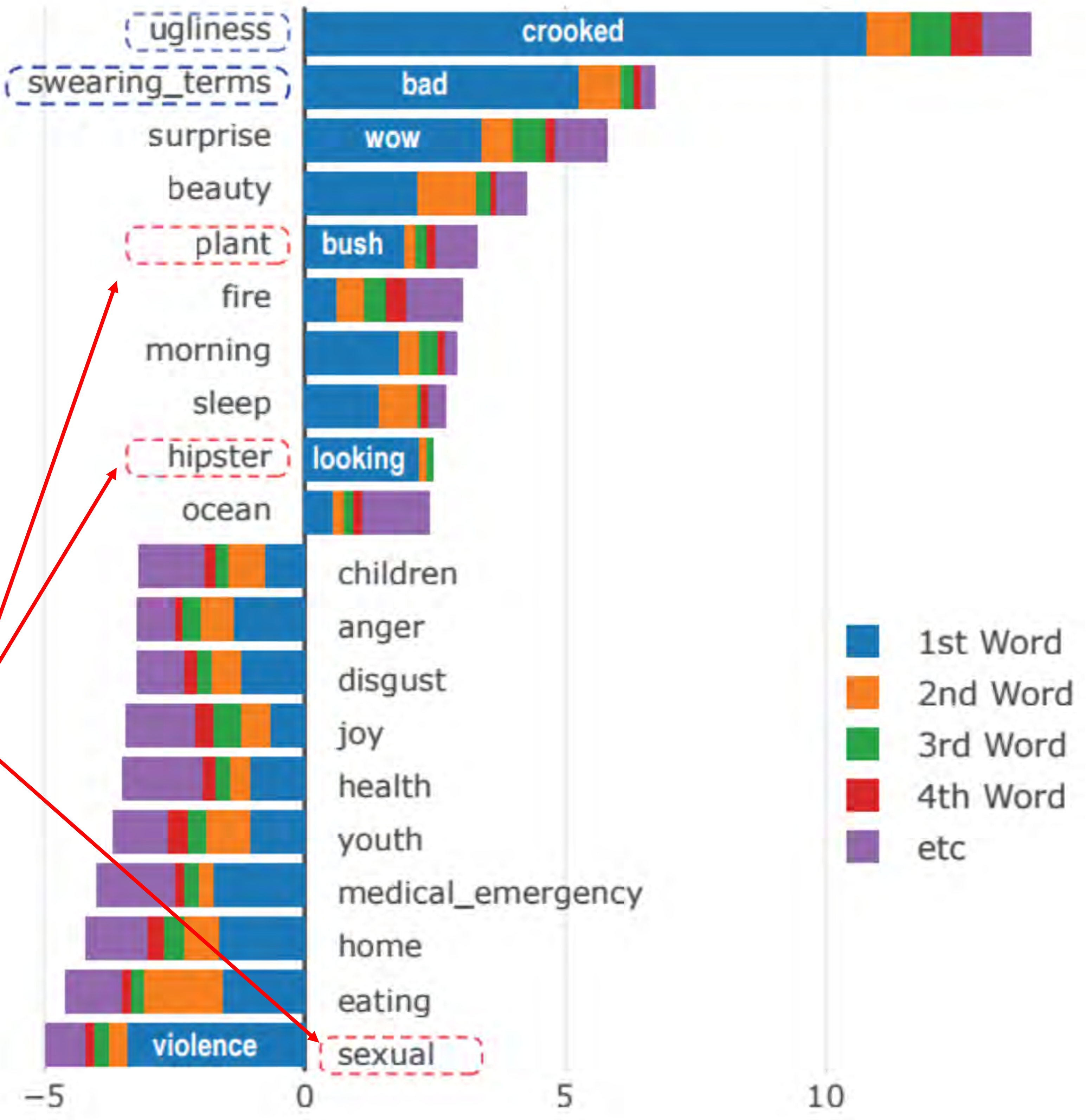
Top 10 categories

Trump vs Hillary



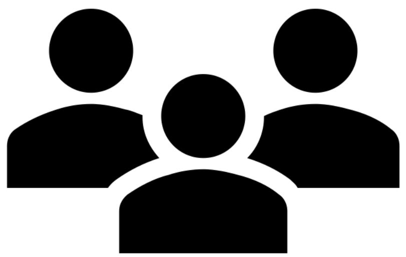
Top 10 categories

Trump vs Hillary

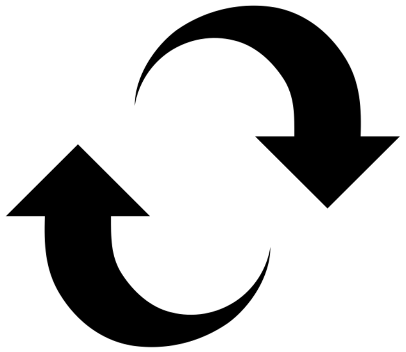


Type II Errors

Design Requirements



D1: Diverse needs

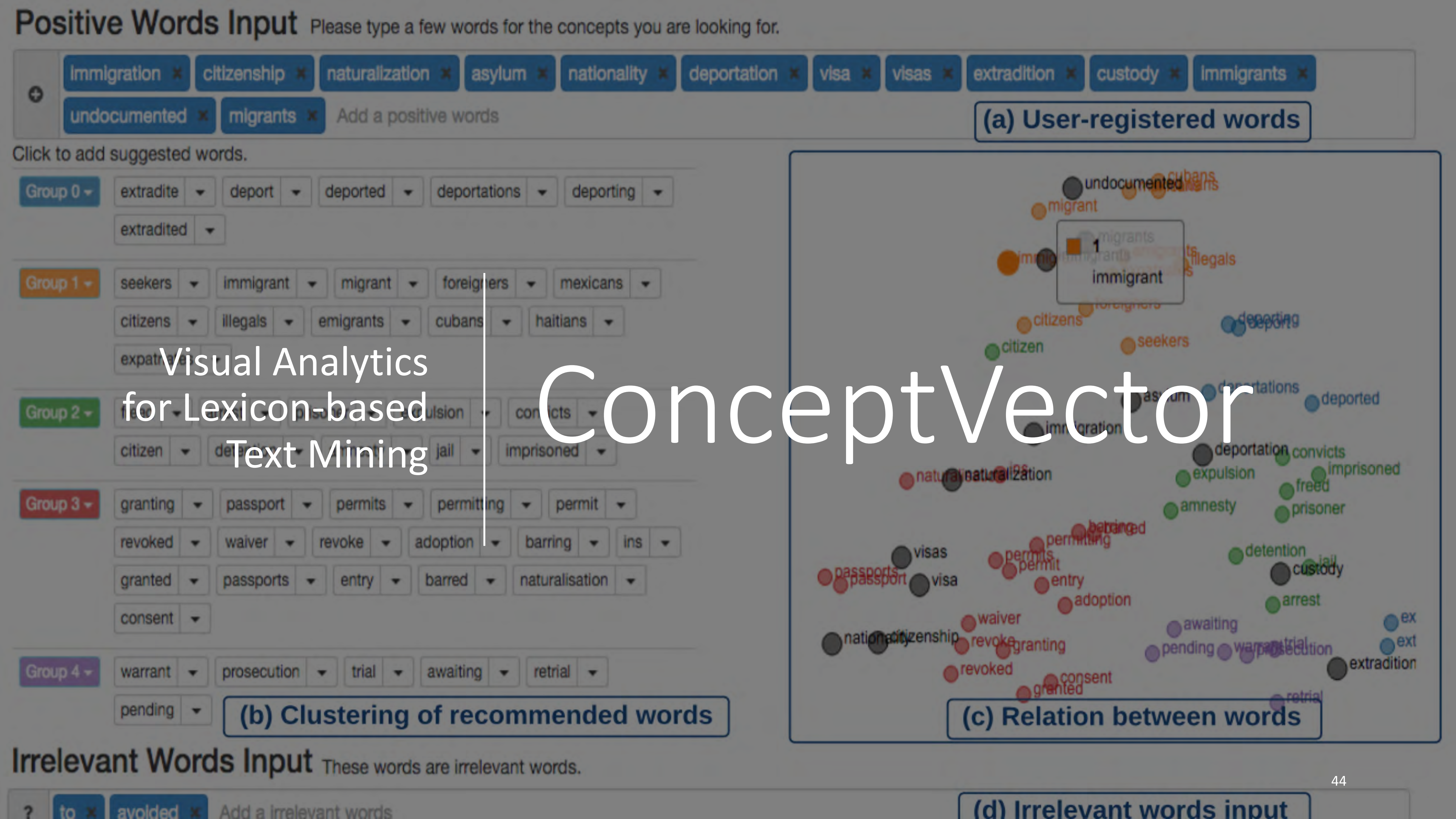


D2: Integrated loop

D3: In-context

Visual Analytics
for Lexicon-based
Text Mining

ConceptVector



Crude oil

Description: This is a concept about crude oil

This is a concept about crude oil

Concept Type: Unipolar vs Bipolar

In unipolar type, the dictionary will be created around single concepts. For example, *crude oil* or *car* can be an unipolar type. In bipolar type, the dictionary will be created between two concepts. For example, *happiness vs sadness* or *Democratic vs Republican* can be an bipolar type.

Unipolar Bipolar

Advanced Settings

Positive Words Input

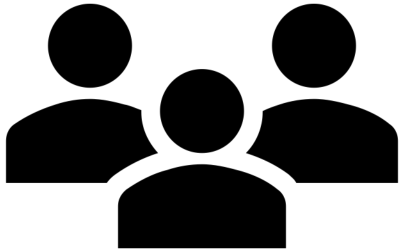
Please type a few words for the concepts you are looking for.

Add a positive words

Click to add suggested words.

Building dictionary

No Data Available.



D1: Diverse needs

Sophisticated concept modeling

Sophisticated concept modeling

Irrelevant words

I am interest in “tidal flooding”, not “storm flooding.”

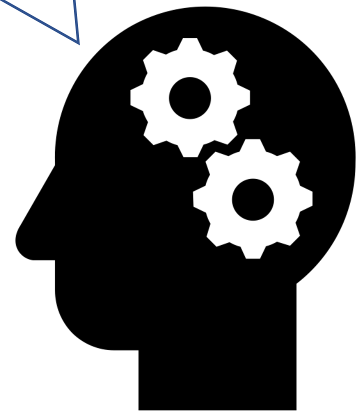
Sophisticated concept modeling

Irrelevant words

I am interest in "tidal flooding", not "storm flooding."

Bipolar Conceps

Can I map the words continuously from "Democratic party" to "Republican party"?



Sophisticated concept modeling

Irrelevant words

I am interest in "tidal flooding", not "storm flooding."

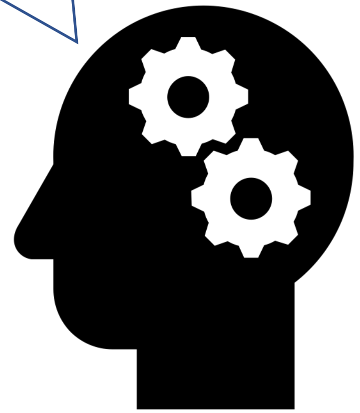
Bipolar Conceps

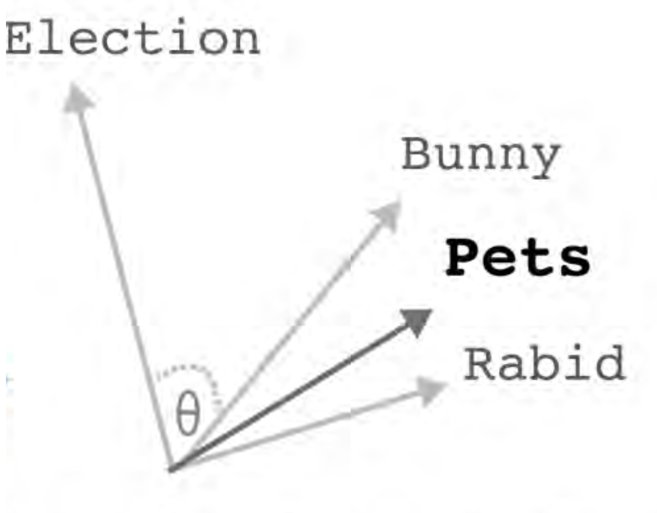
Can I map the words continuously from "Democratic party" to "Republican party"?

Kernel Density Estimation (**KDE**)

using Gaussian Kernel

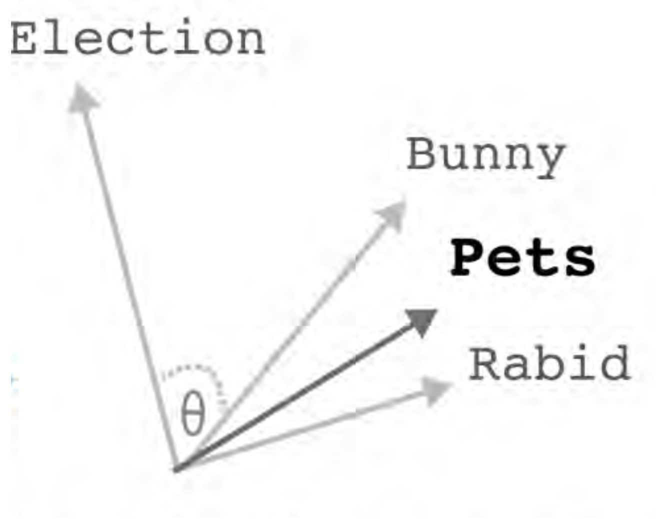
where bandwidth represents selectivity of seed words



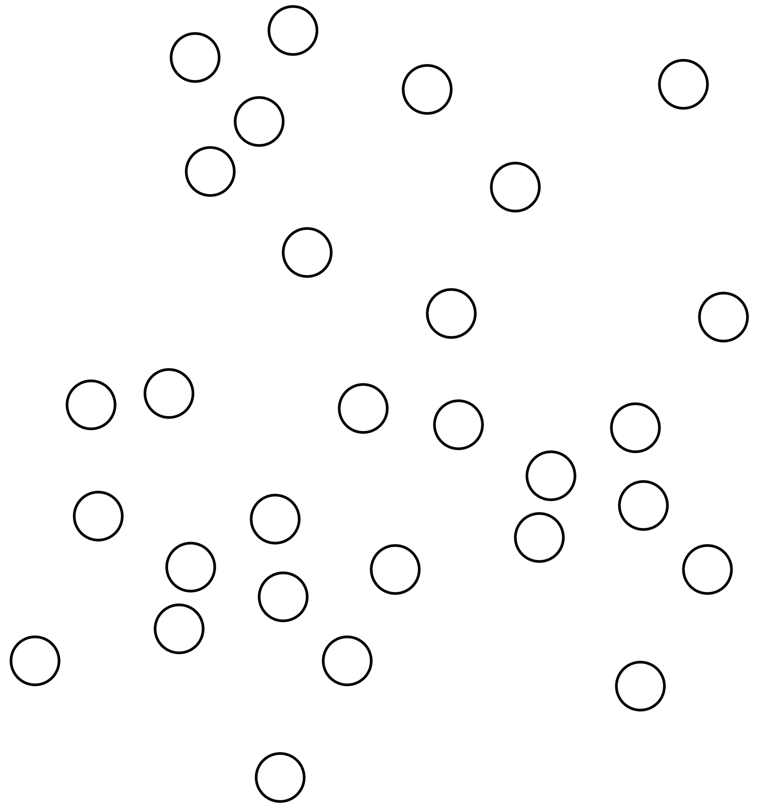


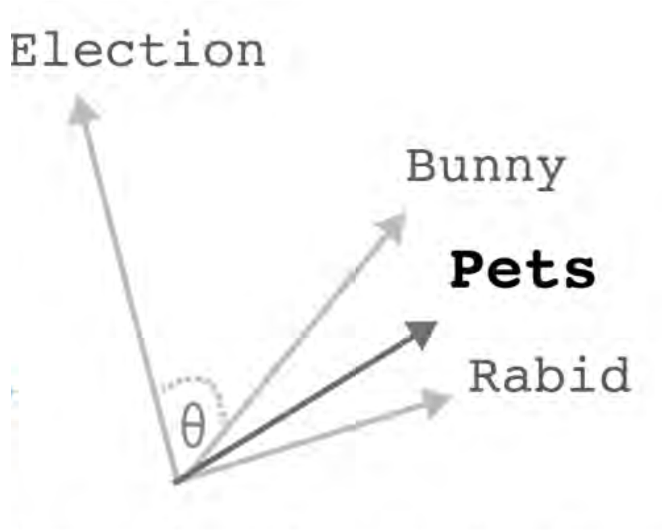
Cosine similarity for
relevance score

Previous approach

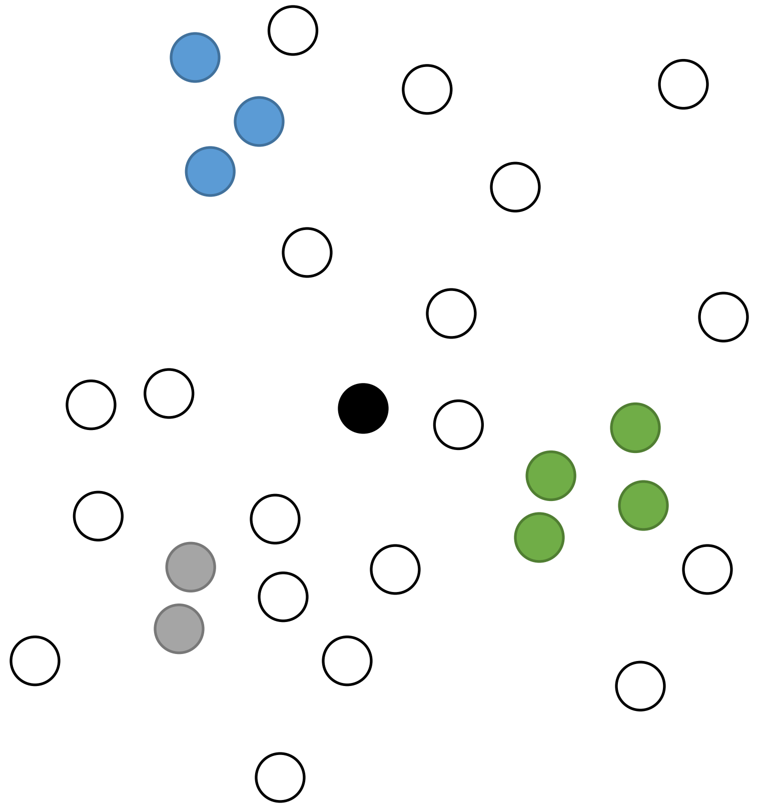


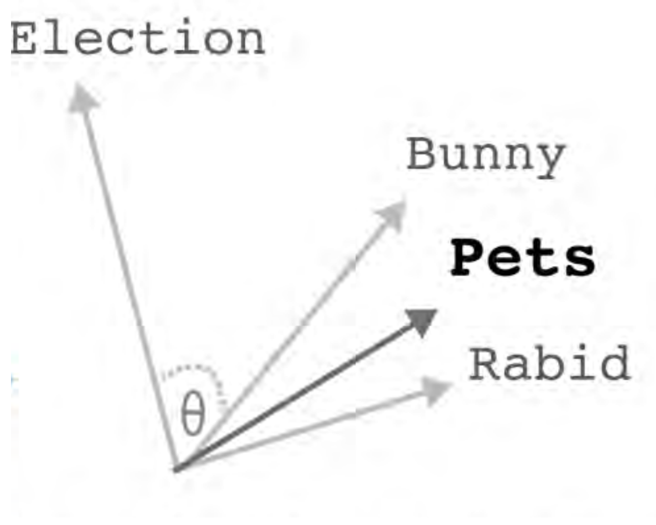
Cosine similarity for relevance score



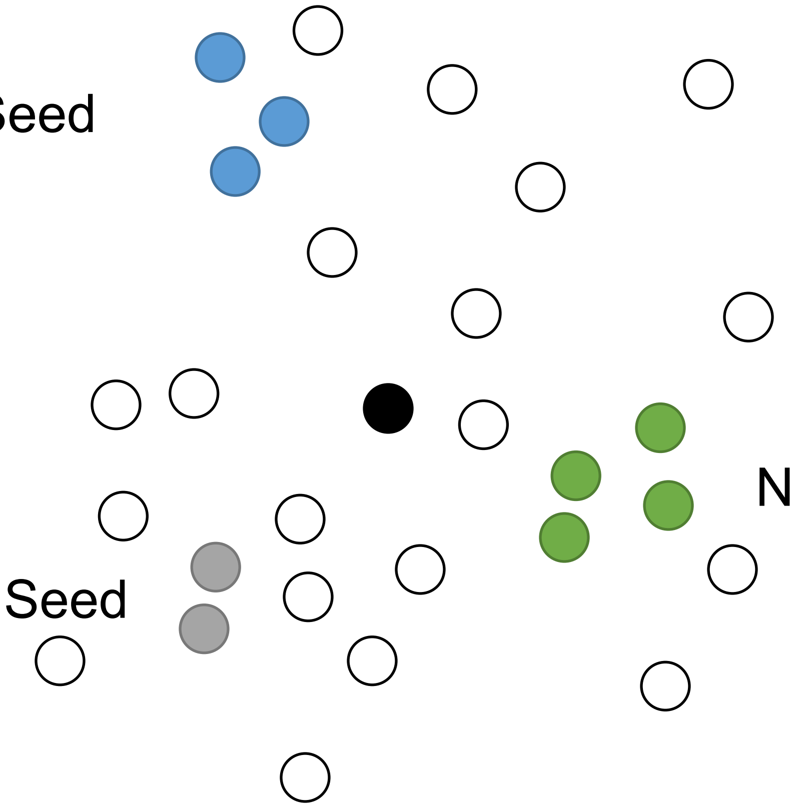


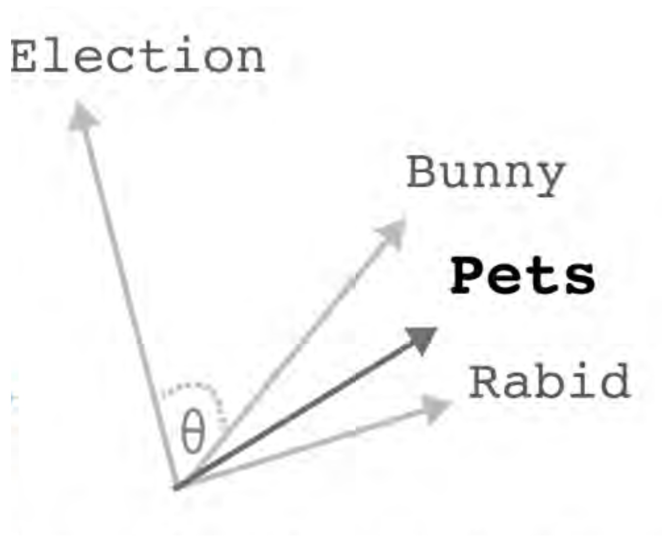
Cosine similarity for relevance score



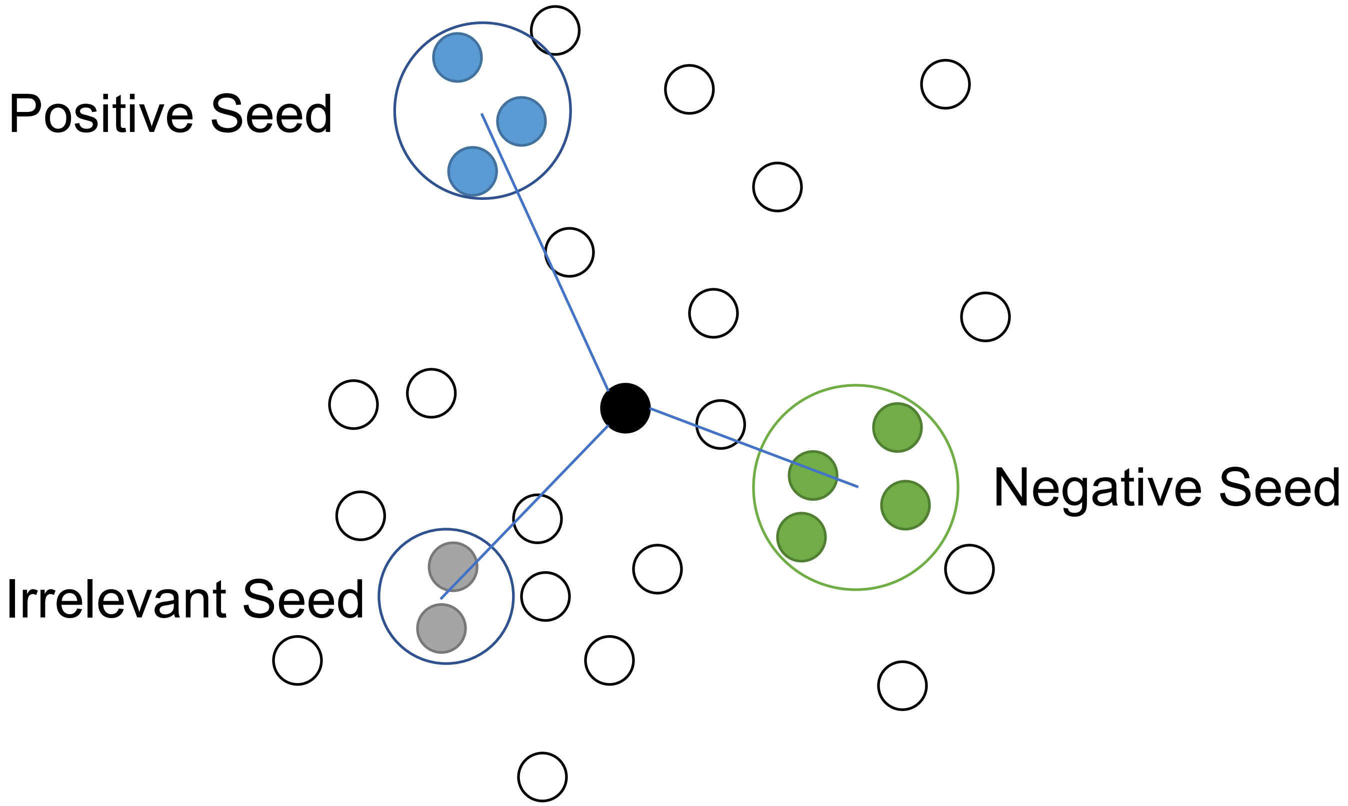


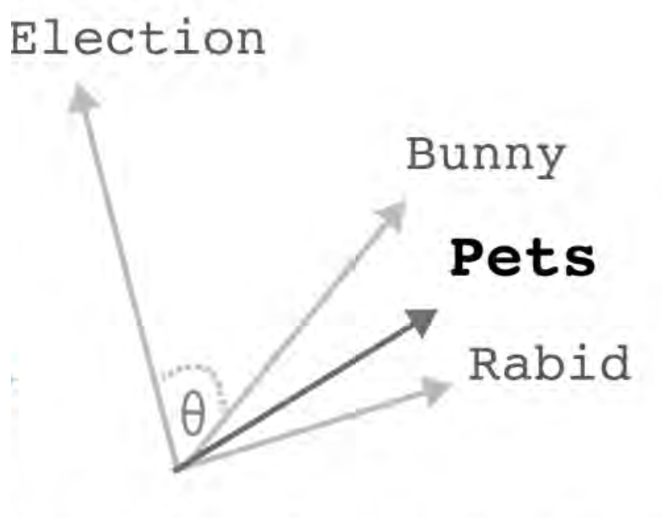
Cosine similarity for relevance score



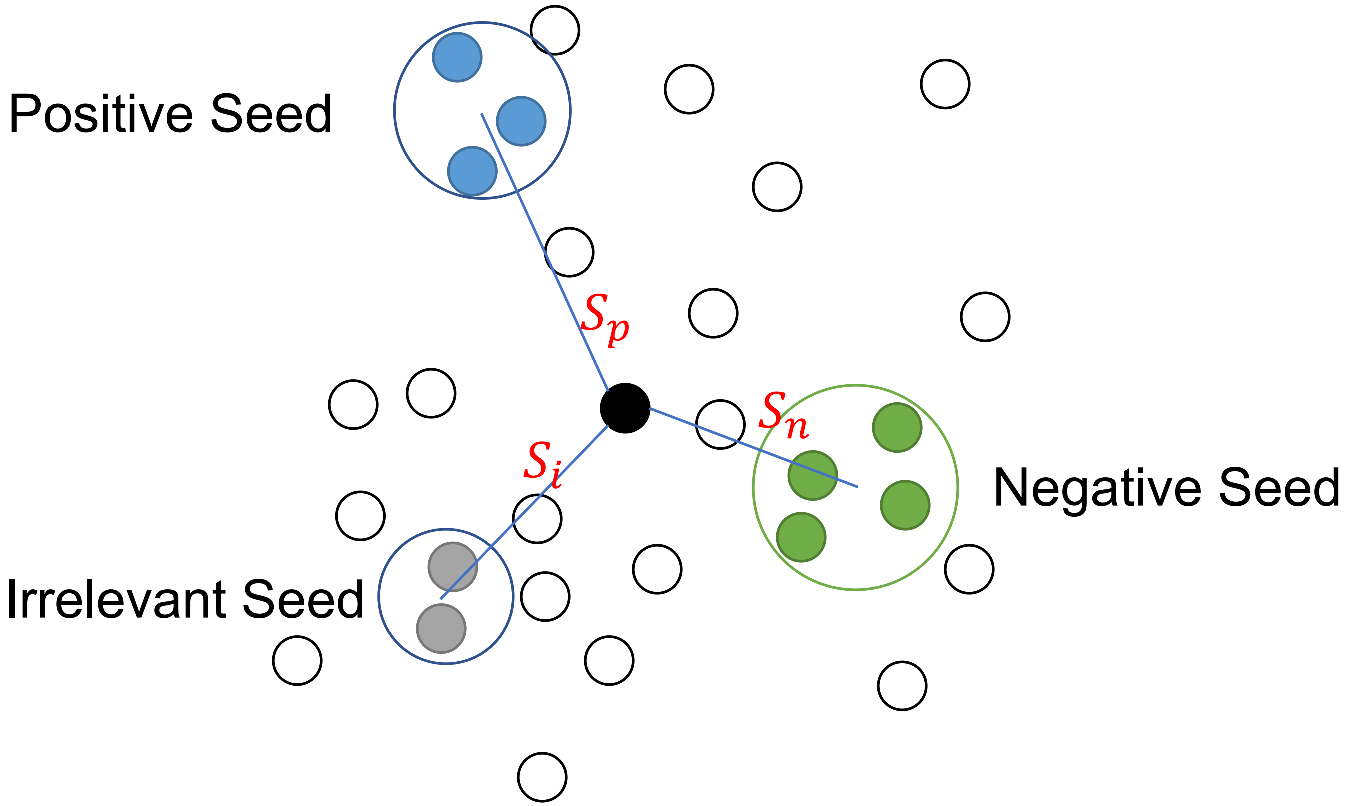


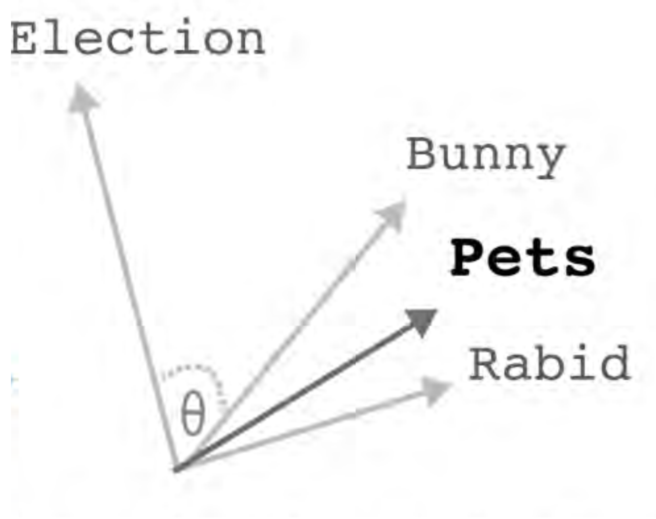
Cosine similarity for
relevance score



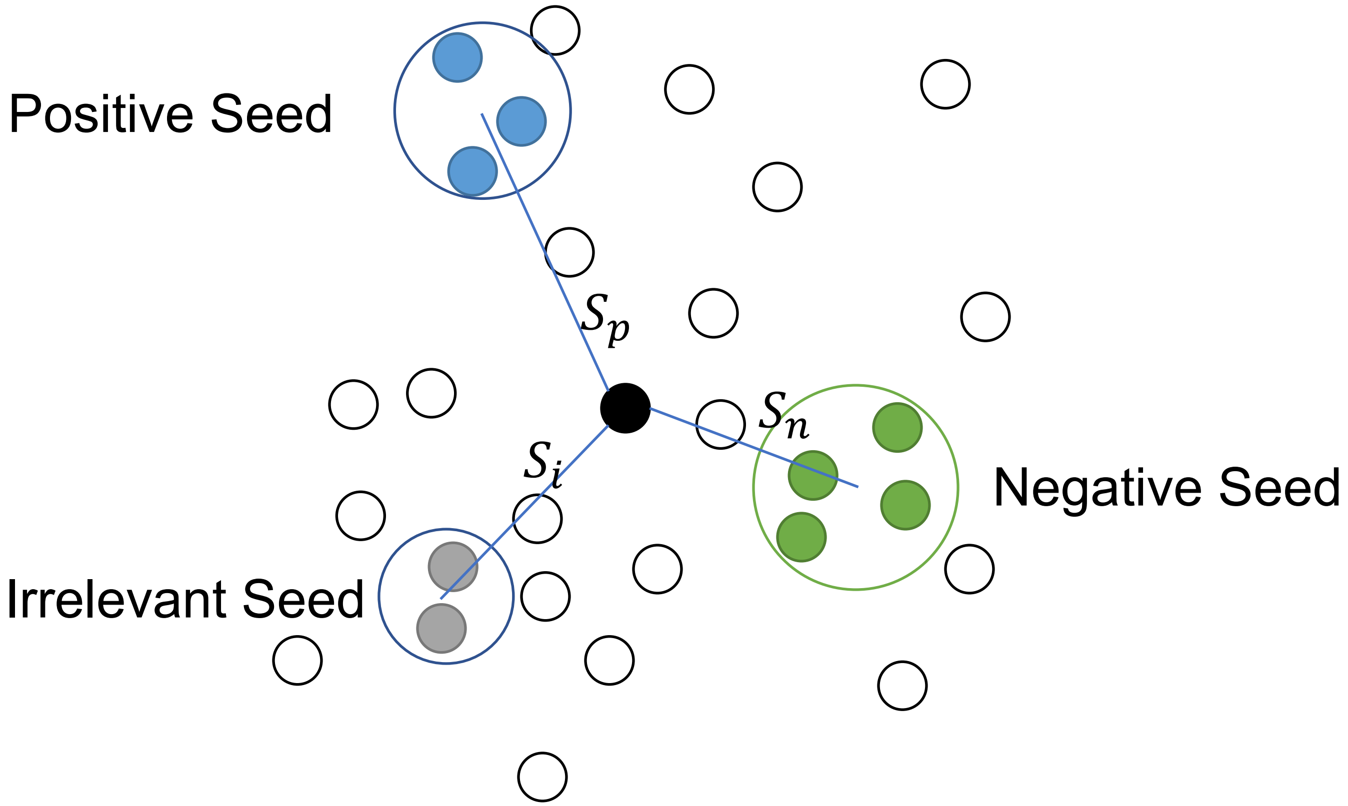


Cosine similarity for relevance score

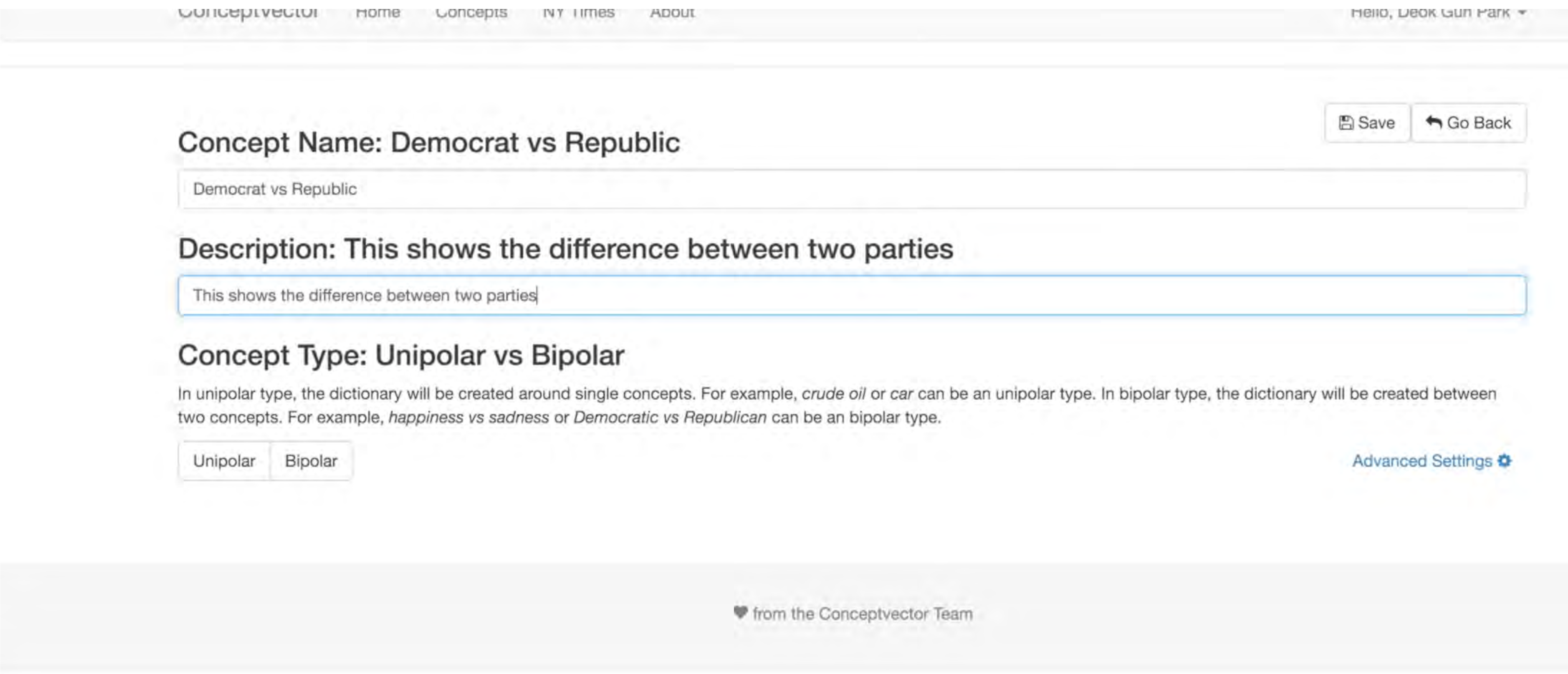




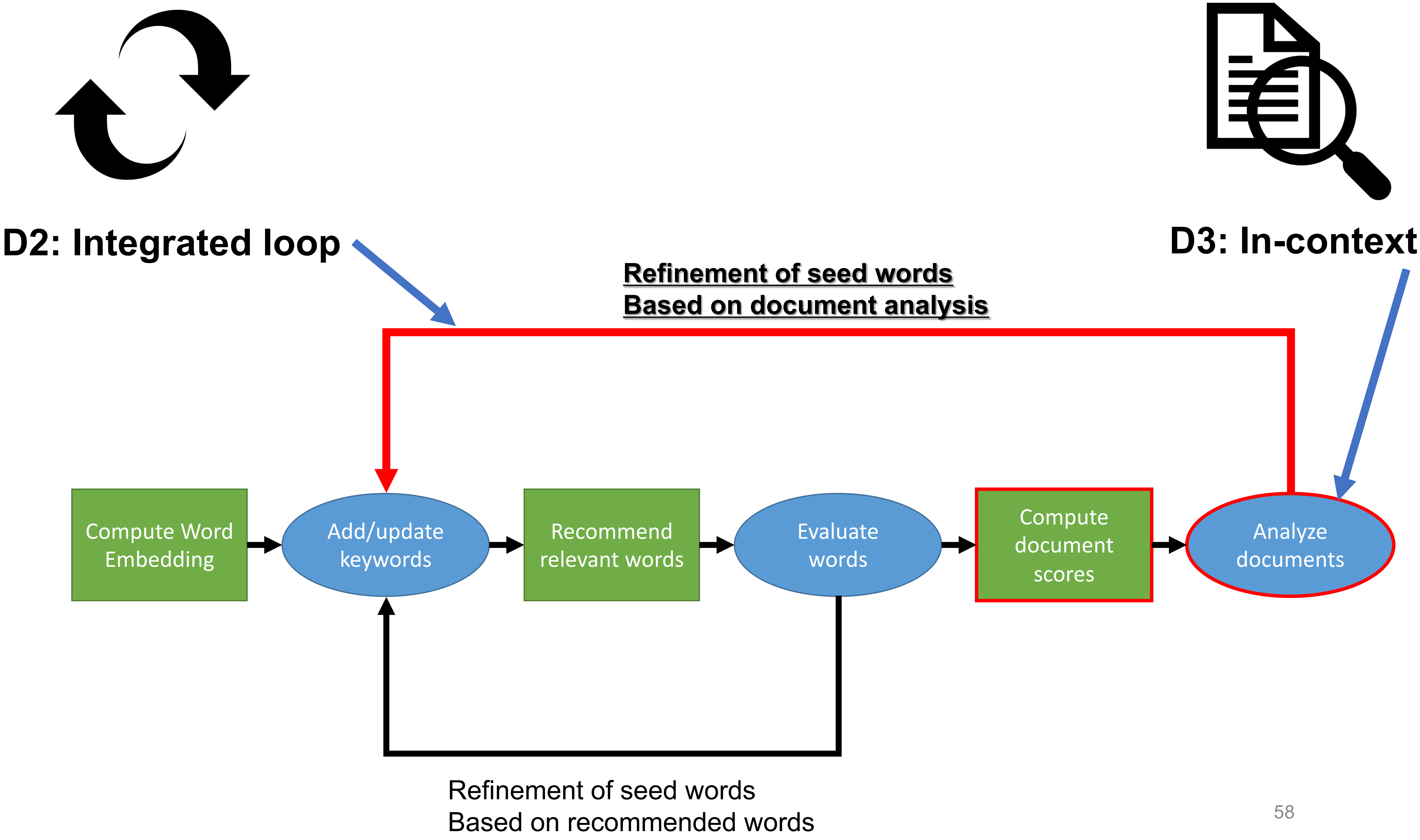
Cosine similarity for
relevance score



$$Relevance = (1 - S_i) \cdot (S_p - S_n)$$



Bipolar concepts


D2: Integrated loop
D3: In-context
Refinement of seed words
Based on document analysis
Compute Word Embedding
Add/update keywords
Recommend relevant words
Evaluate words
Compute document scores
Analyze documents
Refinement of seed words
Based on recommended words

SECTIONS HOME SEARCH **The New York Times**

1.  Top Russian Officials Discussed How to Influence Trump Aides Last Summer

2.  Greg Gianforte, G.O.P. House Candidate, Is Accused of 'Body-Slamming' Guardian...

3.  They Went to Manchester Arena as Homeless Men. They Left as Heroes.

4.

ELECTION 2016 Full Results | Exit Polls | Trump's Cabinet

Inside the Republican Party's Desperate Mission to

By **ALEXANDER BURNS, MAGGIE HABERMAN** and **JONATHAN MARTIN** FEB. 27, 2016

Donald J. Trump speaking in Milford, N.H., a week before the state's primary this month. Some establishment Republicans have been scrambling for a way to prevent him from becoming the party's presidential nominee.
Damon Winter/The New York Times

2763 COMMENTS

Readers shared their thoughts on this article.

The comments section is closed. To submit a letter to the editor for publication, write to letters@nytimes.com.

All 2763 | Readers' Picks 1721 | NYT Picks 21

David Weaver Orlando · February 27, 2016

The political industry does not realize that it is in the middle of a hostile takeover from people who are beyond sick and tired of insiders who use the aspirations of hard working people to line their pockets.

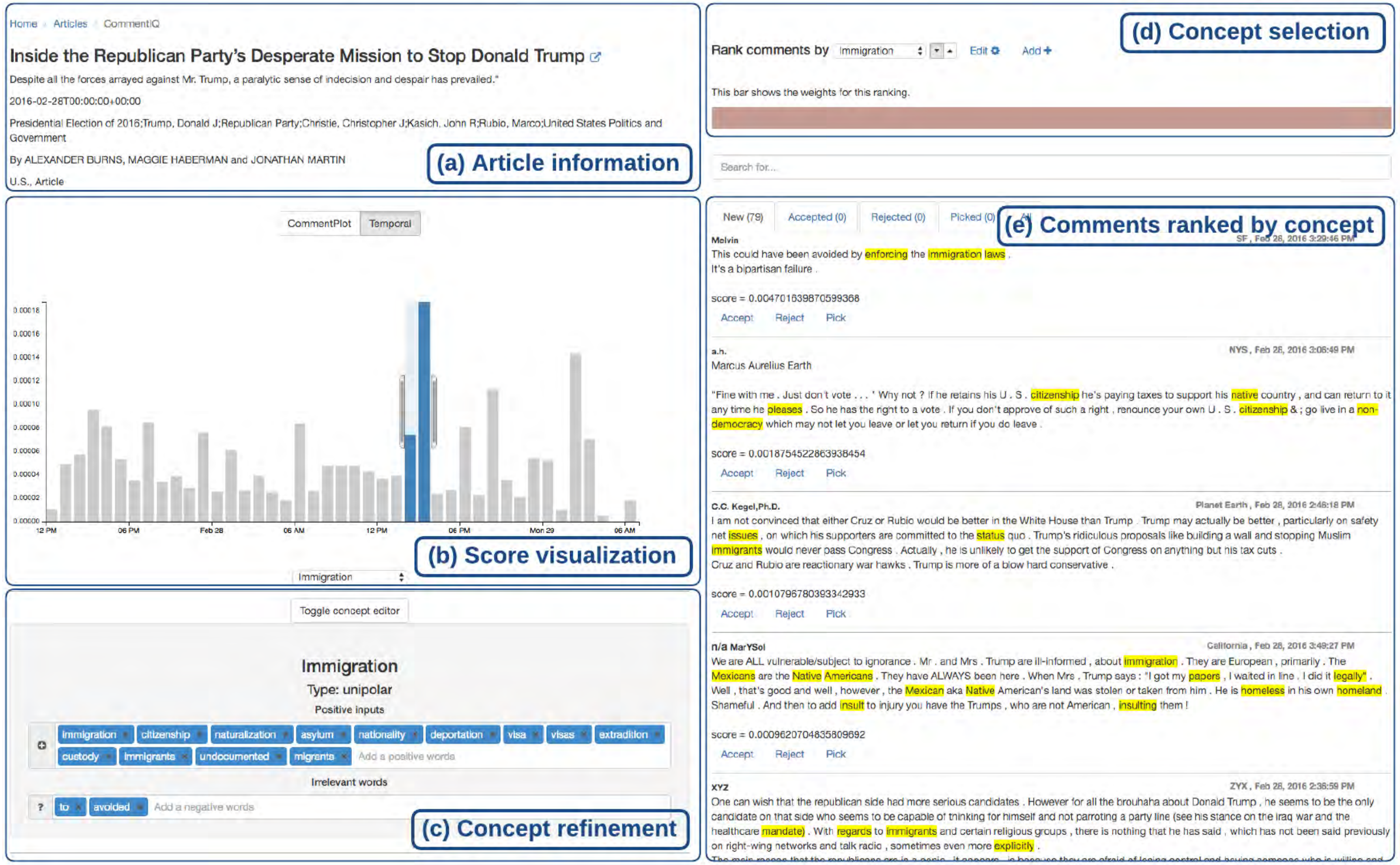
Is Trump perfect? Nope. But the leaders of this nations political indu$try have failed us miserably. Hopefully the "we know what's best for you" crowd has its' days numbered.

1100 Recommend

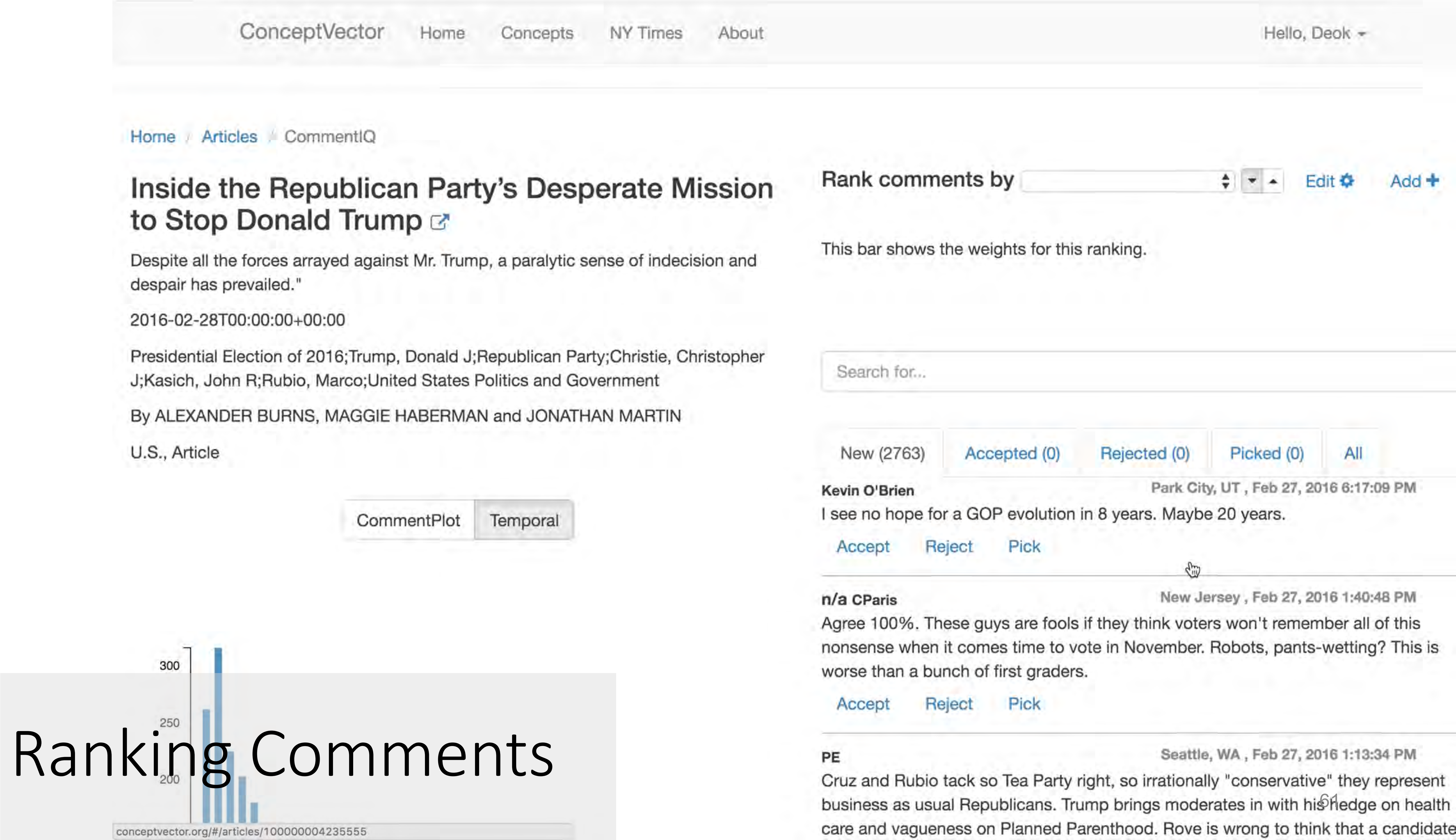
Gfagan PA · February 27, 2016

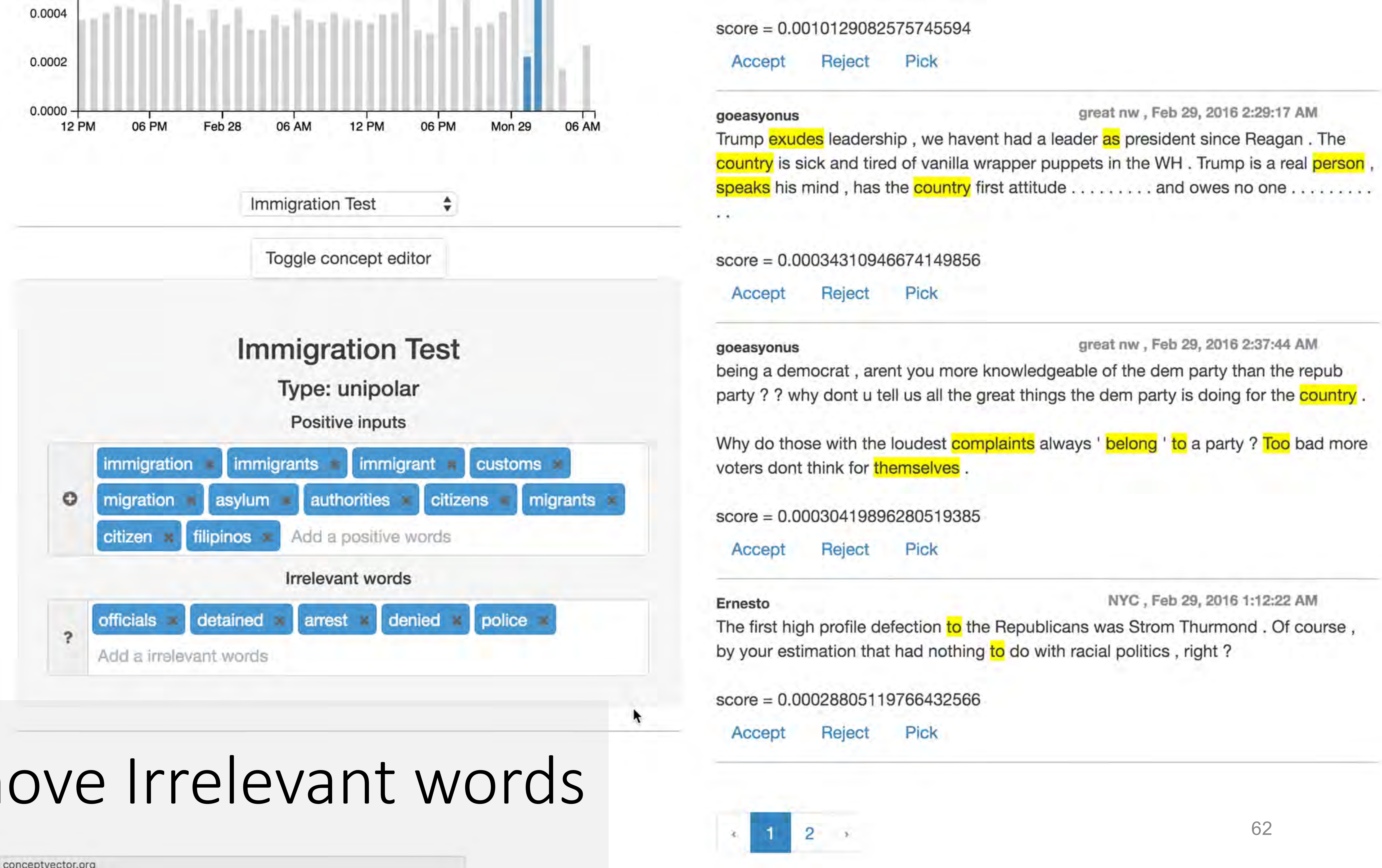
Dr. Frankenstien frantically nailed the planks into place, trying to seal his laboratory off from the rest of the world.
What would happen if the Monster got out? What if people realized he had created it?
The pounding from the inside increased in intensity. The Monster roared. It sounded like "Yuge!"
Dr. Frankenstein wondered what that might mean, then resumed hammering on the nails.
But the door gave way and the boards he was trying to nail in place began to bend under the Monster's fists.
Rage. That is all the doctor saw in the contorted, strangely orange face of his creation.
"Yuge!"
Splinters began flying from the boards.
Dr. Frankenstein turned his back and ran. He'd go to the House. There, at least, he'd be safe doing nothing.
When the questions came, the doctor would deny everything. He'd blame his neighbor, Mr. Obama.
Yes, that would work. That would deflect attention away from his own nefarious activities, spread over the preceding decades, that had led to this catastrophe.
It was now up to others to deal with his Creation.

Document Analysis



Ranking Comments

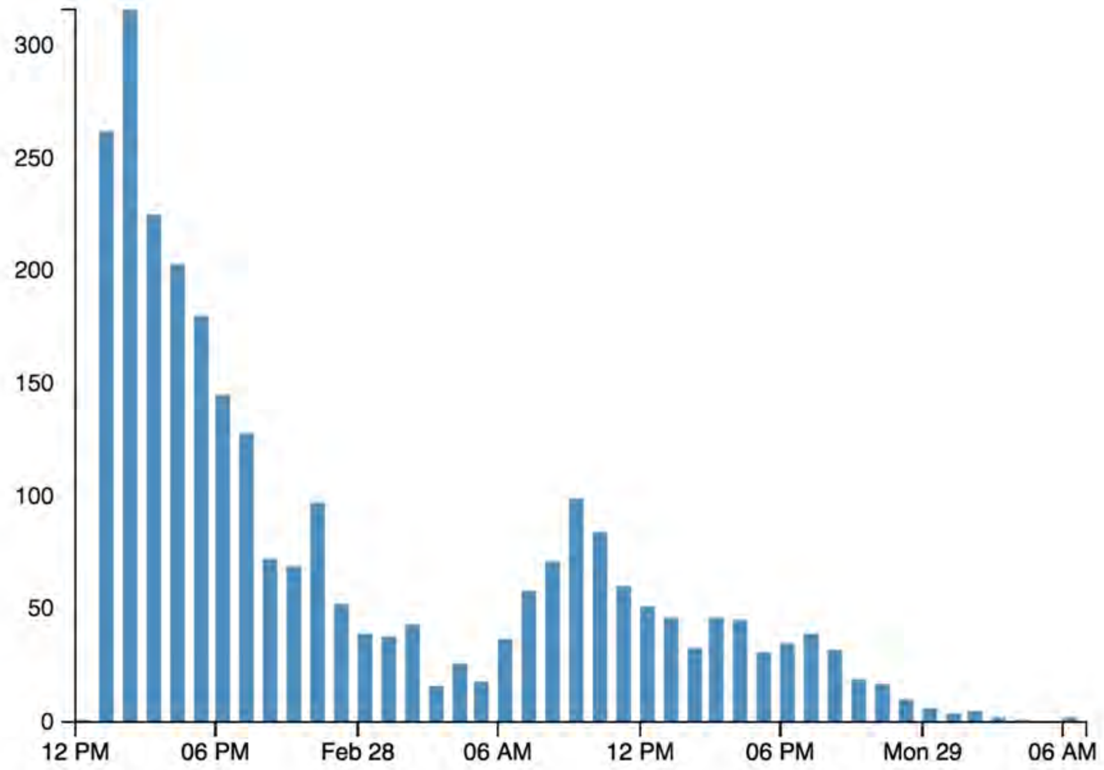




Remove Irrelevant words

Temporal Trend

CommentPlot | Temporal



Number of Comment

Toggle concept editor

You are viewing this concept as a guest. Please login and copy this concept to create a clone and edit.

Immigration-related

and don't want to do anything about illegal immigration . Immigration effects everythi

.

score = 0.009691857174335334

Accept Reject Pick

Melvin SF , Feb 28, 2016 3:29:46 PM

This could have been avoided by enforcing the immigration laws .
It's a bipartisan failure .

score = 0.00565333256500768

Accept Reject Pick

Bahtat San Diego , Feb 27, 2016 4:02:37 PM

The GOP elites have another great love for Illegal Immigrants . As well as being cheap labor , they are not allowed to VOTE .

score = 0.005202291401691696

Accept Reject Pick

Ali Michigan , Feb 27, 2016 7:10:03 PM

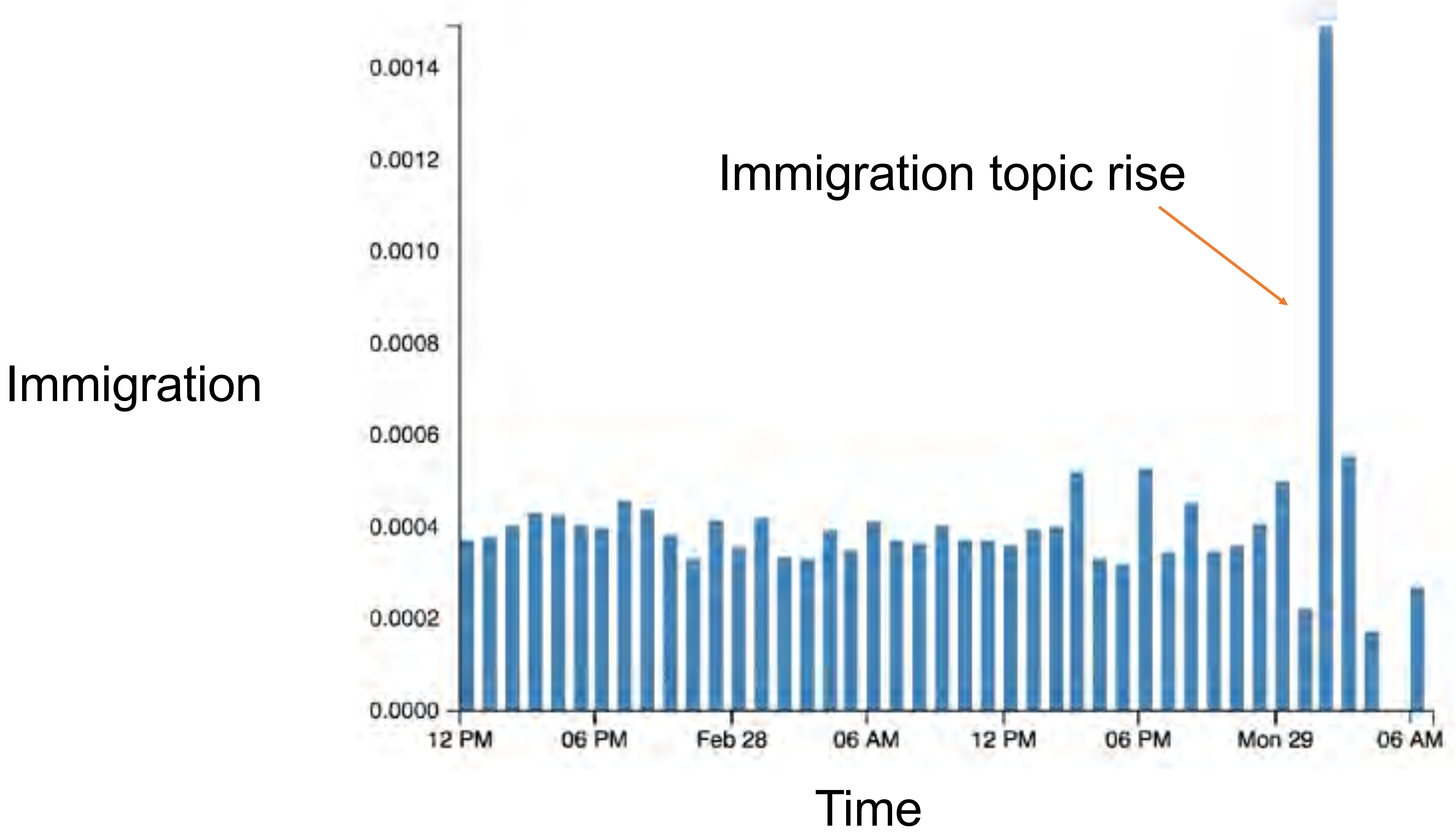
Too bad they simply weren't willing to give up amnesty and more immigration , legal and illegal . They complain about gov't subsidies to business yet that's just what illeg aliens and immigrants are , subsidies . Immigration is the focus of much of what's wrong in this country , to those supporting Trump . from unemployment to under performing schools to high healthcare costs and water shortages .

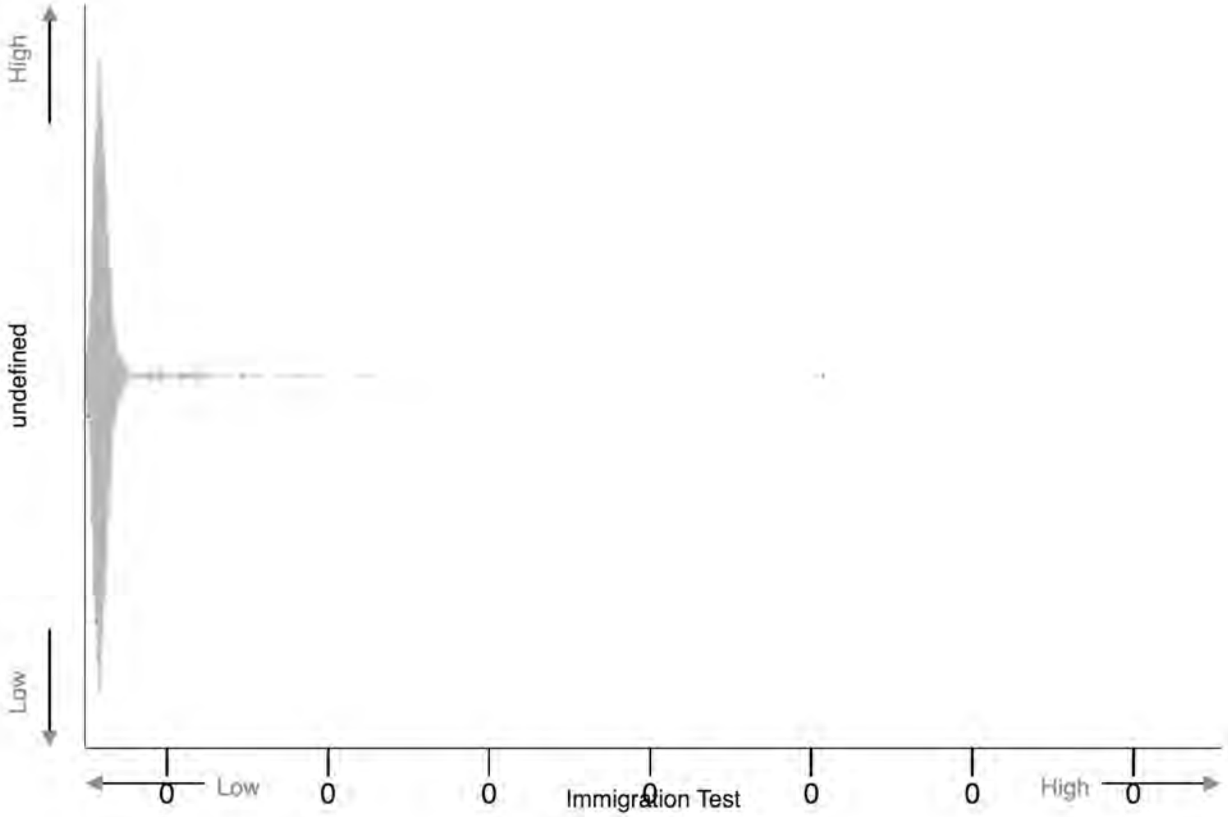
score = 0.004590839907424656

Accept Reject Pick

jmf denver, co , Feb 27, 2016 3:15:43 PM

No its not . Who has stopped all ways to get rid of illegals ? Who has allowed massiv amounts of legal immigration ? Its both of them so that average people cant exist wh


Immigration topic rise
Immigration
0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
12 PM
06 PM
Feb 28
06 AM
12 PM
06 PM
Mon 29
06 AM
Time



Show me Immigration Test vs . -- undefined --

Toggle concept editor

Update success! Score is calculated based on updated concept.

Immigration Test

Type: unipolar

Positive inputs

Comment Plot

want to end this rot in the system ? amend the constitution ! 1) Limit members of Congress to a 4yr single term . that gets them out of the business of endlessly seeking fat cat sponsors , and brings them to the business of true governance . the same applies to senators . let them all step aside after a term to give others also the chance to "serve the American people . " 2) Get the Citizens United judgement reversed . No more big money super-packs . 3) Apportion equal time to all candidates on the media . the media has overplayed the "elections are a spectator/gladiatorial sport" angle , by focusing on their ratings and the 'fun , and the bloodier' candidates . 4) Election day should be declared a holiday , so all can afford to go to vote . Even a far poorer country like India , has this . 5) make voting compulsory . We want a democratic country ? then let's get off our couches .

score = 0.0006589178749339568

Accept Reject Pick

goeasyonus great nw , Feb 29, 2016 2:29:17 AM

Trump exudes leadership , we havent had a leader as president since Reagan . The country is sick and tired of vanilla wrapper puppets in the WH . Trump is a real person , speaks his mind , has the country first attitude . . . . and owes no one . . . .

score = 0.00014680238075992264

Accept Reject Pick

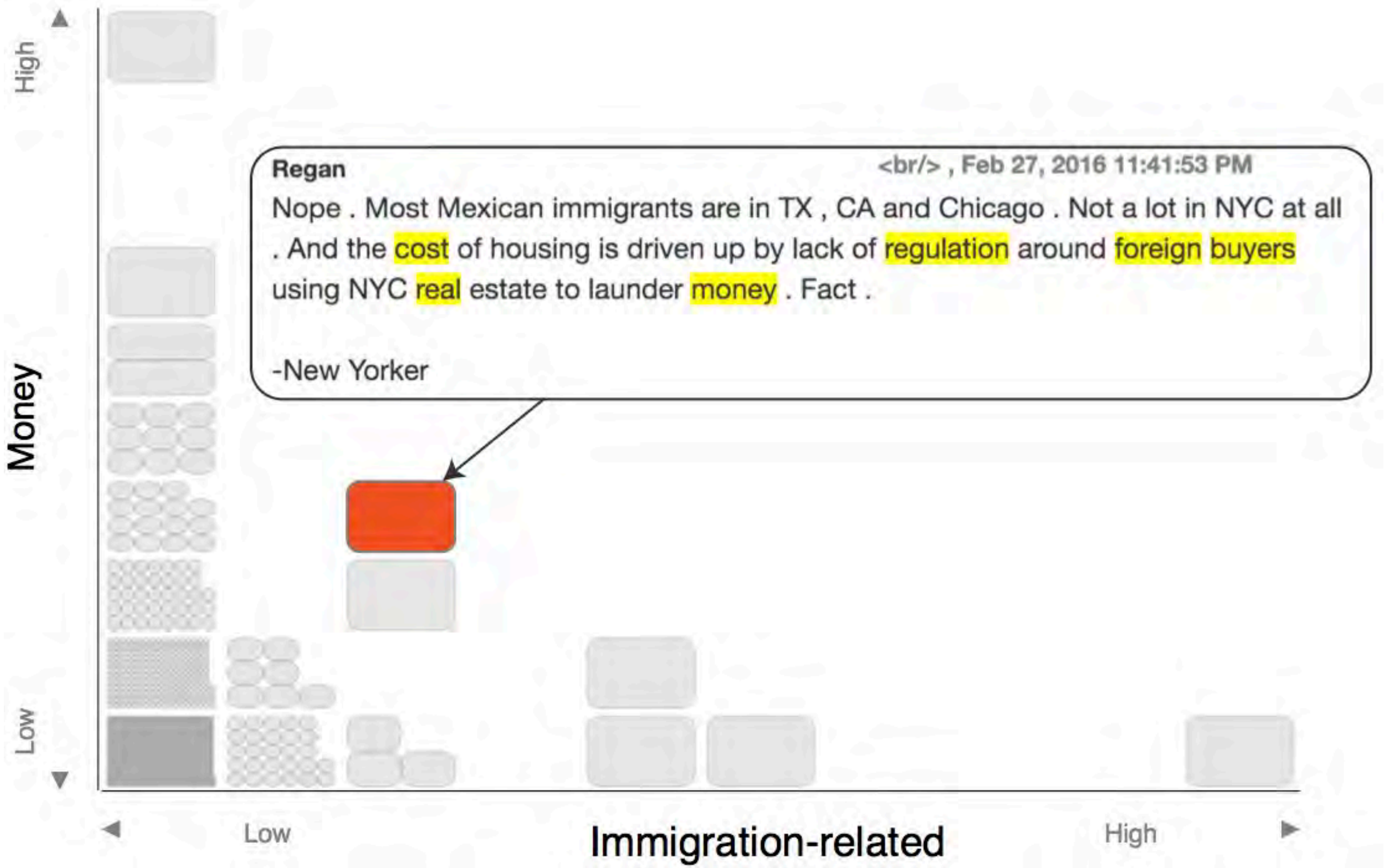
goeasyonus great nw , Feb 29, 2016 2:37:44 AM

being a democrat , arent you more knowledgeable of the dem party than the repub party ? ? why dont u tell us all the great things the dem party is doing for the country .

Why do those with the loudest complaints always ' belong ' to a party ? Too bad more voters dont think for themselves .

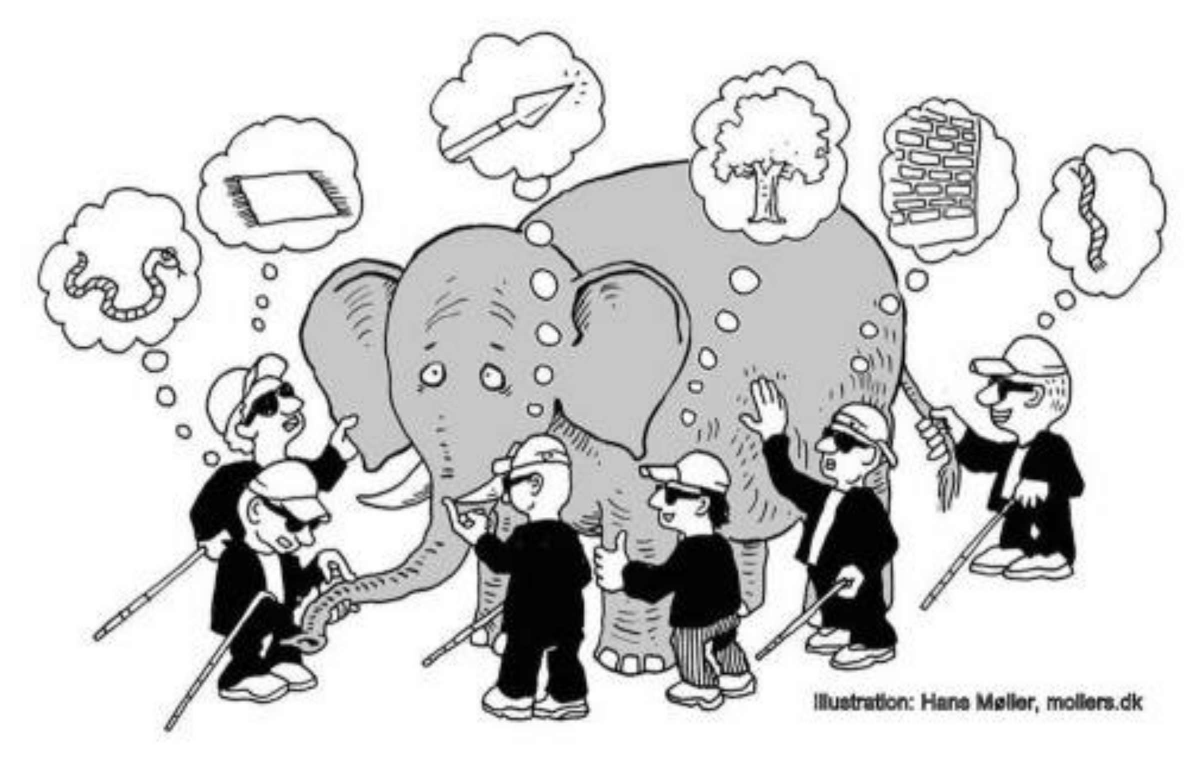
score = 0.00011201416478632322

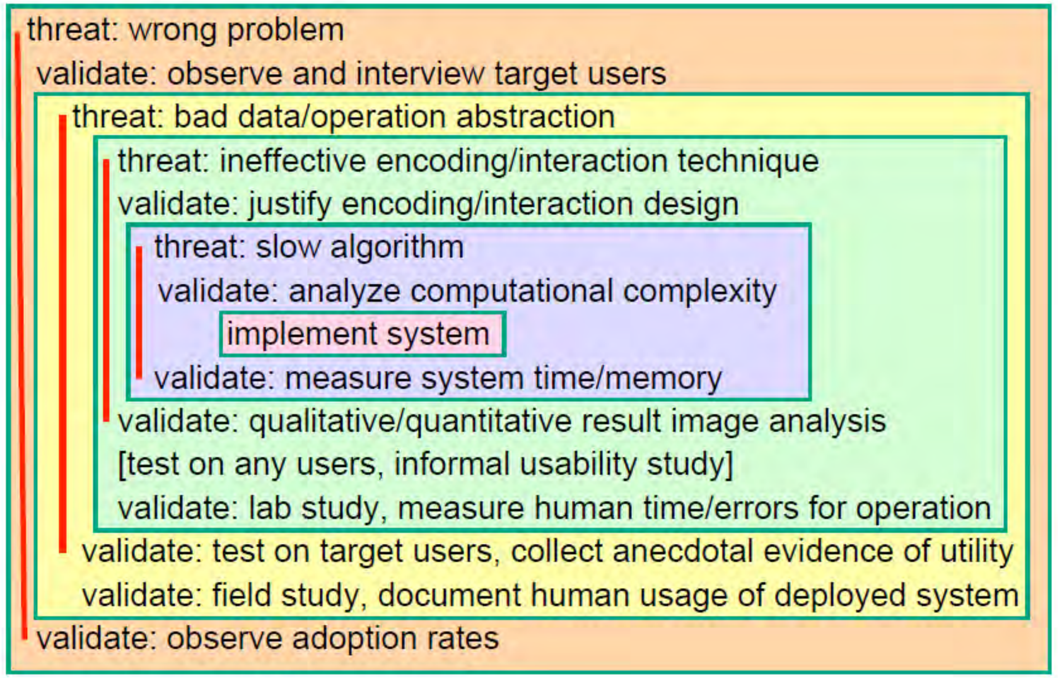
Accept Reject Pick


High
Money
Low
Regan

 , Feb 27, 2016 11:41:53 PM
Nope . Most Mexican immigrants are in TX , CA and Chicago . Not a lot in NYC at all . And the cost of housing is driven up by lack of regulation around foreign buyers using NYC real estate to launder money . Fact .
-New Yorker
Low
Immigration-related
High

Evaluation





Munzner, Tamara. "A nested model for visualization design and validation." *IEEE transactions on visualization and computer graphics* 15.6 (2009).

Evaluation

Lab Experiment

- Lexicon-building interface
- Wordnet and Thesarus.com

Evaluation

Lab Experiment

- Lexicon-building interface
- Wordnet and Thesarus.com

Quantitative Evaluation

- Bipolar concept modeling
- Compared with crowdsourced data

Evaluation

Lab Experiment

- Lexicon-building interface
- Wordnet and Thesaurus.com

Quantitative Evaluation

- Bipolar concept modeling
- Compared with crowdsourced data

Expert Feedback

- System limitation and usage
- Visual analytics expert and NLP expert

Conclusion

Take-away Message

- You can build custom dictionary for your own domain and analyze text.

- Interactive refinement may improve quality of text analysis by reducing false positive errors.


Seungyeon Kim
sylund@gmail.com
Nicholas Diakopoulos
@ndiakopoulos
Deokgun Park
@intuinno
Jaegul Choo
jchoo@korea.ac.kr
Niklas Elmqvist
@NElmqvist
Jurim Lee
jurim0301@korea.
ac.kr
Meet the Team

Questions?

Try demo at:

Conceptvector.org

intuinno@umd.edu