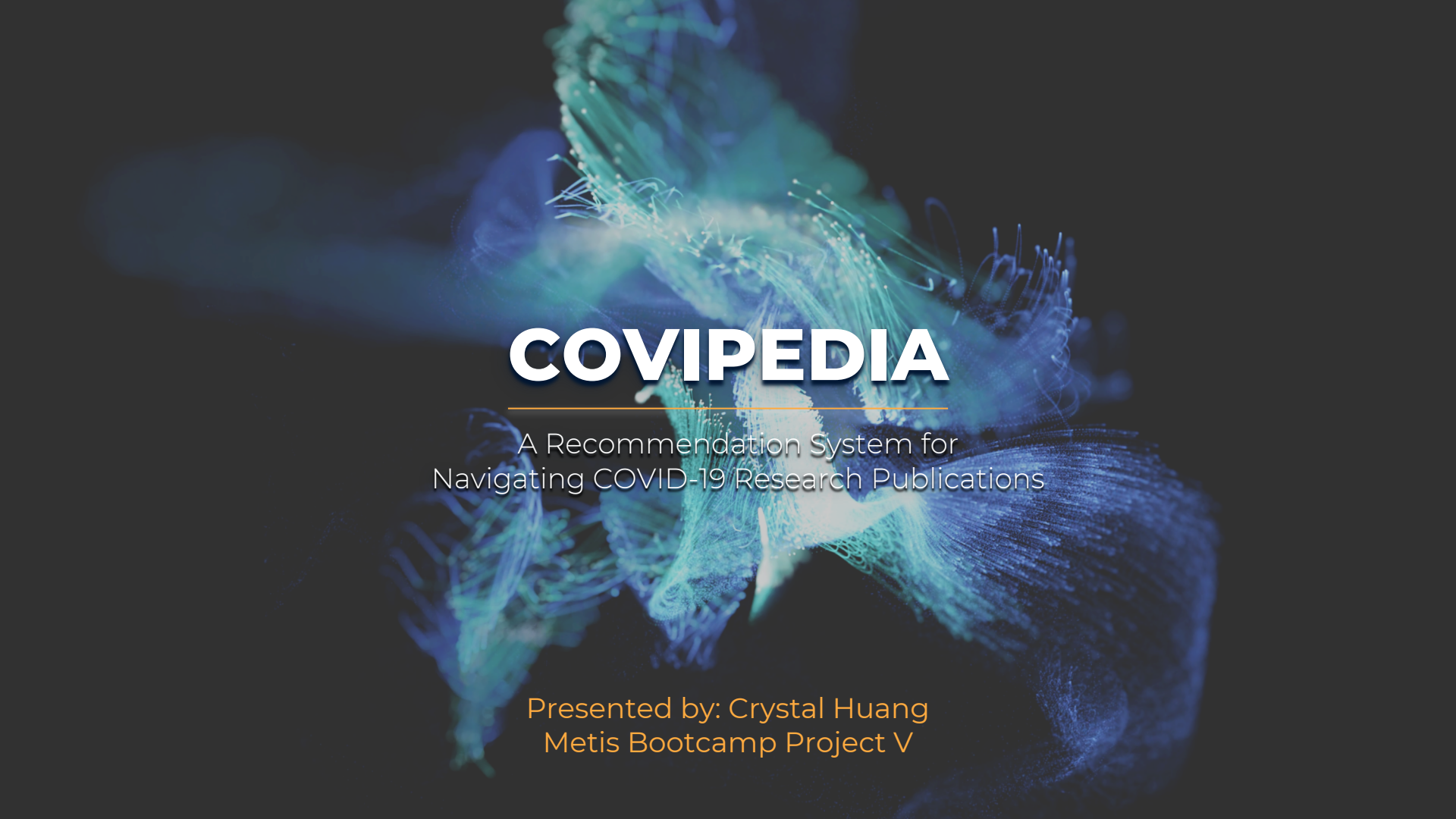


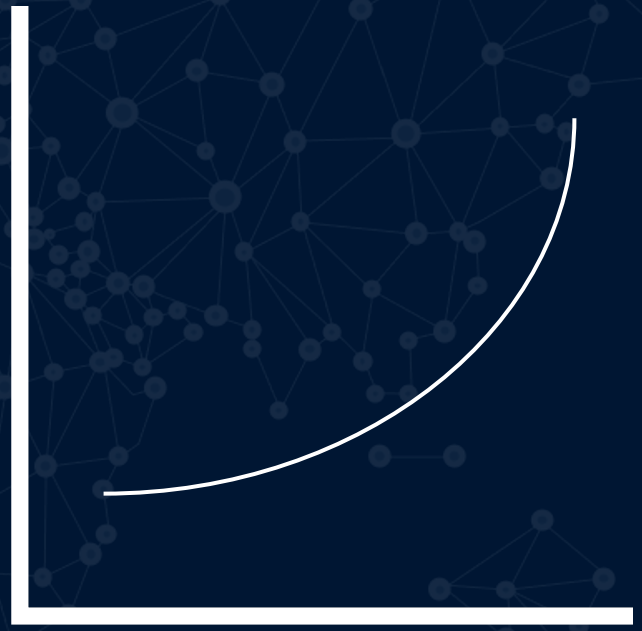
COVIPEDIA

A Recommendation System for
Navigating COVID-19 Research Publications

Presented by: Crystal Huang
Metis Bootcamp Project V

INTRODUCTION

Rapid increase in scientific literature




COVID-19


COVID-19

SARA GIRONI CARNEVALE

Scientists are drowning in COVID-19 papers. Can new tools keep them afloat?

By **Jeffrey Brainard** | May. 13, 2020 , 12:15 PM

PROBLEM

Hard to keep up with the **newest** publication from **various sources**

Hard to **collaborate** and **stay up to date** due to **limitations** of scientific conference

Crucial for scientist to **be aware of ongoing** research to avoid duplicated research effort

GOAL

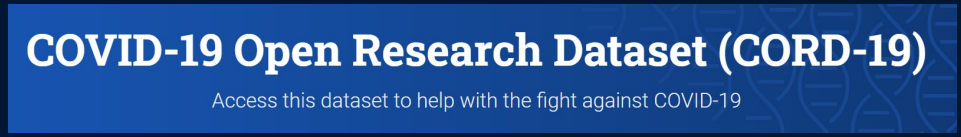
Build a **Recommendation System** to helps researchers

Navigate the current surge of papers about COVID-19

Find **relevant** publications

Uncover the hidden semantic relationships

DATA & PRE-PROCESSING



260k+ scholarly articles
abstract
(Jan 2020 – May 2021)

↓

langdetect Filter non-English articles

↓

(for **biomedical, scientific,** and **clinical** vocabulary)

spaCy scispaCy POS tagging

↓

Gensim Phrase detection

↓

NLTK Stop-word removal
kept nouns, adj, adv, verb
Lemmatization
Tokenization

WORKFLOW

Preprocessed Data	Vector Space	Topic Model	Eval & Vis	RecSys App Deploy
Python Pandas Numpy langdetect NLTK spaCy regex	Gensim.corpora	Gensim LDA	Matplotlib Seaborn WordCloud pyLDAvis	Streamlit Heroku

TOOLS

Evaluating Topic Model

Qualitative Measure

Subjective Inspection

- Challenging for a large dataset

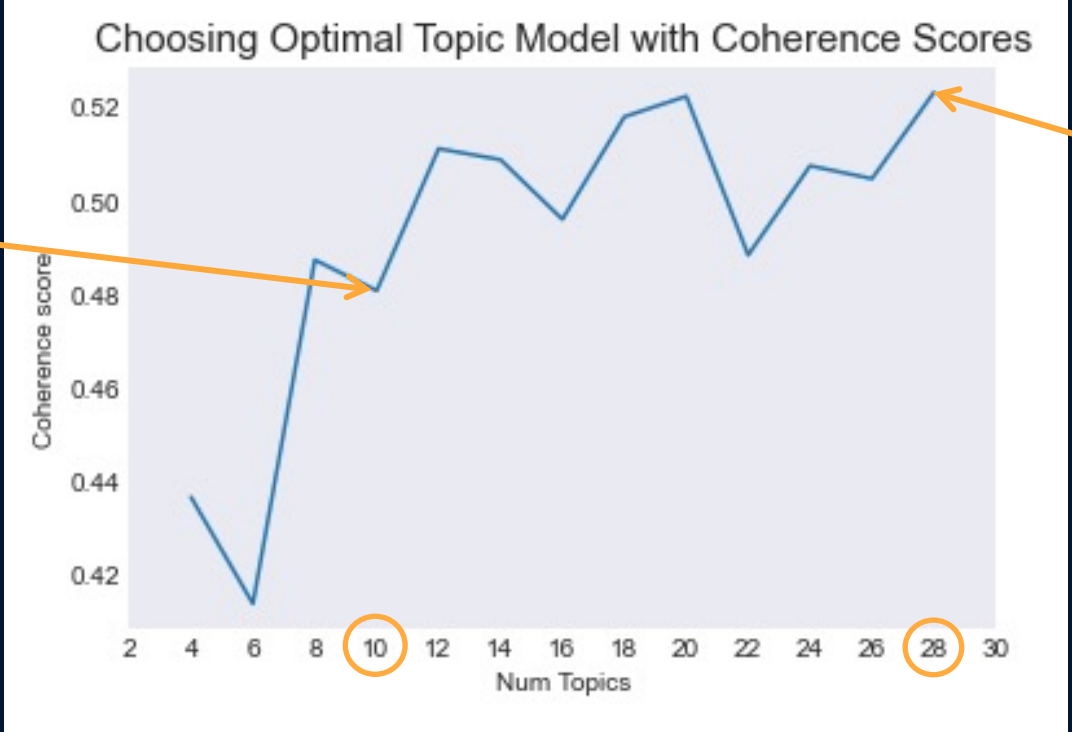
Quantitative Measure

Topic Coherence Score

- Easier to compare

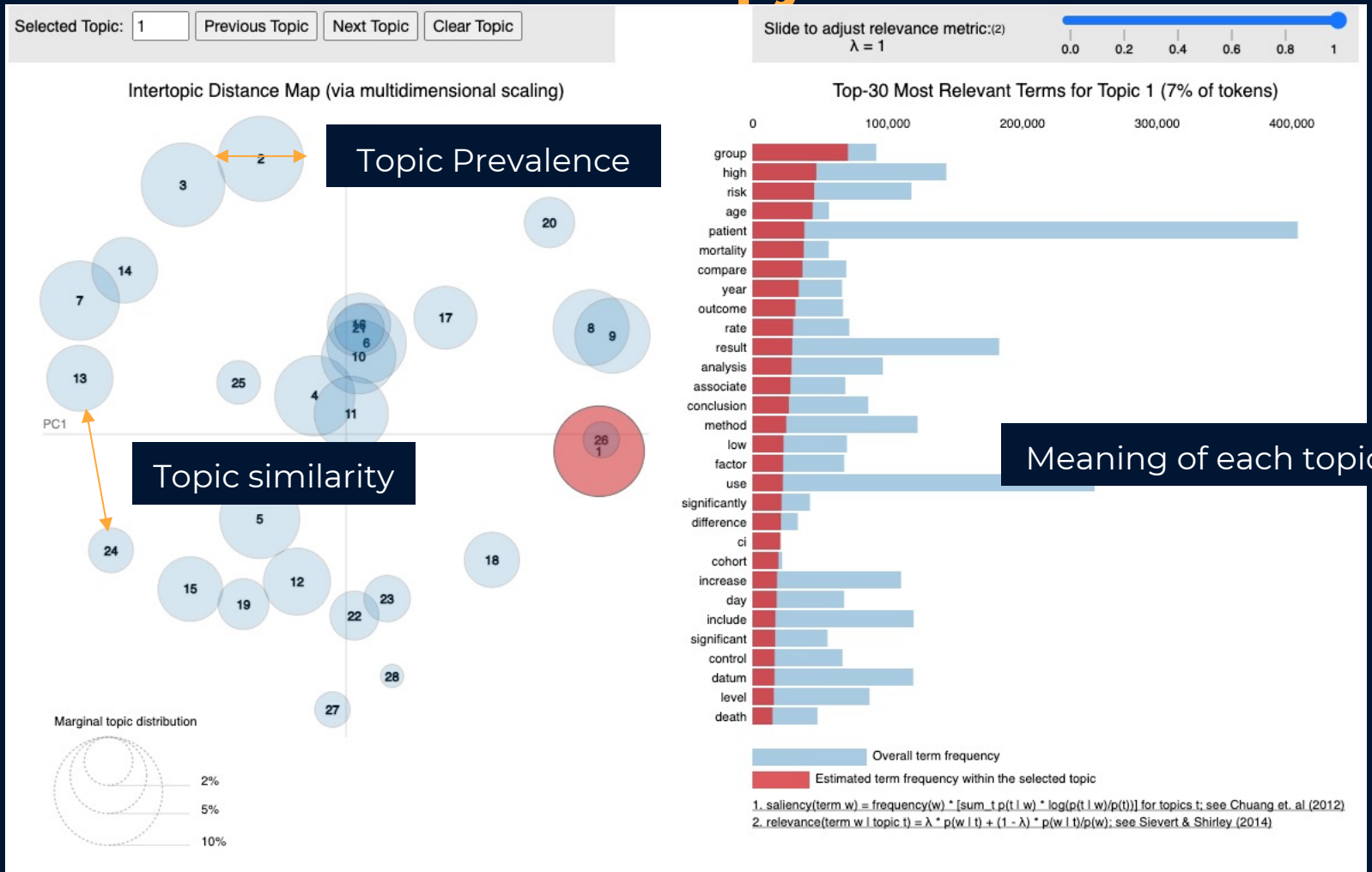
Tuning LDA Model - Number of Topics

Baseline
Coherence Score:
0.483

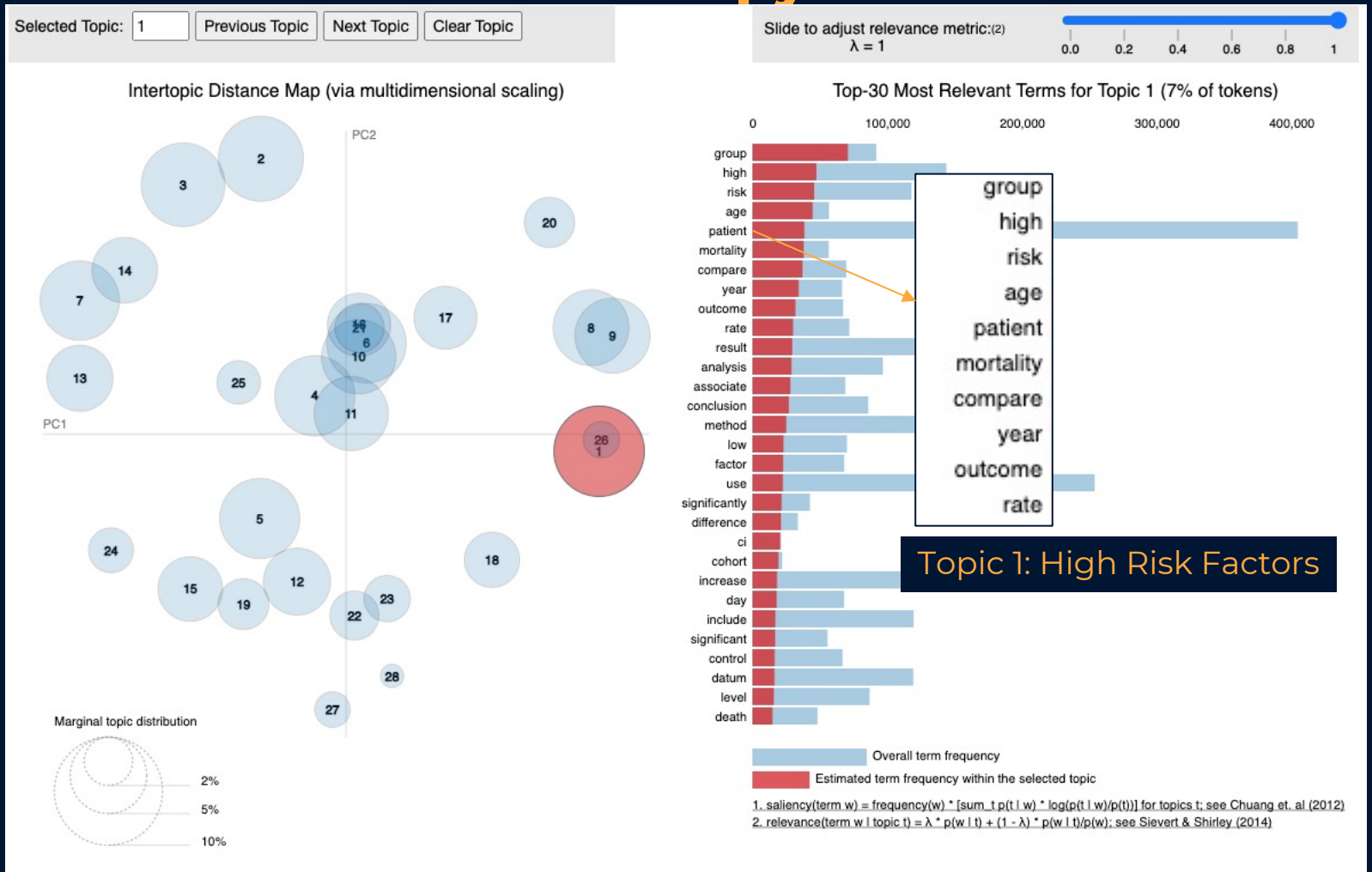


Optimal
Coherence Score:
0.523

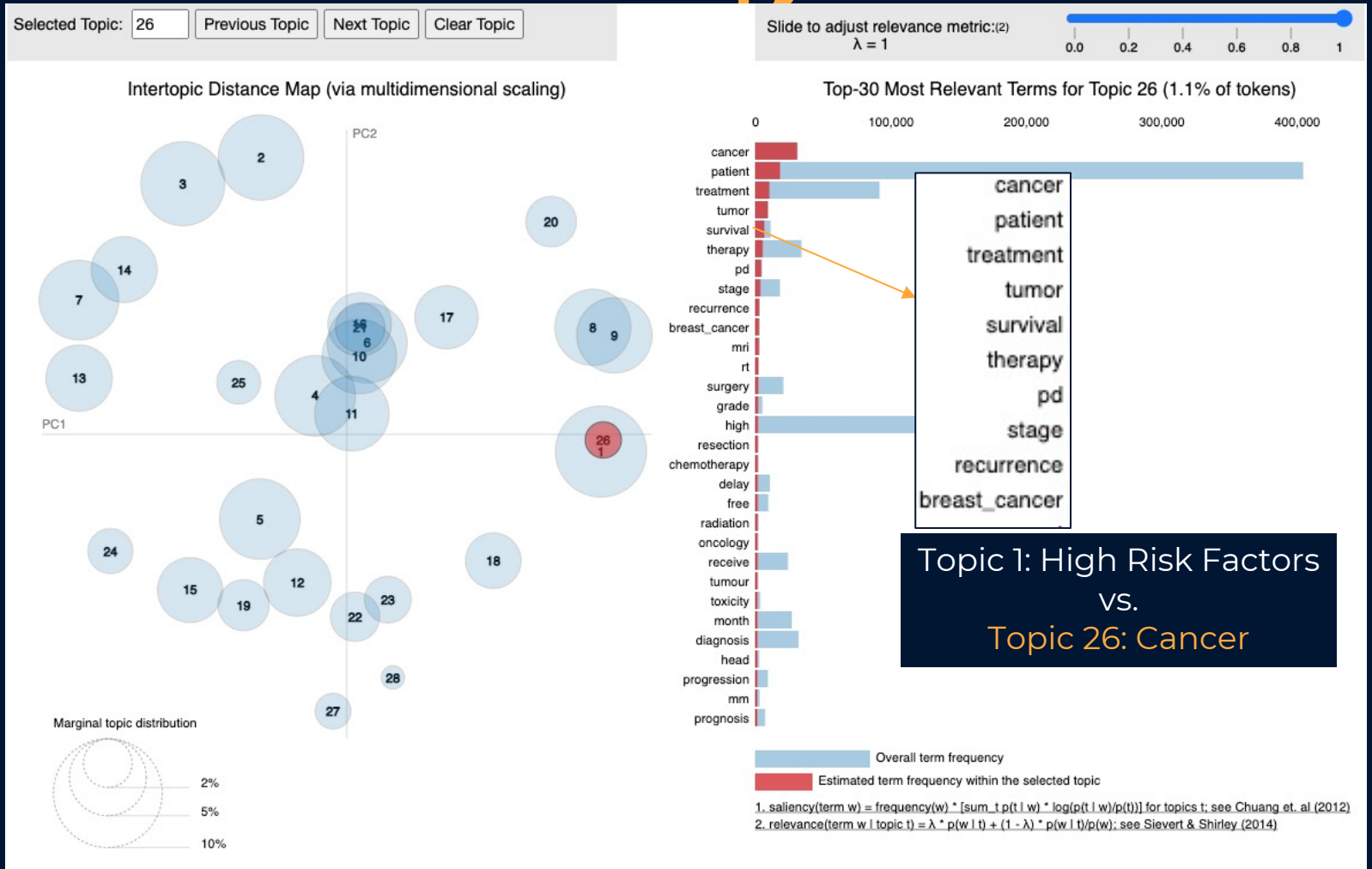
TOPICS VISUALIZATION with pyLDAvis



TOPICS VISUALIZATION with pyLDAvis

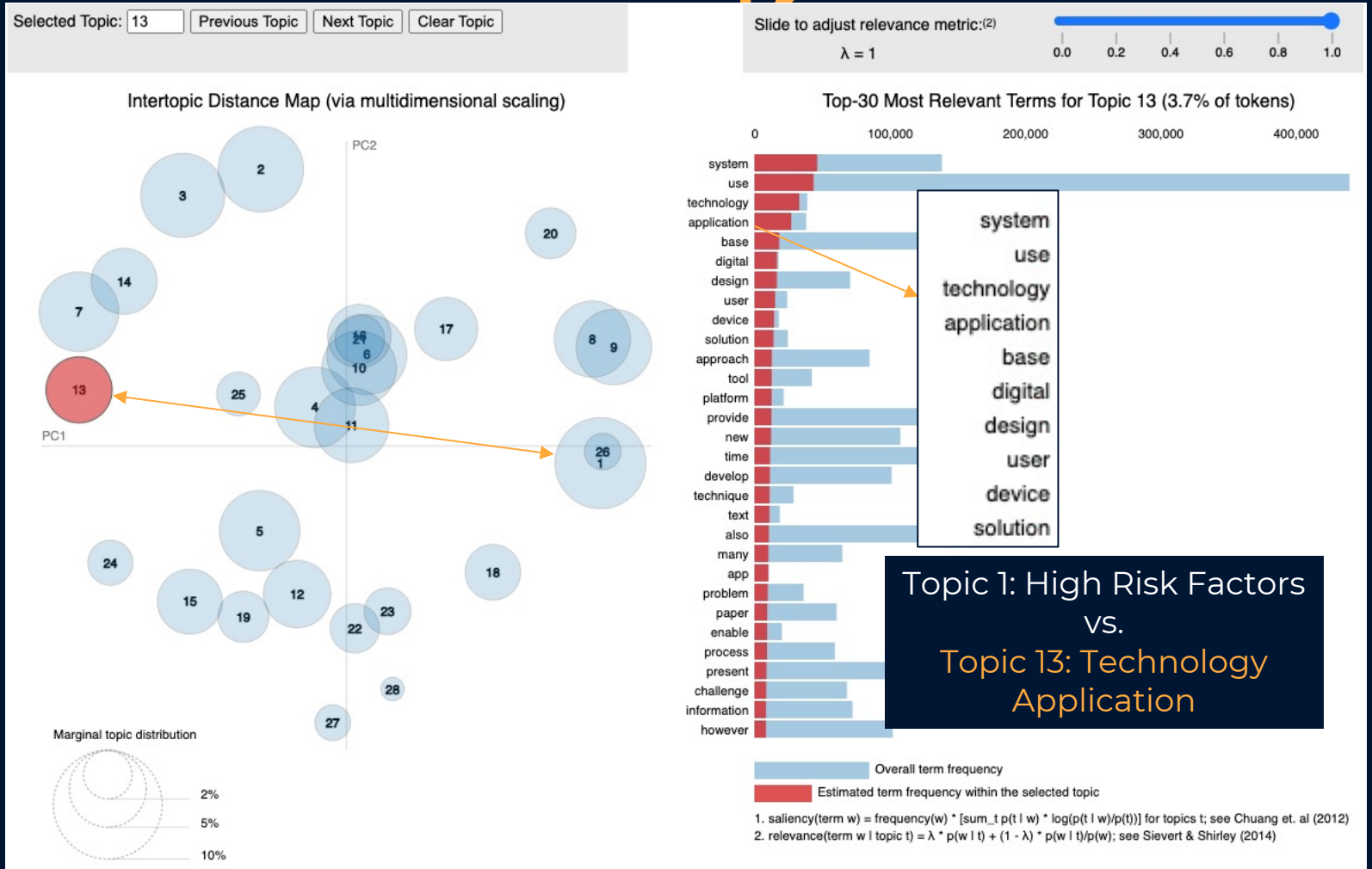


TOPICS VISUALIZATION with pyLDAvis

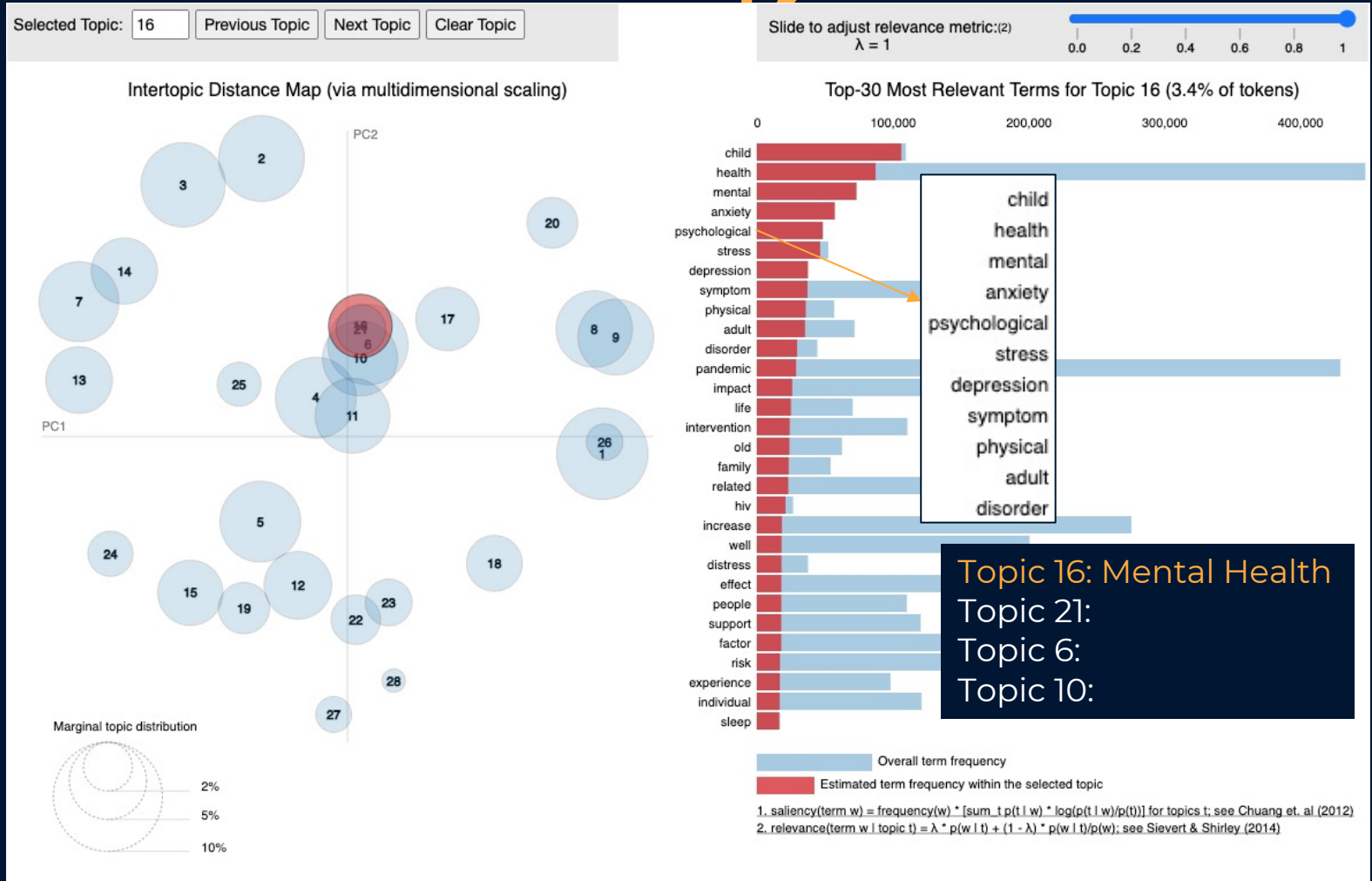


Topic 1: High Risk Factors
vs.
Topic 26: Cancer

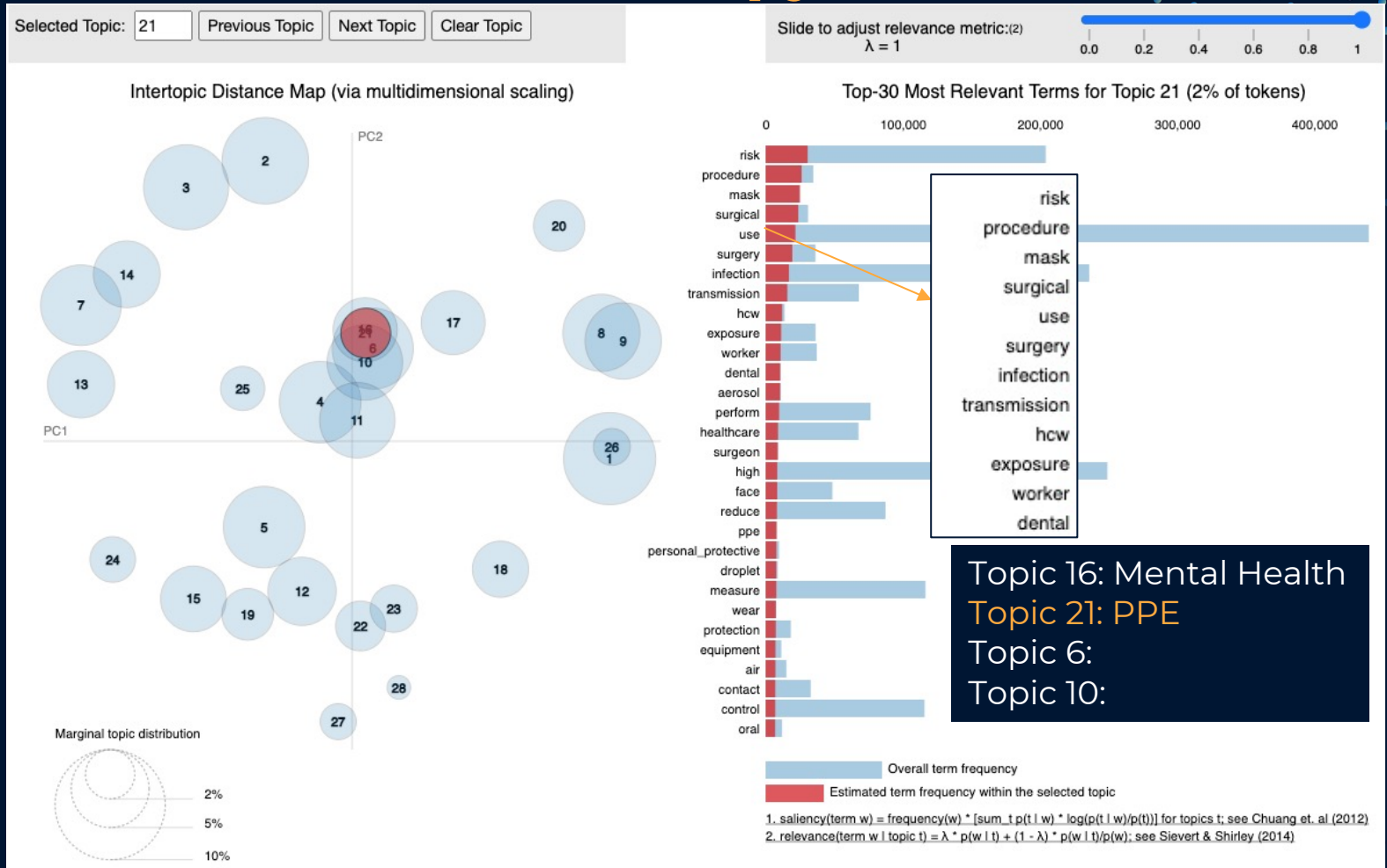
TOPICS VISUALIZATION with pyLDAvis



TOPICS VISUALIZATION with pyLDAvis

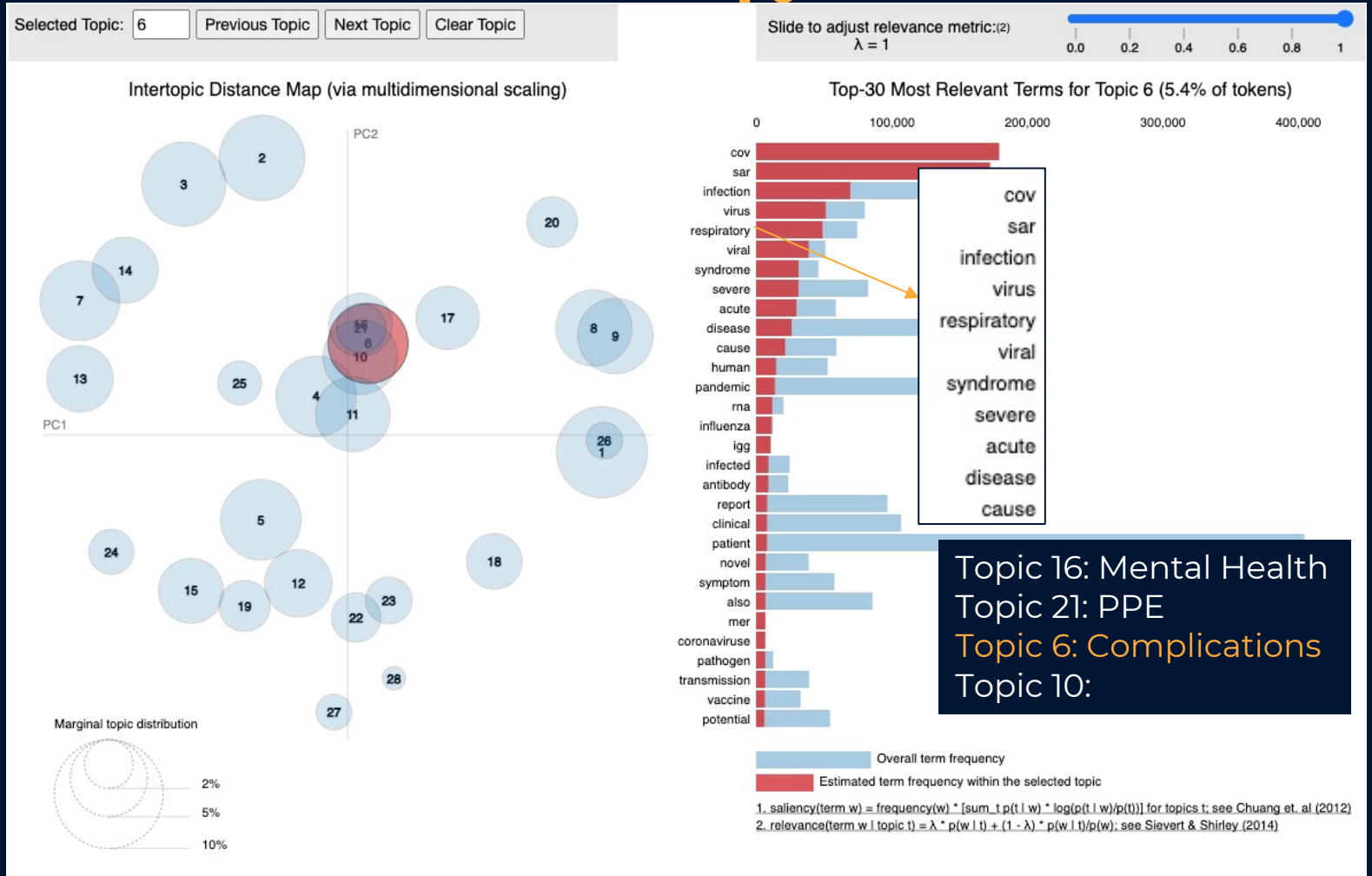


TOPICS VISUALIZATION with pyLDAvis



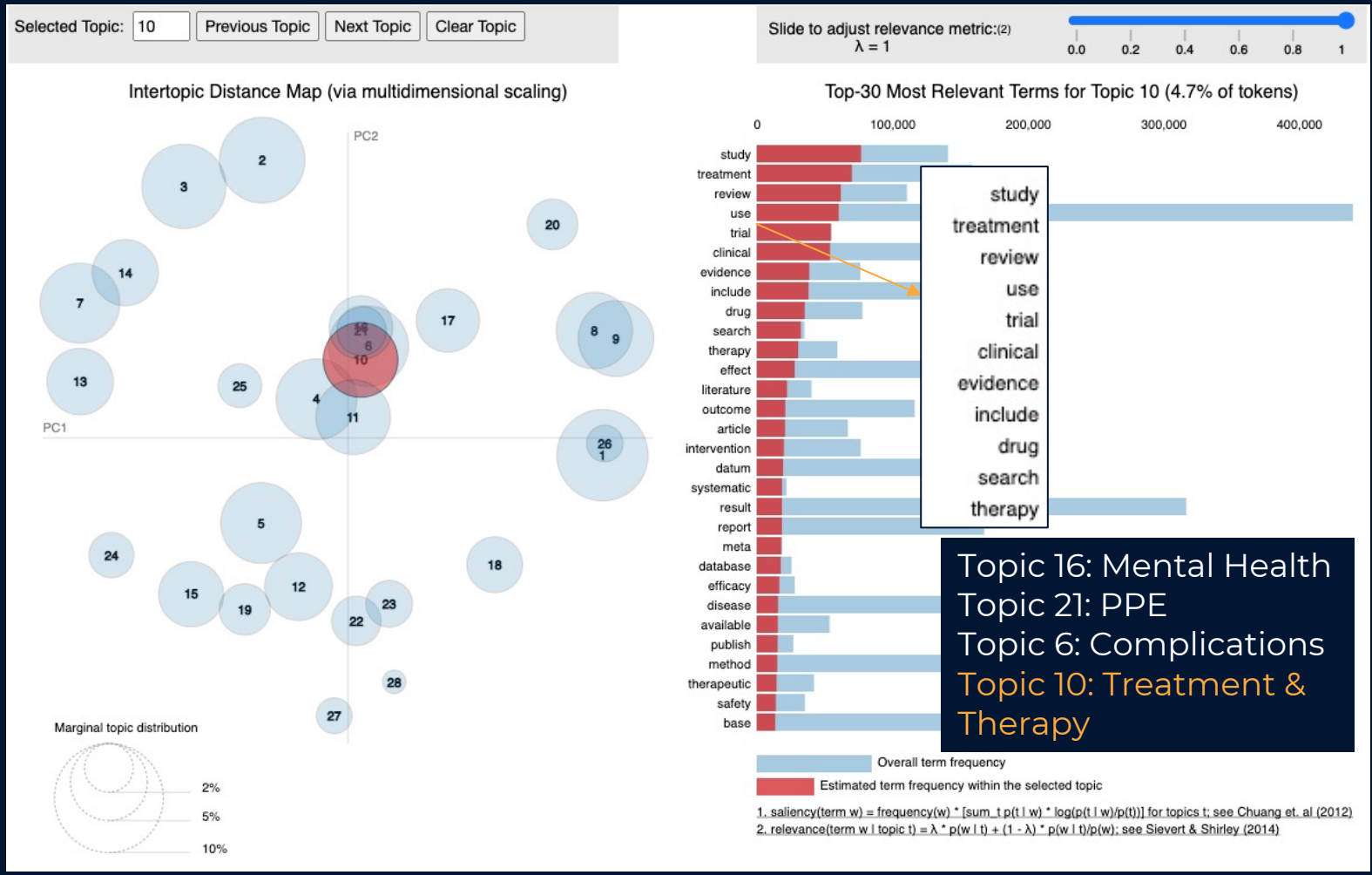
Topic 16: Mental Health
Topic 21: PPE
Topic 6:
Topic 10:

TOPICS VISUALIZATION with pyLDAvis



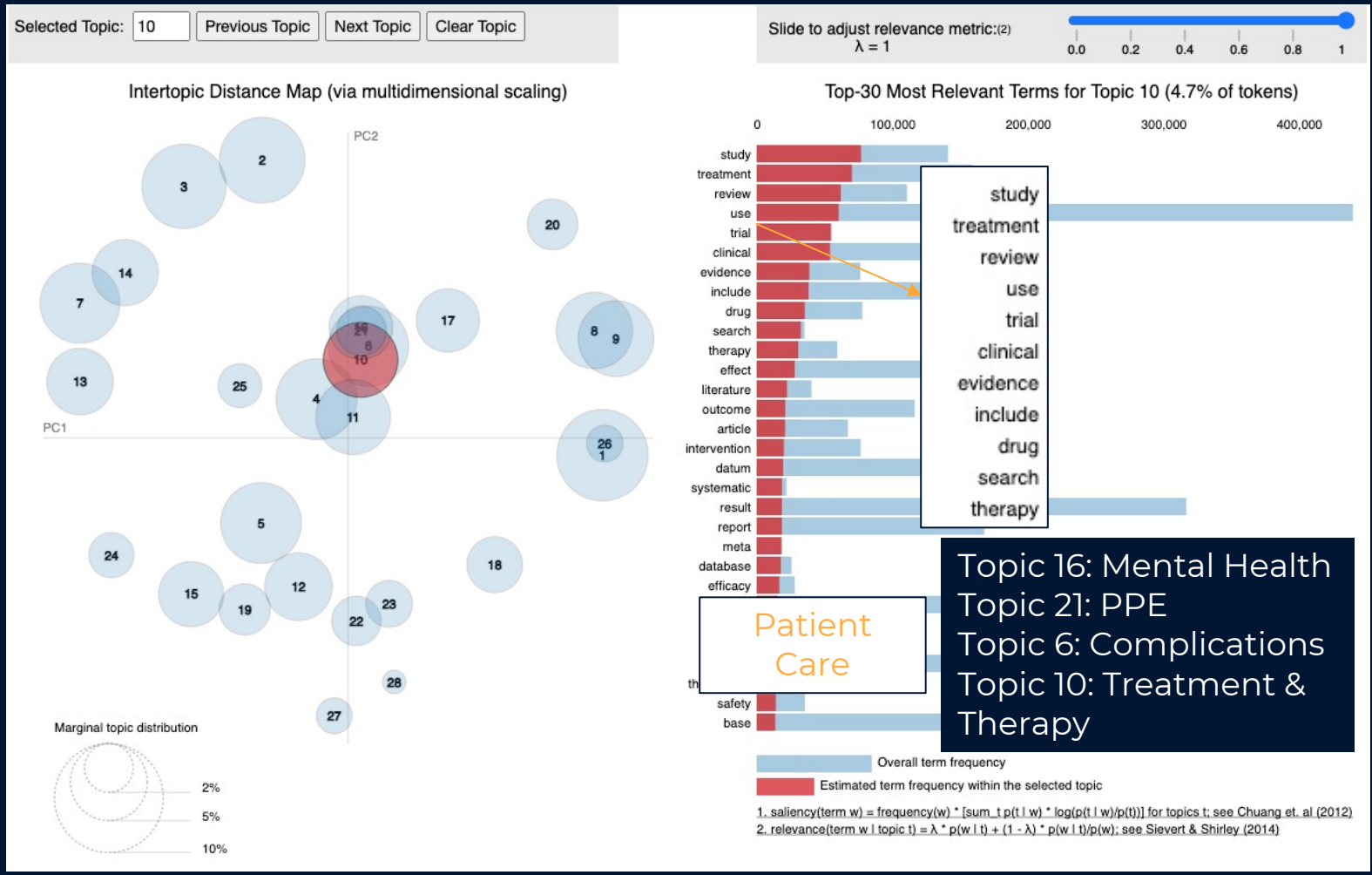
Topic 16: Mental Health
Topic 21: PPE
Topic 6: Complications
Topic 10:

TOPICS VISUALIZATION with pyLDAvis



Topic 16: Mental Health
Topic 21: PPE
Topic 6: Complications
Topic 10: Treatment & Therapy

TOPICS VISUALIZATION with pyLDAvis



APP

myapp · Streamlit

localhost:8501

COVIPEDIA

Navigate COVID-19 scientific literature with **Topics**!

COVID-19 Publications Related to:

Choose the topic to explore:

Education and Training

Here are the articles related to: Education and Training (sorted by relevance)

Perspective from a Teaching and Learning Center During Emergency Remote Teaching.
(https://doi.org/10.5688/ajpe8142; https://www.ncbi.nlm.nih.gov/pubmed/32934391/)

Promote an unexpected online experience through richer content
(https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7323578/)

Do-it-yourself physiology labs: Can hands-on laboratory classes be effectively replicated online?
(https://doi.org/10.1152/advan.00205.2020;
https://www.ncbi.nlm.nih.gov/pubmed/33529143/)

Parasitology education before and after the COVID-19 pandemic
(https://www.ncbi.nlm.nih.gov/pubmed/33191119/;
https://api.elsevier.com/content/article/pii/S1471492220302944;
https://www.sciencedirect.com/science/article/pii/S1471492220302944?v=s5;
https://doi.org/10.1016/j.pt.2020.10.009)

MORE WORK TO DO

1. **More NLP preprocessing**
   Tuning more hyperparameters (ie. max_df)

2. **More model refinement**
   Tuning the number of topics, other hyperparameters

3. **Search engine + Recommendation system**
   Keyword search and related research recommendation

THANKS!

Any Questions?

https://covipedia.herokuapp.com/

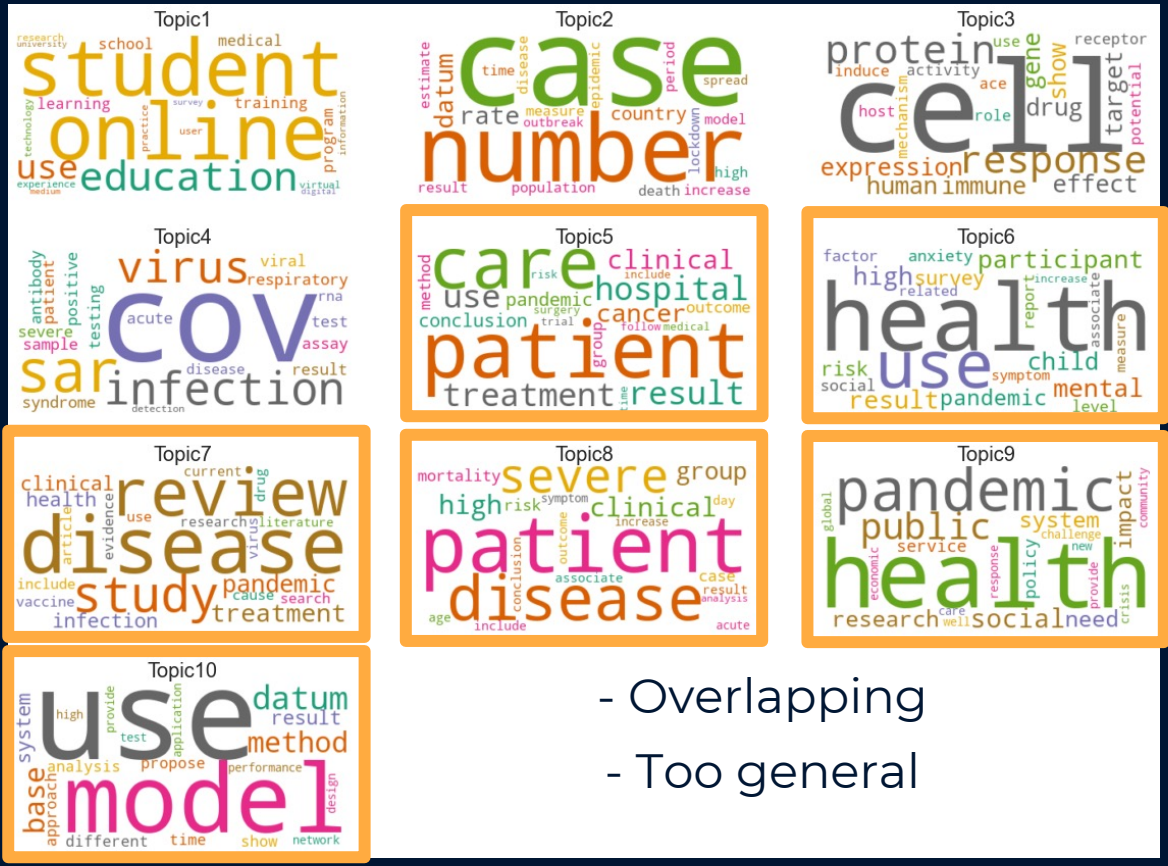
crystal-ctrl

CrystalHuang-ds

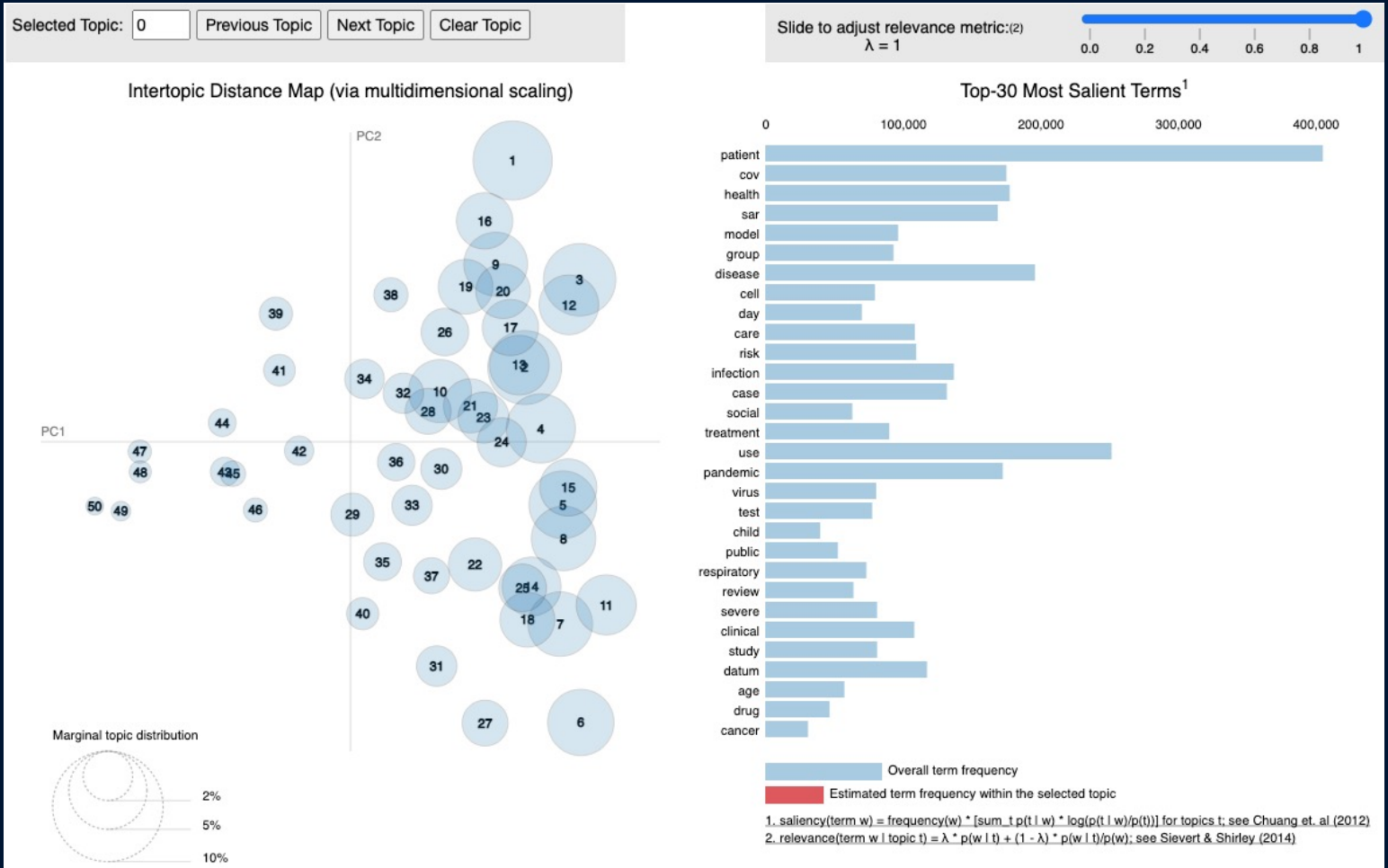
REFERENCES

1. Brainard, J. 2020. “Scientists are Drowning in COVID-19 Papers. Can New Tools Keep Them Afloat.” *Science*. https://www.sciencemag.org/news/2020/05/scientists-are-drowning-covid-19-papers-can-new-tools-keep-them-afloat#
2. Lee, J. 2020. “Benchmarking Language Detection for NLP.” *Towards Data Science, Medium*. https://towardsdatascience.com/benchmarking-language-detection-for-nlp-8250ea8b67c
3. Kapadia, S. 2019. “Evaluate Topic Models: Latent Dirichlet Allocation (LDA)” *Toward Data Science, Medium*. https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0
4. Tran, K. 2021. “pyLDAvis: Topic Modelling Exploration Tool That Every NLP Data Scientist Should Know.” *Neptune Blog*. https://neptune.ai/blog/pyldavis-topic-modelling-exploration-tool-that-every-nlp-data-scientist-should-know

APENDIX 1 - LDA Model with Gensim (10 Topics)



APPENDIX 2


Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
2%
5%
10%
Top-30 Most Salient Terms[1]
0
100,000
200,000
300,000
400,000
patient
cov
health
sar
model
group
disease
cell
day
care
risk
infection
case
social
treatment
use
pandemic
virus
test
child
public
respiratory
review
severe
clinical
study
datum
age
drug
cancer
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)