



SIGNAL:

An online reading recommender to eliminate noise

Justin Wallander

OVERVIEW

01

WHY + HOW

The beginnings of my project

02

EDA

Exploratory Data Analysis

03

NLP/ MODELING

Decisions made during NLP and the modeling process

04

RESULTS + NEXT STEPS

Recommendations + what I will do to expand

01

WHY + HOW

The beginnings of my project

Jul 9 2020

Matt Ridley: How Innovation Works, Part 1

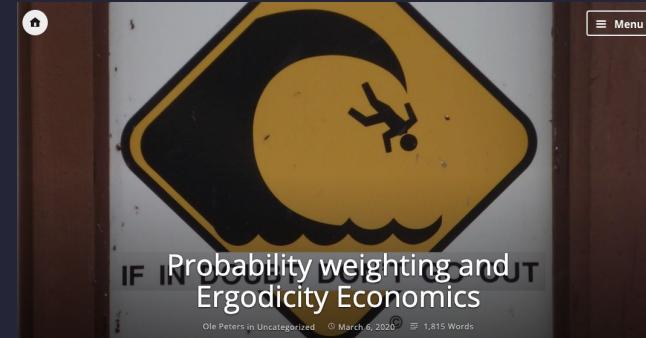
Innovation is the child of freedom

29:49 Get podcast ~

Naval interviews Matt Ridley, the author of *The Red Queen* and, recently, *How Innovation Works*.

Transcript

Naval: I don't have heroes, but there are people who I look up to and have learned a lot from, and Matt Ridley, who is on the line, has got to be near the top of that list. Growing up, I was a voracious reader, especially of science. Matt had a bigger influence on pulling me into science, and a love of science, than almost any other author. His first book that I read was called *Genome*. I must have six or seven dog-eared copies of it lying around in various boxes. It helped me define what life is, how it works, why it's important, and placed evolution as a binding principle in the center of my worldview. That's a common theme that runs across



Probability weighting and Ergodicity Economics

Ole Peters in Uncategorized ⚒ March 6, 2020 1,815 Words

Ergodicity Economics conference
Join us in January 2021:
lmi.org.uk/ee2021/

LECTURE NOTES

One key observation that helped launch the field of behavioral economics into stardom is called probability weighting: a human cognitive bias to assign higher probabilities to extreme events than ... well, than what? Than what someone else thinks the probabilities should be. Below, I will present a very simple mechanistic explanation, most of all for the iconic probability weighting figure in (Tversky and Kahneman, 1992). The result is a now familiar theme: (behavioral) economics expresses a more or less robust observation in psychological terms, as a persistent cognitive error. Ergodicity econ! [Follow](#) ...



WAIT BUT WHY
new post every sometimes

home about archive minis the shed dinner table store support wb

What Makes You You?

December 12, 2014 By Tim Urban

Note: If you want to print this post or read it offline, the PDF is probably the way to go. You can [buy it here](#).

When you say the word "me," you probably feel pretty clear about what that means. It's one of the things you're clearest on in the whole world—something you've understood since you were a year old. You might be working on the question, "Who am I?" but what you're figuring out is the *who am* part of the question—the *I* part is obvious. It's just you. Easy.

But when you stop and actually think about it for a minute—about what "me" really boils down to at its core—things start to get pretty weird. Let's give it a try.

HOME ABOUT NEWSLETTER READING LIST BLOG BEST ARTICLES SPEAKING COURSES CONTACT

RYAN HOLIDAY
MEDITATIONS ON STRATEGY AND LIFE

BLOG The Narrative Fallacy

When I first moved to LA, I didn't have enough money to buy a bed. I borrowed an IKEA futon and slept on the floor for almost two months. Now I know that you can get some really comfortable futons these days, especially if you look on somewhere like ReviewingThis, but this thing I was sleeping on was definitely not that. I was so stressed and scared that I would wake up in the middle of the night just soaked in sweat. My parents practically disowned me.

Here's the thing. I could make that all into some dramatic story —

Ideas/Goals:

How do I decide what to read?

How do I choose out of the seemingly limitless options?

Am I trapped in an echo chamber?

Am I learning anything?

DATA



STORAGE



AWS EC2 Instance



VISUALIZATION

A teal circular logo containing the word "Seaborn" in a white sans-serif font.

A white rectangular logo containing the text "pyLDAvis" in a dark grey sans-serif font.

WORKFLOW



TECH STACK

MACHINE LEARNING



LANGUAGE



02

EDA

Exploratory Data Analysis

23,314

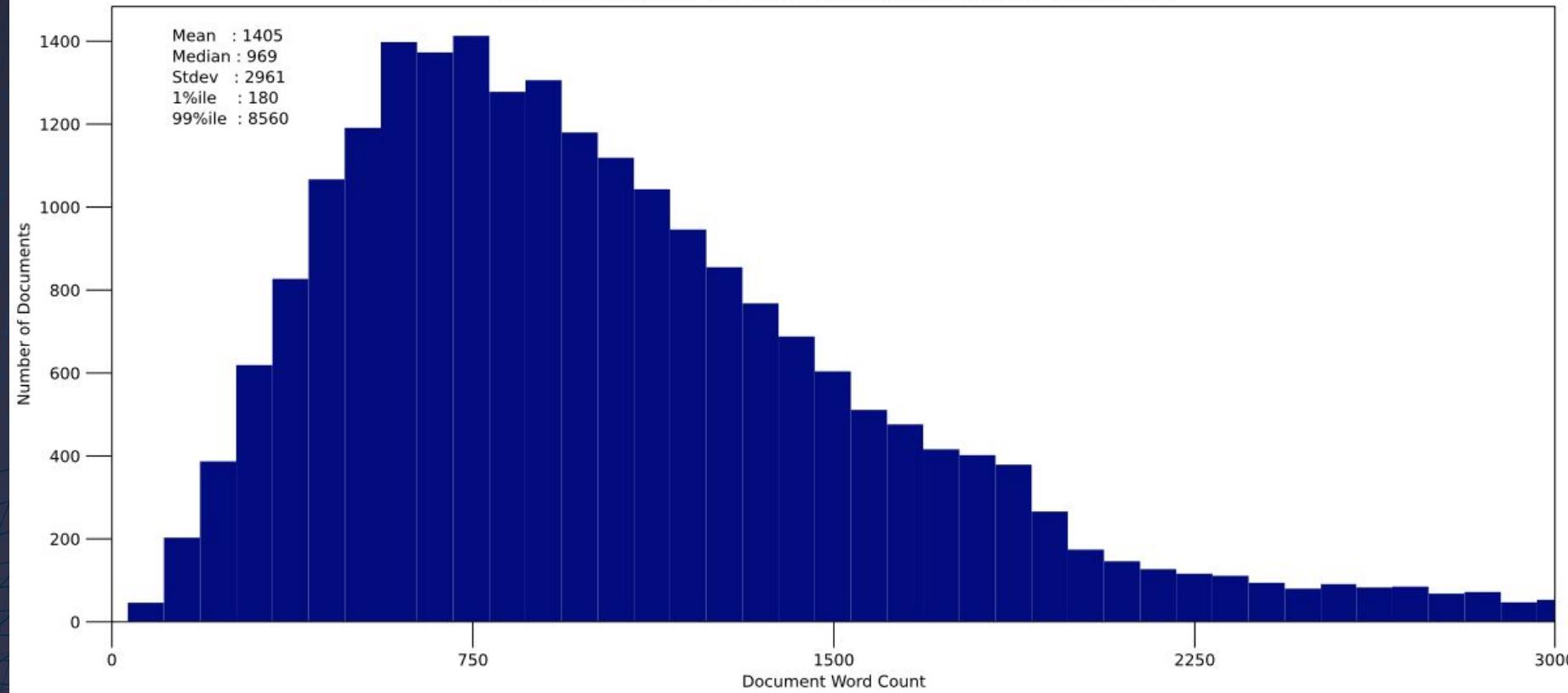
documents in initial corpus

12,000 randomly selected abstracts from arXiv database
(https://arxiv.org/help/bulk_data)

11,314 articles from 20 Newsgroup
(<http://qwone.com/~jason/20Newsgroups/>)

102 articles from fellow students, the majority of which are from Medium

Distribution of Document Word Counts



03

NLP/ MODELING

Decisions made during the
modeling process

Topic Modeling

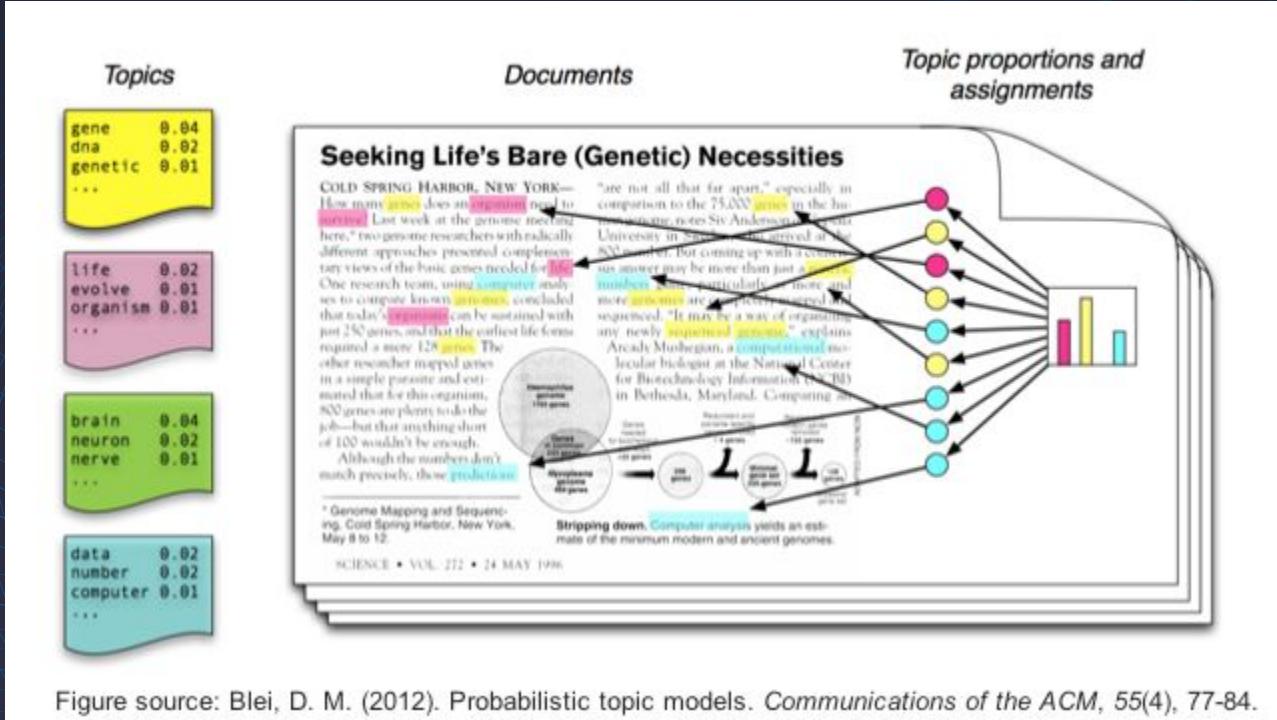
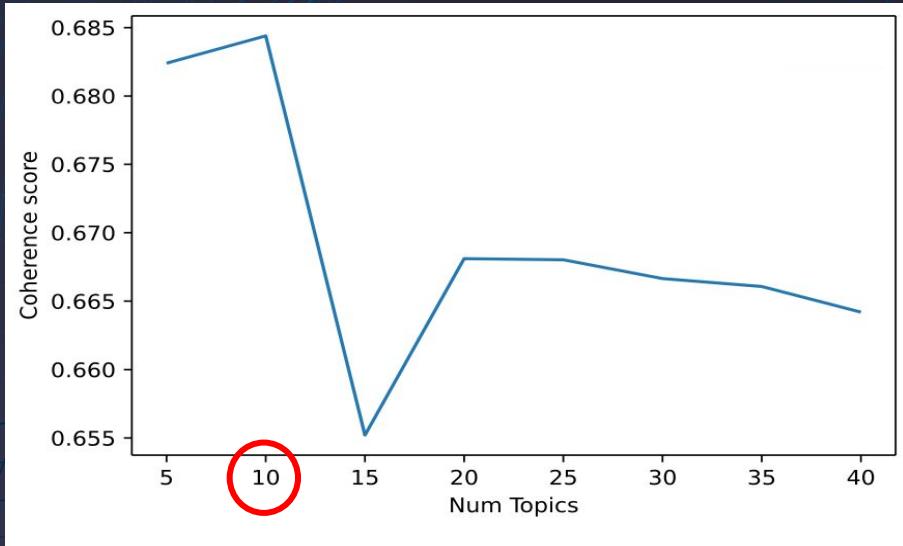


Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.

Latent Dirichlet Allocation (LDA) is a form of topic modeling that takes in a corpus of documents and provides two probability distributions: one for latent topics in documents and one for words in topics.

Model Metric



Coherence Score:

Much like a set of facts is said to be **coherent if they support each other**,
Coherence Scores have 4 supporting factors:

- Segmentation
- Probability calculation
- Confirmation measure
- Aggregation

Top Topics Breakdown

	Keywords	# of Docs	% of Corpus
Science	model, find, energy, result, show, state, high, study, field	5,599	24%
Journalism	write, would, go, people, think, article, be, make, say	4,710	20%
Computers	file, write, system, window, program, need, would, key	4,332	18%
Math	show, give, result, theory, function, space, problem, solution, set, prove	3,620	16%
IT	system, network, model, method, base, propose, time, datum, problem	2,145	9%
Politics	would, say, right, people, government, write, gun, article, go	1,063	5%

04

RESULTS + NEXT STEPS

Recommendations + what I will
do to expand

USER PROFILES



Justin
“Alice in Wonderland”

Top Topics	
Journalism	0.35
IT	0.18
Philosophy	0.18
Learning	0.08
Politics	0.08

Similarity

Rakhi:	0.9385
Lester:	0.9632



Rakhi
“Social Philosopher”

Top Topics	
Journalism	0.42
Philosophy	0.21
Politics	0.17
Learning	0.06
IT	0.05

Similarity

Justin:	0.9385
Lester:	0.9508



Lester
“Contrarian Investor”

Top Topics	
Journalism	0.34
Politics	0.15
IT	0.13
Philosophy	0.11
Learning	0.11

Similarity

Rakhi:	0.9508
Justin:	0.9632

Article:

fs



How to Make Smart Decisions Without Getting Lucky

Few things will change your trajectory in life or business as much as learning to make effective decisions.

The decision-making principles in this article aren't pulled out of thin air. They're the result of many years of experience and experimentation. They draw upon the combined expertise of some of history's deepest thinkers. They summarize the core insights and skills from influential books on decision-making.

(<https://fs.blog/smart-decisions/>)

Corpus recommendations (23,314 documents):

- "Is MSG Sensitivity Superstition?"
 - Topic: sci.med
 - Similarity: **0.9922**

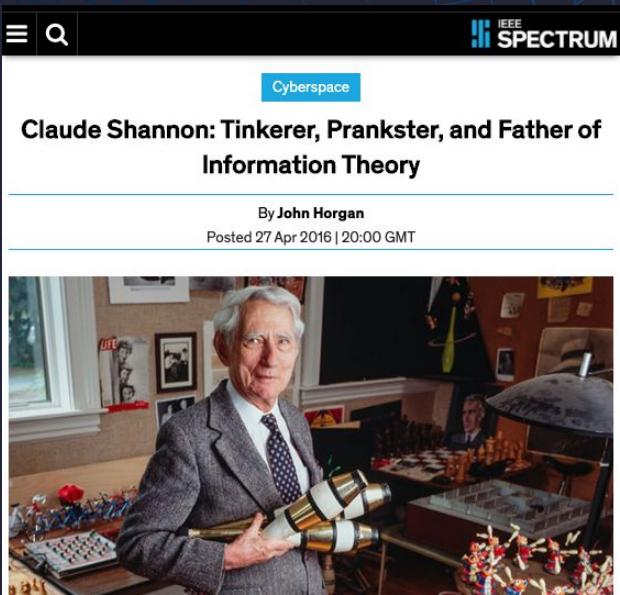
- "Selective Placebo"
 - Topic: sci.med
 - Similarity: **0.9916**

User recommendations (102 documents):

- "The Math Equation That Tried to Stump the Internet"
 - User: Rakhi
 - Topic: Math
 - Similarity: **0.9609**

- "The Fullest Look Yet at the Racial Inequality of Coronavirus"
 - User: Rakhi
 - Topic: Social Justice
 - Similarity: **0.9587**

Article:



The screenshot shows a news article from IEEE SPECTRUM's Cyberspace section. The title is "Claude Shannon: Tinkerer, Prankster, and Father of Information Theory". It was written by John Horgan and posted on April 27, 2016. The author's photo shows him holding a large, complex mechanical device, likely a model of a communication system. In the background, there's a chessboard and various scientific models.

Claude Shannon: Tinkerer, Prankster, and Father of Information Theory

By John Horgan
Posted 27 Apr 2016 | 20:00 GMT



(<https://spectrum.ieee.org/tech-history/cyberspace/clause-shannon-tinkerer-prankster-and-father-of-information-theory>)

Corpus recommendations (23,314 documents):

- "Fractals? What good are they?"
 - Topic: comp.graphics
 - Similarity: 0.9760

- "Eumemics (was: Eugenics)"
 - Topic: sci.med
 - Similarity: 0.9598

User recommendations (102 documents):

- "Preventing Suicide The Modern Way"
 - User: Rakhi
 - Topic: Mental Health
 - Similarity: 0.9096

- "How To Destroy Surveillance Capitalism"
 - User: Lester
 - Topic: Economics/Politics
 - Similarity: 0.8918

Next Steps



MODELING

Experiment with model based recommenders and transfer learning that combine topic distribution with user-inputted rankings of articles



INTEGRATION

Create pipeline for Evernote/ Roam Research that allows users to pull saved articles directly into the recommender



WEB APP

Create a web app where users can input an article's body and receive recommendations for future reading

THANKS!

Shoutouts:

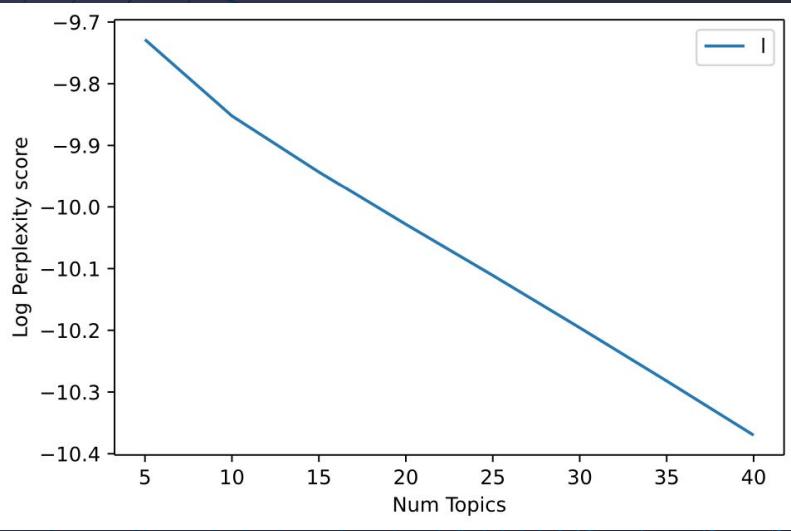
Rakhi, Lester, and Monica for sending me articles. Greatly appreciated, thank y'all!

Does anyone have any questions?
If so, ask away!

justin.wallander@gmail.com

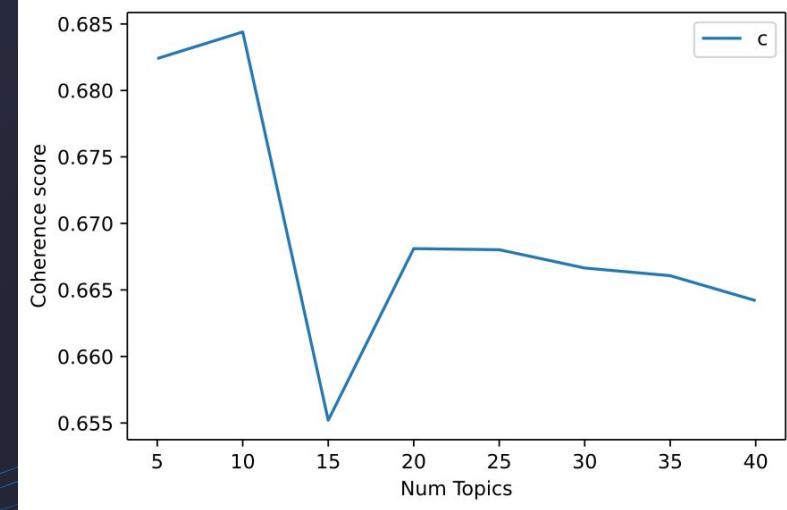
www.github.com/justin-wallander

www.linkedin.com/in/justin-wallander



Perplexity:

Perplexity is how the model reacts to unseen data and is the normalized log likelihood on a hold out test set.

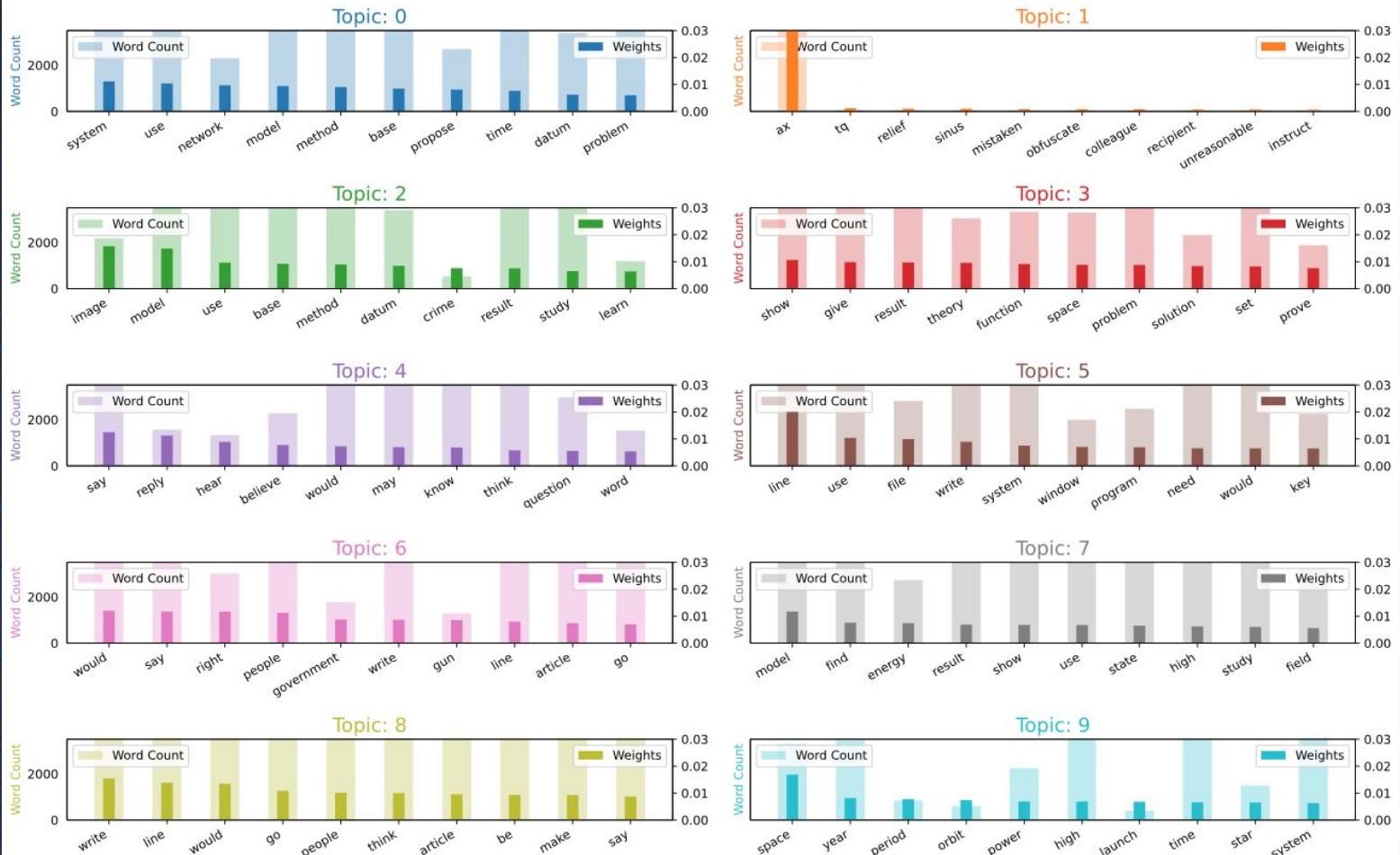


Coherence Scores:

There are quite a few scores to pick from but the two most popular are C_umass and C_v. I went with C_v because it generally leads to the best topics. C_umass is optimized for speed.

C_v uses sliding window segmentation and utilizes normalized pointwise mutual information and cosine similarity as indirect confirmation measures. And in case you are wondering, PMI is the $\log p(x,y) / p(x)p(y)$ and npmi is just with a normalized $h(x)$ which is joint self-information, estimated as $-\log p(X=x, Y=y)$.

Word Count and Importance of Topic Keywords



PyLDAvis

... pretty nifty

