Carnegie Mellon University
Heinz College

# 94-775 Unstructured Data Analytics for Policy

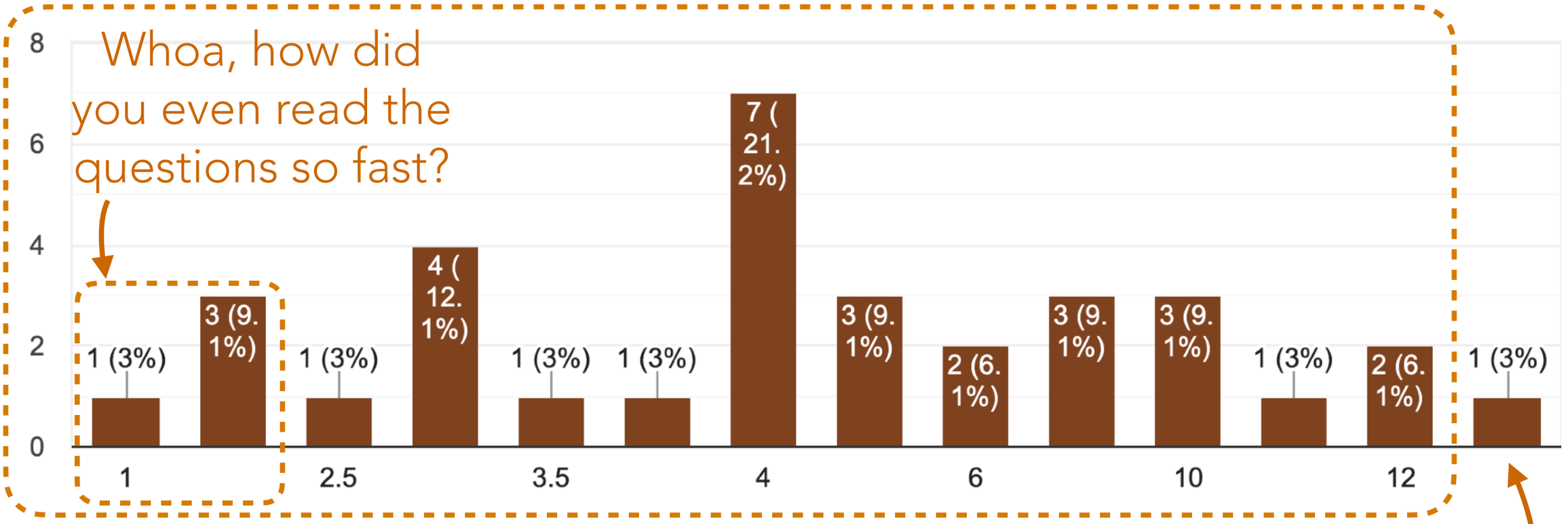
## Lecture 10: Wrap up topic modeling

Slides by George H. Chen

# HW1 Questionnaire (1/n)

How many hours did you take (roughly) to complete homework 1?

33 responses



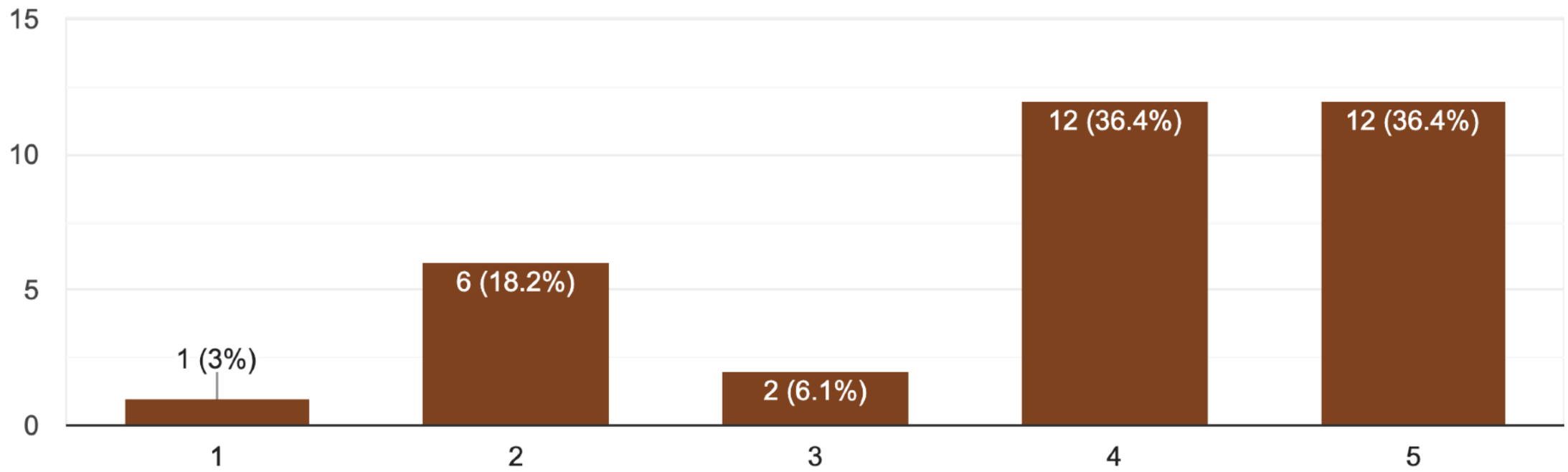

The vast majority of students are finishing within 15 hours
(which is our target of the max time that should be spent *without* AI help)

Spending over 15 hours on each homework is not normal!
Please do swing by OH to chat if you're still spending a lot of time on HW2

# HW1 Questionnaire (2/n)

Are you interested in learning how to use neural network software packages such as PyTorch, TensorFlow, or JAX?

33 responses

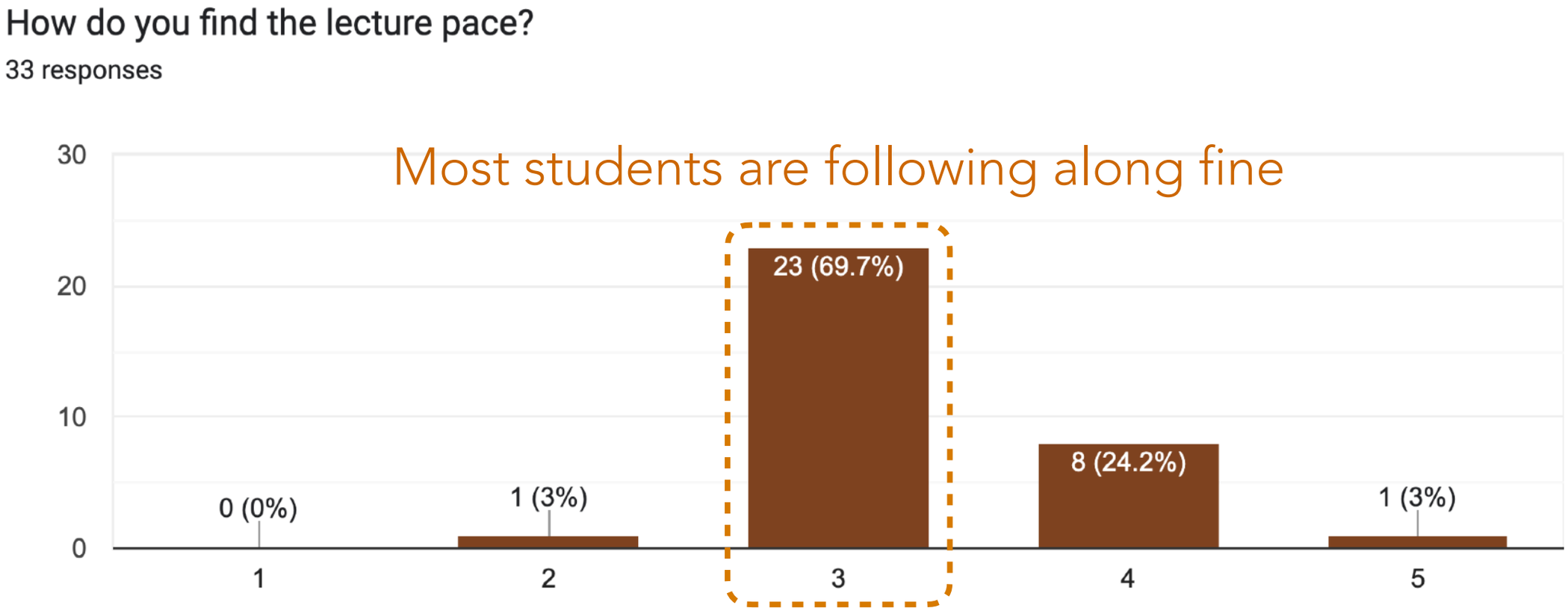



Well, it looks like students are somewhat divided here

If you really, really want to know how modern tools for unstructured data analytics works, it really helps to understand fundamentals of neural net software packages (so that they don't seem like mysterious black boxes)

There is no homework or quiz in this class that will require you to know neural net software, but neural net lectures *will* use PyTorch code

# HW1 Questionnaire (3/n)




Very importantly, I think it's okay to feel that the material takes time to digest & that it doesn't somehow just immediately all makes sense

It's like how when you watch some movies, the first time you see it, it doesn't fully make sense yet but watching a second time helps to clarify

(This is why we have lecture recordings!)

Overall, I'll try to keep the pacing about the same or slightly slower

# HW1 Questionnaire (4/n)

Free response parts:

Hard to make everyone happy!

- Some students wrote that they want more practical examples
- Some students wrote that they want more code examples
- Some students asked for help with refreshing Python, data manipulation, statistics, or math

Hopefully the clustering on text demos & the clustering on images demo showed you how things play out when you're working with real unstructured data (the HW1 questionnaire was due before these demos appeared)

We're already averaging about 1 demo per lecture (hard to squeeze more in)

HW2 & your final projects (+ the example past final projects) also provide good opportunities for seeing more practical examples!

Please make use of office hours to learn more (such as if you want help with Python, data manipulation, stats, or math)! A key reason for going to a university rather than just learning on your own by watching, say, YouTube or talking to an LLM is that you have access to **human profs & TAs with lived experiences**!

# HW1 Questionnaire (5/n)

Suggestion: please actually play with demos yourself after class

- Understand every line of code & whether it can be coded differently
- A major theme in the course is that of *infinite design choices*: identify where design choices appear & change them to see what happens
- Try deleting all code cells except part that loads in data & see if you can redo the exploratory data analysis on your own

We explicitly ask you to try different design choices for the last part of HW2's problem 1

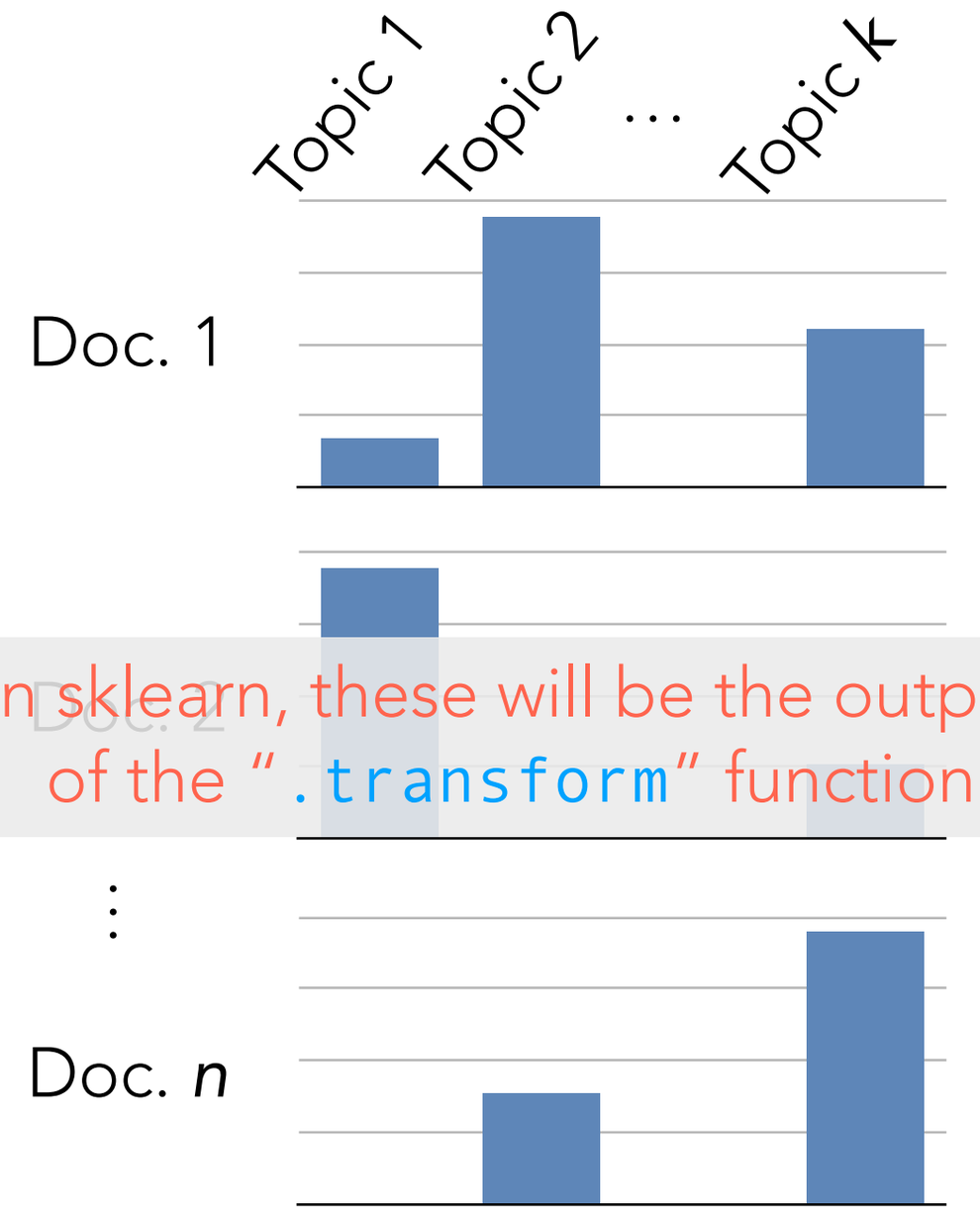
There was some feedback asking for us to provide more comments/ explanation regarding what the code is doing in code demos

I actually think you'll learn more if *you write comments/notes for yourself* ***in your own words*** *in your own copy of lecture demos!*

In many ways, learning the material in this course is like learning how to swim — you can watch other people swim all you want but you're only going to get better if you practice a lot of swimming yourself!

Final project proposals: we're aiming to try to get you feedback by no later than the end of Wednesday

# (Flashback) LDA Generative Model




In sklearn, these will be the output of the "`.transform`" function


Goal: Learn these distributions

In sklearn, these will be in "`.components_`"

LDA models each word in document *i* to be generated as:

1. Randomly choose a topic *Z* (use topic distribution for doc *i*)
2. Randomly choose a word (use word distribution for topic *Z*)

# LDA

Demo

For topic modeling, how do we choose the number of topics *k*?

# How "Coherent" is a Topic?

Let's look at top-20 word lists (the ones from the demo)

```
[Topic 0]
space like just stuff good low probably power use air ground make sure high orbit know deleted time need nasa
```

If we see the word "space", how likely are we to see the word "power"?

P(see word "power" | see word "space")

If this probability is high for every pair of words in the top-20 list, then in some sense the topic is more "coherent"

Coherence of topic:

$$\sum_{\substack{\text{top words } v,w \text{ that} \\ \text{are not the same}}} \log \mathbb{P}(\text{see word } v | \text{see word } w)$$

$$\frac{\#\text{ documents that mention both } v \text{ and } w}{\#\text{ documents that mention } w} \quad \frac{}{+0.1}$$

avoid numerical issues

Focus on a single topic at a time

# How Different is a Topic from the Others?

Let's look at top-20 word lists (the ones from the demo)

If "space" only shows up in the top-20 word list for topic 0, then it is considered a unique top word for topic 0

```
[Topic 0]
space like just stuff good low probably power use air ground make sure high orbit know deleted time need nasa

[Topic 1]
windows thanks use card drive file does know problem using program software help files like pc dos hi need advance

[Topic 2]
edu 1993 10 university 11 93 15 chicago 12 00 20 april 23 apr 13 david cs pittsburgh 18 17

[Topic 3]
people government don just israel think gun state law like right know make time case israeli war said rights use

[Topic 4]
game team players hockey games play armenian year season armenians nhl player teams turkish league turkey win did played la

[Topic 5]
com like just heard know don sorry hear post guy ve does list guess actually think oh news article dave

[Topic 6]
sale price offer mail condition new shipping 00 email address interested sell looking asking thanks used send excellent phone o

[Topic 7]
god people think don jesus believe say just like know good does time bible did christian life church said christians

[Topic 8]
bike car just don cars good know think like dod ve engine got edu road time soon try posting way

[Topic 9]
key encryption clipper chip keys algorithm nsa escrow public security des secure use chips window phone bit secret ax informatio
```

Each topic has a # of unique top words

# How to Choose Number of Topics *k*?

Compute:

Topic 0's coherence score

Topic 1's coherence score

⋮

Topic (k - 1)'s coherence score

Compute average coherence

Topic 0's # unique top words

Topic 1's # unique top words

⋮

Topic (k - 1)'s # unique top words

Compute average # unique top words

Can plot average coherence vs *k*, and average # unique top words vs *k* (for values of *k* you are willing to try)

# How to Choose Number of Topics *k*?

Demo

# Topic Modeling: Last Remarks

- There are score functions aside from coherence & # unique top words (e.g., normalized-mutual-information coherence, (log) lift score, …)
- There are *many* topic models, not just LDA
  (e.g., Hierarchical Dirichlet Process, correlated topic models, SAGE, anchor word topic models, Scholar, embedded topic model, …)
- *Dynamic* topic models can track how topics change *over time*
  - Requires timestamp for every text document we fit the model to
- **Warning:** learning topic models is very sensitive to random initialization
  - Can try fitting data multiple times using different random seeds & seeing which topics consistently show up across random seeds
- On the course webpage, I posted a link to Maria Antoniak's practical guide for using LDA (very helpful if you want to use LDA in the future)

# Course Outline

Part I: Exploratory data analysis

*Identify structure present in "unstructured" data*

- Frequency and co-occurrence analysis
- Visualizing high-dimensional data/dimensionality reduction
- Clustering
- Topic modeling

Part II: Predictive data analysis

*Make predictions using known structure in data*

- Basic concepts and how to assess quality of prediction models
- Neural nets and deep learning for analyzing images and text