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 Visualization
Lexicon-based Text Mining
Lexicon-based Text Mining

Lexicon = A set of keywords that is related to specific concept
Lexicon-based Text Mining

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

Positive = \{ good, great, happy, \ldots \} 

Negative = \{ bad, worst, horrible, \ldots \}
Lexicon-based Text Mining

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

Positive = \{ good, great, happy, \ldots \}  
Negative = \{ bad, worst, horrible, \ldots \}  

Lexicon-based Text Mining

Positive = \{ good, great, happy, \ldots \}
Negative = \{ bad, worst, horrible, \ldots \}
Lexicon-based Text Mining

Positive = \{ good, great, happy, \ldots \}

Negative = \{ bad, worst, horrible, \ldots \}

D1 “The movie was good. I was happy.”

D2 “The movie opens today.”

D3 “The movie is horrible.”
Lexicon-based Text Mining

Positive = \{ good, great, happy, \ldots \}

Negative = \{ bad, worst, horrible, \ldots \}

D1  "The movie was great. I was happy."

D2  "The movie opens today."

D3  "The movie is horrible."

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>D2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D3</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Linguistic Inquiry and Word Count (LIWC)

<table>
<thead>
<tr>
<th>A first</th>
<th>for</th>
<th>oby</th>
<th>und</th>
<th>ac</th>
<th>aces</th>
<th>abs</th>
<th>acc</th>
<th>ad</th>
<th>mape</th>
<th>w/e</th>
</tr>
</thead>
<tbody>
<tr>
<td>above</td>
<td>for</td>
<td>oby</td>
<td>und</td>
<td>ac</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>acclamed</td>
<td>fra*</td>
<td>obys*</td>
<td>under<em>n</em></td>
<td>ac*</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>adminis*</td>
<td>fra*</td>
<td>obys*</td>
<td>under<em>n</em></td>
<td>ac*</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>age</td>
<td>glori*</td>
<td>official*</td>
<td>underprivleg*</td>
<td>achiev*</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>allow*</td>
<td>glori*</td>
<td>official*</td>
<td>underprivleg*</td>
<td>achiev*</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>amateur*</td>
<td>god</td>
<td>oppose*</td>
<td>unqualified</td>
<td>acquir*</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>ambition</td>
<td>god's</td>
<td>oppositi*</td>
<td>unwanted</td>
<td>add</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>ambitions</td>
<td>goddes*</td>
<td>order*</td>
<td>up</td>
<td>add</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>ambitious</td>
<td>good-for-nothing</td>
<td>orders*</td>
<td>upper</td>
<td>add</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>apolo*</td>
<td>govern*</td>
<td>outcast*</td>
<td>upperclass*</td>
<td>add</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>appre*</td>
<td>greatest</td>
<td>outrank*</td>
<td>useless</td>
<td>add</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>approv*</td>
<td>grea*</td>
<td>over</td>
<td>vanquish*</td>
<td>add</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>armies</td>
<td>grea*</td>
<td>overpow*</td>
<td>verb</td>
<td>add</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>army</td>
<td>grea*</td>
<td>overpow*</td>
<td>verb</td>
<td>add</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>aust*</td>
<td>grown-up*</td>
<td>every*</td>
<td>veteran*</td>
<td>add</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>ather*</td>
<td>grownup*</td>
<td>owner*</td>
<td>victim*</td>
<td>add</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>assault*</td>
<td>help</td>
<td>parish*</td>
<td>victor*</td>
<td>add</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>assertive</td>
<td>helper*</td>
<td>parliament*</td>
<td>vip</td>
<td>add</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>attack*</td>
<td>helpless*</td>
<td>passiv*</td>
<td>vip</td>
<td>add</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>attendant</td>
<td>hierarchy</td>
<td>pathetic</td>
<td>vulnerab*</td>
<td>avert*</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>authori*</td>
<td>high</td>
<td>pathetically</td>
<td>war</td>
<td>avoid*</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>authorize</td>
<td>high-ranking</td>
<td>patriarch*</td>
<td>warfare*</td>
<td>avoid*</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>batt*</td>
<td>higher</td>
<td>peasant*</td>
<td>warned</td>
<td>avoid*</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
<tr>
<td>best*</td>
<td>highest</td>
<td>eee*</td>
<td>warning</td>
<td>avoid*</td>
<td>aces</td>
<td>abs</td>
<td>acc</td>
<td>ad</td>
<td>mape</td>
<td>w/e</td>
</tr>
</tbody>
</table>
How to Build a Lexicon?
How to build Lexicon
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)
Crowdsourcing

- Affective Norms for English Words (ANEW)
- Hedonometer

• Scales up for single category
• Costly, limited category
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)
How to build Lexicon

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

**Mixed-initiative**

• Empath

Manual     Automatic
word embedding
word embedding


How to build Lexicon

word embedding

1. Frogs
2. Toad
3. Litoria
4. Leptodactylidae
5. Rana
6. Lizard
7. eleutherodactylus
word embedding

Nearest Neighbor

Frog

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

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

Compute Word Embedding → Add/update keywords → Recommend relevant words

Machine Task

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

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

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

Refinement of seed words
Based on recommended words
Refinement of seed words
Based on document analysis

Refinement of seed words
Based on recommended words

Compute Word Embedding → Add/update keywords → Recommend relevant words → Evaluate words → Compute document scores → Analyze documents

Machine Task
Human Task
<table>
<thead>
<tr>
<th>User</th>
<th>Tweet</th>
<th>Likes</th>
<th>Retweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Donald J. Trump</td>
<td>As President I wanted to share with Russia (at an openly scheduled W.H. meeting) which I have the absolute right to do, facts pertaining....</td>
<td>70,130</td>
<td>16,606</td>
</tr>
<tr>
<td>Hillary Clinton</td>
<td>A man who can be provoked by a tweet should not have his hands anywhere near the nuclear codes. #DebateNight</td>
<td>90,255</td>
<td>46,654</td>
</tr>
</tbody>
</table>

*Tweets from Trump and Hillary during the 2016 election*
{ 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
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

(a) User-registered words

(b) Clustering of recommended words

(c) Relation between words

(d) Irrelevant words input
Description: 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 a bipolar type.

Positive Words Input

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

Click to add suggested words.

Building dictionary
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 Conceans

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 Concepts**

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

Positive Seed

Irrelevant Seed

Negative Seed
Cosine similarity for relevance score
Cosine similarity for relevance score
Cosine similarity for relevance score

\[ \text{Relevance} = (1 - S_i) \cdot (S_p - S_n) \]
Concept Name: Democrat vs Republic

Democrat vs Republic

Description: This shows the difference between two parties

This shows the difference between two parties

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
D2: Integrated loop

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

D3: In-context
Inside the Republican Party’s Desperate Mission to

By ALEXANDER BURNS, MAGGIE HABERMAN and JONATHAN MARTIN

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.

Donato Stata/The New York Times
Document Analysis

(a) Article information

(b) Score visualization

(c) Concept refinement

(d) Concept selection

(e) Comments ranked by concept
Inside the Republican Party’s Desperate Mission to Stop Donald Trump

Despite all the forces arrayed against Mr. Trump, a paralytic sense of indecision and despair has prevailed.

2016-02-28T00:00:00+00:00

Presidential Election of 2016; Trump, Donald J; Republican Party; Christie, Christopher J; Kasich, John R; Rubio, Marco; United States Politics and Government

By ALEXANDER BURNS, MAGGIE HABERMAN and JONATHAN MARTIN

U.S., Article

Rank comments by

This bar shows the weights for this ranking.

Search for...

Kevin O’Brien
Park City, UT, Feb 27, 2016 6:17:09 PM
I see no hope for a GOP evolution in 8 years. Maybe 20 years.

n/a CParis
New Jersey, Feb 27, 2016 1:40:48 PM
Agree 100%. These guys are fools if they think voters won’t remember all of this nonsense when it comes time to vote in November. Robots, pants-wetting? This is worse than a bunch of first graders.

PE
Seattle, WA, Feb 27, 2016 1:13:34 PM
Cruz and Rubio tack so Tea Party right, so irrationally “conservative” they represent business as usual Republicans. Trump brings moderates in with his hedge on healthcare and vagueness on Planned Parenthood. Rove is wrong to think that a candidate
Trump exudes leadership, we haven't 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.

Why do those with the loudest complaints always 'belong' to a party? Too bad more voters don't think for themselves.

The first high profile defection to the Republicans was Strom Thurmond. Of course, by your estimation that had nothing to do with racial politics, right?
Temporal Trend

Immigration-related
Immigration topic rise

Time
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

goeasyonu
Trump exudes leadership, we haven't 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

goeasyonu
Being a democrat, aren't you more knowledgeable of the dem party than the repub party? Why don't you 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 don't think for themselves.

Score = 0.00011201416478632322
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
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 Thesaurus.com
Evaluation

**Lab Experiment**
- Lexicon-building interface
- Wordnet and Thesaurus.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.
Meet the Team

Seungyeon Kim
sylund@gmail.com

Nicholas Diakopoulos
@ndiakopoulos

Jaegul Choo
jchoo@korea.ac.kr

Niklas Elmqvist
@NEElmqvist

Deokgun Park
@intuinno

Jurim Lee
jurim0301@korea.ac.kr
Questions?

Try demo at: Conceptvector.org
intuinno@umd.edu