Machines & Language

Erin E. McCabe
Digital Scholarship Center
UC Libraries
Email: mccabeen@ucmail.uc.edu

bit.ly/dataday_slides

bit.ly/dataday_code

run in Chrome

Today's Agenda

Housekeeping & Context

Get settled in and discuss what in the AI-world we're going to be doing today as well as some of its context.

Classification

We will work through python code together in a web-hosted code notebook to learn the technical steps & logic of classifying text data.

Topic Modeling

Building on what we know about Classification, discuss the concepts behind Topic Modeling.

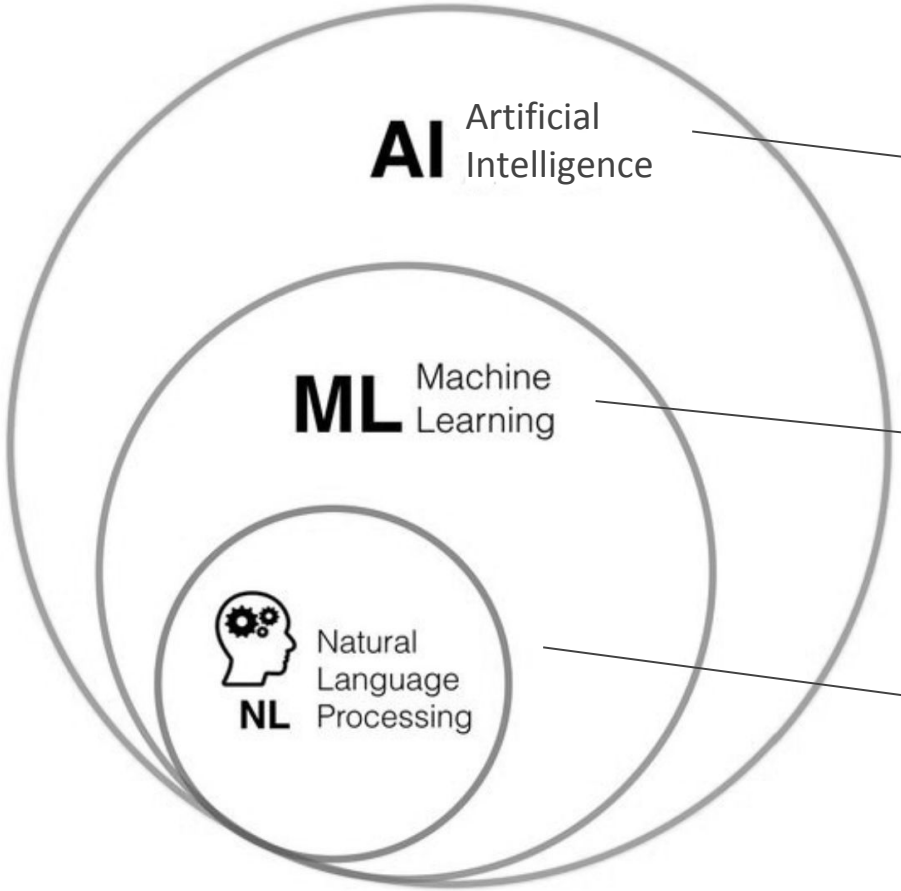
Topic Model Exploration

Interact with a Topic Model visualization based on the text of ~32k Covid-19 research articles.

Share / Review

Discuss the visualization, review what we've learned, and get some other resources.

bit.ly/dataday_slides



Umbrella term for machines that can simulate human intelligence - systems that mimic cognitive capabilities like solving problems.

Computer programs that *automatically* improve with *experience*.

Computer programs that understand human language - can make sense of and perform tasks on text.

bit.ly/dataday_slides

A Non-Exhaustive List of Everyday(ish) NLP

Source: Tableau

- Email Filters
- Smart Assistants
- Predictive Text
- Language Translation
- Text Mining / Analytics
  - e.g. keyword extraction & finding structure / patterns in unstructured text data.

Erin's Python Top 3

Import *Python Library or Package*

Variables:

fname = Erin
age = 34

For Loops:

FOR a_data IN list_o_data:
do some task...

bit.ly/dataday_slides

Web-Hosted Code Notebook: https://bit.ly/dataday_code

Supervised ML

"Supervised Learning is done using a *'ground truth'* … prior knowledge of what the output values for our samples should be..."

e.g. CLASSIFICATION: ML task used to predict a class label (from a predefined list)

Unsupervised ML

"...Unsupervised Learning ... does not have labeled outputs ... its goal is to infer the *natural* structure present within a set of data points."

e.g. TOPIC MODELING: Machine reads a set of docs to detect word patterns & word groups based (from data itself, not a predefined list)

Source: Towards Data Science

bit.ly/cov19_model

learning || machines

bit.ly/cov19_model

Corpus : Covid | Term : severe_acute_respiratory_syndrome_coronavirus_2-OR-covid-19-OR-sars_cov_2 | # Docs : 31,818 | # Topics : 40 | Stopwords : | Years : -

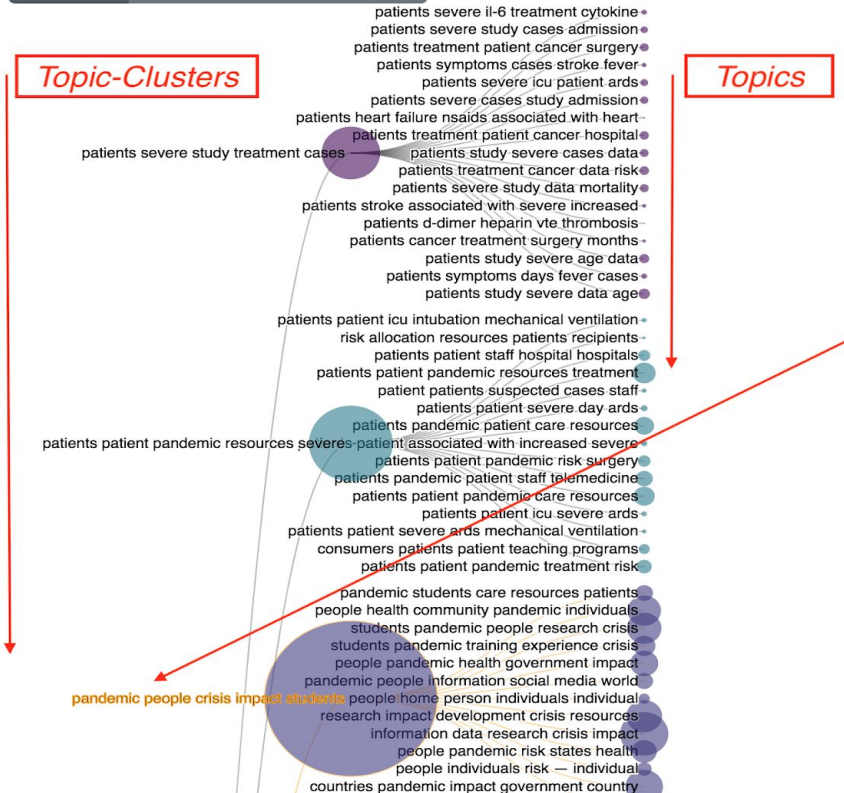
Documents

Title	Score
	53.391784816
	4.406728744
The health-related determinants of politics	4.393991559
Policing the Coronavirus Outbreak: Processes and Prospects for Collective Disorder	4.247873306
Global call to action for inclusion of migrants and refugees in the COVID-19 response	4.230182349
	4.089938640
COVID, food, and the Parable of the Shmoo	4.055532753

The health-related determinants of politics

I was pleased to read Richard Horton's Comment about populist politics.1 However, the problem of political populism and solution that he charts out seem somewhat non-sequitur.If the determinants of health are identified as political, then the remedy must also be political. US President Donald Trump propagates a myth about "forcing American taxpayers to provide unlimited free healthcare to illegal aliens",2 and the responses from academia are merely calls for justice, fairness, and universality. These values are imperative but, in addition to the moral argument, we should also make the political one.Perhaps this political passivity from academics is an indication of a wider problem. Academics are happy to engage with the policy, but afraid to engage with the politics. Research, such as that encompassed in the Lancet Migration, is normative rather than descriptive.3 Such research is done to improve people's quality of life and is therefore political.

Tree | Circle | Network



Clusters

cluster	#para, #docs	# topics	terms
1	0,18905	17	patients severe study treatment cases
16	0,23418	15	patients patient pandemic resources severe
7	0,16931	13	pandemic people crisis impact students
4	0,3887	10	data information users model network
8	0,16355	10	cases data infected model population
10	0,4979	9	patients increased il-6 inflammation cytokine
13	0,3121	9	cells human expression genes data
17	0,4023	8	binding proteins protein interaction interactions
2	0,3976	7	study participants data results survey
3	0,4287	7	positive patients

Corpus : Covid | Term : severe_acute_respiratory_syndrome_coronavirus_2-OR-covid-19-OR-sars_cov_2 | # Docs : 31,818 | # Topics : 40 | Stopwords : | Years : -

Documents

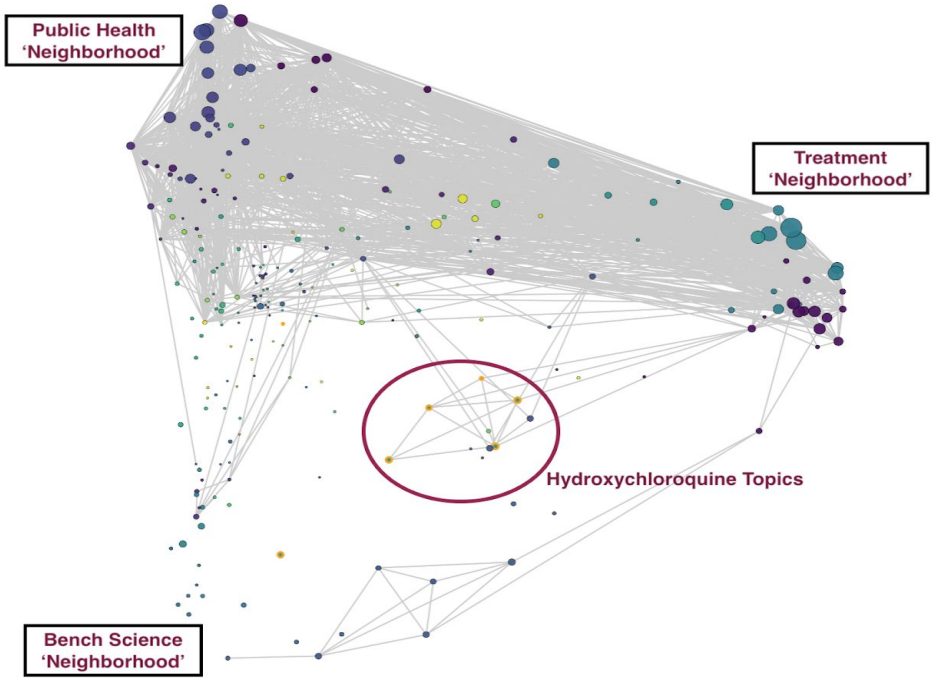
Title	Score
Response to	10.3162498474
Combating Devastating COVID -19 by Drug Repurposing	4.488617509603
Favipiravir versus Arbidol for COVID-19: A Randomized Clinical Trial	3.870159059762
Title: What do we know about remdesivir drug interactions?	3.792886435985
	3.750968873500
Interferon beta-1a for COVID-19: critical importance of the administration route	3.618535131216
Chloroquine and hydroxychloroquine for COVID-19: A word of caution	3.582105875015

Doc View

Tree | Circle | Network

Doc-link limit:4545

Circle – Paragraph Level/Square – Article Level



cluster	#para, #docs	# topics	terms
16	0,23418	15	patients patient pandemic resources severe
1	0,18905	17	patients severe study treatment cases
7	0,16931	13	pandemic people crisis impact students
8	0,16355	10	cases data infected model population
0	0,8368	6	cases china outbreak transmission infected
19	0,6891	7	cases patients patient diagnosis images
10	0,4979	9	patients increased il-6 inflammation cytokine
37	0,4670	5	ppe patients risk patient masks
3	0,4287	7	positive patients negative samples days
20	0,4140	7	treatment drugs patients efficacy hydroxychloroquine

Review

1. Learned about NLP Tasks:
    a. **Classification**
    b. **Topic Modeling**
2. Accessed the **NLTK Brown Corpus** & prepared data for ML tasks
3. Built a **Text Classifier** for sentence category w/ **scikit-learn**
4. Explored an interactive COV-19 **Topic Model**
5. Got a more grounded, conceptual understanding of how machines 'learn' about language & how that can be applied to research.

Resources

1. Introduction to Machine Learning with Python, Andreas C. Muller and Sarah Guido. O'Reilly, 2017.
2. Lena Voita's GitHub Hosted NLP Course
3. Scott Weingart's *Topic Modeling for Humanists: A Guided Tour*

Thank You!

Calvin & Hobbes
by Bill Watterson for June 30, 1993