Carnegie Mellon University
Heinz College

Unstructured Data Analytics for Policy

Lecture 2: Basic text analysis demo, co-occurrence analysis

George Chen

(Flashback)

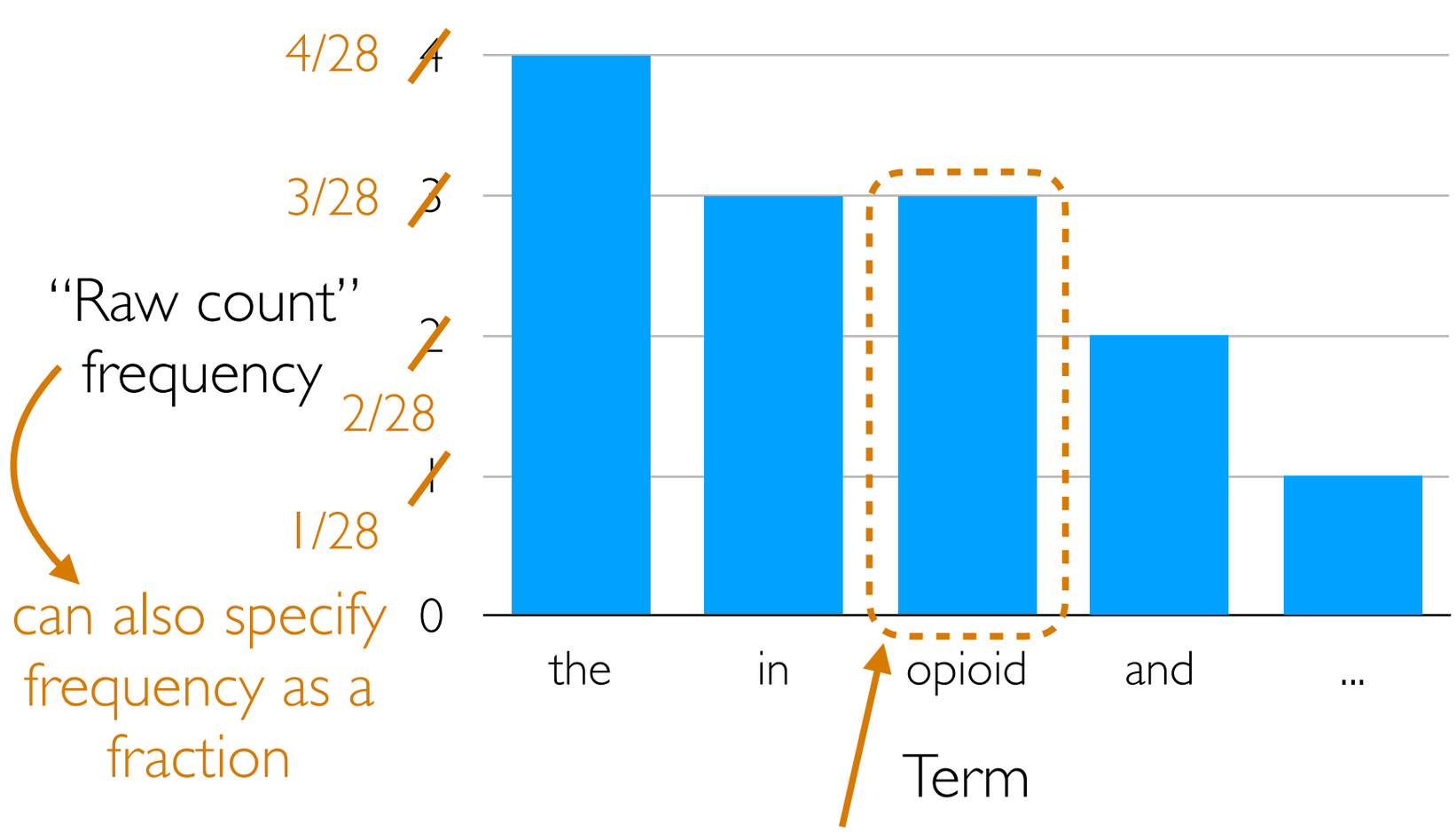
increase the drugs opioid in The States or prescription opioid and of is rapid in opioid crisis the use non-prescription Canada 2010s. in United and the epidemic the

Total number of words in sentence: 28

Term frequencies

The: 1 /28
opioid: 3 /28
epidemic: 1 /28
or: 1 /28
crisis: 1 /28
is: 1 /28
the: 4 /28
rapid: 1 /28
increase: 1 /28
in: 3 /28
use: 1 /28
of: 1 /28
prescription: 1 /28
and: 2 /28
non-prescription: 1 /28
drugs: 1 /28
United: 1 /28
States: 1 /28
Canada: 1 /28
2010s.: 1 /28

Histogram



"Raw count" frequency

can also specify frequency as a fraction

Fraction of words in the sentence that are "opioid"

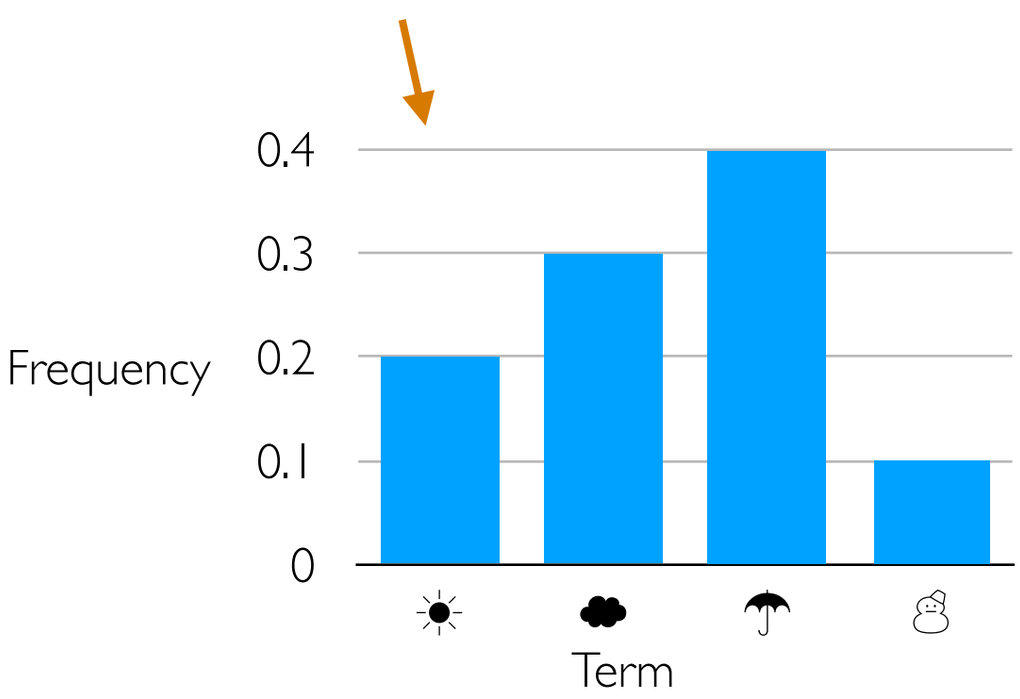
The spaCy Python Package

Demo

Recap: Basic Text Analysis

- Represent text in terms of "features"
  (each feature: how often a specific word/phrase appears)
  - Can repeat this for different documents:
    represent each document as a "feature vector"

"Sentence": ☀☂☁☁☁☂☃☂☂☀



$$\begin{bmatrix} 0.2 \\ 0.3 \\ 0.4 \\ 0.1 \end{bmatrix}$$

This is a point in 4-dimensional space, $\mathbb{R}^4$

dimensions = number of terms

In general (not just text): first represent data as feature vectors

Bigram Model

The opioid epidemic or opioid crisis is the rapid increase in the use of prescription and non-prescription opioid drugs in the United States and Canada in the 2010s.

The opioid	opioid epidemic	epidemic or
or opioid	opioid crisis	crisis is

Ordering of words now matters (a little) ... # unique cards changes dramatically

If using stopwords, remove any phrase with at least 1 stopword

---

1 word at a time: **unigram** model

2 words at a time: **bigram** model

3 words at a time: **trigram** model

n words at a time: ***n*-gram** model

Finding Possibly Related Entities with Co-occurrence Analysis

Elon Musk's Tesla Powerwalls Have Landed in Puerto Rico

How to automatically figure out Elon Musk and Tesla are related?

The solar batteries have reportedly been spotted in San Juan's airport.

By **John Patrick Pullen** October 16, 2017

Exactly one week after Tesla CEO Elon Musk suggested his company could help with Puerto Rico's electricity crisis in the aftermath of Hurricane Maria, more of the company's Powerwall battery packs have arrived on the island, according to a photo snapped at San Juan airport Friday, Oct. 13.

Source: http://fortune.com/2017/10/16/elon-musks-tesla-powerwalls-have-landed-in-puerto-rico/

Co-Occurrences

For example: count # news articles that have different named entities co-occur

	Alphabet	AMD	Tesla
Elon Musk	1500	1000	20000
Sundar Pichai	1000	50	50
Lisa Su	30	700	10

Big values ➔ *possibly* related named entities

Different Ways to Count

- Previous slide: count # doc's in which two named entities co-occur
  - This approach ignores # co-occurrences *within a specific document* (e.g., a single doc might mention Elon Musk & Tesla 20 times)
  - Could instead add # co-occurrences, not just whether a co-occurrence appears at least once in a doc
- Could change the unit of analysis: a "document" could instead be a *sentence*, a *paragraph*, etc

Bottom Line

- There are many ways to count co-occurrences
- You should think about what makes the most sense/is reasonable for the problem you're looking at

Co-Occurrences

For example: count # news articles that have different named entities co-occur

	Alphabet	AMD	Tesla
Elon Musk	1500	1000	20000
Sundar Pichai	1000	50	50
Lisa Su	30	700	10

Big values ➔ *possibly* related named entities

How to downweight "Elon Musk" if there are just way more articles that mention him?

Goal

Elon Musk, Alphabet

Elon Musk, AMD

Elon Musk, Tesla

Sundar Pichai, Alphabet

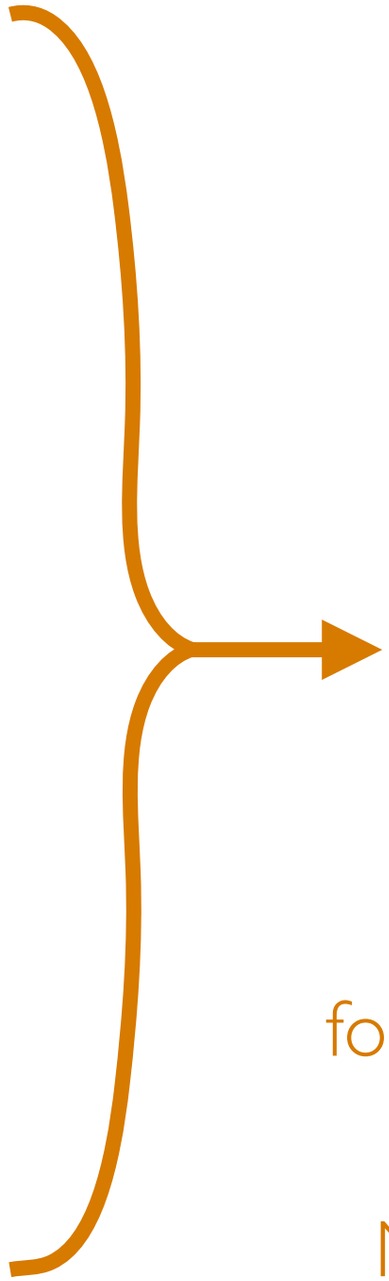
Sundar Pichai, AMD

Sundar Pichai, Tesla

Lisa Su, Alphabet

Lisa Su, AMD

Lisa Su, Tesla



rank these pairs from
"most interesting" to
"least interesting"

For analysis: might want to
focus on most interesting pairs

Need a numerical score for
"interesting"-ness

Key idea: what would happen if people and companies were independent?

	Alphabet	AMD	Tesla
Elon Musk	1500	1000	20000
Sundar Pichai	1000	50	50
Lisa Su	30	700	10

Probability of drawing "Elon Musk, Alphabet"?

Probability of drawing a card that says "Alphabet" on it?

1500 of these cards: Elon Musk, Alphabet

1000 of these cards: Elon Musk, AMD

20000 of these cards: Elon Musk, Tesla

⋮

10 of these cards: Lisa Su, Tesla

Place into bag

Co-occurrence table

	Alphabet	AMD	Tesla
Elon Musk	1500	1000	20000
Sundar Pichai	1000	50	50
Lisa Su	30	700	10

Total: 24340

Joint probability table

	Alphabet	AMD	Tesla
Elon Musk	1500 /24340	1000 /24340	20000 /24340
Sundar Pichai	1000 /24340	50 /24340	50 /24340
Lisa Su	30 /24340	700 /24340	10 /24340

sum to get P(Elon Musk)

Total: 24340

Joint probability table

	Alphabet	AMD	Tesla	
Elon Musk	0.06163	0.04108	0.82169	**0.92440**
Sundar Pichai	0.04108	0.00205	0.00205	**0.04519**
Lisa Su	0.00123	0.02876	0.00041	**0.03040**
	0.10394	**0.07190**	**0.82416**	

Recall: if events A and B are independent, **P(A, B) = P(A)P(B)**

Joint probability table **if people and companies were independent**

	Alphabet	AMD	Tesla	
Elon Musk	0.09609	0.06646	0.76185	**0.92440**
Sundar Pichai	0.0047	0.00325	0.03725	**0.04519**
Lisa Su	0.00316	0.00219	0.02506	**0.03040**
	0.10394	**0.07190**	**0.82416**	

Recall: if events A and B are independent, **P(A, B) = P(A)P(B)**

What we actually observe

	Alphabet	AMD	Tesla
Elon Musk	0.06163	0.04108	0.82169
Sundar Pichai	0.04108	0.00205	0.00205
Lisa Su	0.00123	0.02876	0.00041

What should be the case if people are companies were independent

	Alphabet	AMD	Tesla
Elon Musk	0.09609	0.06646	0.76185
Sundar Pichai	0.0047	0.00325	0.03725
Lisa Su	0.00316	0.00219	0.02506

Pointwise Mutual Information (PMI)

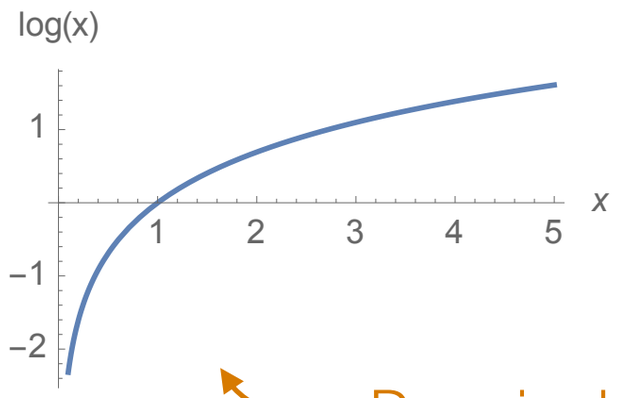
Probability of A and B co-occurring

$$\frac{P(A, B)}{P(A)\ P(B)}$$

if equal to 1
➔ A, B are indep.

Probability of A and B co-occurring *if they were independent*

PMI(A, B) is defined as the log of the above ratio



PMI measures (the log of) a ratio that says how far A and B are from being independent

Reminder: this is what log looks like!

Use PMI as a Numerical Score to Rank *Specific* Person/Company Pairs

$$\text{PMI}(A, B) = \log \frac{P(A, B)}{P(A)\ P(B)}$$

- More positive value means a specific pair appears much more likely than if they were independent
- More negative value means a specific pair appears much less likely than if they were independent
- In practice: need to be careful with named entities that extremely rarely occur
- Sometimes people consider only pairs with positive PMI values to be interesting (called *positive PMI or PPMI*)