LDA Topic Modeling: Techniques and AI Models

Benjamin Reaves

Director of AI @ Metabob
Expert in NLP, Topic Modeling, and Speech Recognition
USC & Stanford Alumnus
Google Scholar

Previous Experience:

ORACLE
Panasonic
NUANCE
ellipsis HEALTH
ATR
TOYOTA

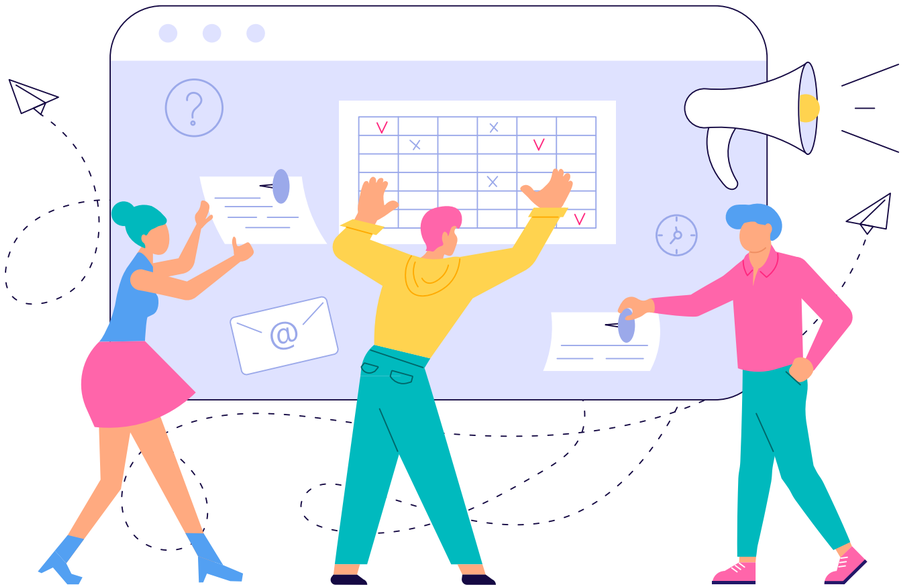
Chapter04
Session 2
Step Up. Speak Up. Go Up

June 6th 10AM PST

zoom.goupaz.com

Hosted by
Metabob

Agenda

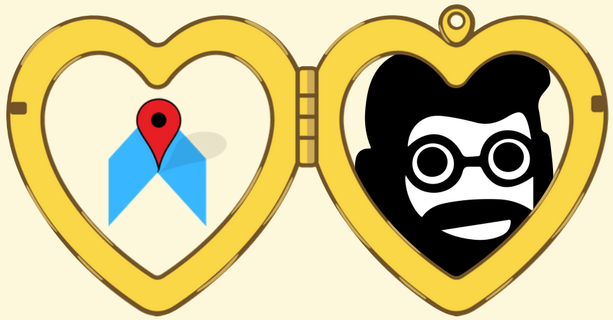


QUOTE OF THE SESSION

In open source, you have the right to control your destiny.

- **10:00–10:10 | Kickoff**
  - Introduction
  - Chit Chat
- **10:10-11:10 | Presentation**
  - Speaker
- **11:10-11:15 | Break**
  - Pre Q&A
- **11:15-11:30 | Q&A**
  - Discussion
- **11:30-11:35 | Wrap Up**
  - Contact Speaker

Chapter04
Session 02



Hosted by

Axel Lönnfors

Digital Marketing Specialist | Community Manager

Enthusiastic sports fan originally from Finland, passionate about technologies enhancing employee productivity

Metabob

It's the fast, easy, and visual way of debugging code.

https://metabob.com

GOUP

Community driven Open source accelerator!
https://goupaz.com

Are you ready?

Let's Begin!

POLL

How experienced are you with AI/ML models and concepts overall?

1 Not experienced

2 Somewhat experienced

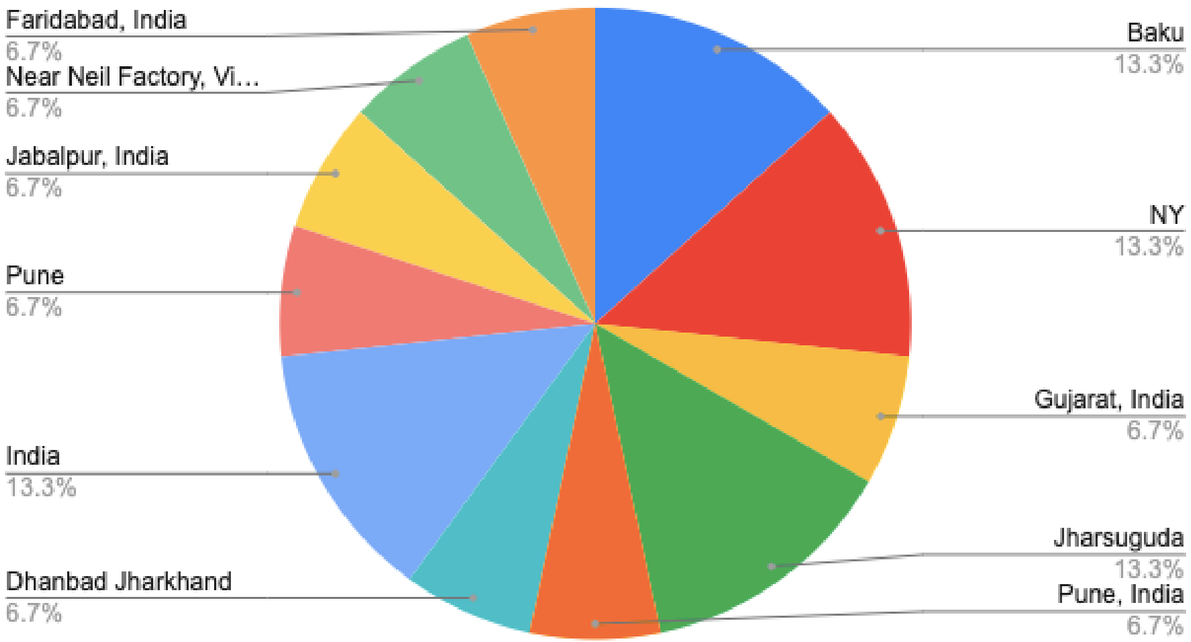
3 Semi-experienced

4 Experienced

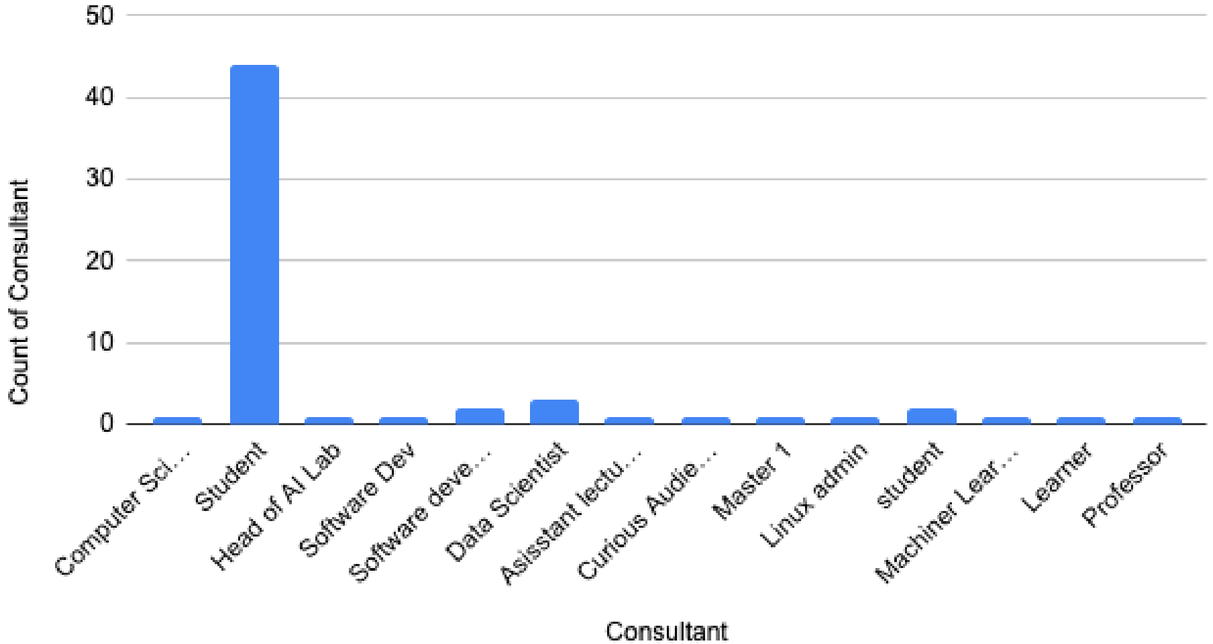
5 Very experienced

Audience

Locations



Roles



Code of Conduct

1. Learn, benefit, contribute
2. No marketing, selling, competing
3. Equality despite roles & bg

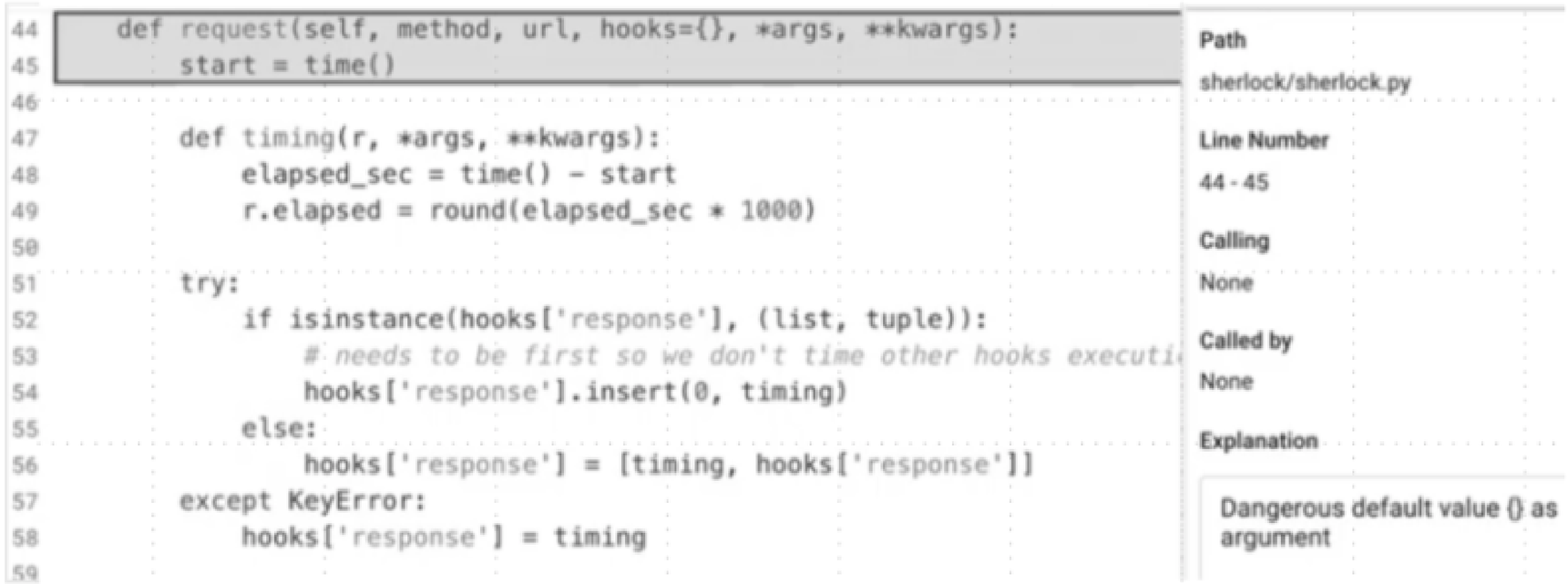
Presentation

Who is Ben?

- Stanford University – Statistical Signal Processing
- Speech Recognition using HMM models at Panasonic
- Speech Translation Japanese-English 1997 using Python
- Speech controlled navigation at Toyota ITC 2002 (used in 2005 Prius)
- Statistical Language Modeling at Oracle in 2002 and 2012
- Detection of clinical depression, 2017 by CNN using Keras
- Text processing and NLP using LDA at Metabob

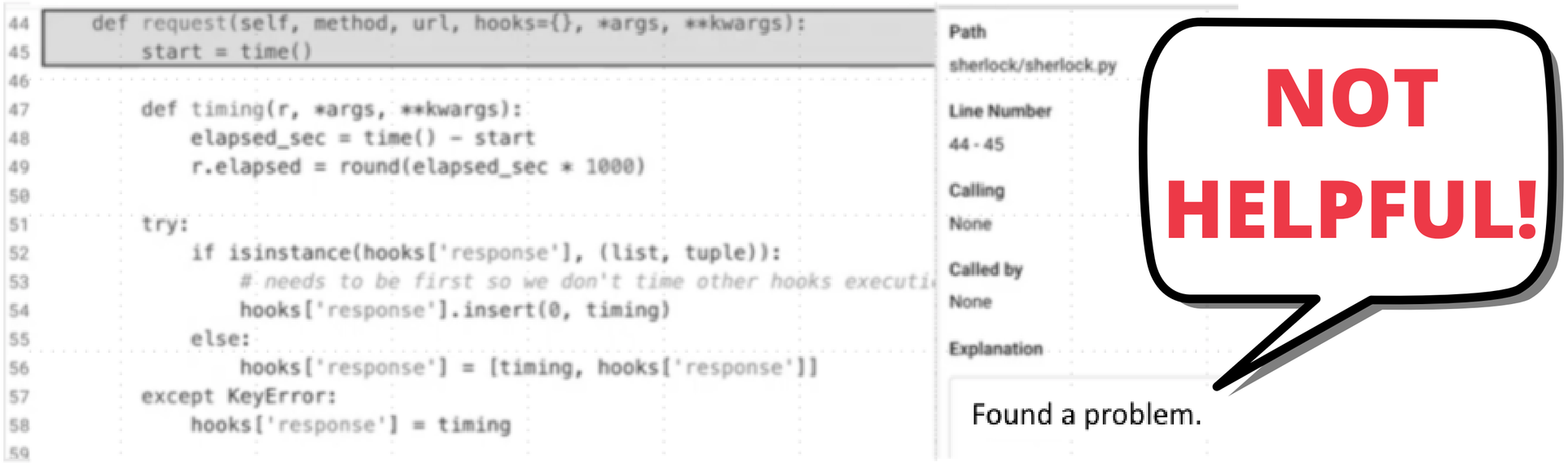
What Metabob Does?

- Show code snippet and generate an explanation

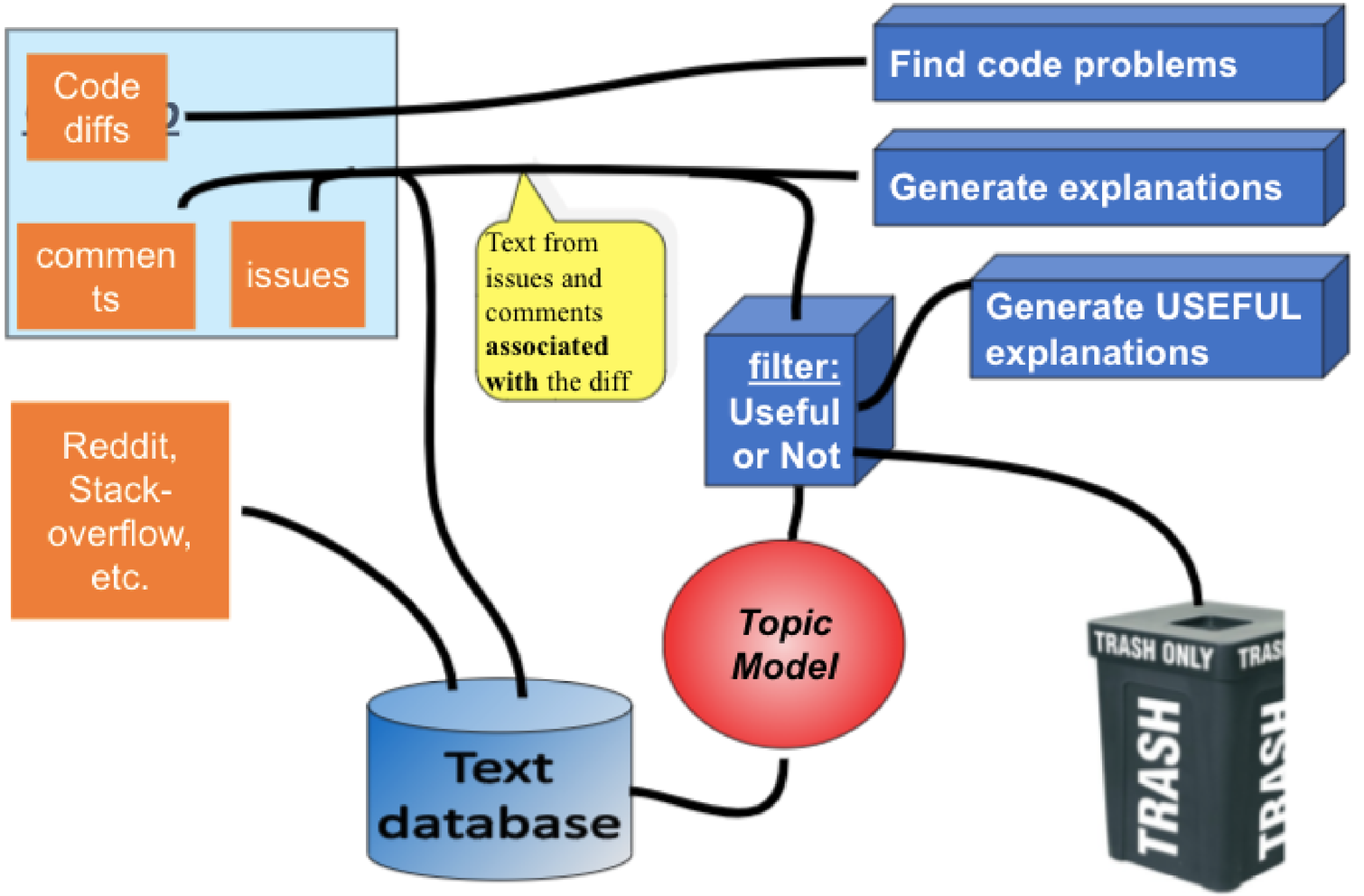


Problem: Garbage in, Garbage out.

- That model for generating an explanation must be trained on good text data! Garbage in, garbage out.
- Example of poorly trained explanation model:



Under the hood...



What's our data look like?

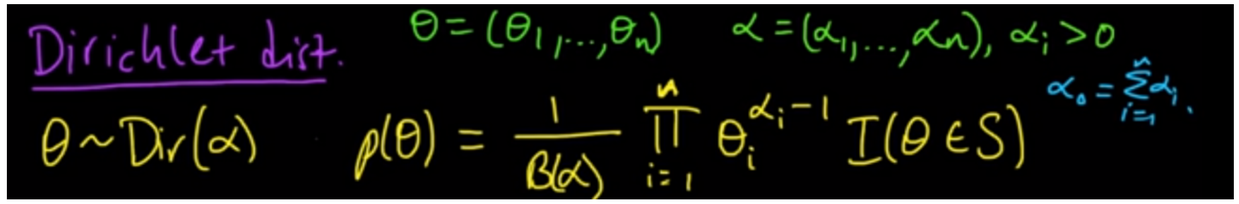
	content	label
2	fixed outdated npm modules causing security warnings in GitHub, added more learn resources	U
3	Add test case for security schema and definition for openapi 3 and 2	U
4	Update security vulnerabilities	U
5	Fix security vulnerabilities.	U
6	Good catch, thanks!	N
7	Cool. I can't test it on my machine, but I'm sure it works great.	N
8	Thanks for your contribution!	N
9	no comment.	N
10	Thanks :)	N
11	LGTM :)	N
12	The point of the regex is _only_ to match comments. This will now match all lines.	U
13	Fixes #458	U
14	å›¾ç‰‡æœ€å¥½è¶…é“¾æŽ¥	N

How to filter it? Topic Modeling

- Topic Modeling is used for Spam filtering, etc.
- Many Labelled → inputs → train a model → model.
- New input → run that model → [('U',0.8), ('N', 0.2)]
- Problem: we have many inputs but not yet labelled.

Cool. I can't test it on my machine, but I'm sure it works great.
Thanks for your contribution!
no comment.
Thanks :)
LGTM :)
The point of the regex is _only_ to match comments. This will now match all lines.
Fixes #458

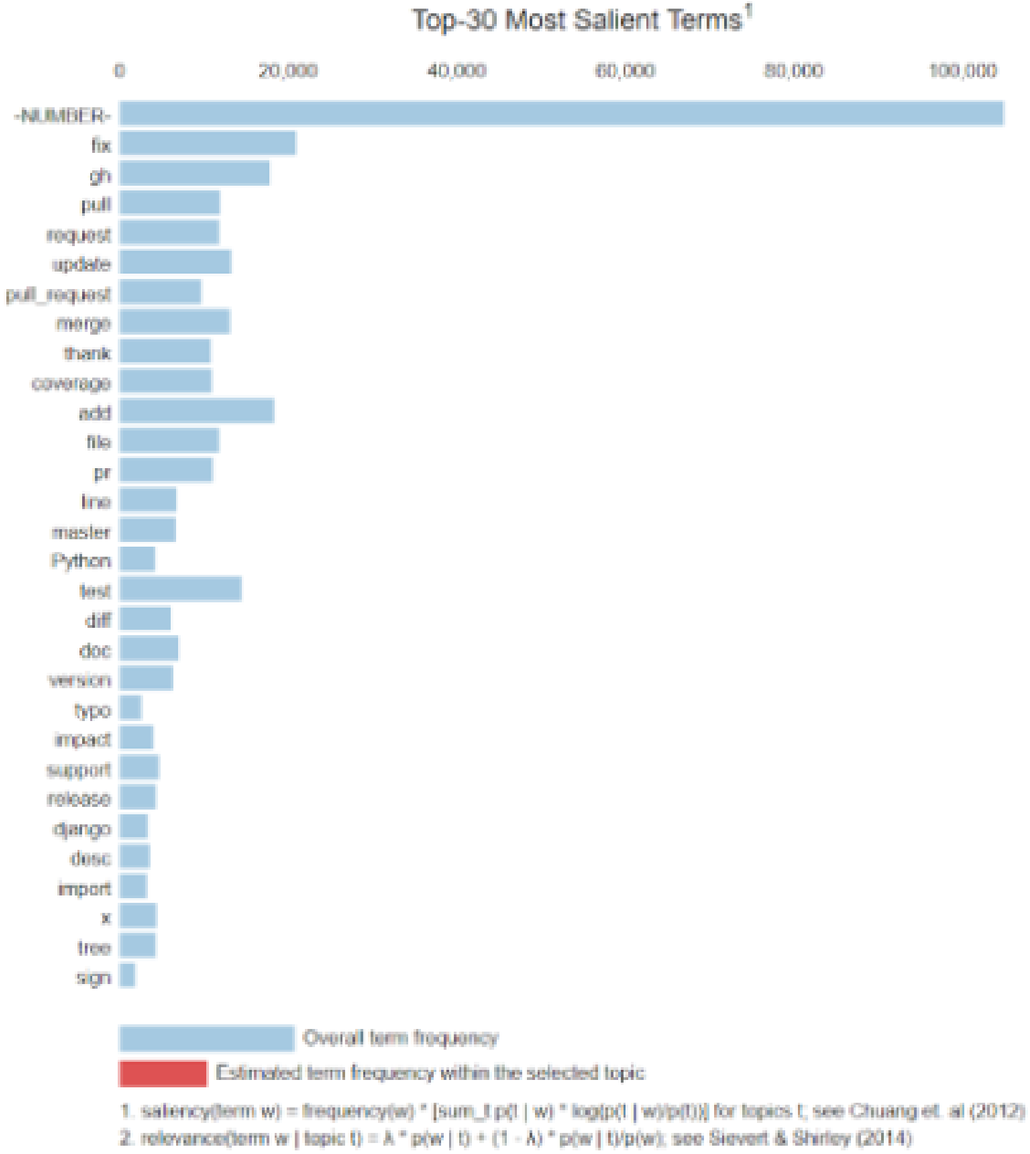
how does LDA work?



Why did we choose LDA?

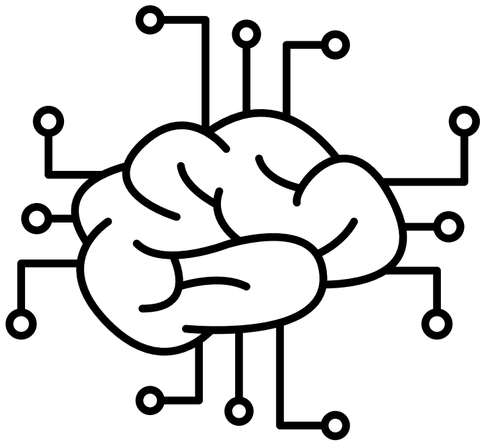
- It is unsupervised, but it can be semi-supervised (some labeled data)
- Python has a framework for LDA (genism and tomotopy) and they have active user communities, so I need not write from scratch :)
- It performs better than TF-IDF (1983) and LSA
  - TF-IDF maximizes the frequency of a word / (frequency of that word in all documents)
  - LSA (Latent Semantic Allocation) maximizes the frequency of a combination of words close to each other
  - LDA (Latent Dirichlet Allocation) does the same but the words need not be close to each other
  - Example: a document with “virus” and “covid” are probably about the pandemic,
    but a document with “virus” and “2fa” is probably about computer security.
  - TF-IDF would look at “virus” and “covid” separately.
  - LSA would be confused that “virus” has 2 meanings.
  - LDA just cares that “virus” and “covid” are in the same document, even if they are not right next to each other.

Typical LDA result: k=16



Why not a Neural Network?

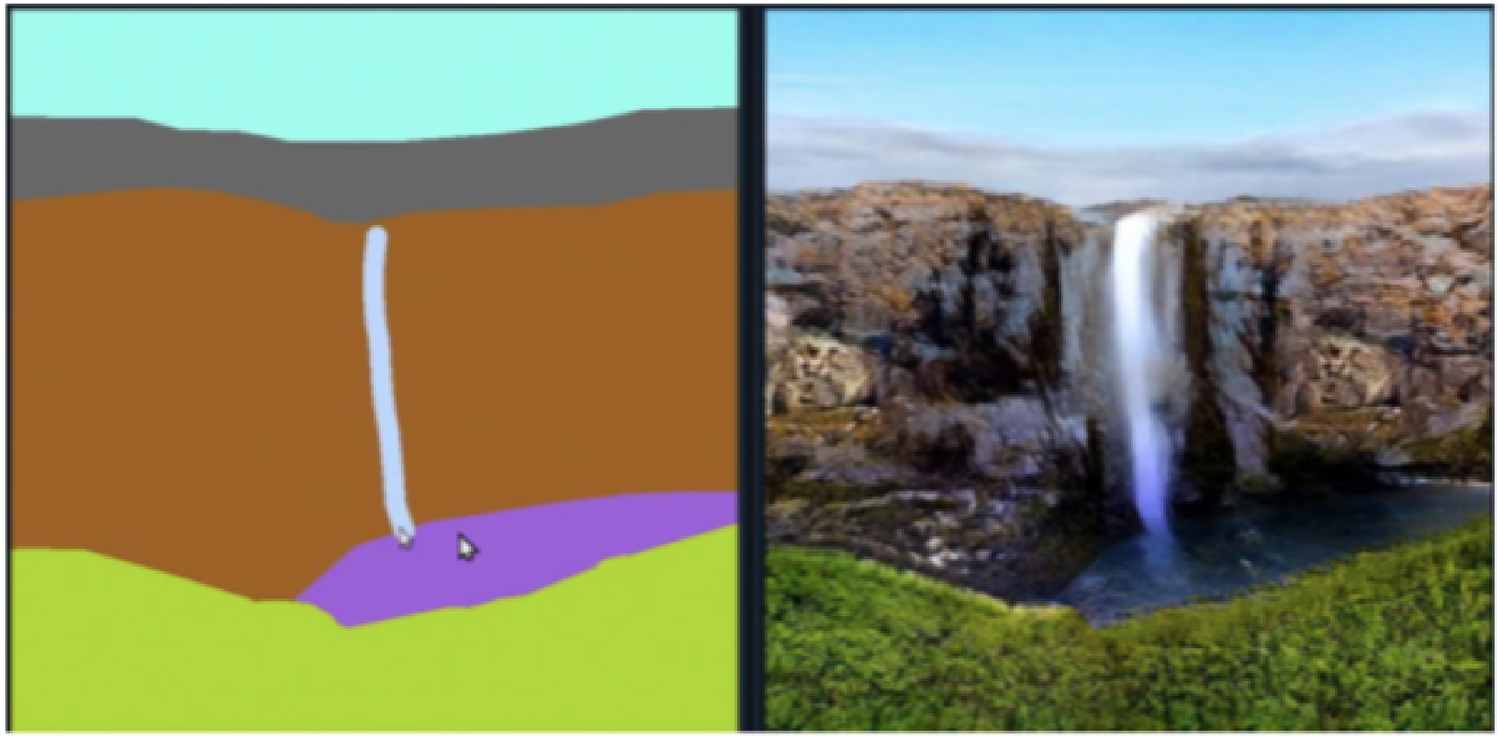
- NNs require a lot more data, labeled.
- NNs usually cannot accept any “prior knowledge” for example:
  I want topic#1 to contain the words oauth, login, virus, security;
  I want topic#2 to contain words version, release, update;
  I want topic#3 to contain words doc, spell, typo, readme
- It can be difficult to interpret why a NN made a wrong decision. Can’t debug it.
- But they usually have high accuracy, given enough data



What’s the result? How does this help?

- We’re generating language/text but I’ll give you an example from image generation. This is from Nvidia
- They train a NN to generate a beautiful picture from a rough picture.
- But the data used for generating it, must come from only beautiful pictures.
- Result is on the next page

Example: generating a picture

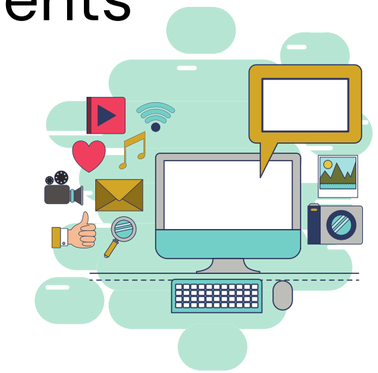


Their goal is to generate beautiful pictures;
ours is generate beautifully helpful suggestions

Lessons Learned using ML and AI

- I first used NN's in 1993 using Stuttgart University Neural Network
  → installation is painful, training is slow, running is faster.
- More robust models (DTW → HMM → NN) require more data
- Many startups underestimate time and cost to acquire and label data
- Does this explain how the brain works?
  - No. The mind is way more complex.

"Biology creates robust systems from unreliable components, but computers require reliable components"

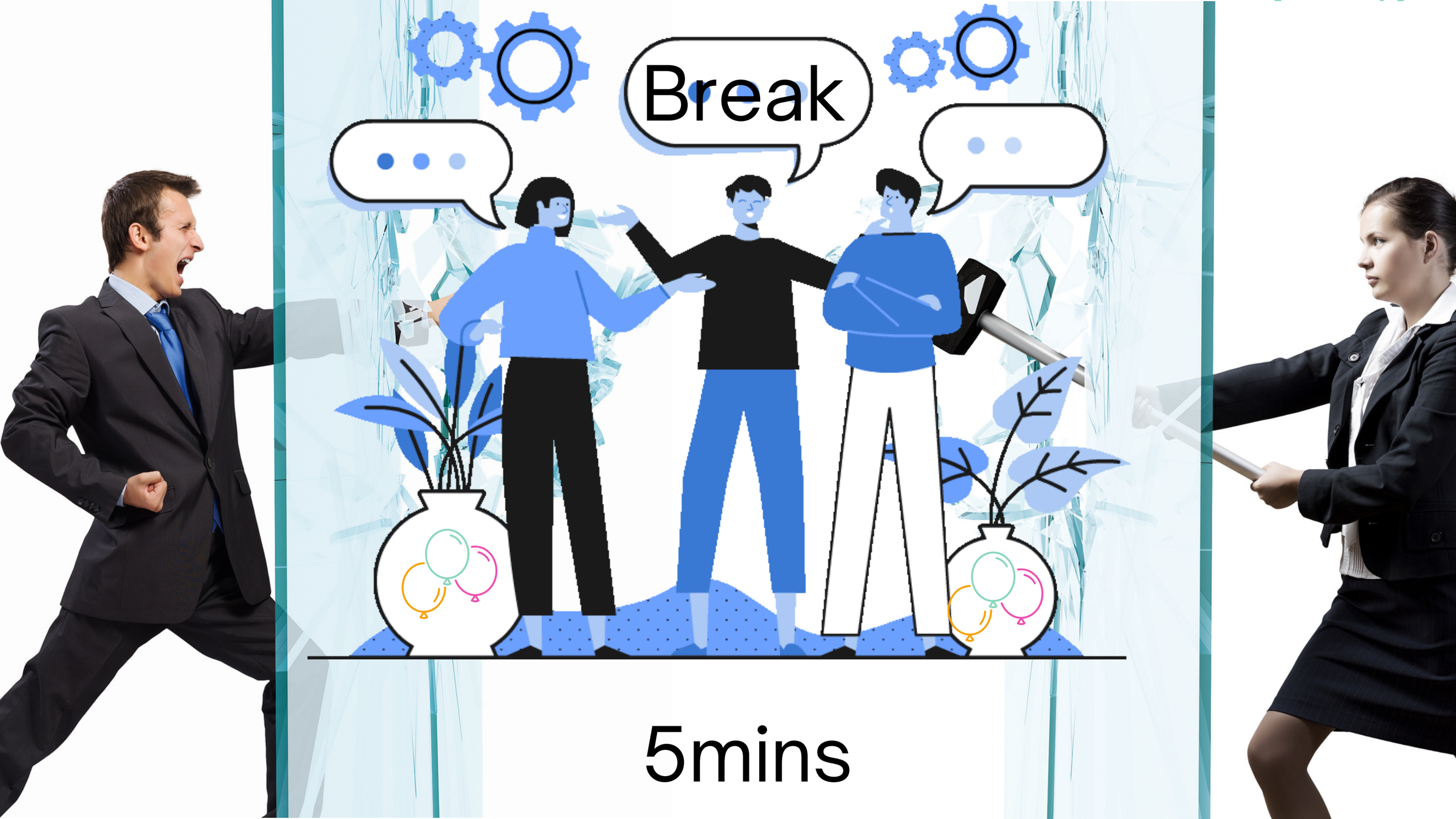


Q&A | Discussion

Type your thought here

Type your thought here

Type your thought here


Break
5mins

Wrap Up

Remember!

- Re-Search
- Be short and clear
- Re-mind
- Q&A over Slack

Linkedin: @benreaves
Email: benreaves@stanfordalumni.org
GOUP Slack: @Ben Reaves

Benjamin Reaves

Metabob

Join Us!

1. Community Managers
2. Tech Writers
3. In/Out Ambassadors
4. Marketing Creators & Editors
5. Course Creators
6. Project Creators

HTTPS://GOUPAZ.COM

HTTPS://METABOB.COM

Thank You

Culture
#egoless #collaborative #competent #decentralized #scalable #fun

Open source
#creator #contributor

How can we do better?

Diversity
#age #gender #location #economics #religion #politicalview