

US-SERBIA & WEST BALKAN DATA SCIENCE WORKSHOP

Belgrade, Serbia | August 26-28, 2018

A pilot cognitive computing system to understand immunization programs

Marija Stanojevic¹, Fang Zhou¹, Sarah Ball², William Campbell², Alison Thaung², Jason Brinkley², Stacie Greby³, Alexandra Bhatti³, Allison Fisher³, Yoonjae Kang³, Cynthia Knighton³, Pamela Srivastava³, Zoran Obradovic¹

Abstract

US national, state, local, and territorial immunization programs use quantitative and qualitative data to ensure vaccinations are provided to prevent diseases. The results of qualitative data analysis are not always available to improve vaccination coverage because of analysis is labor intensive. The Immunization Program Cognitive Computing System (IPCCS) was developed to analyze qualitative data for the Centers for Disease Control and Prevention (CDC).

Data

Text from a variety of formal and informal sources was collected to develop the IPCCS lexicon. Formal data included policy documents, vaccine-related websites, scientific journals, textbooks, and state vaccination-related laws. Informal data included Sysomos searches of Twitter, online forums news, and social media feeds from November 2016 to May 2018.

Problem and results description

Main challenges to IPCCS development included: 1) collection and matching spatio-temporal information of formal and informal data, 2) data cleaning and pre-processing to remove references to external documents, non-relevant data, jargon, typos, and misspellings, and 3) fast and well-performing word and paragraph searching. To address these challenges, data were iteratively and thoroughly cleaned and filtered, the best existing algorithms for text understanding were used, and a new algorithm for paragraph searching was developed. Customized features for the lexicon were developed to ensure that the results of the IPCCS are useful to vaccine domain researchers (e.g., ranking of US states to show representation extent of search phrase).

Lexicon creation

Glove algorithm builds matrix of words co-occurrences X_{ii} and then optimizes cost function:

$$\hat{f} = \sum_{i,j} f(X_{ij}) (w_i^T \bar{w}_j - \log X_{ij})^2$$
 where $w \in \mathbb{R}^d$ are word vectors

and $\bar{w} \in \mathbb{R}^d$ are context word vectors and $f(x) = \{ (x/x_{max})^{3/4}, if x < x_{max}, 1, otherwise \} \}$

This function prevents influence from large co-occurrence of common words (e.g., "on the", "at that", "he is"). The glove algorithm generally shows better results than word2vec [3], but we found that algorithm was not always stable.

WTM model is a purely statistical model which combines ideas from latent Dirichlet allocation (LDA) and word2vec algorithms to learn both representation of words and topics simultaneously. The WTM model generally shows better results than word2vec [4], but it is more complex model and slower at producing output.

Conclusions

Evaluation of word2vec/doc2vec algorithms showed the best outcomes were given by continuous bag of word with negative sampling followed by PV-DM on both datasets. References used different vectors lengths – number of features to describe words. For

Data preprocessing

Only textual information were used for creating lexicons, even though quantitative data such as location, time, and other features were available. Text was cleaned from links, user-names, special characters (except dash and apostrophe symbols), and numbers and then split into words/tokens. Stop-words were removed from tokens. Jargon terms, connected words, and misspellings were kept in the data to allow describing routinely used language immunization to be used in future analysis.

Lexicon creation

Several algorithms and related versions were used to develop the lexicon: word2vec [1] and its variation doc2vec [2], glove [3] and WTM [4]. Word2vec algorithm is a small neural network with very little parameters that allows fast

Qualitative (textual) data preprocessing Interpunction removal Stop word r<u>e</u>moval Lower case transformation Stemming **Tokenization**

PROJECTION

SUM

w(t-2)

w(t-1)

OUTPUT

versions of word2vec algorithm, vectors of length 50 gave best results, even though recommended setting in the references are different. Change in vector size for glove algorithm wasn't that crucial as in word2vec algorithms.

Analysis showed that different aspects of similarity are brought to the top by different algorithms. The word2vec result was usually better in understanding relationships between query and paragraph, while glove algorithm results were more influenced by frequency of occurrence of the most significant words in query and paragraph.

References

[1] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119). [2] Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning (ICML-14) (pp. 1188-1196).

[3] Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Formal lexicon - words results

Social media lexicon - words Social media lexicon - queries Formal data lexicon - words Formal data lexicon - queries Logout Formal data lexicon - words Query: vaccine OR exemption OR requirements Results number: 10

Location: United States Search WTM Word2vec Glove 1. nonmedical: 0.83 1. et: 3.48 1. religious: 0.89 2. philosophical: 0.8 2. al: 3.47 2. required: 0.86 3. religious: 0.8 3. vaccination: 3.07 3. philosophical: 0.85 4. waiver: 0.73 4. required: 3.05 4. nonmedical: 0.84 5. religious: 2.62 5. objection: 0.73 5. philosophic: 0.84 6. philosophic: 0.72 6. regulation: 2.58 6. conscientious: 0.83 7. 1: 2.54 7. conscientious: 0.71 7. must: 0.83 8. process: 0.7 8. health: 2.46 8. homeschooling: 0.82 9. 2: 2.44 9. homeschool: 0.82 10. 2009: 2.42 10. timetable: 0.82

Informal lexicon - paragraphs results nformal lexicon - words Informal lexicon - paragraphs Formal lexicon - words Formal lexicon - paragraphs Logo Informal lexicon - paragraphs uery: mumps or measles cases and outbre Results number: 10 End date: 05/02/201 Start date: 06/01/201 Word2ved 1. RT @GodlessApeMan: Dumbest? #YEC #MAGA #ResearchFlatEarth . The Flu Vaccine Is Working Better Than Expected, C.D.C. Find SOURCE New York Blood Center NYBC Urges the Public to Donate to #AntiVax: 0.9679; tps://t.co/w0apyR12lv: 1.3159 munity's Critically Low Blood Supply NEW YORK To Link; Time: 2018-02-19 21:17:41; Sentiment: NEUTRAL; Auth ink; Time: 2018-02-16 16:09:22 ation on how to organize a blood drive: Call Tol J5 Project: Reach: 504 uckyball360; Reach: 33 enter.org/blood (Please see . RT @D_J_Oz4: @CDCgov @OANN STOP telling ppl to get their flu shots a safe blood supply, a seven-day inventory of all types must be continually It's only going to #KILL people who HAVE the flu! They've ALREADY Link; Time: 2017-04-05 16:41:26; Sentiment: NONE; Author: eplenished. Companies, organizations, and community groups are also the point? #RefugeeCrisis Side Effects: 1.3171 build the blood supply. Hosting a blood drive is easy, and NYBC staff will RT @bengoldacre: Just to reiterate: Andrew Wakefield, struck-off fraudulent

iuLab4Virology; Reach: 168

naMorphis; Reach: 8231

RAdultImmDoc; Reach: 1642

learning of vector representation of each word. It has four different versions distinguished by sampling of output (hierarchical softmax or negative sampling) and by learning architecture (continuous bag of words which is learning word from context and skip-gram which is learning context from word).

Doc2vec is also able to learn vectors of documents. It has two versions based on architecture that can learn word from paragraph and words (PV-DM) or can learn words from paragraph (PV-DBOW). All of those six versions were evaluated on both formal and informal datasets using different number of vector sizes.

[4] Fu, X., Wang, T., Li, J., Yu, C., & Liu, W. (2016, November). Improving Distributed Word **Representation and Topic Model** by Word-Topic Mixture Model. In Asian Conference on Machine Learning (pp. 190-205). Acknoledgment This research was supported by

the Centers for Disease Control through project Pilot to Assess **Cognitive Computing to Analyze** Immunization Program Data.

p you every step of the way. O negative blood donors are considered ed most readily in trauma situa ss the country. Due to its high demand. C e blood is in short supply, and NYBC encourages individuals with this ally low level, with under a two-day supply of O tive. B negative, and A negative blood. As we head into the summe nation. The best preparation for an ng blood on hospital shelves in advance. This is By spreading the word or even hosting your blood drive, inviting friends, family, and community organizations, you save lives in your community," said Andrea Cefarelli Historically, during s have had to focus on building up the ty's blood supply, as it tends to diminish due to seasonal factors

Word2ved

1. Virginia: 1.214

2. Georgia: 1.19

3. Illinois: 1.164

4. Iowa: 1.141

5. Minnesota: 1.08

District of Columbia: 1.0.

8. West Virginia: 0.95

10. Philadelphia: 0.544

6. Nevada: 1.06

9. Alabama: 0.9

9. safe: 0.69

10. pbe: 0.69

Link: Time: 2018-02-18 04:31:55: Sentiment: NEGATIVE: Author: anti-vaccine godfather, is at Trump's inaugural ball ...: 0.9741; wlove9: Reach: 2335 ink; Time: 2017-01-21 10:10:28; Sentiment: NEGATIVE; Author RT @ChelseaClinton: Ironic that George Washington is on an anti-science

RT @doritmi: #BeHPVfreeFL #FLImmysummit Dr. Michael Brown: we car uarantee no flu if get #vaccines. Working on it - none yet. But better the

Link: Time: 2017-02-03 21:15:37; Sentiment: NEUTRAL; Author ARAdultimmDoc: Reach: 1642

. I added a video to a @YouTube playlist https://t.co/VWXRyuHfwl Undying Mythology of Tetanus - Dr Tim O'Shea: 1.3242; ink; Time: 2018-01-03 23:01:15; Sentiment: NEUTRAL; Author SteveCherry3000; Reach: 124

. RT @MicrobiomDigest: The Flu Vaccine Is Working Better C.D.C. Finds https://t.co/rFiXQ0Rxil: 1.3259; Link: Time: 2018-02-16 00:10:44: Sentiment: POSITIVE: Autho

Formal lexicon - results aggregated by state

anti-vax banner - he had the Continental Army vaccinated agains ...: 0.9758;

. RT @NFIDvaccines: #FF During #HeartMonth #GetVaccinated to #FightFlu

@Texas_Heart @H_eHA @everettclinic @DrBGellin @iCubed_URI...

Link; Time: 2017-02-10 20:03:30; Sentiment: NEUTRAL; Author

Spies, Scandals, and naughty liaisons: A RECKLESS REDEMPTIC

Link; Time: 2017-03-31 21:16:10; Sentiment: NEGATIVE; Author

Social media lexicon - queries Formal data lexicon - words Formal data lexicon - queries Log

Formal data lexicon - queries Results number: 5 • Location: United States WTM 1. Virginia: 1.177 1. Georgia: 1.14 2. Iowa: 1.132 2. Iowa: 1.109 8. Nevada: 1.118 District of Columbia: 1.0 4. Virginia: 1.11 4. Nevada: 1.067 Minnesota: 1.0 5. Georgia: 1.044 6. West Virginia: 1.04 7. Minnesota: 1.04 . West Virginia: 0.96 8. Illinois: 0.964 8. Illinois: 1.029 9. Alabama: 0.937 9. Alabama: 0.922 10. Philadelphia: 0.672 10. Philadelphia: 0.445