Carnegie Mellon University
Heinz College

94-775: Unstructured Data Analytics for Policy

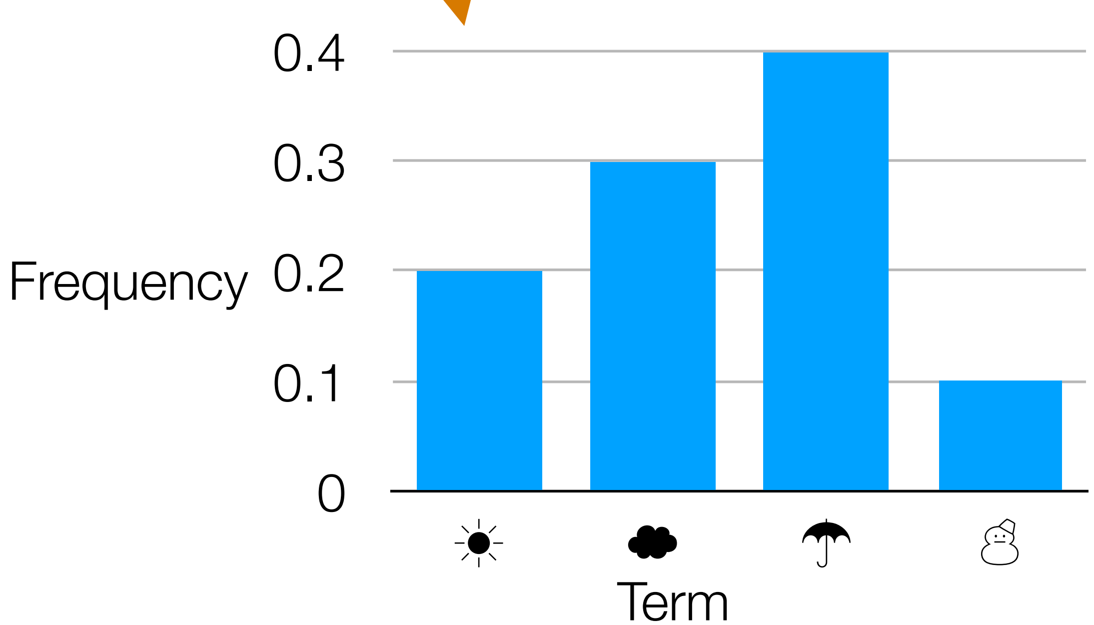
Lecture 3: Finding possibly related entities

George Chen

Recap: Basic Text Analysis

- Represent text in terms of “features”
  (such as how often each 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

So far: look at how frequently individual "flashcards"/terms appear

What about how two different terms *co-occur*?

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

	Apple	Facebook	Tesla
Elon Musk	10	15	300
Mark Zuckerberg	500	10000	500
Tim Cook	200	30	10

Big values ➔ *possibly* related named entities

Different Ways to Count

- Just saw: for all doc's, count # of doc's in which two named entities co-occur
    - This approach ignores # of co-occurrences *within a specific document* (e.g., if 1 doc has "Elon Musk" and "Tesla" appear 10 times, we count this as 1)
    - Could instead add # co-occurrences, not just whether it happened in a doc
- Instead of looking at # doc's, look at co-occurrences within a *sentence*, or 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

	Apple	Facebook	Tesla
Elon Musk	10	15	300
Mark Zuckerberg	500	10000	500
Tim Cook	200	30	10

Big values ➔ *possibly* related named entities

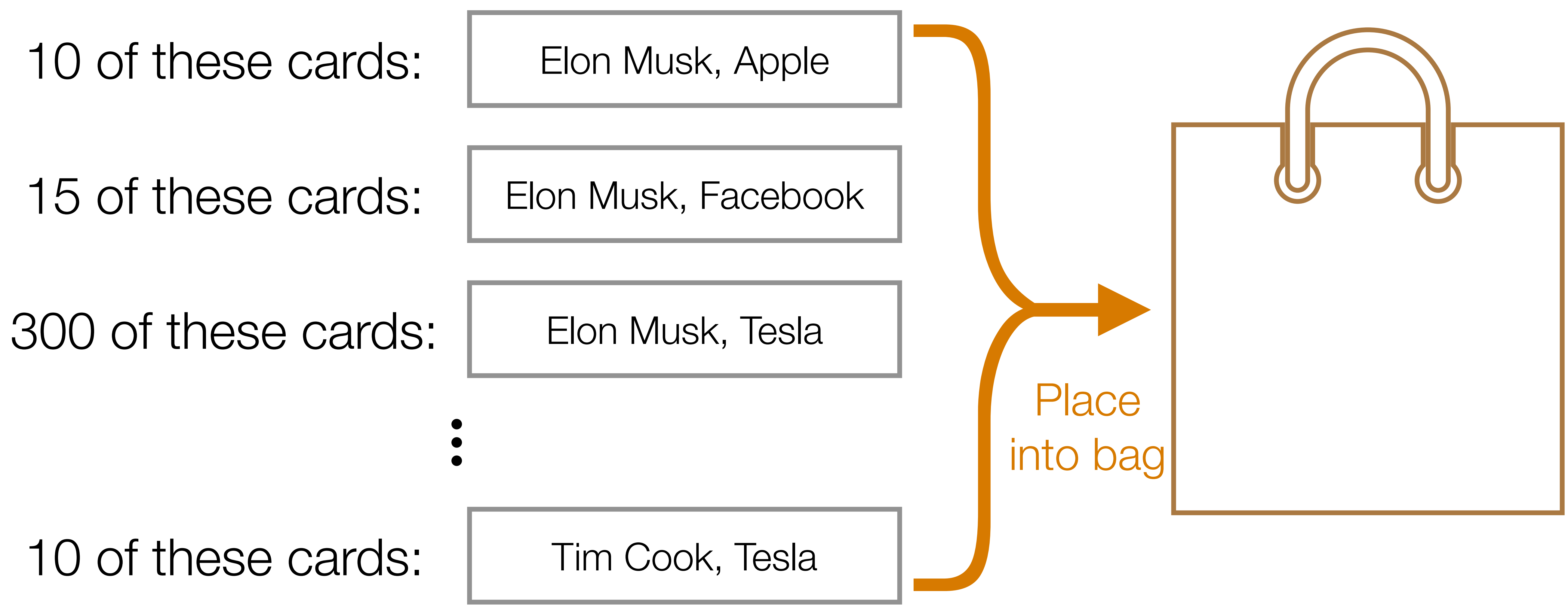
How to downweight "Mark Zuckerberg" if there are just way more articles that mention him?

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

	Apple	Facebook	Tesla
Elon Musk	10	15	300
Mark Zuckerberg	500	10000	500
Tim Cook	200	30	10

Probability of drawing “Elon Musk, Apple”?

Probability of drawing a card that says “Apple” on it?



Co-occurrence table

	Apple	Facebook	Tesla
Elon Musk	10	15	300
Mark Zuckerberg	500	10000	500
Tim Cook	200	30	10

Total: 11565

Joint probability table

	Apple	Facebook	Tesla
Elon Musk	10 /11565	15 /11565	300 /11565
Mark Zuckerberg	500 /11565	10000 /11565	500 /11565
Tim Cook	200 /11565	30 /11565	10 /11565

sum to get P(Elon Musk)

Total: 11565

Joint probability table

	Apple	Facebook	Tesla	
Elon Musk	0.00086	0.00130	0.02594	**0.02810**
Mark Zuckerberg	0.04323	0.86468	0.04323	**0.95115**
Tim Cook	0.01729	0.00259	0.00086	**0.02075**
	0.06139	**0.86857**	**0.07004**	

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

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

	Apple	Facebook	Tesla	
Elon Musk	0.00173	0.02441	0.00197	**0.02810**
Mark Zuckerberg	0.05839	0.82614	0.06662	**0.95115**
Tim Cook	0.00127	0.01802	0.00145	**0.02075**
	0.06139	**0.86857**	**0.07004**	

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

What we actually observe

	Apple	Facebook	Tesla
Elon Musk	0.00086	0.00130	0.02594
Mark Zuckerberg	0.04323	0.86468	0.04323
Tim Cook	0.01729	0.00259	0.00086

What should be the case if people are companies were independent

	Apple	Facebook	Tesla
Elon Musk	0.00173	0.02441	0.00197
Mark Zuckerberg	0.05839	0.82614	0.06662
Tim Cook	0.00127	0.01802	0.00145

Pointwise Mutual Information (PMI)

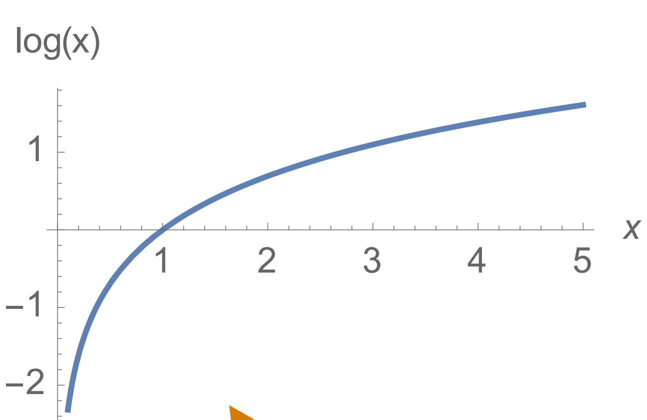
Probability of A and B co-occurring

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

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

if equal to 1
➔ A, B are indep.

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!

Recap: Use PMI 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*)

What about figuring out if people (as a whole)/ companies (as a whole) is an "interesting" relationship?

For example, perhaps we want to understand how different entity types are related (e.g., people/companies, people/locations, people/dates, companies/locations, etc)

There can be many such pairings, and we may only want to focus on a few

Looking at All Pairs of Outcomes

- PMI measures how P(A, B) differs from P(A)P(B) using a **log ratio**
- **Log ratio** isn't the only way to compare!
- Another way to compare:

$$\frac{[\ P(A, B) - P(A)\ P(B)\ ]^2}{P(A)\ P(B)}$$

Phi-squared is between 0 and min(#rows, #cols)-1

0 ➔ pairs are all indep.

$$\text{Phi-squared} = \sum_{A,\,B} \frac{[\ P(A, B) - P(A)\ P(B)\ ]^2}{P(A)\ P(B)}$$

$$\text{Chi-squared} = N \times \text{Phi-square}$$

Measures how close *all* pairs of outcomes are close to being indep.

N = sum of all co-occurrence counts

There's also a variant of these that is always between 0 and 1:

Cramér's V = Sqrt(Phi-squared / [min(#rows, #cols)-1])

Phi-Squared/Chi-Squared/Cramér’s V Calculation

Demo

Co-occurrence Analysis Applications

- If you're an online store/retailer:
  anticipate *when* certain products are likely to be purchased/rented/consumed more
  - Products & dates
- If you have a bunch of physical stores:
  anticipate *where* certain products are likely to be purchased/rented/consumed more
  - Products & locations
- If you're the police department:
  create "heat map" of where different criminal activity occurs
  - Crime reports & locations

Co-occurrence Analysis Applications

- If you're an online store/retailer:
  anticipate *when* certain products are likely to be purchased/
  re
  -
- If y
  an   sed/
  re
  -
- If y
  cr   curs
  - Crime reports & locations

Examples of data to take advantage of:
- data collected by your organization
- social networks
- news websites
- blogs

Web scraping frameworks can be helpful:
- Scrapy
- Selenium (great with JavaScript-heavy pages)