Tweet Topics and Sentiments Analysis

Analyze tweet topics and sentiment using Unsupervised machine learning

Cindy Su 11.12.2021

Motivation

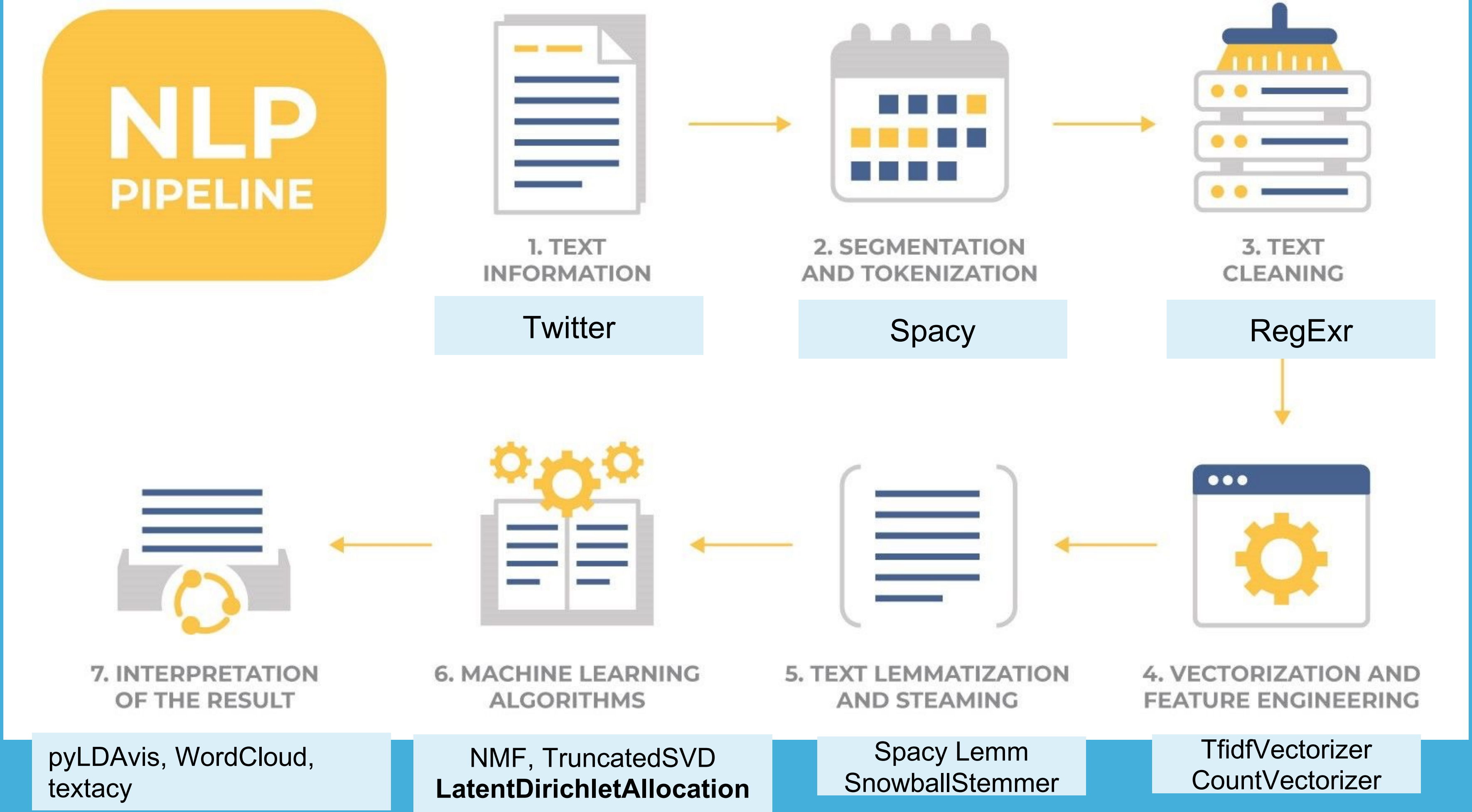
Twitter with 192 million daily active users worldwide.

It contains rich source of data to analyze and understand social behavior.

Goal

- How do people think about the the overall economic environment during the pandemic?

- Whether the sentiment in social media relates to the stock market?



NLP PIPELINE

1. TEXT INFORMATION

Twitter

2. SEGMENTATION AND TOKENIZATION

Spacy

3. TEXT CLEANING

RegExr

4. VECTORIZATION AND FEATURE ENGINEERING

TfidfVectorizer
CountVectorizer

5. TEXT LEMMATIZATION AND STEAMING

Spacy Lemm
SnowballStemmer

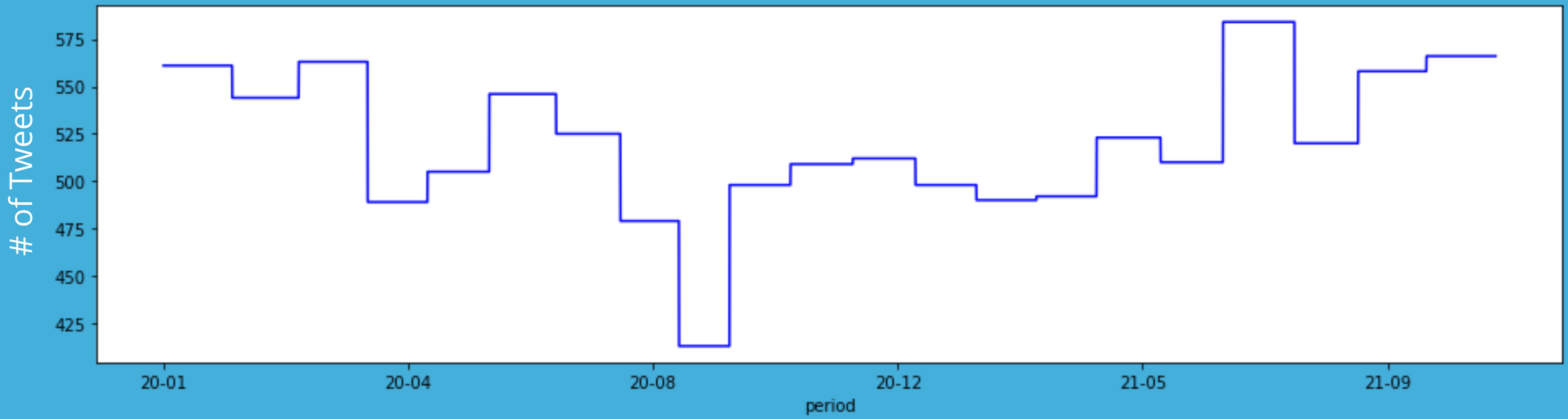
6. MACHINE LEARNING ALGORITHMS

NMF, TruncatedSVD
LatentDirichletAllocation

7. INTERPRETATION OF THE RESULT

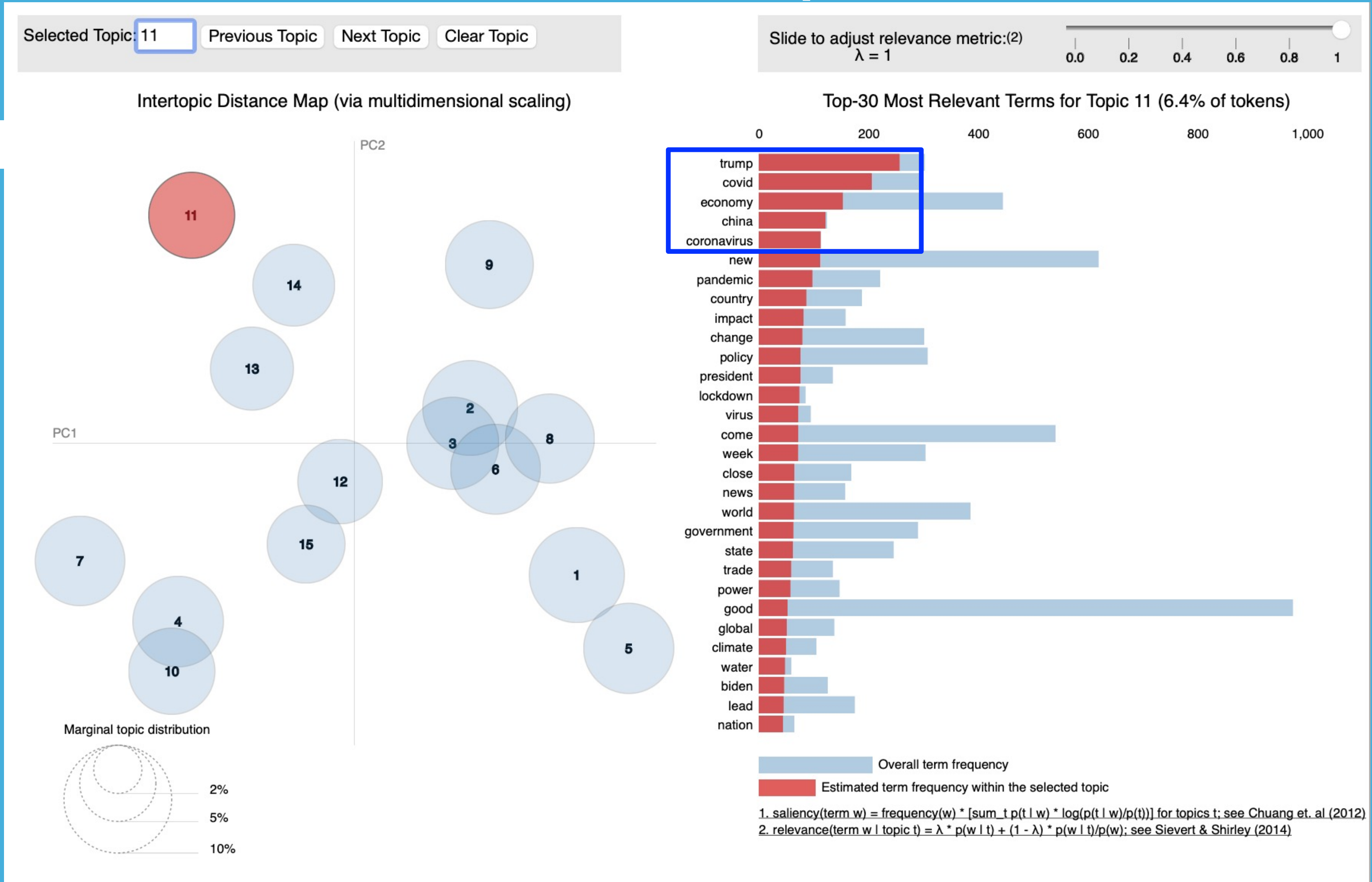
pyLDAvis, WordCloud, textacy

Exploratory Data Analysis



- ○ 2020.01-2021.10 : 69,060 data
- ○ LikeCount > 2
- ○ Keyword/hashtag: economics, business, financ

Latent Dirichlet Allocation – Topic 11



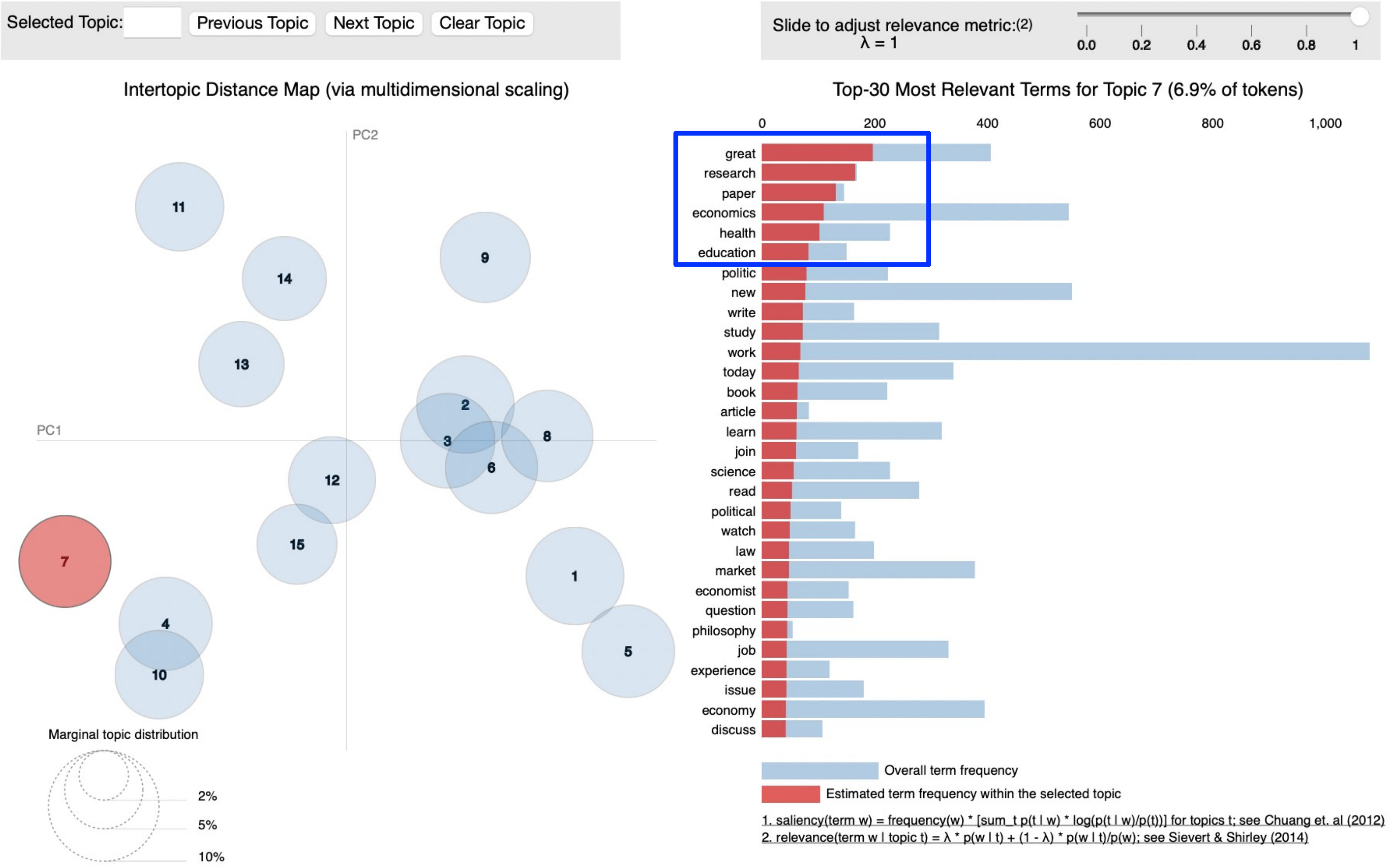
Selected Topic: 11

Slide to adjust relevance metric:(2)
λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

○ **Politics/ coronavirus/ country'**

Latent Dirichlet Allocation – Topic 7



Selected Topic:

Slide to adjust relevance metric:(2)
λ = 1

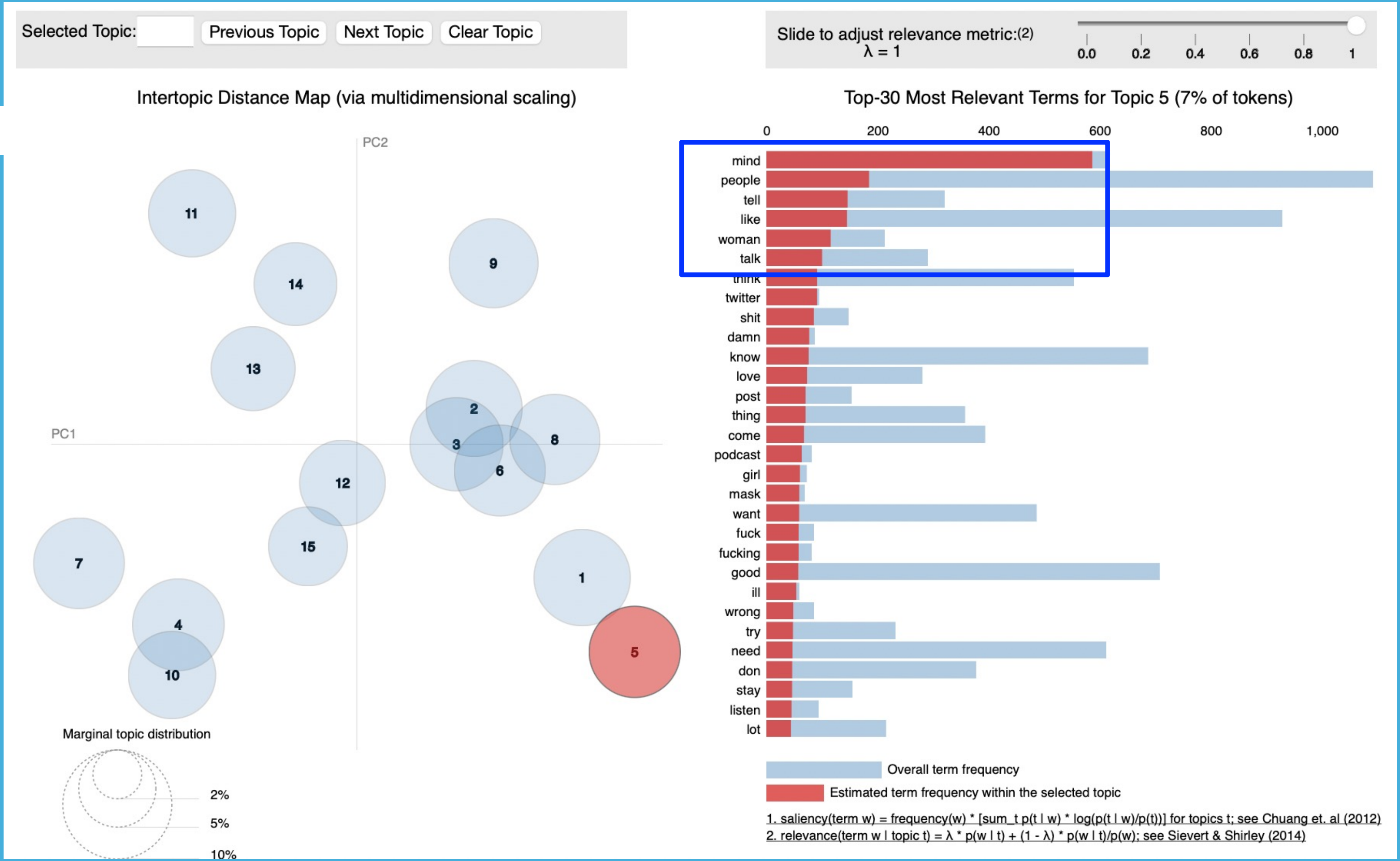
Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 7 (6.9% of tokens)

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

- ○ Education /research /school /study

Latent Dirichlet Allocation – Topic 5



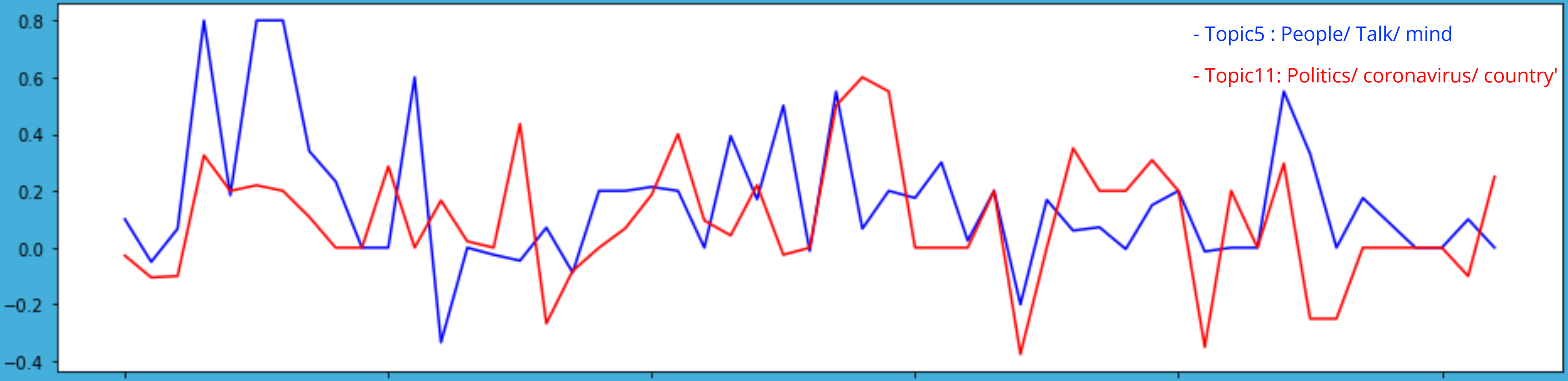
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

- ○ People /Talk /mind

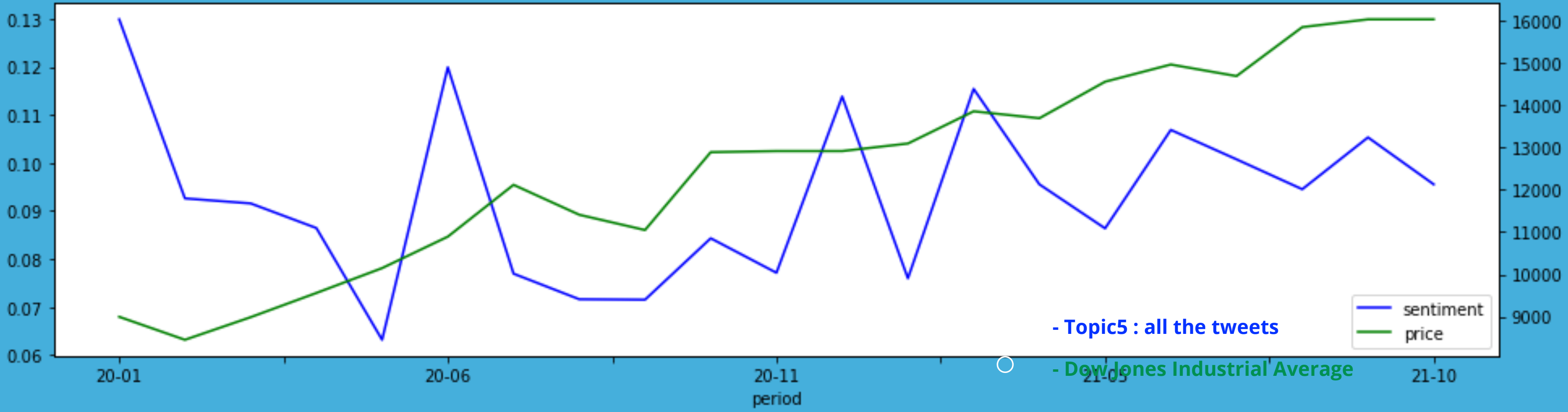
Latent Dirichlet Allocation

- **Categories:**
- Education/ research/ school/ study: 4, 7, 10
- Wealth/ work: 2, 9
- People/talk/mind: 1, 5
- Politics/coronavirus/country : 11
- Market, tax, government: 3

Sentiment Analysis



Sentiment Analysis



- Topic5 : all the tweets

○ - Dow Jones Industrial Average

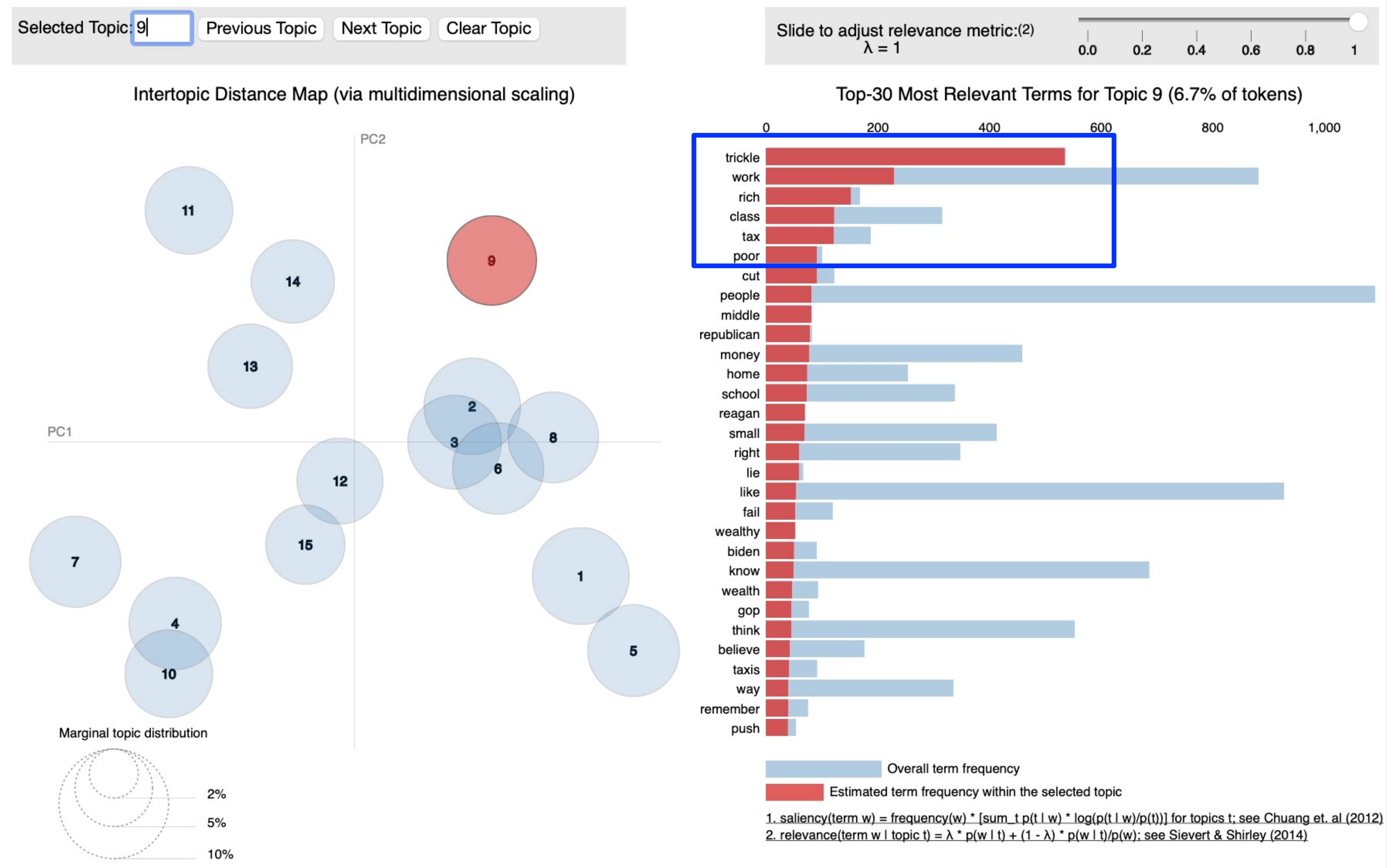
Future work

- Explore more on top tweets in each topics
- May need more data/months to observe better result (300-500/months)

THANKS!

Any questions?

Latent Dirichlet Allocation – Topic 9



Selected Topic: 9

Slide to adjust relevance metric:(2) λ = 1

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 9 (6.7% of tokens)

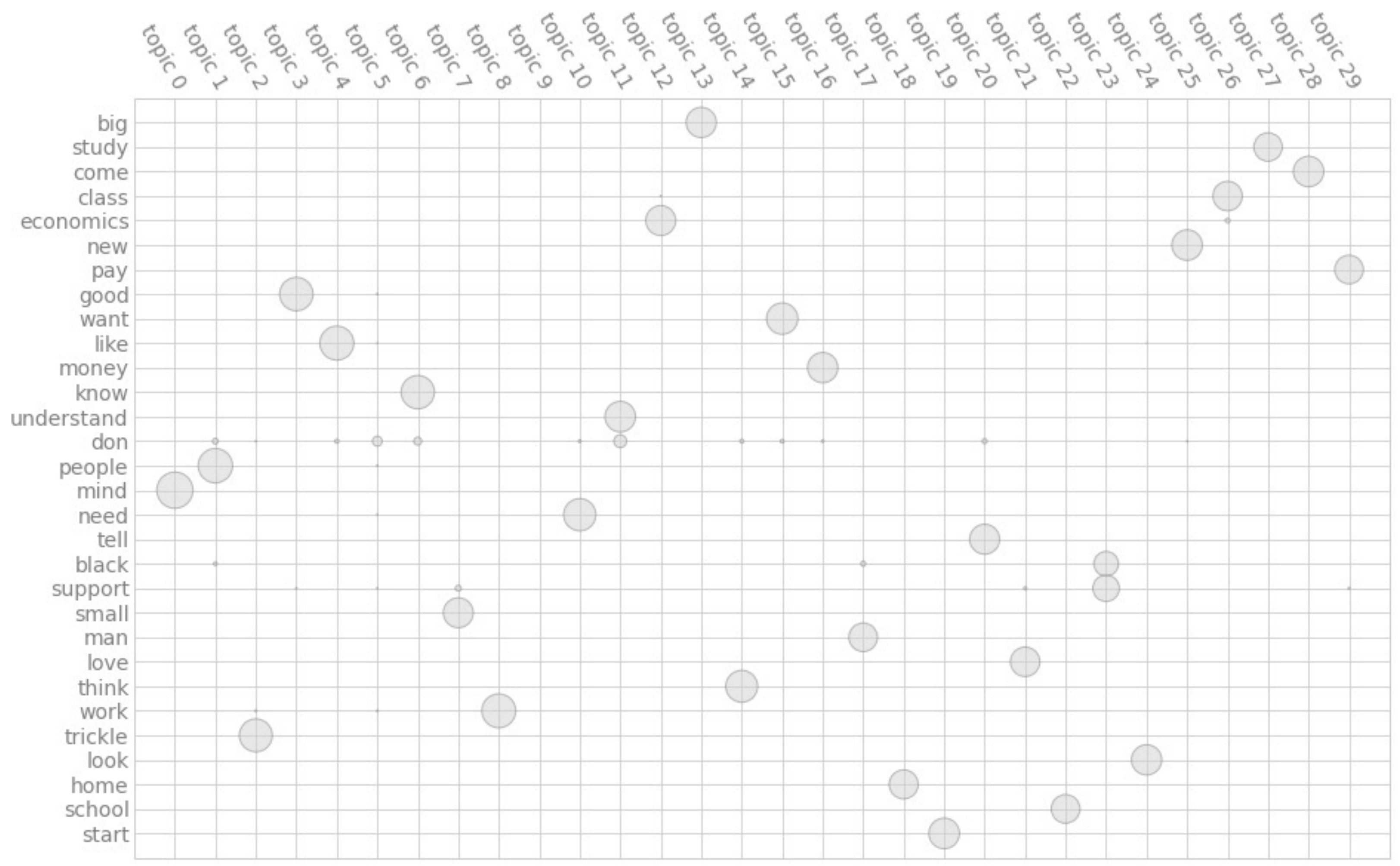
🟦 Overall term frequency

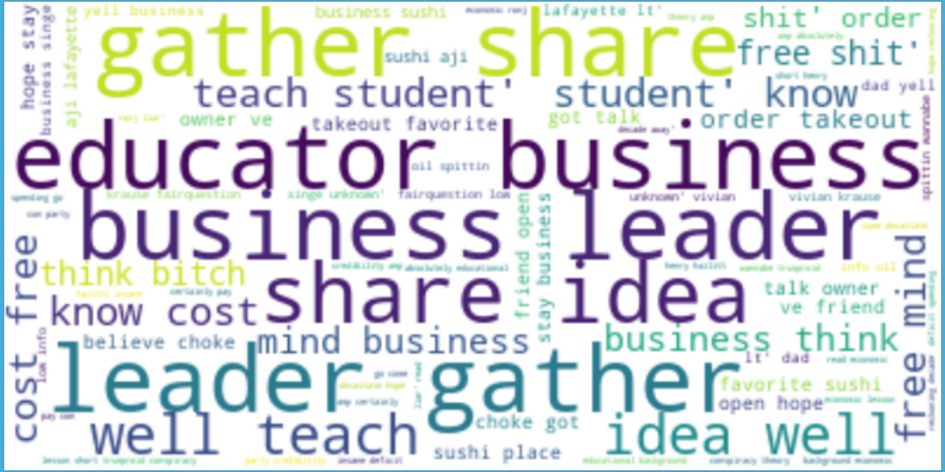
🟥 Estimated term frequency within the selected topic

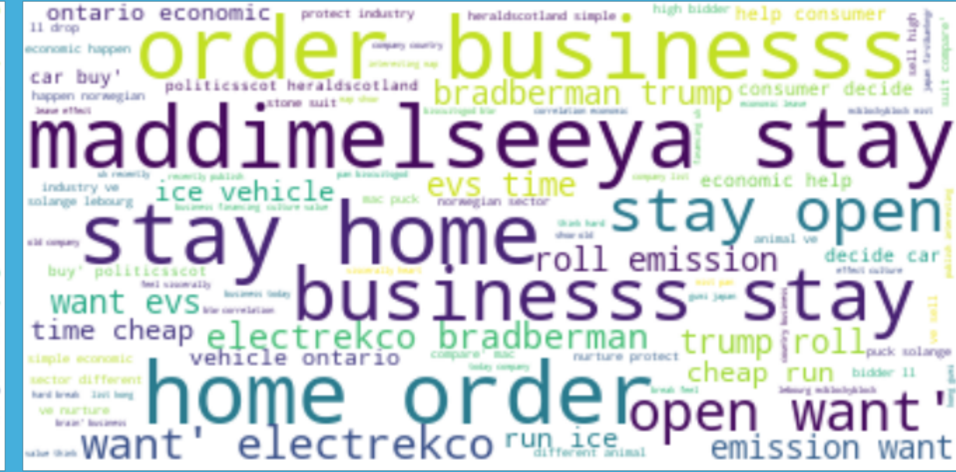
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

- Wealth/ Tax/ Trickle-down economics

textacy


topic 0
topic 1
topic 2
topic 3
topic 4
topic 5
topic 6
topic 7
topic 8
topic 9
topic 10
topic 11
topic 12
topic 13
topic 14
topic 15
topic 16
topic 17
topic 18
topic 19
topic 20
topic 21
topic 22
topic 23
topic 24
topic 25
topic 26
topic 27
topic 28
topic 29
big
study
come
class
economics
new
pay
good
want
like
money
know
understand
don
people
mind
need
tell
black
support
small
man
love
think
work
trickle
look
home
school
start


gather share
educator business
business leader
share idea
leader gather
well teach
idea well
teach student' student' know
free shit'
shit' order
order takeout
takeout favorite
think bitch
know cost
mind business
business think
free mind
cost free
favorite sushi
open hope
choke got
sushi place
believe choke
talk owner
ve friend
friend open
stay business
owner ve
got talk
sushi aji
aji lafayette
lafayette lt'
yell business
business sushi
business singe
hope stay
oil spittin
spittin wannabe
krause fairquestion
singe unknown'
fairquestion low
unknown' vivian
vivian krause
info oil
lt' dad
dad yell


order businesss
maddimelseeya stay
stay home
stay open
businesss stay
home order
open want'
want' electrekco
electrekco bradberman
bradberman trump
trump roll
roll emission
emission want
want evs
evs time
time cheap
cheap run
run ice
ice vehicle
vehicle ontario
ontario economic
economic help
help consumer
consumer decide
decide car
car buy'
buy' politicsscot
politicsscot heraldscotland
heraldscotland simple
protect industry
high bidder
economic happen
happen norwegian
norwegian sector
industry ve
solange lebourg
mac puck
puck solange
bidder 11
11 drop
simple economic
sector different
different animal
ve nurture
nurture protect
compare' mac
stone suit
suit compare'
sell high
animal ve