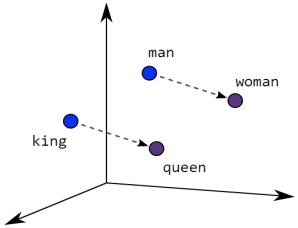
Analyzing East Asian Biases During the COVID-19 Pandemic through Word Embeddings

Lok Chi Hon
lhon@oswego.edu

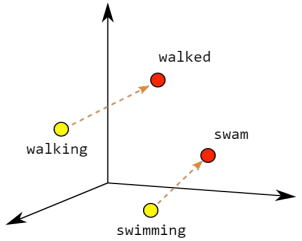
Background

- Anti-Asian hate crimes and incidents have spiked during the COVID-19 pandemic.
- In the United States alone, hate crimes have risen by nearly 150%.
- In NYC, hate crimes have become so frequent that the NYPD had to create an Asian Hate Crime Task Force.
- In January, President Joe Biden highlighted the issue by signing an executive order to direct federal agencies to combat this xenophobia.
- However, it is important to note that Anti-Asian racism and discrimination did not start with COVID-19.
  - 1882: The Chinese Exclusion Act
  - 1942: Executive Order 9066
  - 1965: The Hart-Oellar Act / The 1965 Immigration and Nationality Act
- Even though this increase in hate crimes was concerning, but it took some time for major news outlets to provide proper coverage on this issue.

Word Embeddings

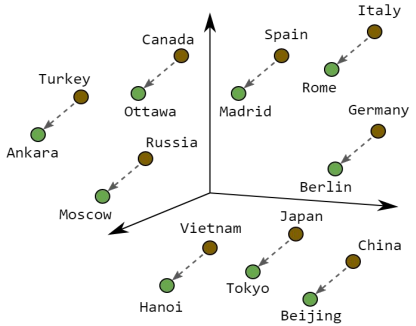


Male-Female



Verb Tense

- In simplistic terms, word embeddings are:
  - Texts converted into numbers
  - Mappings of words to vectors
  - A form of word representation that bridges the human understanding of language to that of a machine language
- Goal of Word Embeddings
  - To reduce dimensionality
  - To use a word to predict the words around it
  - Inter-word semantics must be captured
- Applications:
  - Natural language processing problems
  - Machine learning / Deep learning



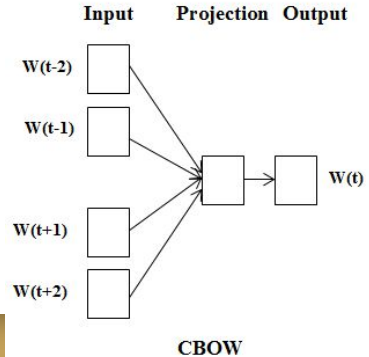
Country-Capital

Gensim Word2Vec Model

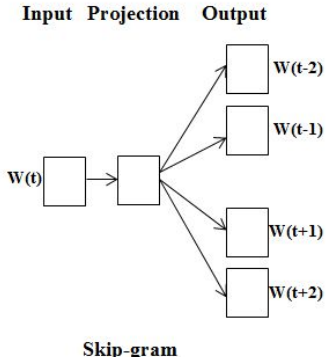
- One of the most popular algorithms/techniques used to learn word embeddings
- Uses a shallow neural network model to learn word associations from a large corpus of text
- Input: a large corpus of text

  Output: a vector space with each unique word in the corpus being assigned a corresponding vector in the space

- CBOW
  - Learns by using each of these contexts to predict the current word
- Skip-gram
  - Learns by using the current word in order to predict its neighbors



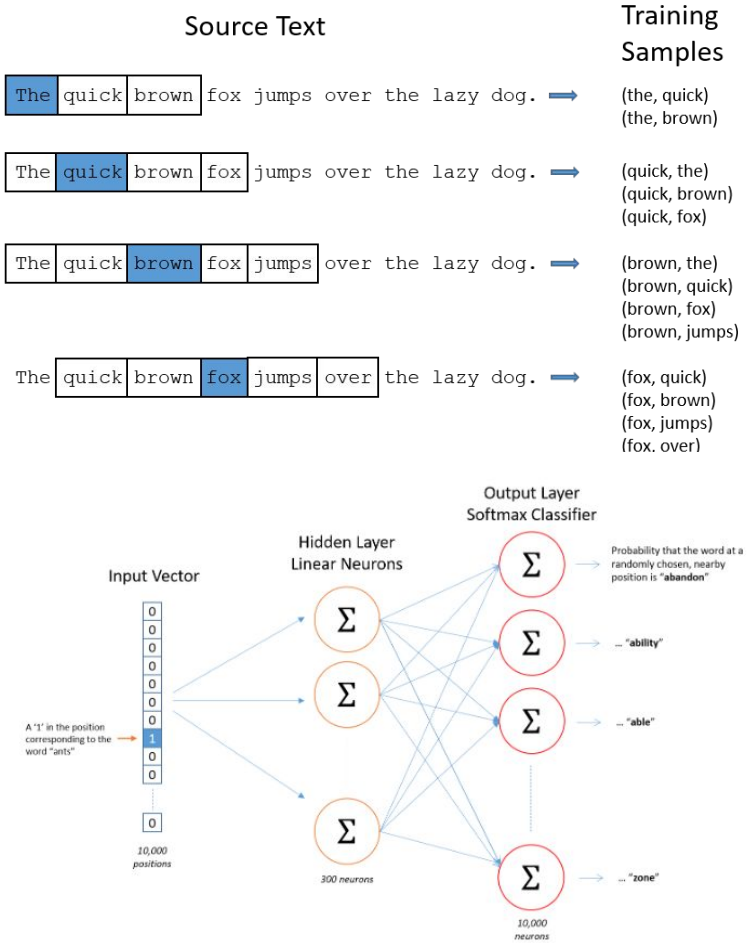
CBOW



Skip-gram

Skipgram Model

1. Takes every word in a text
2. Takes one by one the words surrounding it within a defined “window” to feed to a neural network
3. Neural network assigns weights to each word
4. Output after training is prediction of the probability for each word to appear in the window around the targetword



Project

- Responsibilities:
  - Web scrape and collect data from news sources (New York Times, New York Post, CNN, The Atlantic)
  - Clean and organize data
  - Create and train a skip-gram model
  - Using a predefined list of words, analyze output vectors and their interpretations

- Libraries Used:
  - Nytimesarticle
  - Json
  - Bs4 (BeautifulSoup)
  - Selenium / Chrome Driver Manager
  - Pandas
  - NLTK
  - Gensim.models import Word2Vec
  - Numpy
  - MatPlotLib

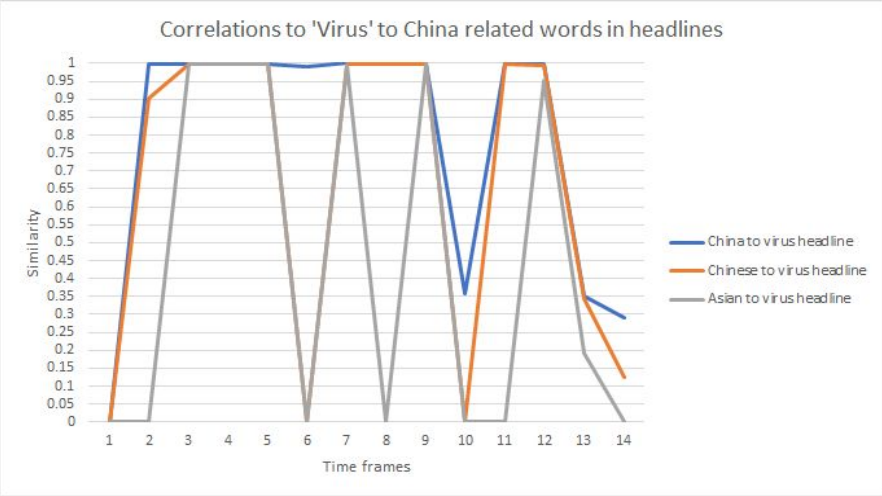
- Number of articles collected:
  - Total: 31493
  - New York Times: 6018
  - New York Post: 10161
  - CNN: 14424
  - The Atlantic: 893
  - Time Frame 1: 75
  - Time Frame 2: 848
  - Time Frame 3: 2187
  - Time Frame 4: 4486
  - Time Frame 5: 6460
  - Time Frame 6: 1188
  - Time Frame 7: 3460
  - Time Frame 8: 3576
  - Time Frame 9: 3644
  - Time Frame 10: 548
  - Time Frame 11: 2364
  - Time Frame 12: 1233
  - Time Frame 13: 595
  - Time Frame 14: 507

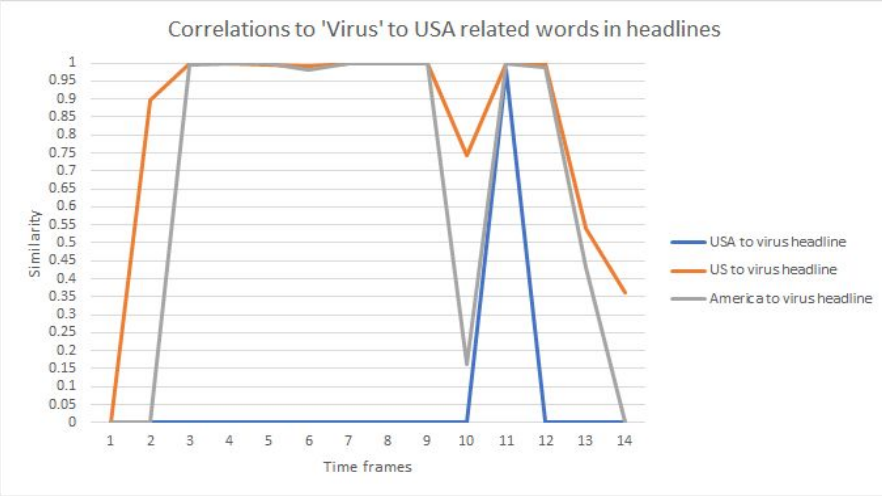
Time Frames

Time Frame	Date	Event
1	1/1/20-1/23/20	
2	1/24/20-2/24/20	Wuhan starts lockdown
3	2/25/20-3/16/20	Italy starts lockdown
4	3/17/20-4/11/20	‘China Virus’ tweet
5	4/12/20-5/28/20	US reported most cases throughout the world
6	5/29/20-6/10/20	US death count passes 100,000
7	6/11/20-7/7/20	US cases reach 2 million

Time Frame	Date	Event
8	7/8/20-8/8/20	US cases reach 3 million
9	8/9/20-9/22/20	US cases reach 5 million
10	9/23/20-10/2/20	US death count passes 200,000
11	10/3/20-11/9/20	Outbreak in White House
12	11/10/20-12/2/20	US cases reach 10 million
13	12/3/20-12/12/20	US record highs in daily deaths, new infections, and hospitalizations
14	12/13/20-12/31/20	Vaccines are approved for emergency use by FDA

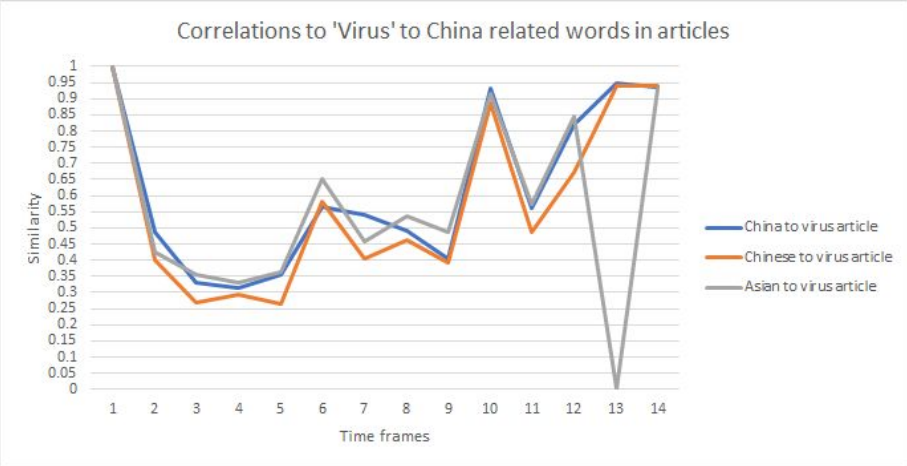
Results: Correlations to the word ‘virus’

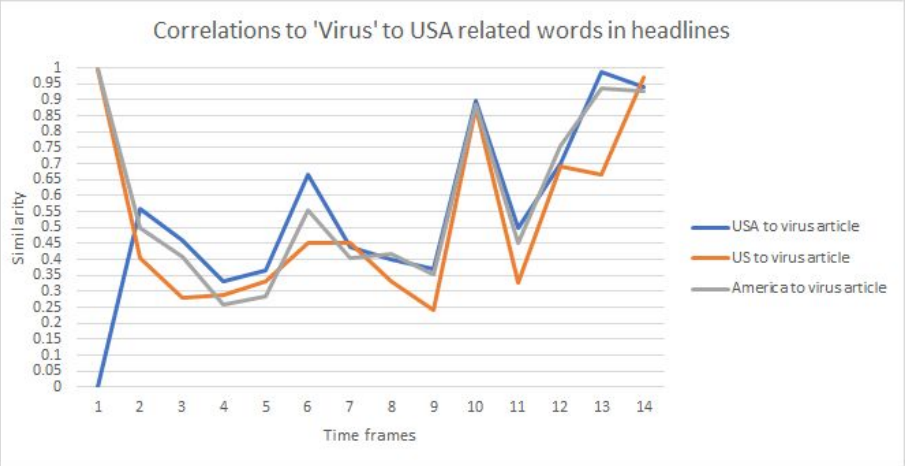




- There seems to be a dip during the 10th time frame.
  - Error in data collection
  - Large event that directed the news attention (less attention/news coverage on COVID-19 related events)

Results: Correlations to the word ‘virus’

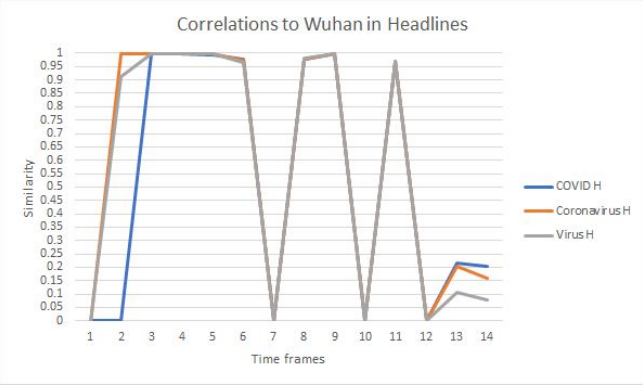


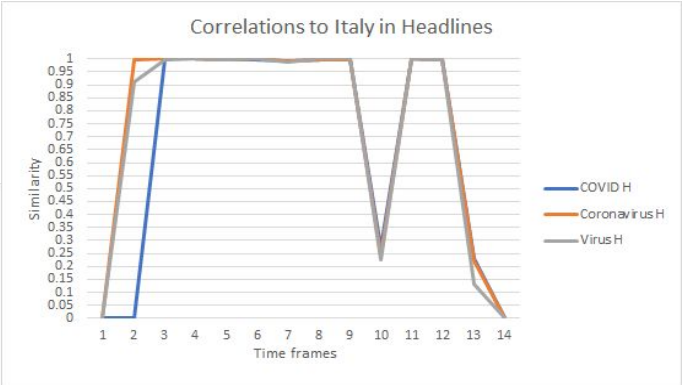


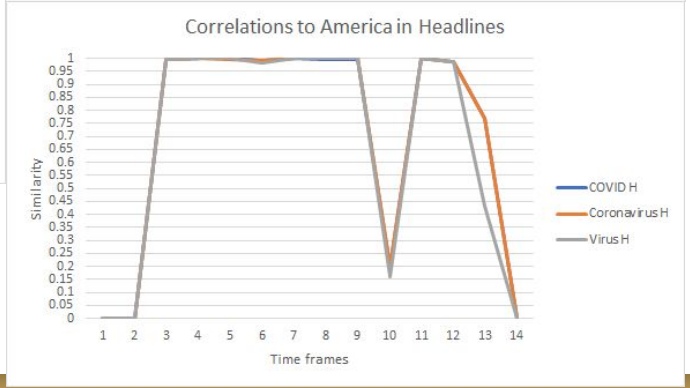
- The articles contained more text (and more words) for the model to train on.
  - In headlines, cosine similarities were mostly close to 0 or close to 1.
  - In articles, cosine similarities varied between 0 and 1.
- Even though the United States was hit with the pandemic a little later than China, trends to the word “virus” seem to be similar both in the headlines and in the articles
- There were slightly higher correlations in relation to the “China related words.”

Results:

Headline correlations to the words ‘Wuhan’, ‘Italy’, ‘America’



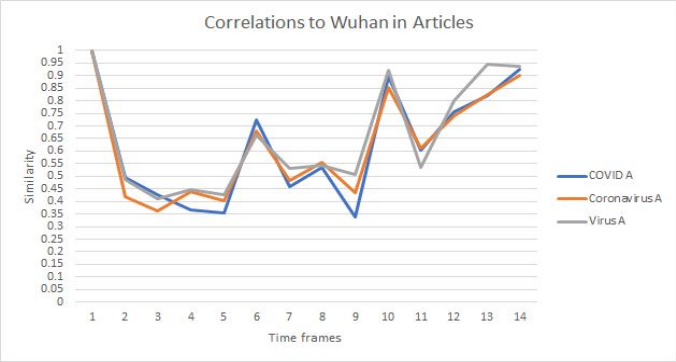


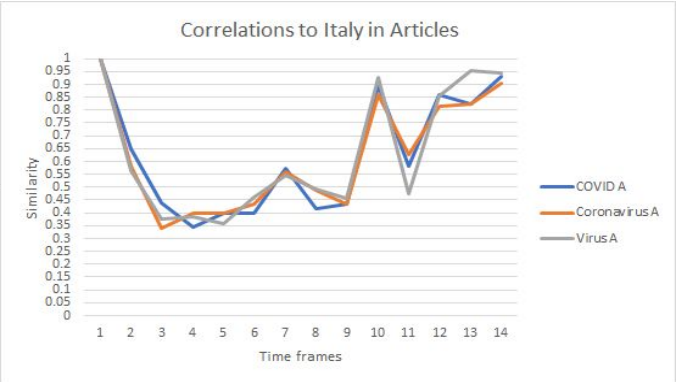


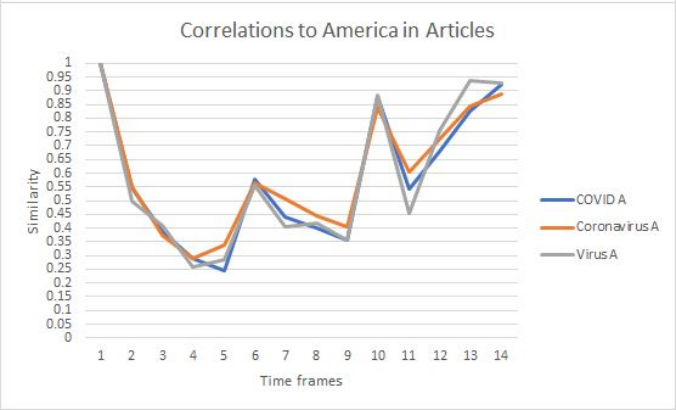
- As mentioned previously, the cosine similarities in headlines tended to be close to 0 or 1.
- Wuhan’s and Italy’s correlation to virus-related words advance to 1 earlier, which is expected because the pandemic hit those places first.
- It seems that near the end of the 2020, Wuhan is still mentioned more than Italy and America in the news (in correlation with the virus).

Results:

Headline correlations to the words ‘Wuhan’, ‘Italy’, ‘America’







- As mentioned previously, the cosine similarities in articles varied more between 0 and 1.
- Wuhan’s correlation is higher earlier on, which is expected as Wuhan was the first place to deal with COVID-19.
- Despite different timelines in the severity of the pandemic in these different locations, the trend seems to be similar. This suggests the American coverage on news of other countries.

Results (continued)

- This research is meant to highlight how biases and prejudices are present everywhere, including what a lot of people may believe to be an impartial and fair.
- It is important to be aware of ingrained biases in society and be active in acknowledging and dealing with them.
- These findings are not complete as I am still in the process of writing my thesis, which I will be submitting before (or on) May 14th.
- There findings are also not 100% accurate and my analysis/observations are also not 100% accurate.

Future Work

- Compare different models to the Google's Word2Vec model (vs Stanford's GloVE model or Facebook's FastText model)
- Gather resources from more news sources
- Gather resources from a larger variety of sources (left-wing / right wing / international)
- Expand research to other major events depicted in the news
    - Black Lives Matter movement
- Expand research to include analysis of images used in news media
- Look at news coverage
    - Look at amount of coverage
    - Look at language used to describe Asian hate crimes

Discussion

Questions?
Comments?

Analogical Findings

- Virus is to China as _____ is to America

all_w2v_hd	predicts	all_w2v_ar	coronavirus
nyt_w2v_hd	congress	nyt_w2v_ar	coronavirus
nypost_w2v_hd	permits	nypost_w2v_ar	bug
cnn_w2v_hd	caribbean	cnn_w2v_ar	said
ta_w2v_hd	coronavirus	ta_w2v_ar	coronavirus

df0_w2v_hd	N/A	df0_w2v_ar	touching
df1_w2v_hd	N/A	df0_w2v_ar	japan
df2_w2v_hd	nyc	df0_w2v_ar	nation
df3_w2v_hd	man	df0_w2v_ar	nation
df4_w2v_hd	adviser	df0_w2v_ar	bug
df5_w2v_hd	nyc	df0_w2v_ar	seeing
df6_w2v_hd	asks	df0_w2v_ar	latin
df7_w2v_hd	linked	df0_w2v_ar	badly
df8_w2v_hd	spike	df0_w2v_ar	terms
df9_w2v_hd	nyc	df0_w2v_ar	every
df10_w2v_hd	government	df0_w2v_ar	afraid
df11_w2v_hd	start	df0_w2v_ar	spread
df12_w2v_hd	daily	df0_w2v_ar	positivity
df13_w2v_hd	N/A	df0_w2v_ar	uk

Analogical Findings

- Virus is to China as _____ is to US

all_w2v_hd	see	all_w2v_ar	said
nyt_w2v_hd	states	nyt_w2v_ar	coronavirus
nypost_w2v_hd	rate	nypost_w2v_ar	bug
cnn_w2v_hd	top	cnn_w2v_ar	said
ta_w2v_hd	coronavirus	ta_w2v_ar	person

df0_w2v_hd	N/A	df0_w2v_ar	four
df1_w2v_hd	N/A	df0_w2v_ar	yet
df2_w2v_hd	coronavirus	df0_w2v_ar	ultimately
df3_w2v_hd	nyc	df0_w2v_ar	fortunate
df4_w2v_hd	lowest	df0_w2v_ar	bug
df5_w2v_hd	coronavirus	df0_w2v_ar	asymptomatic
df6_w2v_hd	could	df0_w2v_ar	covid
df7_w2v_hd	another	df0_w2v_ar	says
df8_w2v_hd	covid	df0_w2v_ar	covid
df9_w2v_hd	covid	df0_w2v_ar	covid
df10_w2v_hd	states	df0_w2v_ar	cdc
df11_w2v_hd	covid	df0_w2v_ar	include
df12_w2v_hd	reports	df0_w2v_ar	friday
df13_w2v_hd	N/A	df0_w2v_ar	percent