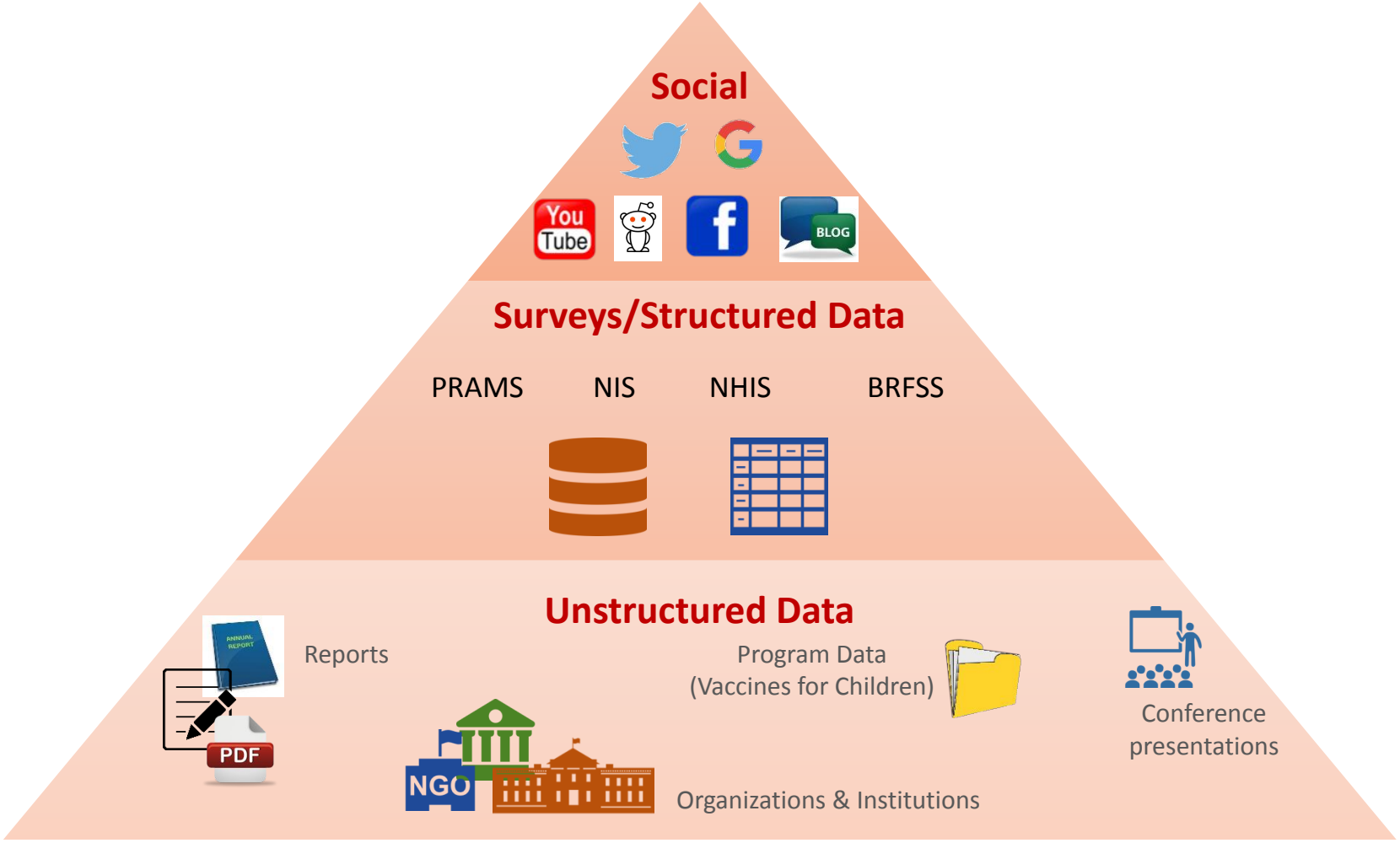
PILOT TO ASSESS COGNITIVE COMPUTING TO ANALYZE IMMUNIZATION PROGRAM DATA

Marija Stanojevic
marija.stanojevic@temple.edu

Data Analytics and Biomedical Informatics Center,
Computer and Information Sciences Department,
Temple University, Philadelphia

Data sources



Data

Informal data

- Twitter
- News articles
- Forums
- Other social media

Formal data

- VFC policies and procedures
- NIS data
- VaxVIew data
- School vaccination requirements
- Westlaw data
- Books and papers

Quantitative data preprocessing

Generalized logistic (GL) normalization (Cao et al., 2016)

Structured feature selection (Ghalwash et al., 2016)

Qualitative (textual) data preprocessing

Interpunction removal

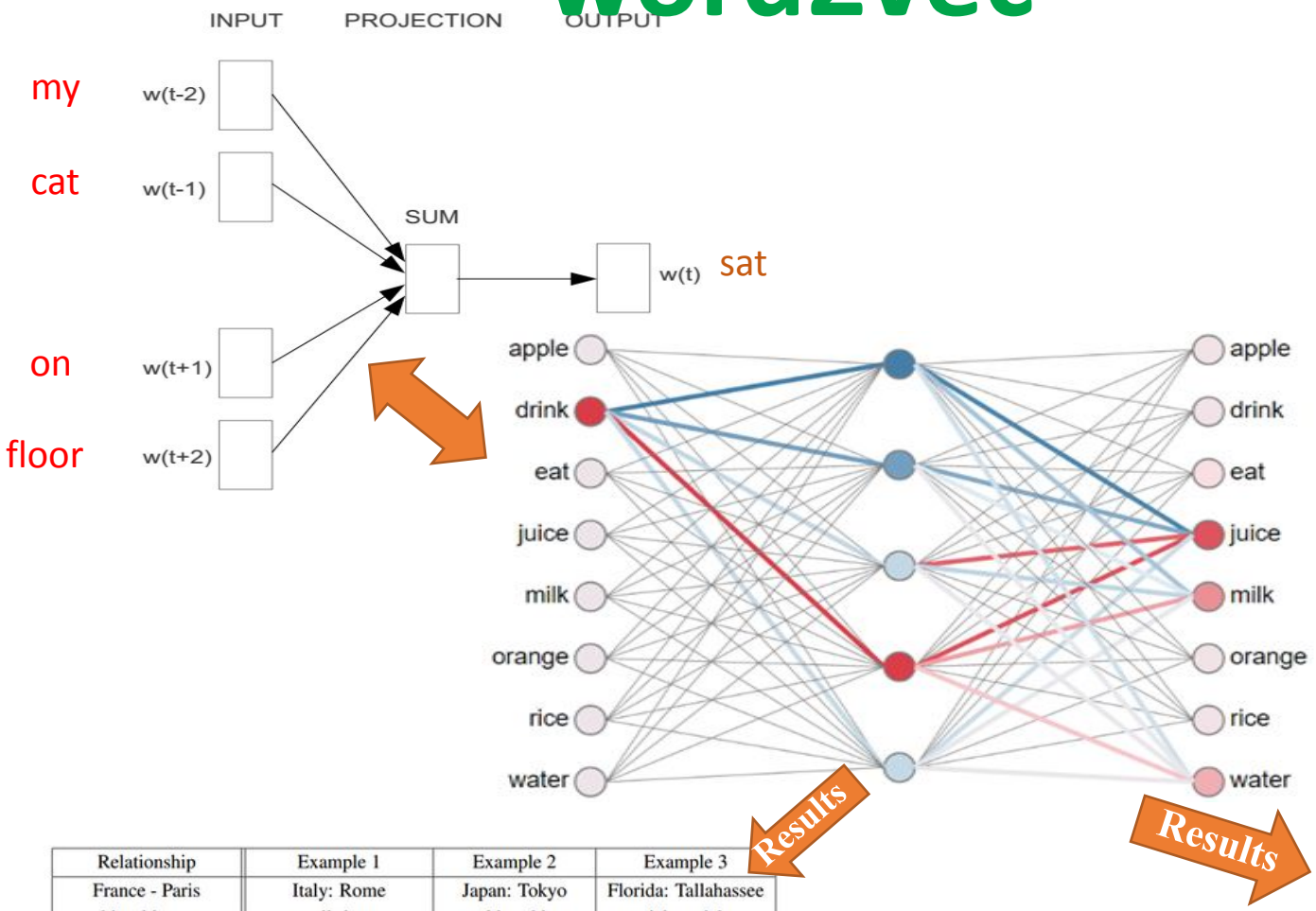
Stop word removal

Lower case transformation

Stemming

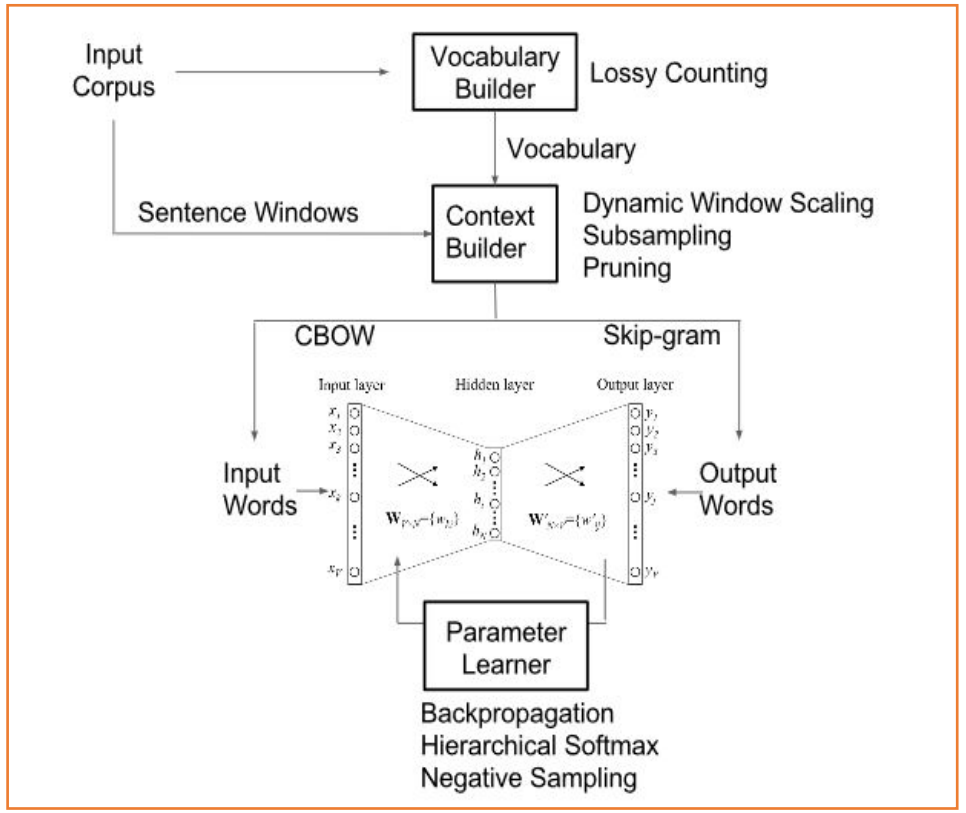
Tokenization

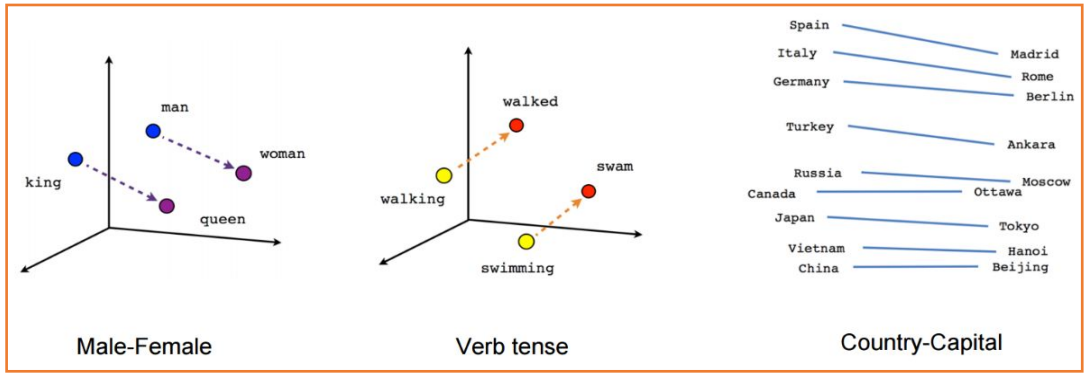
Lexicon Creation – word2vec



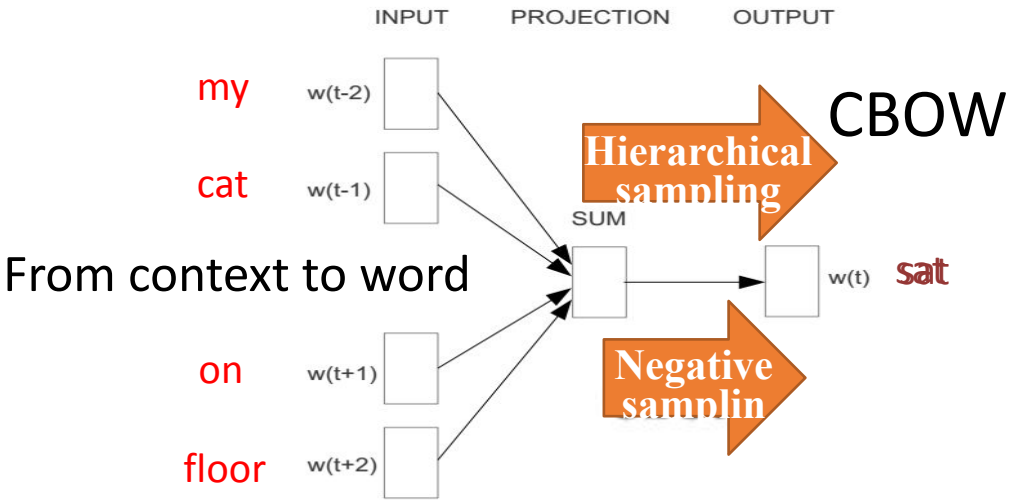
Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

word2vec algorithm details



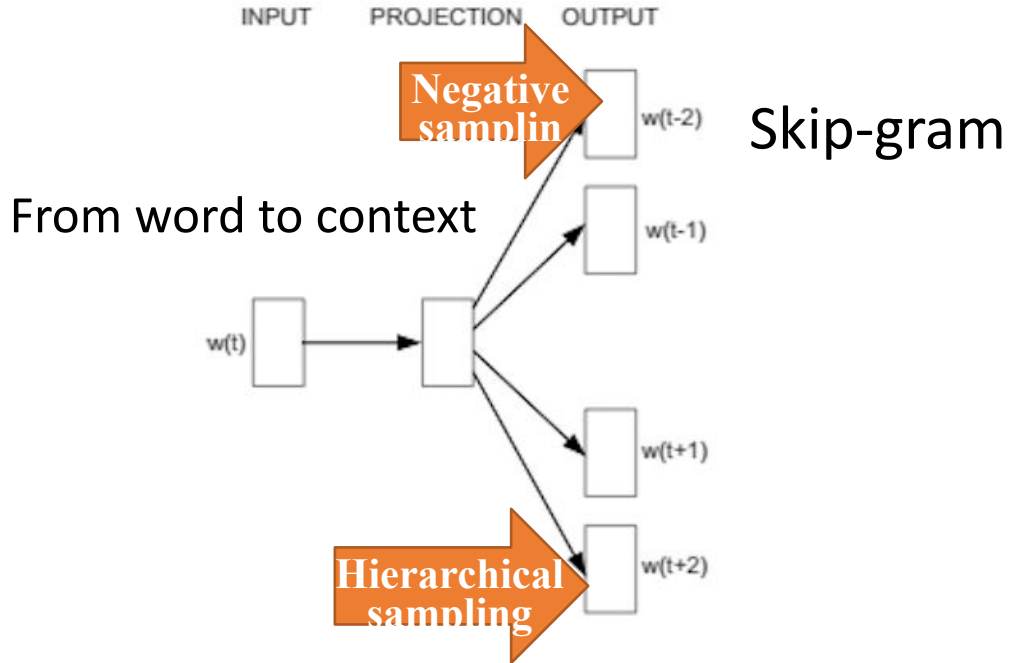


Word2vec algorithm developed 4 types



CBOW

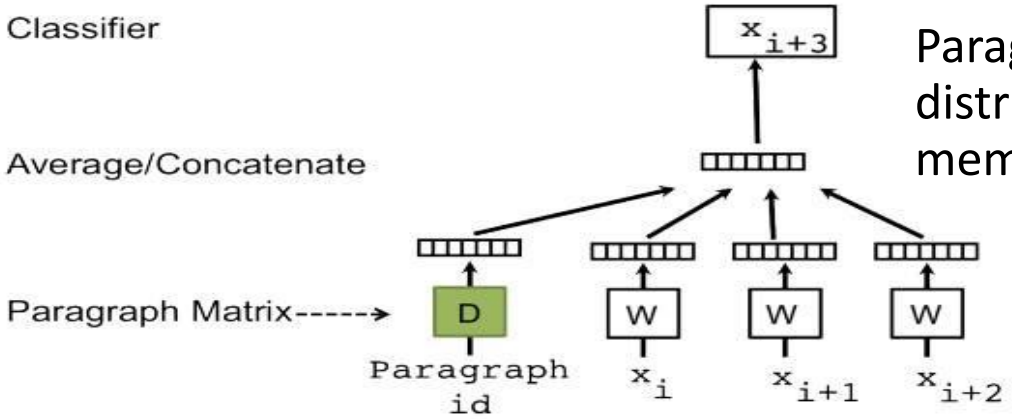
From context to word



Skip-gram

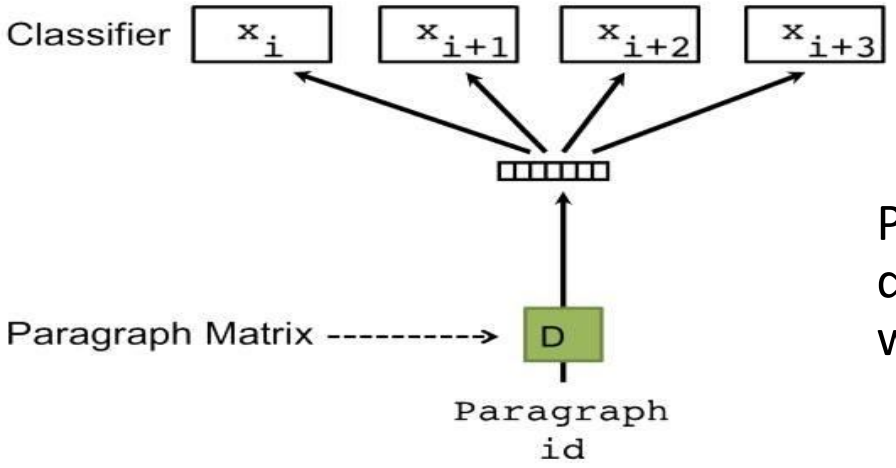
From word to context

Doc2vec algorithm developing 2 types



Paragraph vector: distributed memory (PV-DM)

From paragraph and previous words to next word



words

Paragraph vector: distributed bag of words (PV-DBOW)

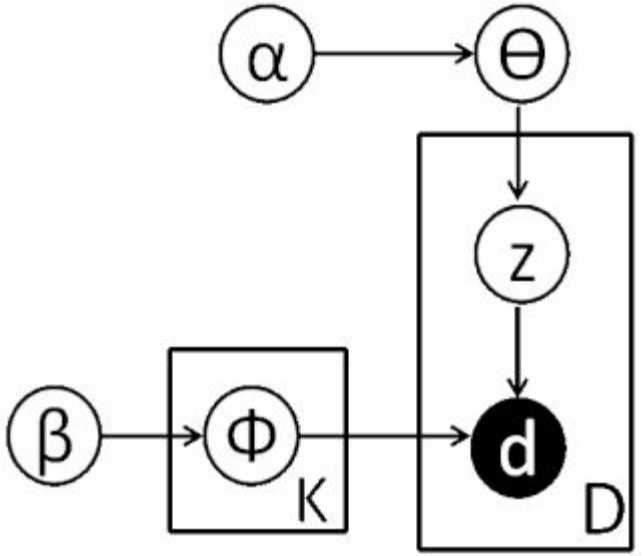
From paragraph to words

Lexicon Creation - GloVe

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

- Statistical method based on idea that word that make sense in the same context will have higher probability of mutual appearance in text
- Better results than word2vec, but less stable
- Combining ideas from word2vec and matrix decomposition
- Will be applied on formal and informal textual data
- Optimizes function: $\hat{J} = \sum_{i,j} f(X_{ij})(w_i^T \tilde{w}_j - \log X_{ij})^2$

Lexicon Creation
DMM – Dirichlet Multinomial Mixture



- Fully statistical method
- Automatically clusters documents
- Very efficient for large volume of data
- Will be applied on both formal and informal data

- Based on the mixture of Dirichlet and multinomial distributions, ie. multinomial distribution parameters are determined using Dirichlet distribution, then multinomial distribution is used to model document

Welcome to the lexicon

Please, enter username and password below to be able to access the lexicons

Username:

Password:

Social media lexicon - words

Formal data lexicon - words

Social media lexicon - queries

Formal data lexicon - queries

Social media lexicon - queries

New York: Philosophical exemption; 15th September 2017

Word2vec

1. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed eu euismod mi. Donec tortor est, elementum eget tempus feugiat, maximus vitae purus. Nulla ut quam a nisi rhoncus imperdiet. (New York, 15th September 2017, positive, 5 retweets)
2. Pellentesque pulvinar sapien sollicitudin enim luctus, nec bibendum magna pulvinar. Sed at egestas risus. (New York, 15th September 2017, positive, 1000 retweets)
3. Nam porta sagittis ullamcorper. Nam sollicitudin sapien sed elementum feugiat. Sed venenatis urna at ligula volutpat malesuada. (New York, 15th September 2017, negative, 0 retweets)
4. Nunc accumsan nisl est, eu pharetra neque pellentesque eu. Praesent elementum maximus eleifend. Sed quis fermentum sapien. Curabitur at tellus mattis, rhoncus lacus eget, molestie neque. (New York, 15th September 2017, negative, 2349 retweets)
5. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Phasellus ultrices, sem non tristique ultricies, lorem ligula finibus mauris, at aliquam ex diam at urna. (New York, 15th September 2017, positive, 5248 retweets)
6. Praesent tincidunt, tellus id imperdiet rutrum, velit justo cursus enim, sed lacinia nisl dui in magna. (New York, 15th September 2017, negative, 56 retweets)

Glove

1. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed eu euismod mi. Donec tortor est, elementum eget tempus feugiat, maximus vitae purus. Nulla ut quam a nisi rhoncus imperdiet. (New York, 15th September 2017, positive, 5 retweets)
2. Pellentesque pulvinar sapien sollicitudin enim luctus, nec bibendum magna pulvinar. Sed at egestas risus. (New York, 15th September 2017, positive, 1000 retweets)
3. Nam porta sagittis ullamcorper. Nam sollicitudin sapien sed elementum feugiat. Sed venenatis urna at ligula volutpat malesuada. (New York, 15th September 2017, negative, 0 retweets)
4. Nunc accumsan nisl est, eu pharetra neque pellentesque eu. Praesent elementum maximus eleifend. Sed quis fermentum sapien. Curabitur at tellus mattis, rhoncus lacus eget, molestie neque. (New York, 15th September 2017, negative, 2349 retweets)
5. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Phasellus ultrices, sem non tristique ultricies, lorem ligula finibus mauris, at aliquam ex diam at urna. (New York, 15th September 2017, positive, 5248 retweets)
6. Praesent tincidunt, tellus id imperdiet rutrum, velit justo cursus enim, sed lacinia nisl dui in magna. (New York, 15th September 2017, negative, 56 retweets)

Dirichlet Multinomial Mixture

1. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed eu euismod mi. Donec tortor est, elementum eget tempus feugiat, maximus vitae purus. Nulla ut quam a nisi rhoncus imperdiet. (New York, 15th September 2017, positive, 5 retweets)
2. Pellentesque pulvinar sapien sollicitudin enim luctus, nec bibendum magna pulvinar. Sed at egestas risus. (New York, 15th September 2017, positive, 1000 retweets)
3. Nam porta sagittis ullamcorper. Nam sollicitudin sapien sed elementum feugiat. Sed venenatis urna at ligula volutpat malesuada. (New York, 15th September 2017, negative, 0 retweets)
4. Nunc accumsan nisl est, eu pharetra neque pellentesque eu. Praesent elementum maximus eleifend. Sed quis fermentum sapien. Curabitur at tellus mattis, rhoncus lacus eget, molestie neque. (New York, 15th September 2017, negative, 2349 retweets)
5. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Phasellus ultrices, sem non tristique ultricies, lorem ligula finibus mauris, at aliquam ex diam at urna. (New York, 15th September 2017, positive, 5248 retweets)
6. Praesent tincidunt, tellus id imperdiet rutrum, velit justo cursus enim, sed lacinia nisl dui in magna. (New York, 15th September 2017, negative, 56 retweets)

Formal data lexicon - queries

Philosophical exemption

☑ Aggregate per state

Word2vec

1. New York, 1
2. Pennsylvania, 0.96
3. Michigen, 0.96
4. West Virginia, 0.95
5. WIsconsin, 0.93
6. Virginia, 0.91
7. Mississippi, 0.89
8. D.C., 0.89
9. Ohio, 0.32
10. Oklahoma, 0.30
11. New Jersey, 0.30
12. New Mexico, 0.22
13. Califormia, 0.19
14. ...

Glove

1. New York, 1
2. Pennsylvania, 0.96
3. Michigen, 0.96
4. West Virginia, 0.95
5. WIsconsin, 0.93
6. Virginia, 0.91
7. Mississippi, 0.89
8. D.C., 0.89
9. Ohio, 0.32
10. Oklahoma, 0.30
11. New Jersey, 0.30
12. New Mexico, 0.22
13. Califormia, 0.19
14. ...

Dirichlet Multinomial Mixture

1. New York, 1
2. Pennsylvania, 0.96
3. Michigen, 0.96
4. West Virginia, 0.95
5. WIsconsin, 0.93
6. Virginia, 0.91
7. Mississippi, 0.89
8. D.C., 0.89
9. Ohio, 0.32
10. Oklahoma, 0.30
11. New Jersey, 0.30
12. New Mexico, 0.22
13. Califormia, 0.19
14. ...

Formal data lexicon - words

phylosophical AND exemption

Word2vec

1. Personal
2. School
3. Medical
4. Religious
5. Allowing
6. Form
7. Signature
8. Signed
9. Forms
10. Vaccine

Glove

1. Personal
2. School
3. Medical
4. Religious
5. Allowing
6. Form
7. Signature
8. Signed
9. Forms
10. Vaccine

Dirichlet Multinomial Mixture

1. Personal
2. School
3. Medical
4. Religious
5. Allowing
6. Form
7. Signature
8. Signed
9. Forms
10. Vaccine

Future research interest

- **Formal data**:
  - Find references in text that mention internal/external source (document, link, image, table)
  - Understand meaning of numbers and dates from context
  - Learn changes law from text
- **Informal data**:
  - Understand spreading of news geographically and temporally
  - Modeling public opinion about certain topic
- **Both**:
  - Automatic evaluation of lexicon
  - Model relation between changes in law and public opinion