SIGNAL:

An online reading recommender to eliminate noise

Justin Wallander

OVERVIEW

01 **WHY + HOW**
The beginnings of my project

02 **EDA**
Exploratory Data Analysis

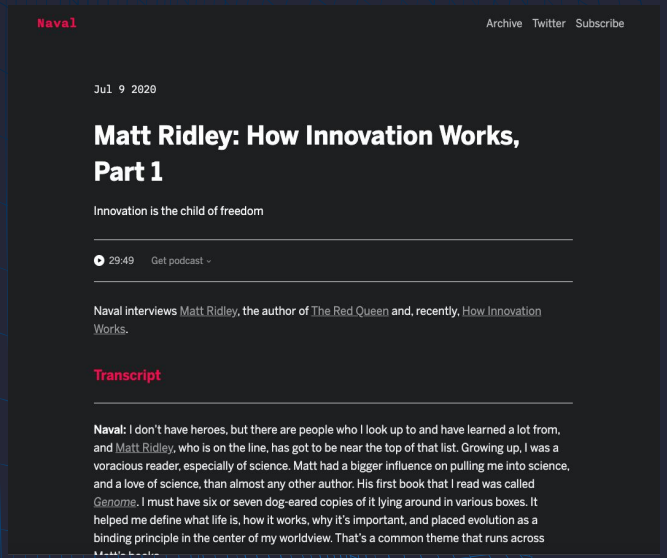
03 **NLP/ MODELING**
Decisions made during NLP and the modeling process

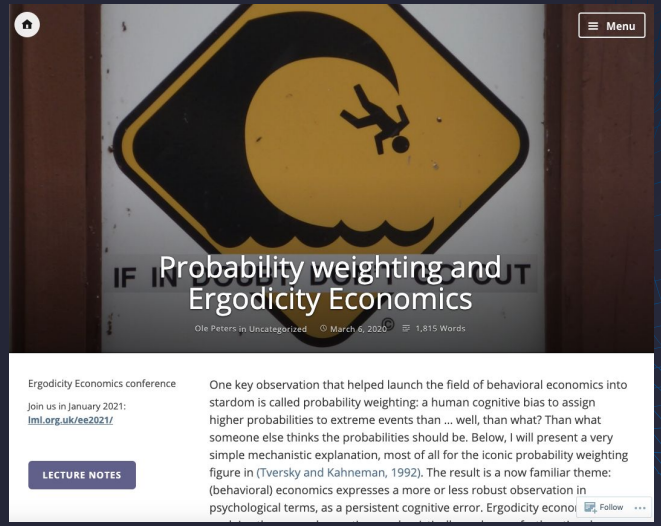
04 **RESULTS + NEXT STEPS**
Recommendations + what I will do to expand

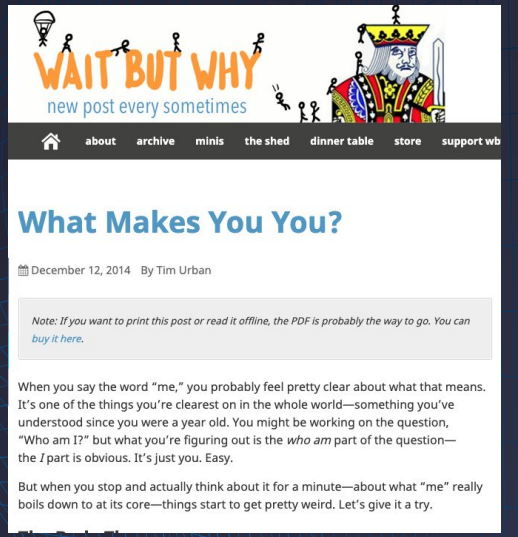
01

WHY + HOW

The beginnings of my project







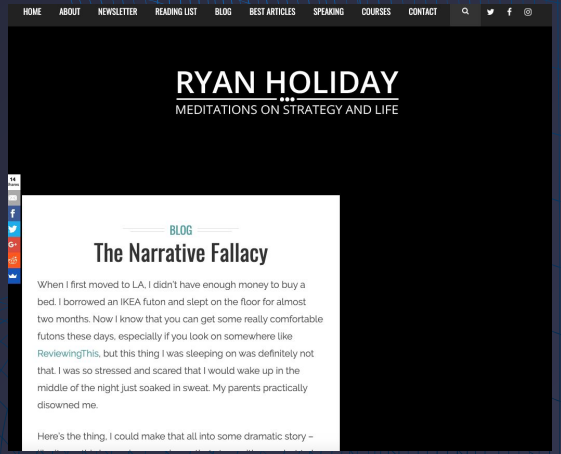
Ideas/Goals:

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?



TECH STACK

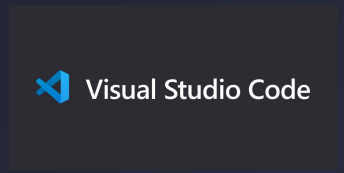
DATA

VISUALIZATION

MACHINE LEARNING

STORAGE

WORKFLOW



LANGUAGE

02

EDA

Exploratory Data Analysis

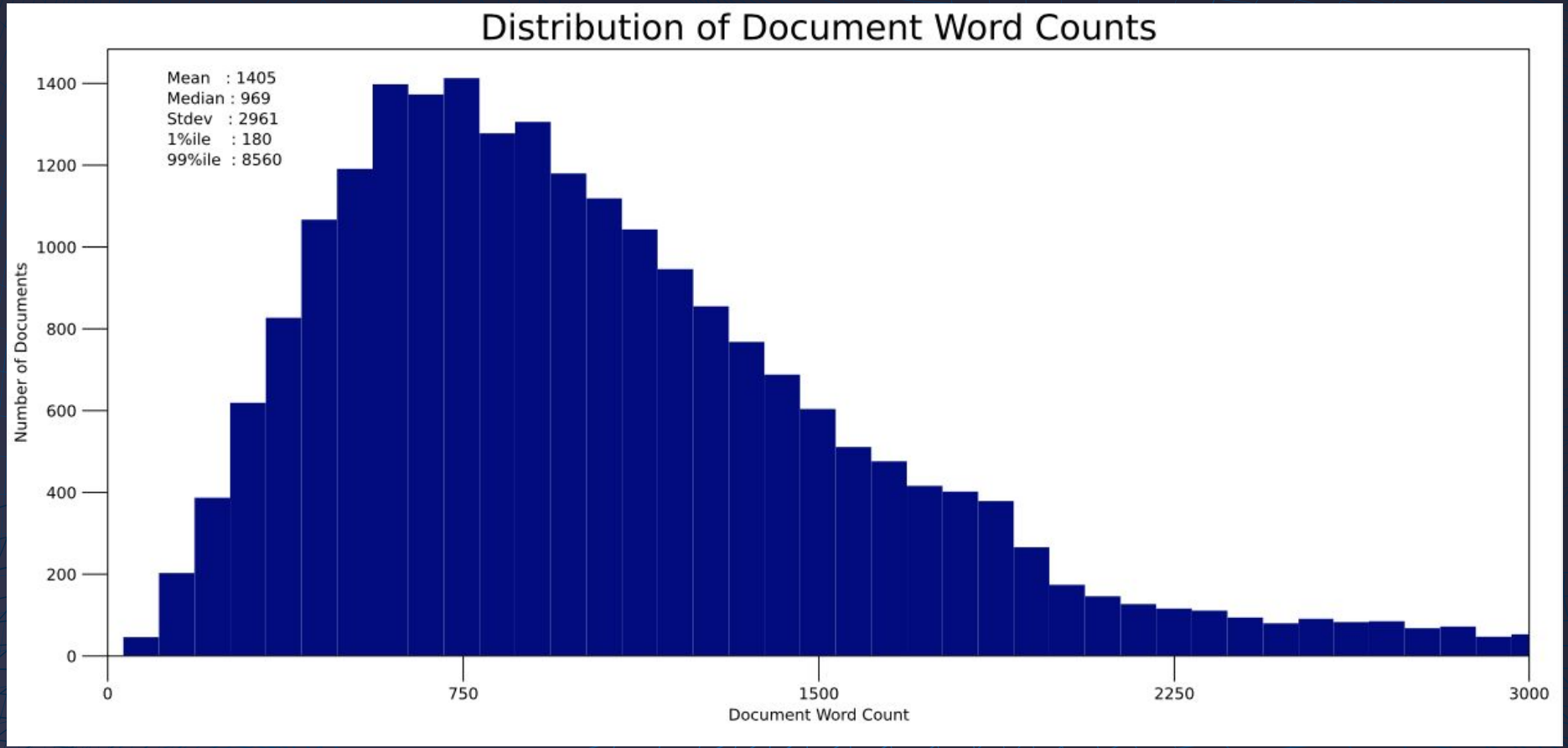
23,314

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
Mean : 1405
Median : 969
Stdev : 2961
1%ile : 180
99%ile : 8560
Number of Documents
0
200
400
600
800
1000
1200
1400
0
750
1500
2250
3000
Document Word Count

03

NLP/ MODELING

Decisions made during the modeling process

Topic Modeling

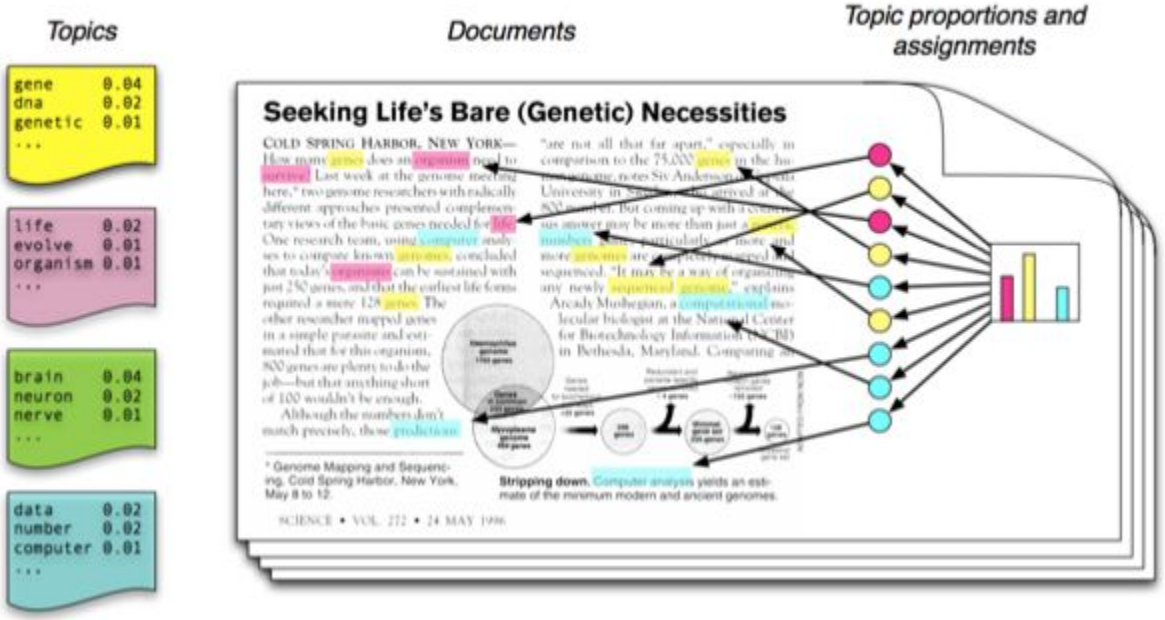
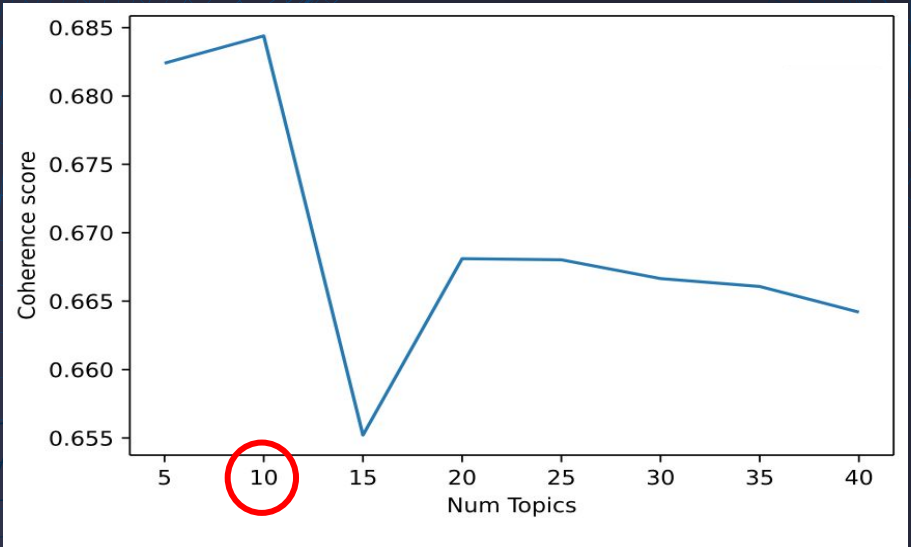


Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

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

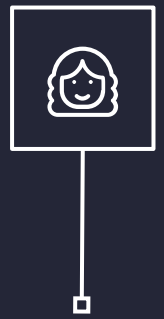
“Alice in Wonderland”

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/)

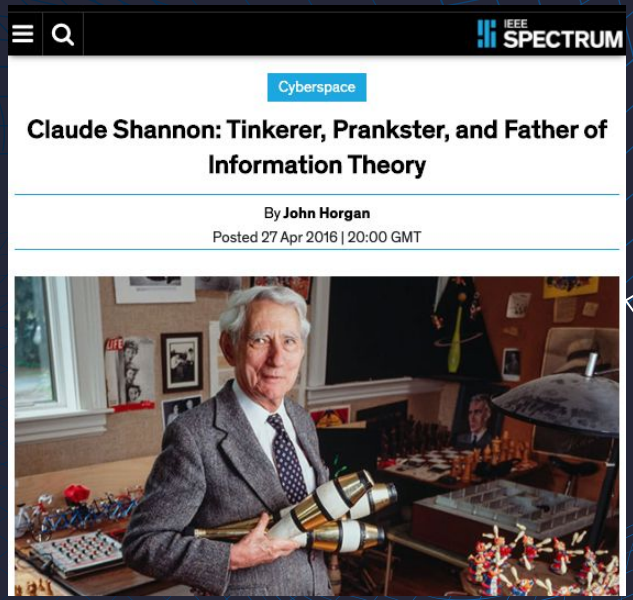
Corpus recommendations (23,314 documents):

- **"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:



(https://spectrum.ieee.org/tech-history/cyberspace/claude-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

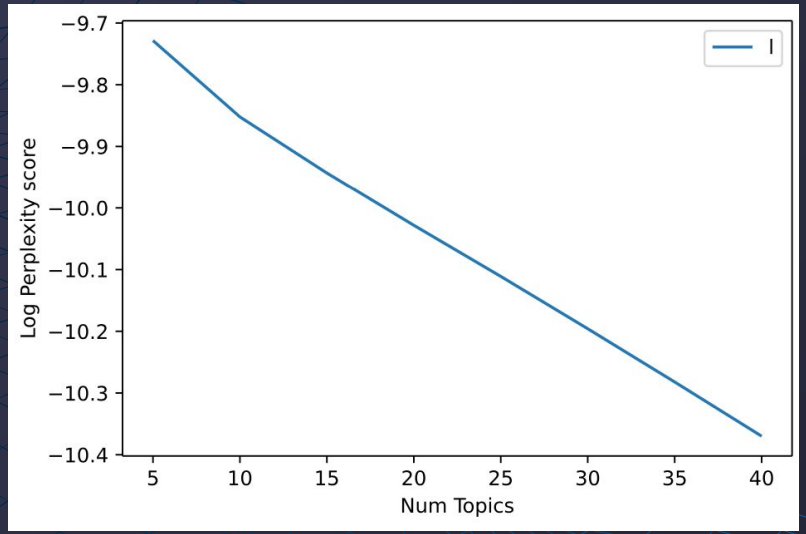
Create a web app where users can input an article's body and receive recommendations for future reading

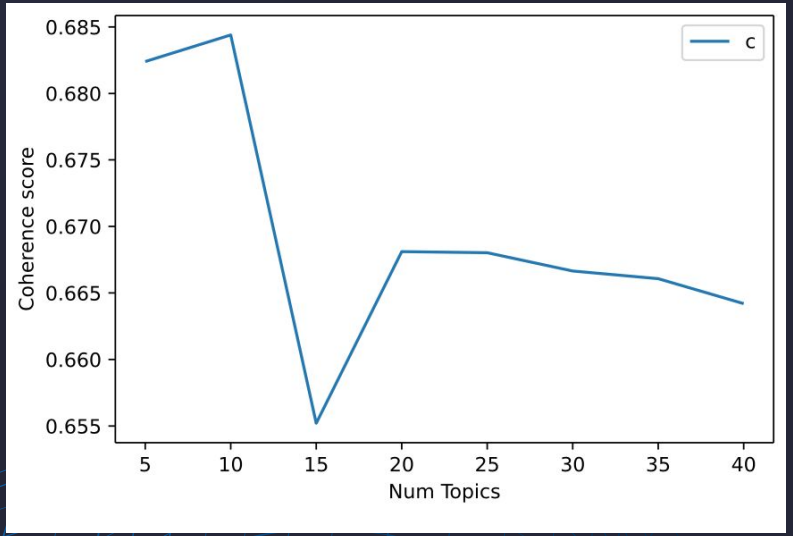
THANKS!

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





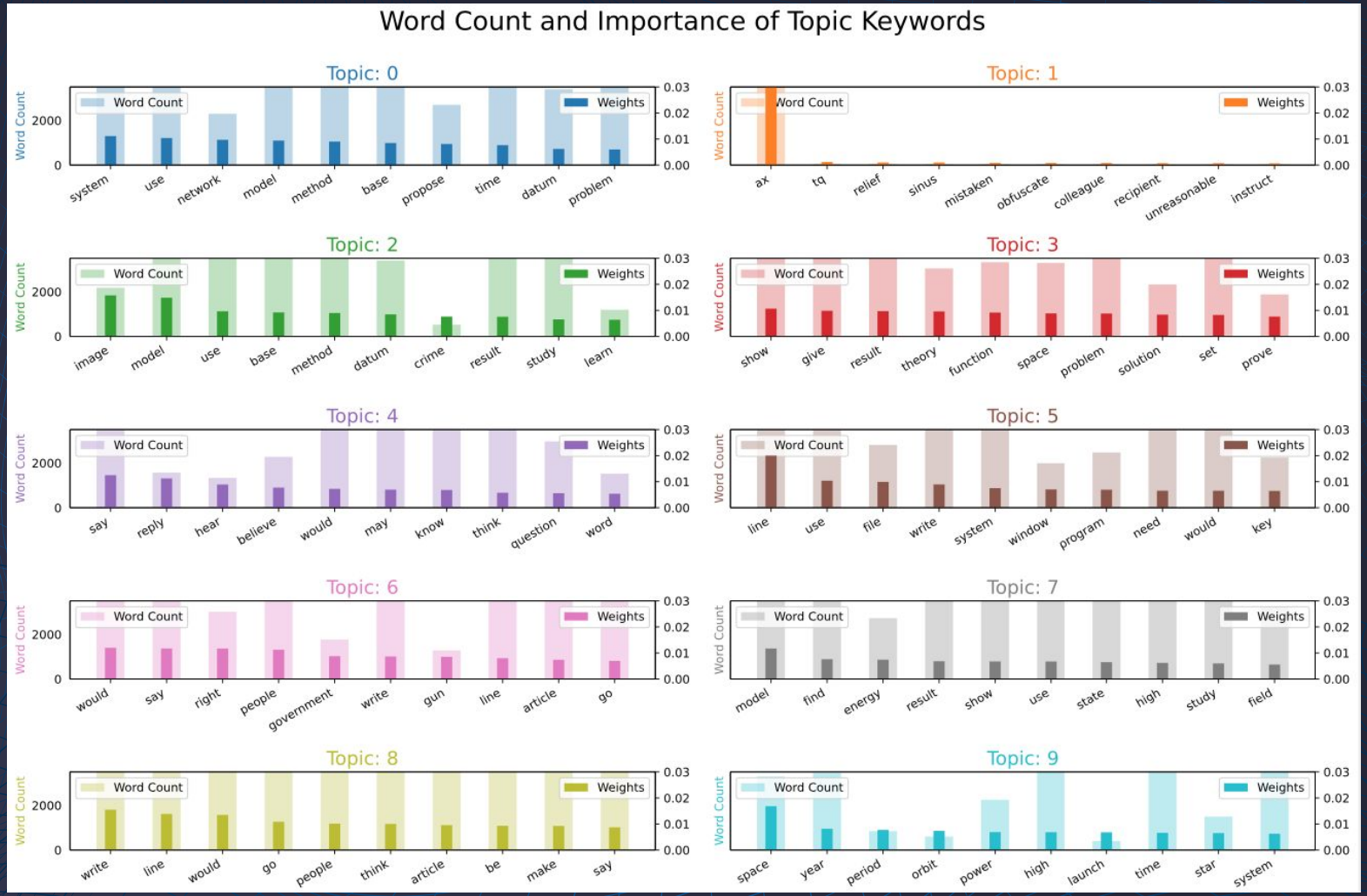
Perplexity:

Perplexity is how the model reacts to unseen data and is the 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
Topic: 0
Word Count
Weights
system
use
network
model
method
base
propose
time
datum
problem
Topic: 1
ax
tq
relief
sinus
mistaken
obfuscate
colleague
recipient
unreasonable
instruct
Topic: 2
image
model
use
base
method
datum
crime
result
study
learn
Topic: 3
show
give
result
theory
function
space
problem
solution
set
prove
Topic: 4
say
reply
hear
believe
would
may
know
think
question
word
Topic: 5
line
use
file
write
system
window
program
need
would
key
Topic: 6
would
say
right
people
government
write
gun
line
article
go
Topic: 7
model
find
energy
result
show
use
state
high
study
field
Topic: 8
write
line
would
go
people
think
article
be
make
say
Topic: 9
space
year
period
orbit
power
high
launch
time
star
system

PyLDAvis

. . . pretty nifty