Indexing an Outbreak

Using Natural Language Processing to Consolidate Scientific Journal Articles

By Cynthia Chiang, Jobeth Muncy, and Clay Carson

Setting the Stage

Problem Statement:

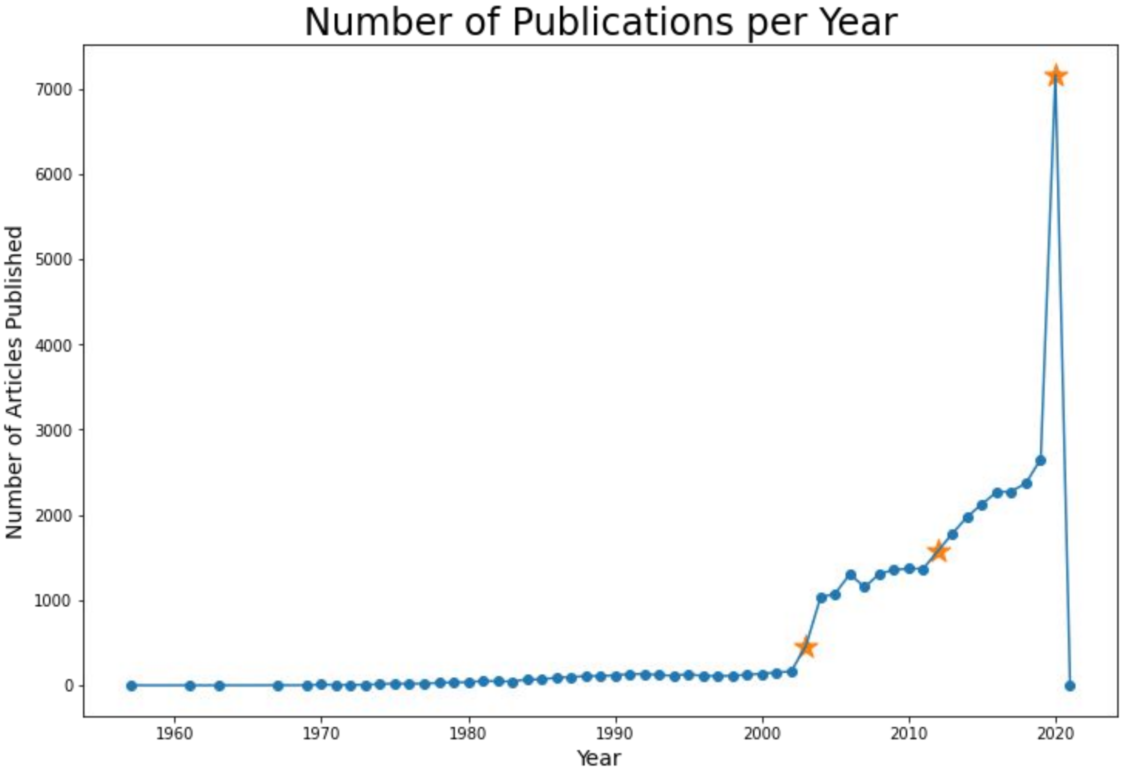
How can we organize decades worth of coronavirus articles in a meaningful and accessible way for researchers to use in their race to create a vaccine?

Data source:

- Compilation of over 63,000 scientific articles about coronaviruses
- Created by the Allen Institute for AI in collaboration with various other institutions
    - Chan Zuckerberg Foundation
    - Georgetown University
    - Microsoft Research
    - National Library of Medicine (NIH)
    - The White House

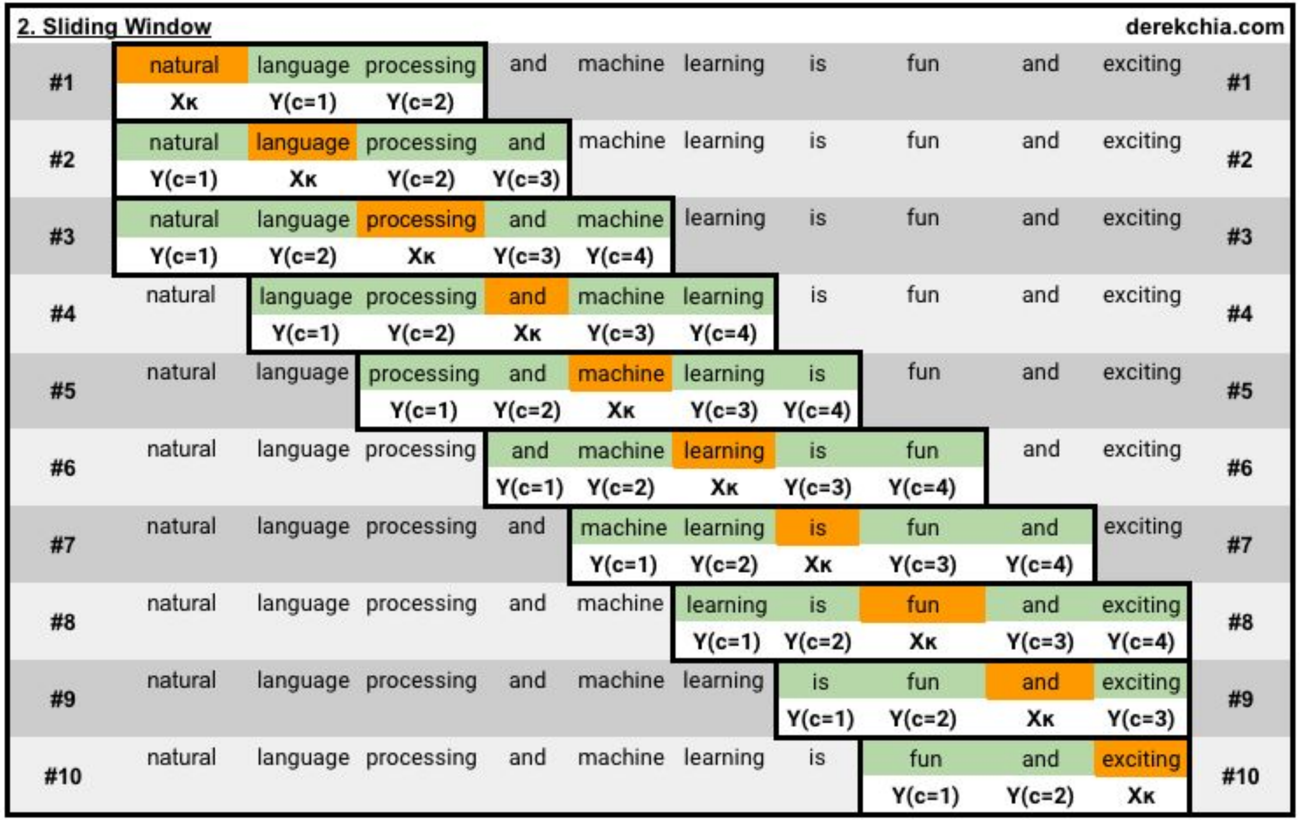
Study of Coronaviruses

- 2003 SARS outbreak
- 2012 MERS outbreak
- Spike in number of publications in the years following a outbreak
- Interest dwindles after about 5 years



Word2Vec

Uses context of the surrounding words and mathematical distances to map out similarity between words

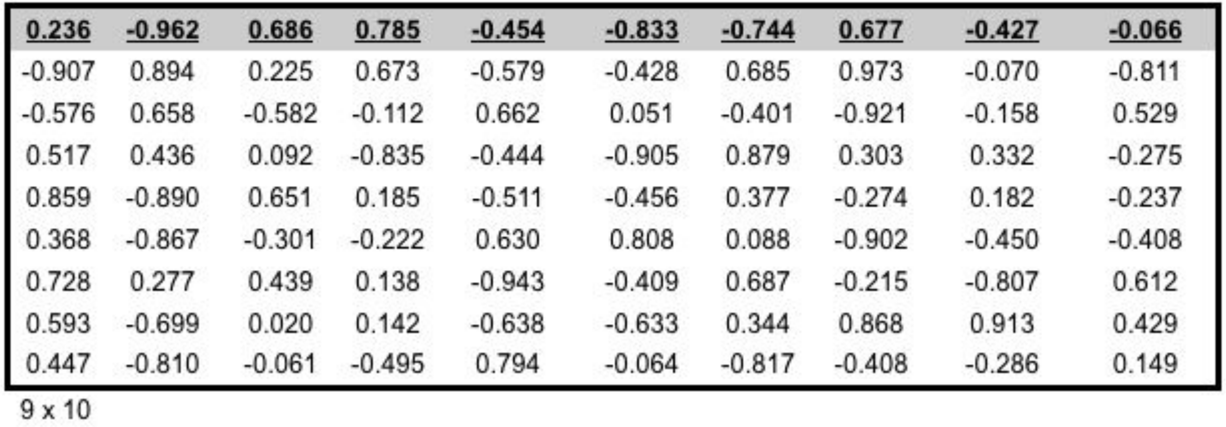


2. Sliding Window derekchia.com

#1	natural Xκ	language Y(c=1)	processing Y(c=2)	and	machine	learning	is	fun	and	exciting	#1
#2	natural Y(c=1)	language Xκ	processing Y(c=2)	and Y(c=3)	machine	learning	is	fun	and	exciting	#2
#3	natural Y(c=1)	language Y(c=2)	processing Xκ	and Y(c=3)	machine Y(c=4)	learning	is	fun	and	exciting	#3
#4	natural	language Y(c=1)	processing Y(c=2)	and Xκ	machine Y(c=3)	learning Y(c=4)	is	fun	and	exciting	#4
#5	natural	language	processing Y(c=1)	and Y(c=2)	machine Xκ	learning Y(c=3)	is Y(c=4)	fun	and	exciting	#5
#6	natural	language	processing	and Y(c=1)	machine Y(c=2)	learning Xκ	is Y(c=3)	fun Y(c=4)	and	exciting	#6
#7	natural	language	processing	and	machine Y(c=1)	learning Y(c=2)	is Xκ	fun Y(c=3)	and Y(c=4)	exciting	#7
#8	natural	language	processing	and	machine	learning Y(c=1)	is Y(c=2)	fun Xκ	and Y(c=3)	exciting Y(c=4)	#8
#9	natural	language	processing	and	machine	learning	is Y(c=1)	fun Y(c=2)	and Xκ	exciting Y(c=3)	#9
#10	natural	language	processing	and	machine	learning	is	fun Y(c=1)	and Y(c=2)	exciting Xκ	#10

Word2Vec

Resulting matrix

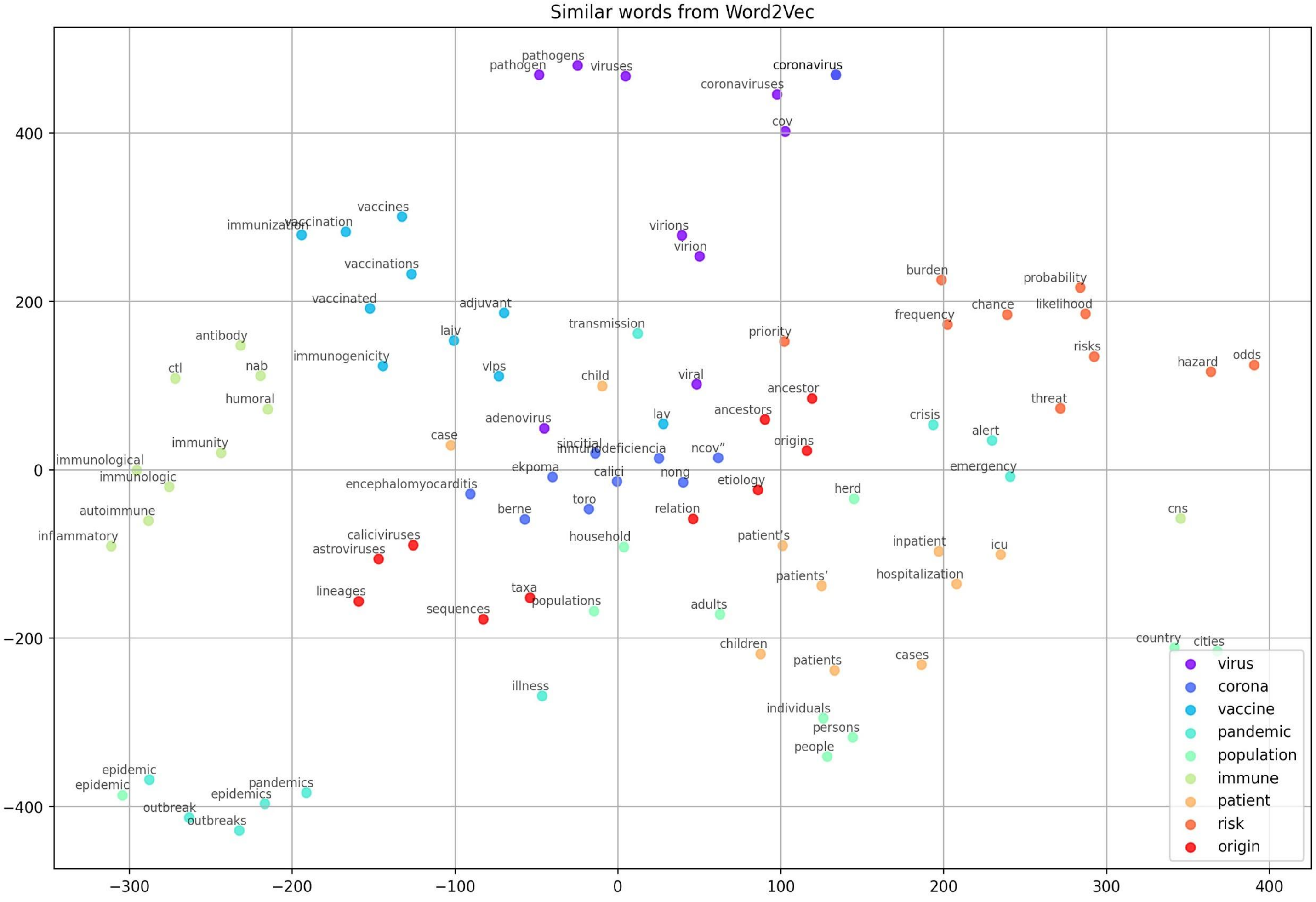


Natural	language	processing	and	machine	learning	is	fun	and	exciting
0.236	**-0.962**	**0.686**	**0.785**	**-0.454**	**-0.833**	**-0.744**	**0.677**	**-0.427**	**-0.066**
-0.907	0.894	0.225	0.673	-0.579	-0.428	0.685	0.973	-0.070	-0.811
-0.576	0.658	-0.582	-0.112	0.662	0.051	-0.401	-0.921	-0.158	0.529
0.517	0.436	0.092	-0.835	-0.444	-0.905	0.879	0.303	0.332	-0.275
0.859	-0.890	0.651	0.185	-0.511	-0.456	0.377	-0.274	0.182	-0.237
0.368	-0.867	-0.301	-0.222	0.630	0.808	0.088	-0.902	-0.450	-0.408
0.728	0.277	0.439	0.138	-0.943	-0.409	0.687	-0.215	-0.807	0.612
0.593	-0.699	0.020	0.142	-0.638	-0.633	0.344	0.868	0.913	0.429
0.447	-0.810	-0.061	-0.495	0.794	-0.064	-0.817	-0.408	-0.286	0.149

9 x 10

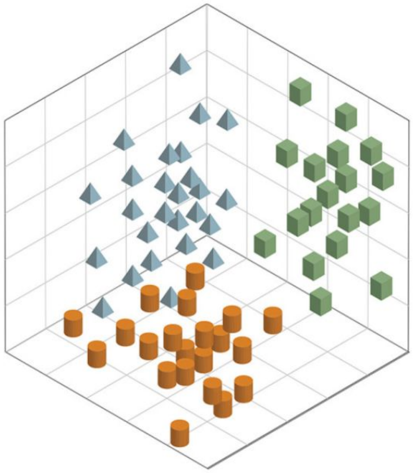
Word2Vec

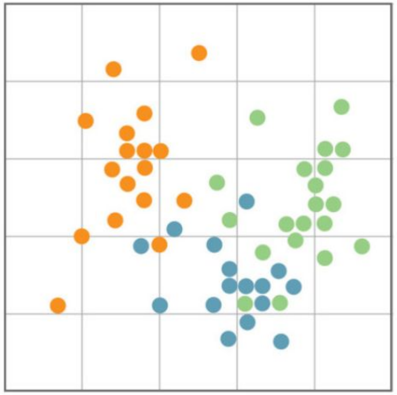
Two dimensional representation of word vectors

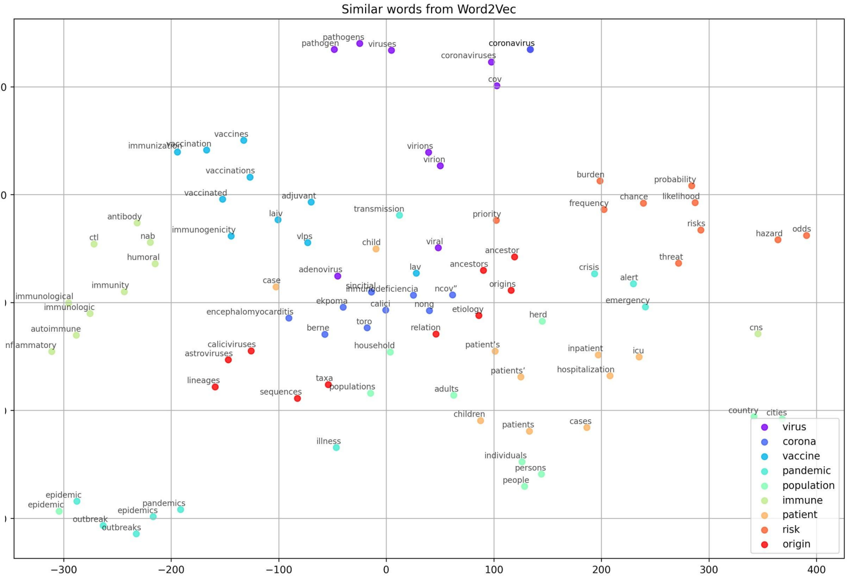


Word2Vec

The dangers of an oversimplified dimension reduction

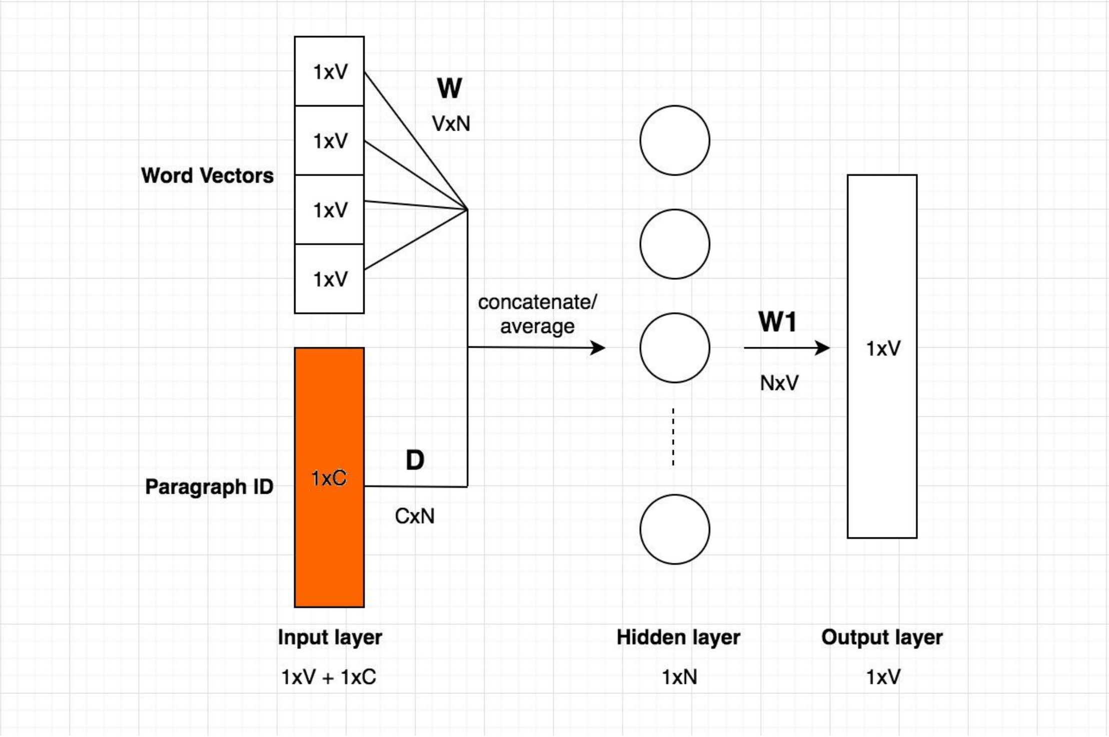


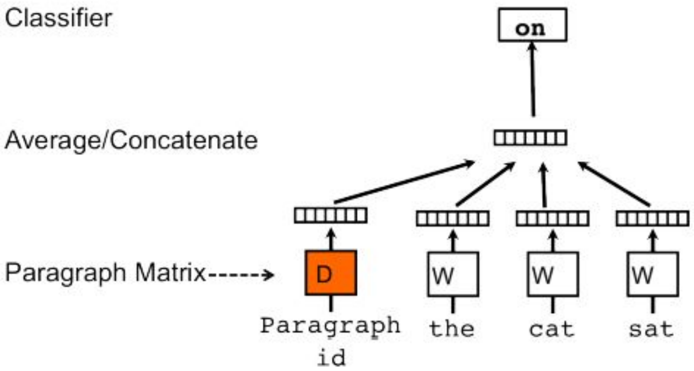




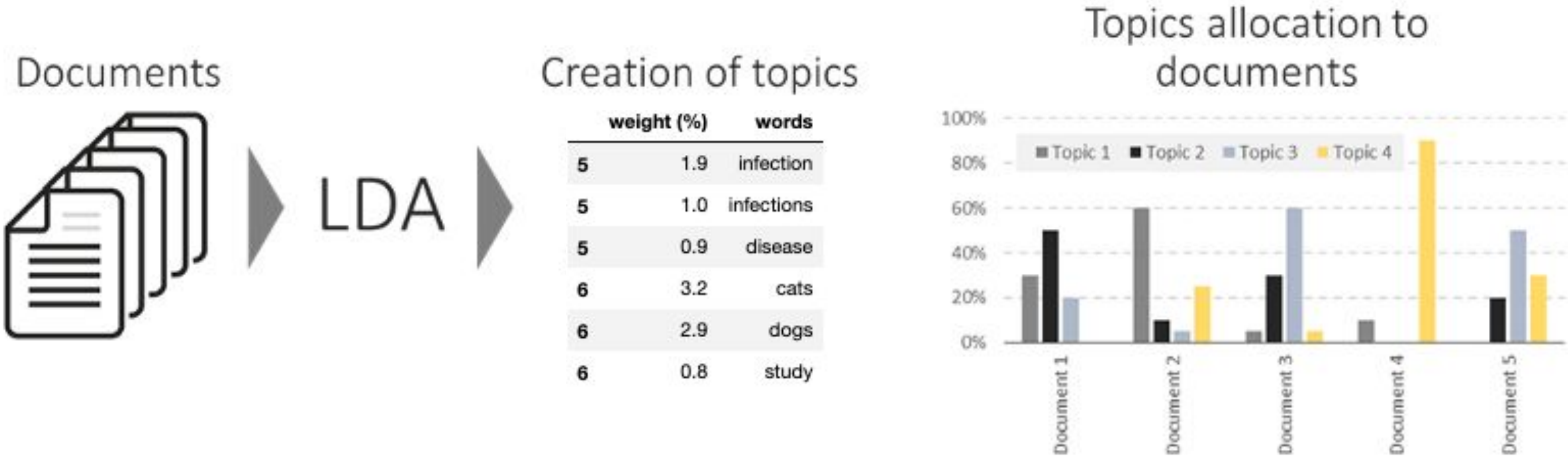
Similar words from Word2Vec

Doc2Vec


1xV
1xV
1xV
1xV
Word Vectors
W
VxN
concatenate/ average
W1
NxV
1xV
Paragraph ID
1xC
D
CxN
Input layer
1xV + 1xC
Hidden layer
1xN
Output layer
1xV


Classifier
on
Average/Concatenate
Paragraph Matrix-----→
D
W
W
W
Paragraph id
the
cat
sat

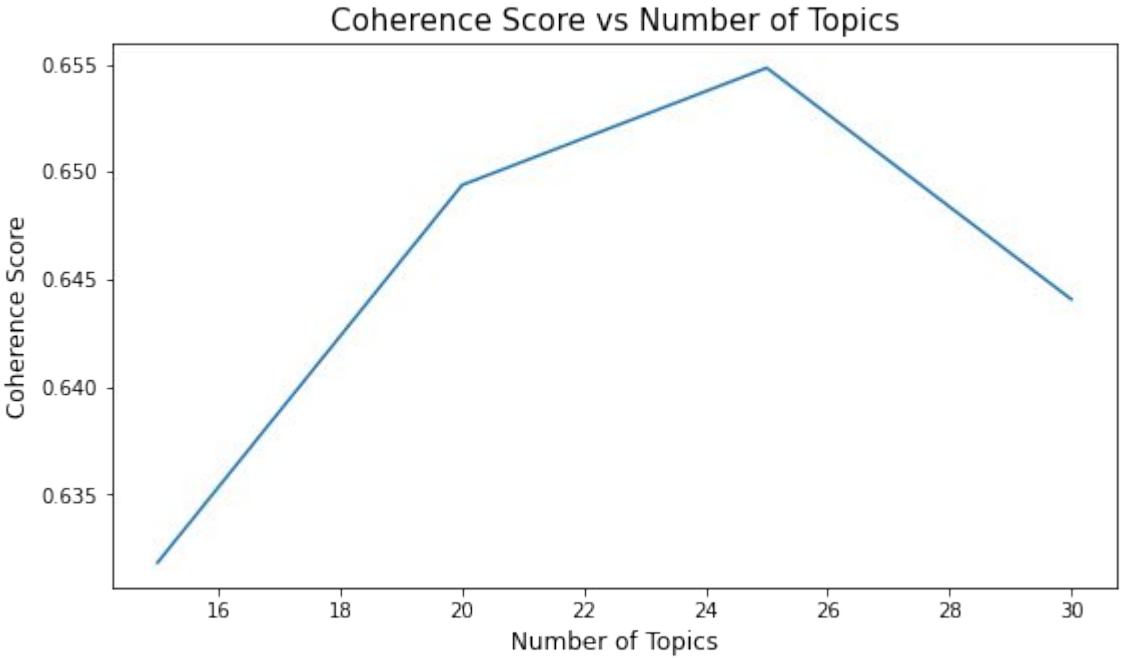
Latent Dirichlet Allocation



	weight (%)	words
5	1.9	infection
5	1.0	infections
5	0.9	disease
6	3.2	cats
6	2.9	dogs
6	0.8	study

Optimal Number of Topics

- Single Topic
- Semantic Similarity



Demonstration

Demonstration

- Show topic pyLDAvis visualization first
- Keyword: transmission (~12 seconds)
- Topic: Severe Outbreaks (~12 seconds, this combo gives the most results)

Conclusion

- Combining different NLP models, we were able to create an application that allows you to search through the articles by keyword or topic
- Researchers can use this resource to more easily find related scientific articles and aid in furthering their research
- Next steps:
    - Increase the vocabulary of our models to include more words and allow searches using more than one keyword
    - Remove stopwords in other languages before training and translating the articles
    - Optimize the search engine

Picture Credits

- https://shuzhanfan.github.io/2018/08/understanding-word2vec-and-doc2vec/
- https://www.kdnuggets.com/2019/09/overview-topics-extraction-python-latent-dirichlet-allocation.html
-