

CLARITY & REACH IN SCIENTIFIC COMMUNICATION

Rachel Z. Insler

Capstone Project Presentation

DSIR – 22221E

May 13, 2021

“ Clear communication should always be a goal in science. It's important to step back and always remind yourself as a scientist:

how do I describe what I'm doing to someone who is not doing this 24/7 like I am?"

-Sabine Stanley, planetary scientist at Johns Hopkins University

[The New York Times, Trilobites Column; April 9, 2021](#)

BACKGROUND

As data scientists, our ability to transform information into actionable insights can have a profound impact on the organizations we support.

But if no one understands what we are talking about...our influence will be severely limited.

BACKGROUND

A recent exploration into the literature surrounding cave science found that research papers containing lots of specialized terminology are less likely to be cited by other researchers.

Does that finding hold for the highly technical field of machine learning?

PART I: REPRODUCTION ATTEMPT

- Establish a machine learning 'jargon' dictionary
- Develop a corpus of abstracts from peer-reviewed scientific journals that reference 'machine learning'
- Calculate proportion of jargon words in each abstract
- Does using a greater proportion of jargon lead to fewer citations?

ESTABLISH JARGON

MACHINE LEARNING DICTIONARY

- 412 words and phrases scraped from Google Developers' Machine Learning Glossary
- From "A/B testing" to "wide model"

Machine Learning Courses Practica Guides Glossary ▾ 🔍 Search English ▾

Rectified Linear Unit (ReLU)


An **activation function** with the following rules:

- If input is negative or zero, output is 0.
- If input is positive, output is equal to input.

recurrent neural network

A **neural network** that is intentionally run multiple times, where parts of each run feed into the next run. Specifically, hidden layers from the previous run provide part of the input to the same hidden layer in the next run. Recurrent neural networks are particularly useful for evaluating sequences, so that the hidden layers can learn from previous runs of the neural network on earlier parts of the sequence.

For example, the following figure shows a recurrent neural network that runs four times. Notice that the values learned in the hidden layers from the first run become part of the input to the same hidden layers in the second run. Similarly, the values learned in the hidden layer on the second run become part of the input to the same hidden layer in the third run. In this way, the recurrent neural network gradually trains and predicts the meaning of the entire sequence rather than just the meaning of individual words.



CREATE CORPUS

- Gather 3,151 abstracts
 - Query machine learning PubMed IDs using **pmidcite** library and NCBI API
 - “machine learning” in abstract or title
 - from peer-reviewed, English-language journals
 - publication date range: 2010 -2020
 - Scrape* corresponding **abstracts**, **titles**, **publication dates**, and **citation counts** using the **requests** library and **BeautifulSoup**

*legally

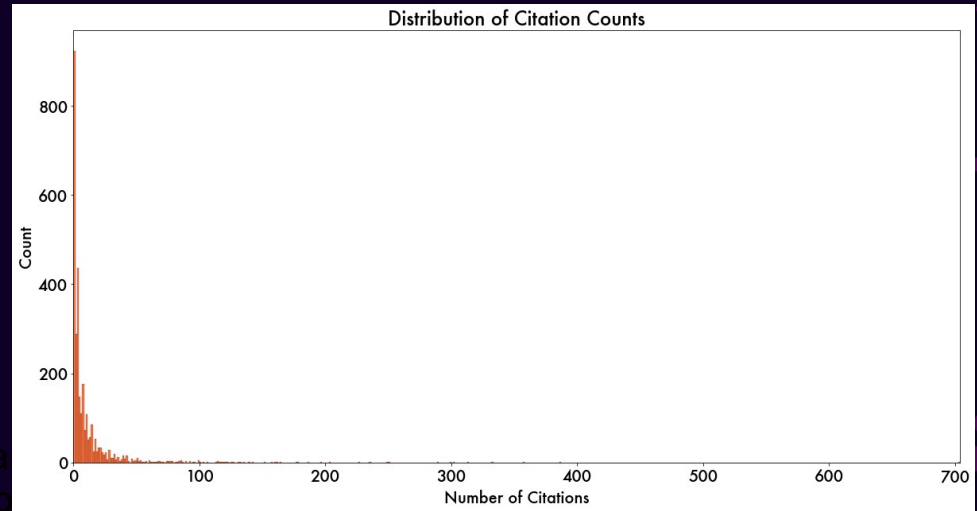
CLEANING

	pmid	citations	title	date	text
0	31229078	6.0	Alternative data mining/machine learning metho...	2019 Aug;122:25-39.	In recent years, the variety and volume of dat...
1	31226833	2.0	Is It Possible to Predict the Odor of a Molecu...	2019 Jun 20;20(12):3018.	The olfactory sense is the dominant sensory pe...
2	31222562	3.0	Overview of image-to-image translation by use ...	2019 Sep;12(3):235-248.	Since the advent of deep convolutional neural ...
3	31222375	6.0	Machine learning approaches for pathologic dia...	2019 Aug;475(2):131-138.	Machine learning techniques, especially deep l...
4	31221831	9.0	Trends and challenges in robot manipulation	2019 Jun 21;364(6446):eaat8414.	Dexterous manipulation is one of the primary g...
5	31220370	2.0	Machine learning and statistical models for pr...	2019 Sep;29(5):704-726.	Indoor air quality (IAQ), as determined by the...
6	31219658	8.0	Current status of artificial intelligence appl...	2019 Jun 20.	Objective: To investigate t...
7	31217702	6.0	Critical Care, Critical Data	2019 Jun 12;10:1179597219856564.	As big data, machine learning, and artificial ...
8	31215361	1.0	Recent Advances in Machine Learning Based Pred...	2019;26(8):601-619.	The interactions between RNAs and proteins pla...
9	31214847	6.0	Structural Imaging in Parkinson's Disease: New...	2019 Jun 18;19(8):50.	Purpose of review: To revie...

- Text and title were cleaned during scraping
- Month and Year extracted from 'date' post-scraper
- Dropped any rows with missing data → 2,986 remaining

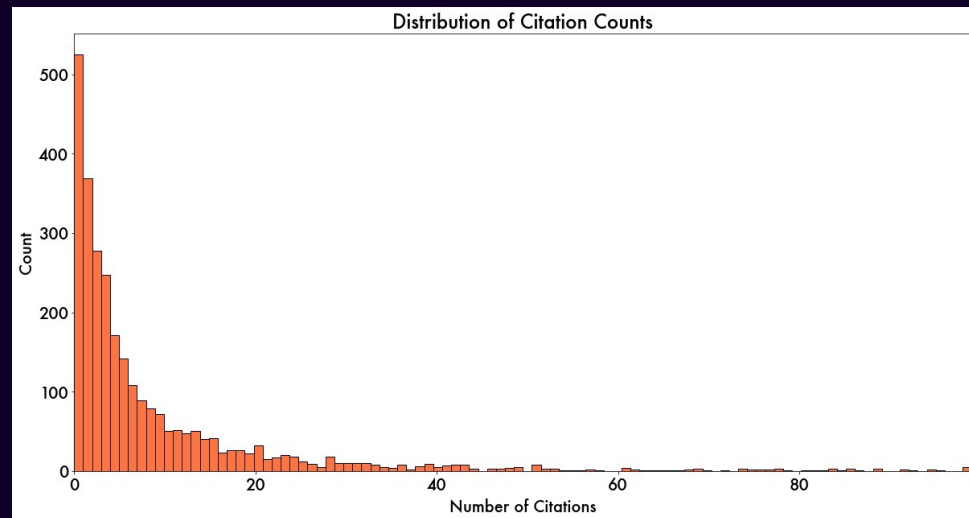
DATA EXPLORATION

- 'citations'
distribution very
right-skewed
(skew = 9.1);
several outliers
- range: (0, 704)
- Text file created during scrapping
- Month and Year extracted from 'date' parameter
- Dropped any rows with missing data



DATA EXPLORATION

- Removed outliers
(skew = 3.26)
 - 2,809 rows remain
- citations range : (0, 99)
- Explored correlations between citations &:
 - Length of title
 - Length of abstract
 - Time since publication



TEXT PREPROCESSING

- Combine title & abstract
- Remove special chars & punctuation
- Lowercase
- Remove English stopwords
- Stemming? Two versions:
 - None: jargon analysis
 - PorterStemmer: standard frequency analysis

Sample abstract

"Applying Fourier-transform infrared (FTIR) spectroscopy (or related technologies such as Raman spectroscopy) to biological questions (defined as biospectroscopy) is relatively novel. Potential fields of application include cytological, histological and microbial studies. This potentially provides a rapid and non-destructive approach to clinical diagnosis. Its increase in application is primarily a consequence of developing instrumentation along with computational techniques. In the coming decades, biospectroscopy is likely to become a common tool in the screening or diagnostic laboratory, or even in the general practitioner's clinic. Despite many advances in the biological application of FTIR spectroscopy, there remain challenges in sample preparation, instrumentation and data handling. We focus on the latter, where we identify in the reviewed literature, the existence of four main study goals: Pattern Finding; Biomarker Identification; Imaging; and, Diagnosis. These can be grouped into two frameworks: Exploratory; and, Diagnostic. Existing techniques in Quality Control, Pre-processing, Feature Extraction, Clustering, and Classification are critically reviewed. An aspect that is often visited is that of method choice. Based on the state-of-art, we claim that in the near future research should be focused on the challenges of dataset standardization; building information systems; development and validation of data analysis tools; and, technology transfer. A diagnostic case study using a real-world dataset is presented as an illustration. Many of the methods presented in this review are Machine Learning and Statistical techniques that are extendable to other forms of computer-based biomedical analysis, including mass spectrometry and magnetic resonance."



```
['extract', 'biolog', 'inform', 'comput', 'analysi', 'fouriertransform', 'infrar', '(ftir)', 'biospectroscopi', 'datasets:', 'current', 'practic', 'futur', 'perspect', 'appli', 'fouriertransform', 'infrar', '(ftir)', 'spectroscopi', '(or', 'relat', 'technolog', 'raman', 'spectroscopy)', 'biolog', 'question', '(defin', 'biospectroscopy)', 'rel', 'novel', 'potenti', 'field', 'applic', 'includ', 'cytolog', 'histolog', 'microbi', 'studi', 'potenti', 'provid', 'rapid', 'nondestruct', 'approach', 'clinic', 'diagnosi', 'increas', 'applic', 'primarili', 'consequ', 'develop', 'instrument', 'along', 'comput', 'techniqu', 'come', 'decad', 'biospectroscopi', 'like', 'becom', 'common', 'tool', 'screen', 'diagnost', 'laboratori', 'even', 'gener', 'practition', 'clinic', 'despit', 'mani', 'advanc', 'biolog', 'applic', 'ftir', 'spectroscopi', 'remain', 'challeng', 'sampl', 'prepar', 'instrument', 'data', 'handl', 'focu', 'latter', 'identifi', 'review', 'literatur', 'exist', 'four', 'main', 'studi', 'goals:', 'pattern', 'finding;', 'biomark', 'identification;', 'imaging;', 'diagnosi', 'group', 'two', 'frameworks:', 'exploratory;', 'diagnost', 'exist', 'techniqu', 'qualiti', 'control', 'preprocess', 'featur', 'extract', 'cluster', 'classification', 'critic', 'review', 'aspect', 'often', 'visit', 'method', 'choic', 'base', 'stateofart', 'claim', 'near', 'futur', 'research', 'focus', 'challeng', 'dataset', 'standardization;', 'build', 'inform', 'systems', 'develop', 'valid', 'data', 'analysi', 'tools;', 'technolog', 'transfer', 'diagnost', 'case', 'studi', 'use', 'realworld', 'dataset', 'present', 'illustr', 'mani', 'method', 'present', 'review', 'machin', 'learn', 'statist', 'techniqu', 'extend', 'form', 'computerbas', 'biomed', 'analysi', 'includ', 'mass', 'spectrometri', 'magnetic', 'reson']
```

Title + abstract post-preprocessing

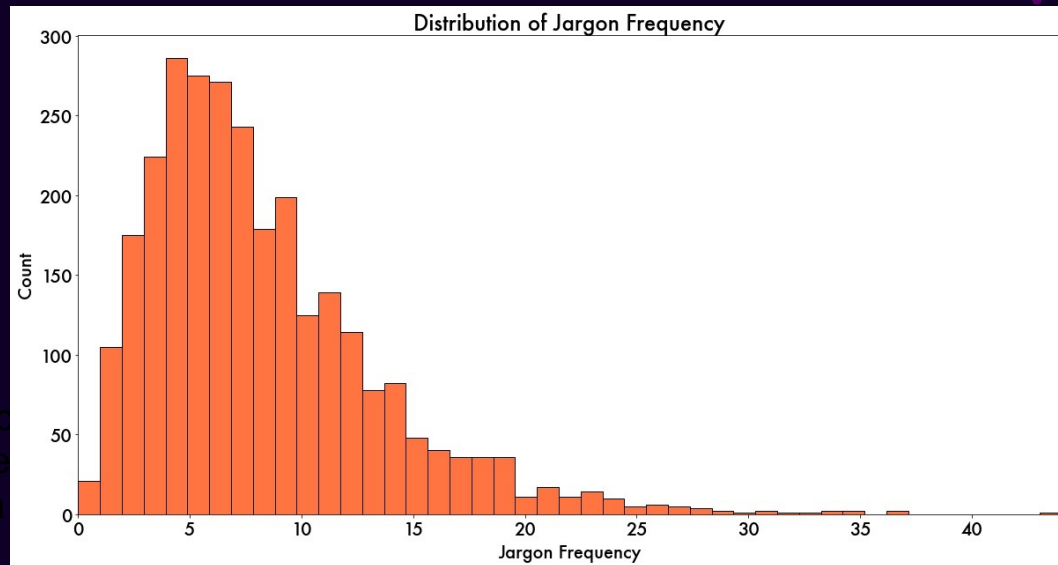
JARGON ANALYSIS

- Constructed sparse matrix of jargon vectors for each document

- **Total jargon frequency**

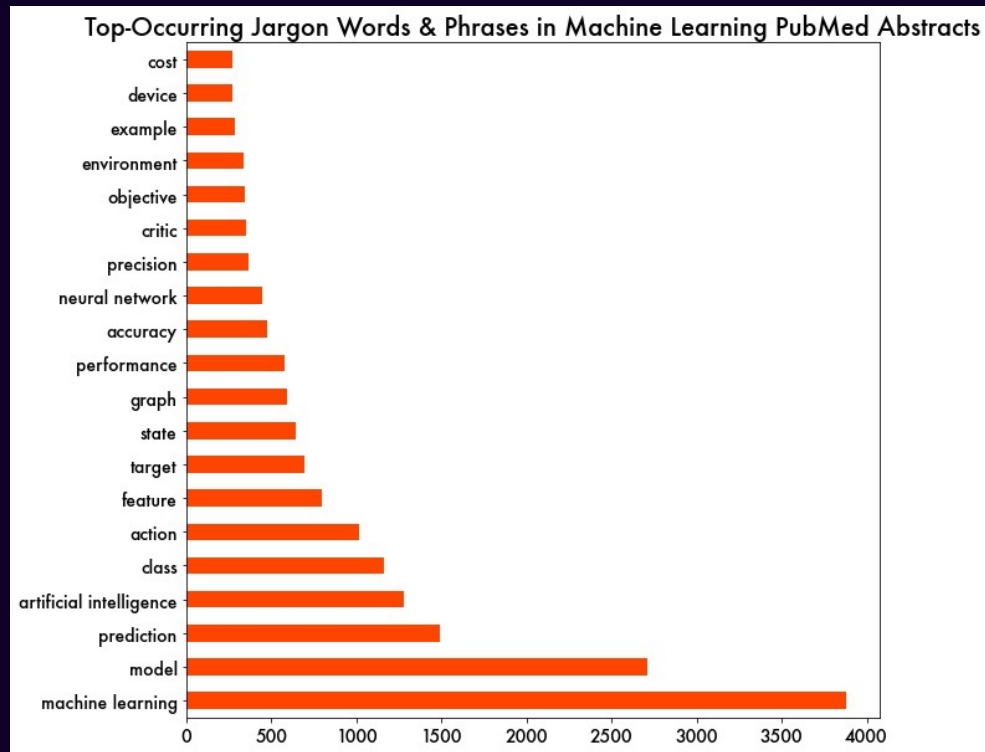
- **Individual jargon term frequency**

- Text and title were cleaned during scraping
- Month and Year extracted from 'date' column
- Dropped any rows with missing data



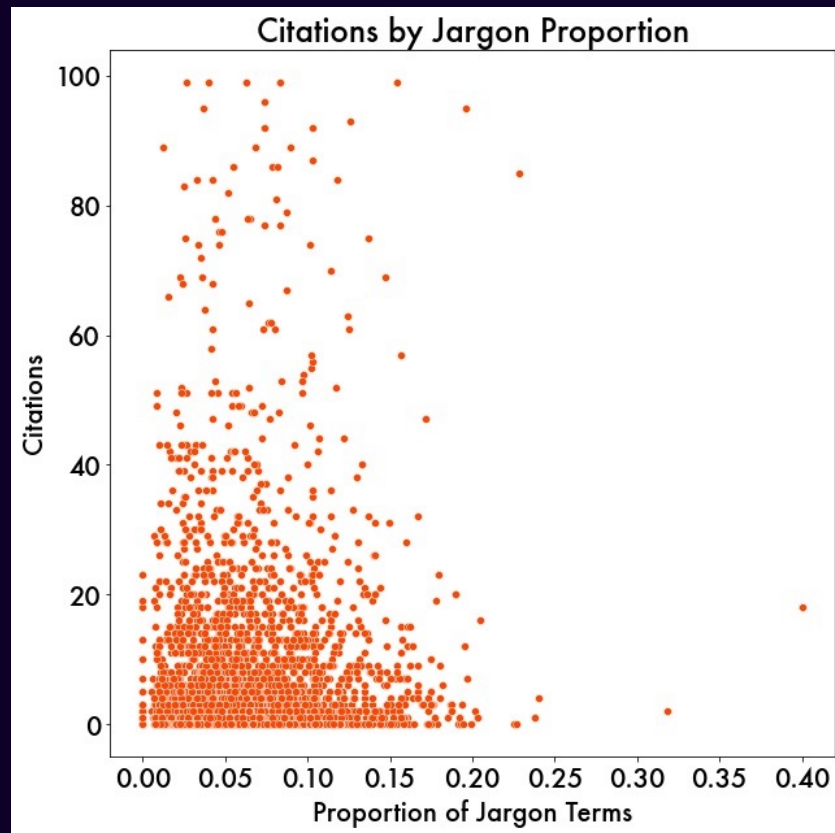
JARGON ANALYSIS

- Constructed sparse matrix of jargon vectors for each document
 - Total jargon frequency
 - Individual jargon term frequency



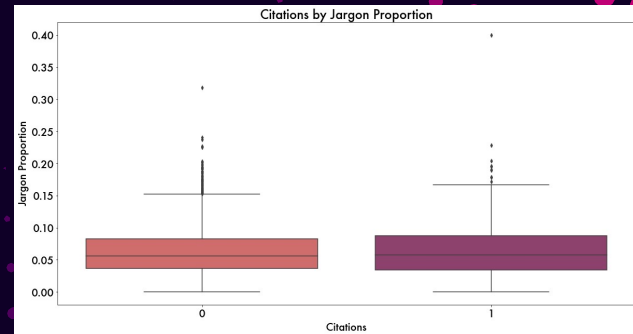
JARGON ANALYSIS

- No correlation between number of citations and proportion of jargon
 - $r = .03, p = .08$
- Failure to replicate Martinez & Mammola findings.



JARGON ANALYSIS

- Jargon frequency?
 - No correlation between # citations and jargon count ($r = 0.03$, $p = .15$)
- Specific jargon terms?
 - Regression analyses showed jargon terms have no predictive power.
- Classification model?
 - Created “highly cited” class
 - Citations > 10 (mean + 1 sd)
 - Models never above baseline accuracy



PART I CONCLUSIONS

FAILURE TO REPLICATE

- Neither jargon proportion, nor jargon frequency, nor specific jargon terms were found to be predictive of the number of times that a paper was cited.

POSSIBLE LIMITATIONS

- Insufficient data – limited by access
- Journals spanned many disciplines

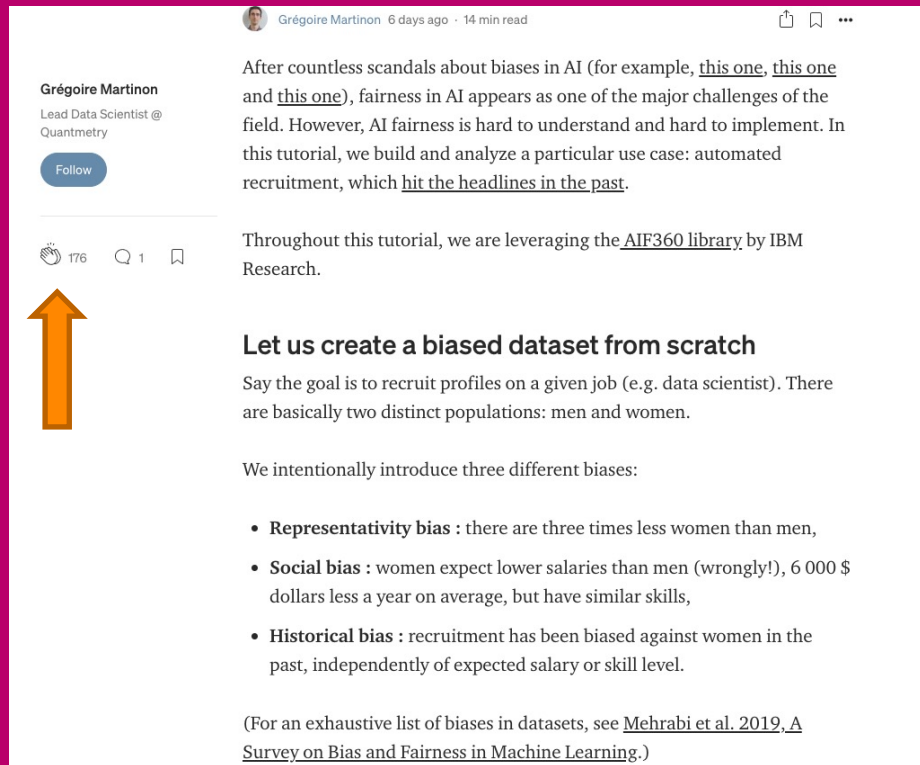
PART II: TOWARDS DATA SCIENCE

- Establish a machine learning 'jargon' dictionary ✓
- Develop a corpus of posts from Medium's Toward Data Science blog that reference 'machine learning'
- Calculate proportion of jargon words in each post
- Does using more jargon lead to fewer citations?
 - Establish appropriate (available) metric

ESTABLISH PROXY METRIC

CLAPS

- "Clapping is a way readers can show appreciation for a Medium story and recommend it to their followers."
 - The more you like something, the more you can clap (up to 50X per post).
 - Clapping plays a role in determining how much authors can earn per story.



CREATE CORPUS

- Gather 9,804 posts
 - Query posts from Towards Data Science blog
 - Scrape* **date**, **title**, **subtitle**, **section titles**, **paragraph text**, and number of **claps** from each post using the **requests** library and **BeautifulSoup**

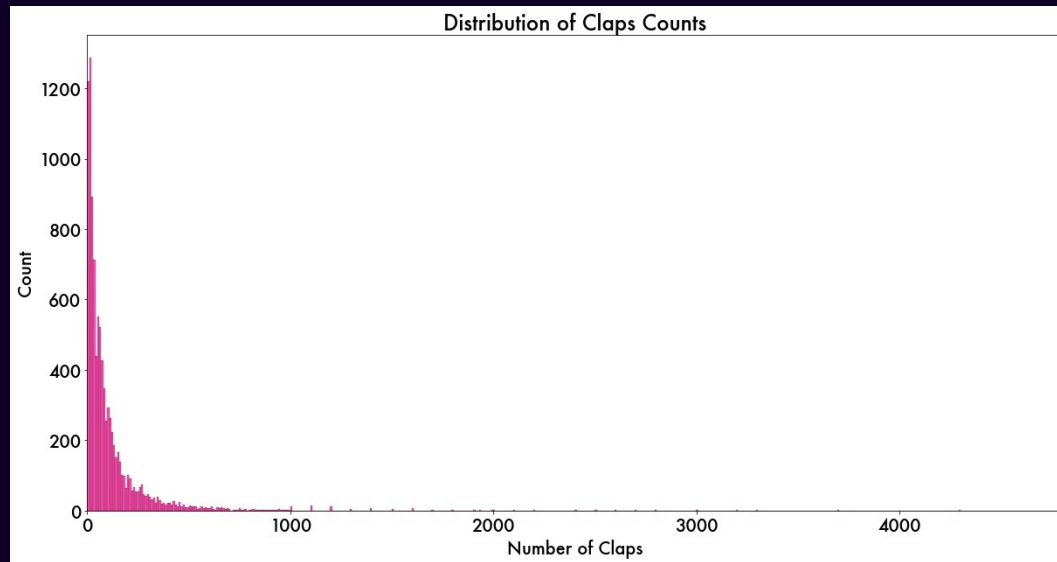
CLEANING

	date	title	subtitle	claps	responses	reading_time (mins)	number_sections	section_titles	number_paragraphs	paragraphs
0	01/01/2021	7 Most Recommended Skills to Learn in 2021 to ...	Recommended by some of the largest...	1K	10	6	11	['7 Most Recommended Skills to Learn in 2021 t...	36	['Terence Shin', 'Jan 1-8 min read', 'Happy Ne...
1	01/01/2021	The Ultimate Guide to Acing Coding Interviews ...	Data Science Interview	489	4	11	12	['The Ultimate Guide to Acing Coding Interview...	42	['Emma Ding', 'Jan 1-11 min read', 'Written by...
2	01/01/2021	Shakespeare versus Eminem—who's the better ly...	He is known for his poetry, his writings on life...	139	2	9	13	['Shakespeare versus Eminem—who's the better l...	64	['Jeroen van Zeeland', 'Jan 1-9 min read', 'Da...
3	01/01/2021	Customer Segmentation in Online Retail	A detailed step-by-step explanation on perform...	159	1	19	15	['Customer Segmentation in Online Retail', 'Un...	93	['Rahul Khandelwal', 'Jan 1-19 min read', 'In ...
4	01/01/2021	Implementing VisualTransformer in PyTorch	Hi guys, happy new year! Today we are going to...	133	2	6	6	['Implementing Vision Transformer (ViT) in PyT...	60	['Francesco Zuppicchini', 'Jan 1-6 min read', '...
5	01/01/2021	Stock Price Analysis with Pandas and Altair	Practical guide for Pandas and Altair	92	0 responses	5	1	['Stock Price Analysis with Pandas and Altair']	29	['Soner Yildirim', 'Jan 1-5 min read', 'Stock ...

- Text and title were cleaned during scraping
- Month and Year were extracted from 'date' and stored separately
- Dropped any rows with missing data
- Converted 'claps' and 'responses' to integers
- No missing data

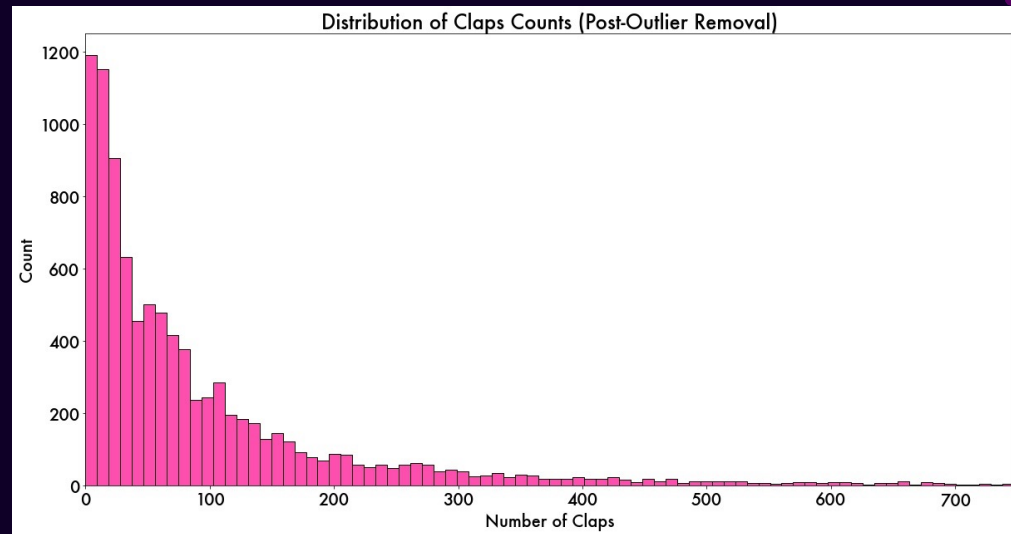
DATA EXPLORATION

- 'claps'
distribution very
right-skewed
(skew = 7.73);
several outliers
- range: (0, 4,800)



DATA EXPLORATION

- Removed outliers (skew = 2.41)
 - 9,290 posts remain
- claps range : (0, 747)
- Explored correlations between claps &:
 - Days live
 - Length of post



TEXT PREPROCESSING

- Combine text fields
- Remove special chars & punctuation
- Lowercase
- Remove English stopwords
- Stemming? Two versions:
 - None: jargon analysis
 - PorterStemmer: standard frequency analysis

'those working with neural networks know how complicated object detection techniques can be it is no wonder there is no straight forward resource for training them you are always required to convert your data to a cocolike json or some other unwanted format it is never a plug and play experience moreover no diagram thoroughly explains faster rcnn or yolo as there is for unet or resnet there are just too many details while these models are quite messy the explanation for their lack of simplicity is quite straightforward it fits in a single sentence: neural networks have fixed sized outputs in object detection you can't know a priori how many objects there are in a scene there might be one two twelve or none the following images all have the same resolution but feature different numbers of objects the one million dollar question is: how can we build variable sized outputs out of fixed sized networks plus how are we supposed to train a variable number of answers and loss terms how can we pen'

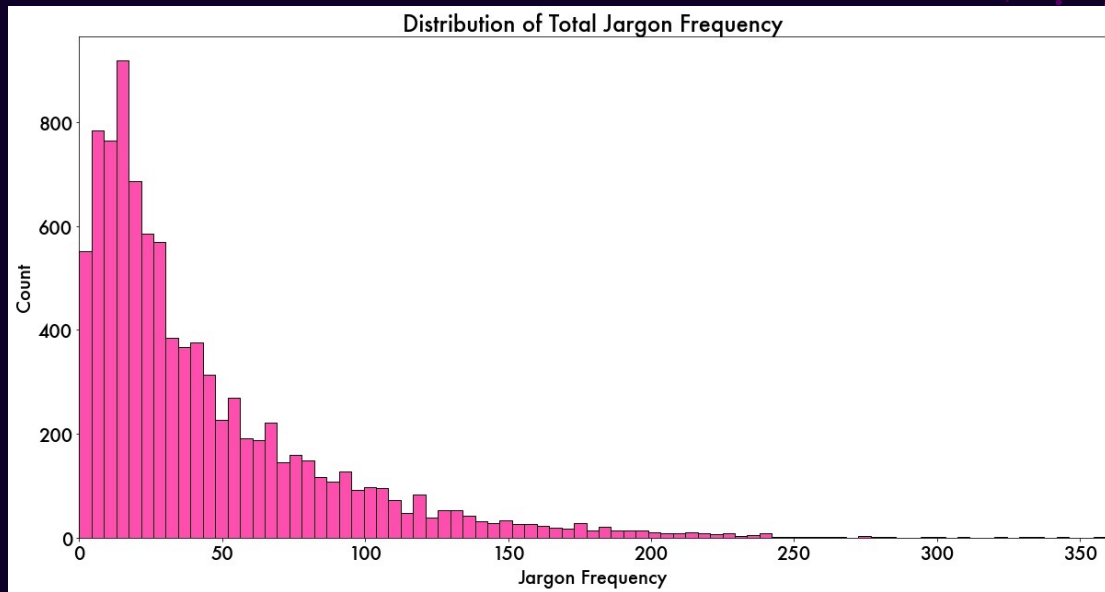
First 1,000 characters of a sample post

'work neural network know complic object detect techniqu wonder straight forward resourc train alway requir convert data cocolik json unwanted format never plug play experi moreov diagram thoroughli explain faster rcnn yolo unet resnet mani detail model quit messi explan lack simplic quit straight forward fit single sentence: neural network fixeds output object detect can't know priori mani object scene might one two twelve none follow imag resolut featur differ number object one million dollar question is: build variables output fixeds network plu suppos train variabl number answer loss term penal wrong predict creat output vari size two approach domin literature: "one size fit all" approach output broad suffic applic "lookahead" idea search regionsofinterest classifi made term 😊 practic known "onestage" "twostage" approach had less selfexplanatori overfeat yolo ssd retinanet etc can't variables output shall return output large alway larger need prune excess whole idea take greedy rout '

Post-preprocessing

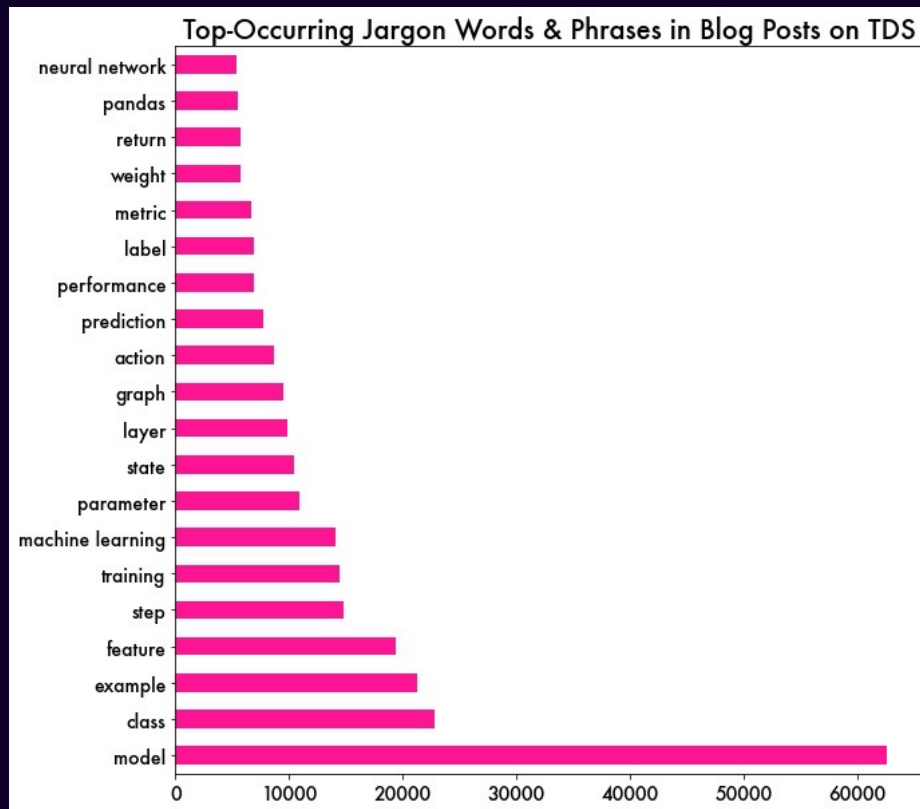
JARGON ANALYSIS

- Constructed sparse matrix of jargon vectors for each document
 - Total jargon frequency
 - Individual jargon term frequency



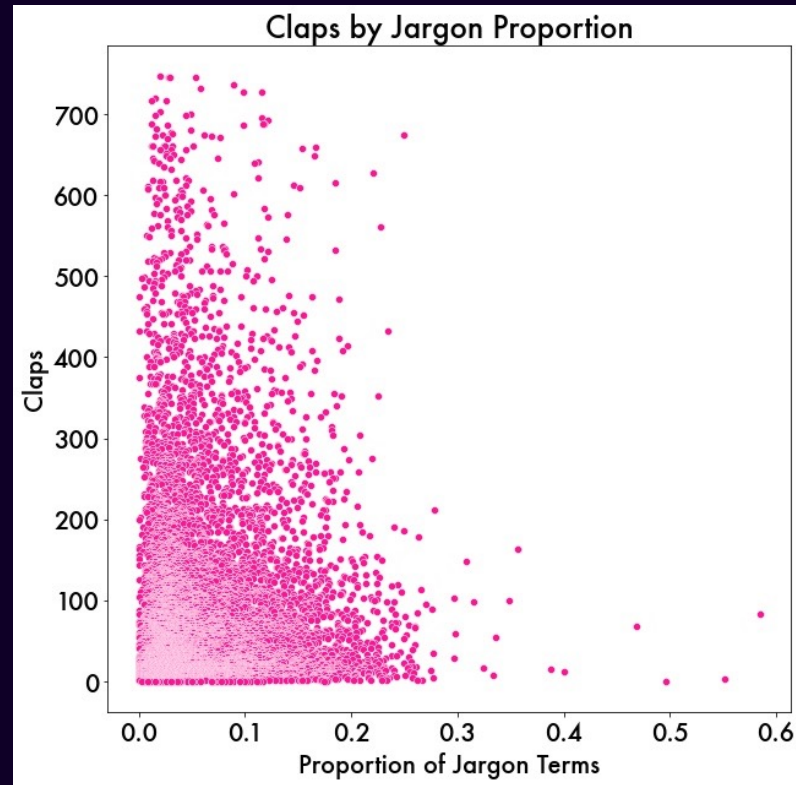
JARGON ANALYSIS

- Constructed sparse matrix of jargon vectors for each document
 - Total jargon frequency
 - Individual jargon term frequency



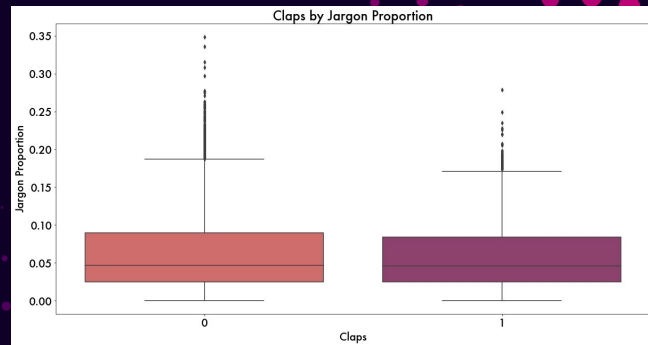
JARGON ANALYSIS

- **Extremely weak** negative correlation between number of claps and proportion of jargon
 - $r = -.02, p = .03$
- Subsequent regression analysis confirmed that the proportion of jargon accounted for practically none of the variance in 'claps'



JARGON ANALYSIS

- Jargon frequency?
 - No correlation between # claps and jargon count ($r = 0.02$, $p = .06$)
- Specific jargon terms?
 - Regression analyses showed jargon terms have no predictive power.
- Classification model?
 - Created “many claps” class
 - Claps > 208 (mean + 1 sd)
 - Models never above baseline accuracy



PART II CONCLUSIONS

FAILURE TO REPLICATE (X 2!)

- Neither jargon proportion, nor jargon frequency, nor specific jargon terms were found to be predictive of the number of claps received by a Towards Data Science post.

POSSIBLE LIMITATIONS

- Perhaps the jargon isn't the right jargon
- 'Claps' are perhaps not the ideal metric

FUTURE DIRECTIONS

JARGON ANALYSIS:

- Develop a custom jargon dictionary
- PubMed Papers:
 - Gain access to a machine learning journal
 - Within a specific discipline (i.e. cancer detection), does the phrase 'machine learning' (or 'deep learning') affect number of citations?
 - If so, how does that effect change over time?
- Towards Data Science Posts:
 - Gain access and explore other engagement metrics, like pageviews or referral links

PART III: TOPIC MODELING

- Unsupervised learning technique
- Probabilistic model that discovers the abstract “topics” in a collection of documents.
 - Using contextual clues, topic models can connect words with similar meanings and distinguish between uses of words with multiple meanings
- Identify, then classify.
- Is there a specific topic that gets more ‘claps’ on Towards Data Science?

TOPIC MODELING APPROACH

- LDA (Latent Dirichlet Allocation)*
 - Generative probabilistic model that represents documents as mixtures of topics
- Implement in python with Gensim, Spacy, and pyLDAvis libraries
 - Plus a Gensim wrapper for MALLET (MAchine Learning for Language Toolkit) LDA topic modeling, which uses Gibbs sampling – an MCMC algorithm!
- Determine topics present in corpus and dominant topic in each document

**Source: [Journal of Machine Learning Research](#)*

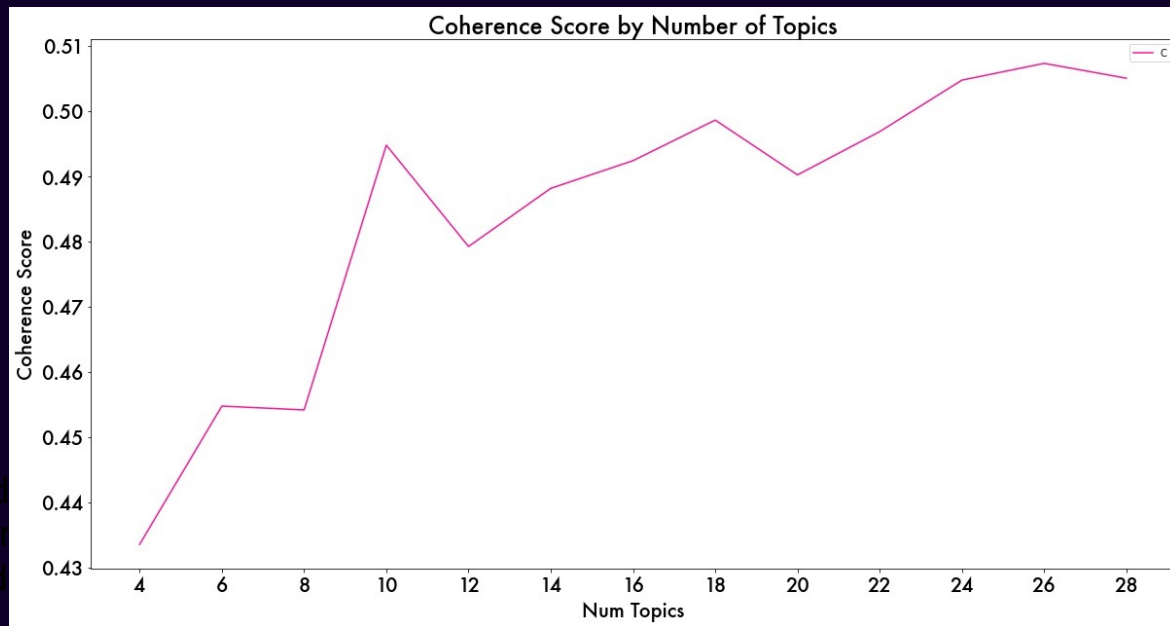
WHAT DOES A TOPIC LOOK LIKE?

- Collection of keywords:
 - probability, distribution, state, function, sample, number, algorithm, give, problem, time
 - word, text, model, language, topic, document, article, sentence, nlp, search
 - data, database, system, service, cloud, process, pipeline, tool, query, store

TOPIC MODEL SELECTION

- Key Hyperparameters
 - Bigrams vs Trigrams
 - Number of Topics
- Selection criteria
 - Metrics
 - **Coherence** vs Perplexity
 - Human judgment
 - Qualitative evaluation of topic keywords
 - Qualitative visual evaluation of topic spread/overlap

COHERENCE SCORE



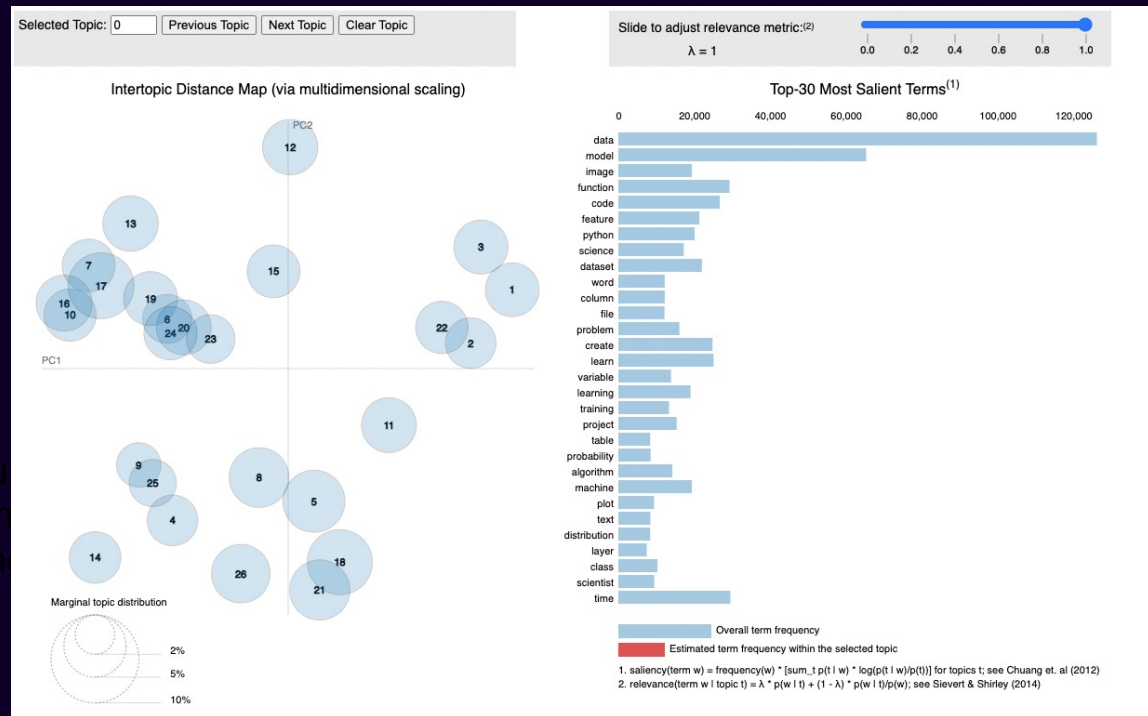
- Peaking at 26 topics, followed by 24, with 18 and 10 close behind.

TOPIC MODEL: 26 TOPICS

- Coherence score: 0.507

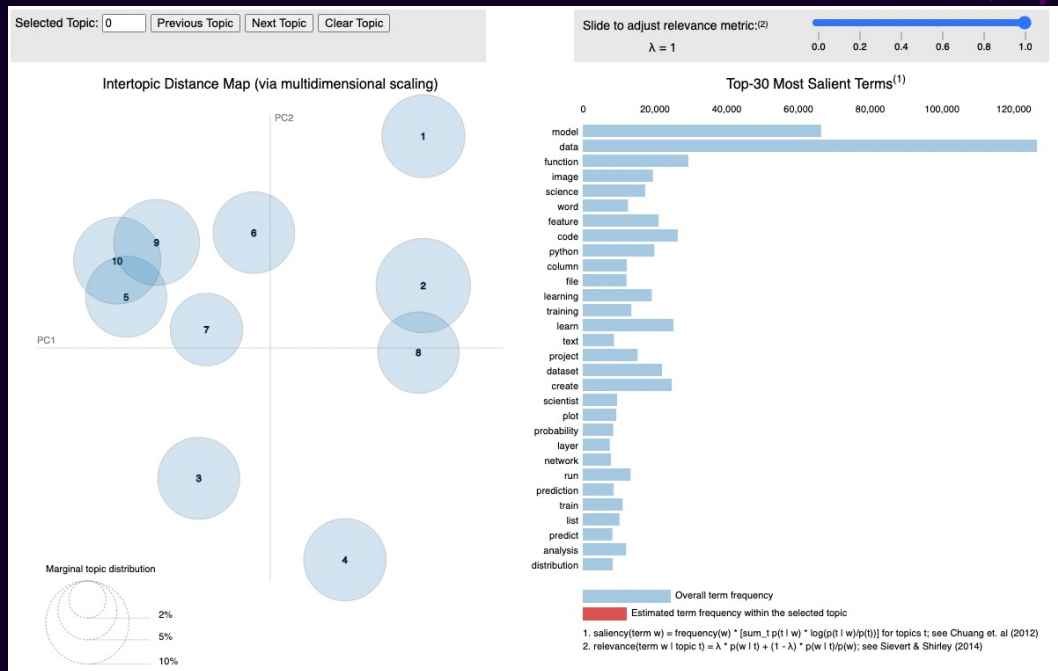
- Good distribution

- Text and title were cleaned during preprocessing
 - Removed any rows with missing values
 - Dropped any rows with missing values



TOPIC MODEL: 10 TOPICS

- Coherence score: 0.495
- This model provides the best balance of spread, disambiguation, and coherence score.



10-TOPIC KEYWORDS

ai, game, system, people, human, world, team, player, make, decision	data, science, learn, work, scientist, project, machine, good, time, question	function, column, python, data, code, list, method, create, table, type	code, file, create, run, python, project, follow, package, app, api	image, model, network, layer, training, learning, input, deep, learn, train
data, plot, analysis, time, show, cluster, visualization, number, dataset, customer	word, text, model, language, topic, document, article, sentence, nlp, search	data, database, system, service, cloud, process, pipeline, tool, query, store	probability, distribution, state, function, sample, number, algorithm, give, problem, time	model, data, feature, dataset, prediction, machine, predict, set, learning, test

APPLY TOPIC MODEL TO DOCUMENTS

FIND DOMINANT TOPIC FOR EACH DOCUMENT

GETTING STARTED

The Ultimate Guide to Hypothesis Testing and Confidence Intervals in Different Scenarios

a step-by-step tutorial for one-sample and two-sample mean, proportion statistical inference

Zijing Zhu Jan 7 · 13 min read

Statistical inference is the process of making reasonable guesses about the population's distribution and parameters given the observed data. Conducting hypothesis testing and constructing confidence interval are two examples of statistical inference. Hypothesis testing is the process of calculating the probability of observing sample statistics given the null hypothesis is true. By comparing the probability (P-value) with the significance level ($1-\alpha$), we make reasonable guesses about the population parameters from which the sample is taken. With a similar process, we can calculate the confidence interval with a certain confidence level. A confidence interval is an interval estimation for a population parameter, which is point estimation plus and minus the critical value times sample standard error. This article will discuss the standard procedure of conducting hypothesis testing and estimating confidence intervals in the following different scenarios:

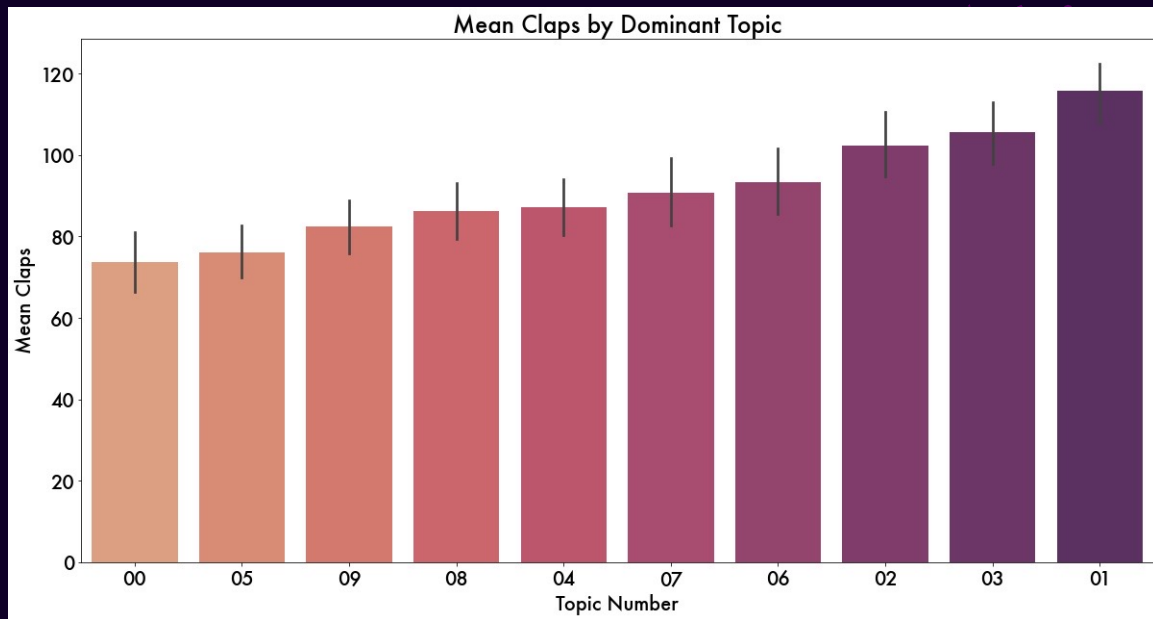


Topic 8:

- probability, distribution, state, function, sample, number, algorithm, give, problem, time
- 0.87 percent contribution

IS THERE A TOPIC THAT GETS MORE 'CLAPS'?

- 115.8 clap mean (Topic 1):
 - model, data, feature, dataset, prediction, machine, predict, set, learning, test
- 73.8 clap mean (Topic 0)
 - data, database, system, service, cloud, process, pipeline, tool, query, store
- But topic accounts for almost none of the variance in claps ($R^2 = 0.013$)



PART III: CONCLUSIONS & FUTURE DIRECTIONS

POST TOPIC IS NOT PREDICTIVE OF CLAPS

- The "Ultimate Guide to Machine Learning Prediction Models with Data & Features" posts may get more claps, on average, than "Using Cloud Service Tools to Process & Store Your Pipeline Data" – but not due to subject matter.

FUTURE DIRECTION

- Revisit with a better metric; control for other effects, like author popularity

KEY TAKEAWAYS FOR DATA SCIENCE COMMUNICATION

- Write what you know.
- Put yourself in the shoes of your intended audience.
 - Is the language appropriate?
 - Is your message clear to the audience?
- “Before you leave the house, look in the mirror and take one thing off.”

RESOURCES

- [Distill Research Journal](#)
- [Up Goer Five Challenge](#)

THANKS!

Acknowledgements:

With much gratitude to Chuck, Varun, & Grant for your knowledge, guidance, & support.

Special thanks to DSIR-2221E for your educational, entertaining, and overall excellent company on this journey!

CREDITS

Special thanks to all the people who made and released these awesome resources for free:

- ⊗ Presentation template by [SlidesCarnival](#)
- ⊗ Photographs by [Unsplash](#)