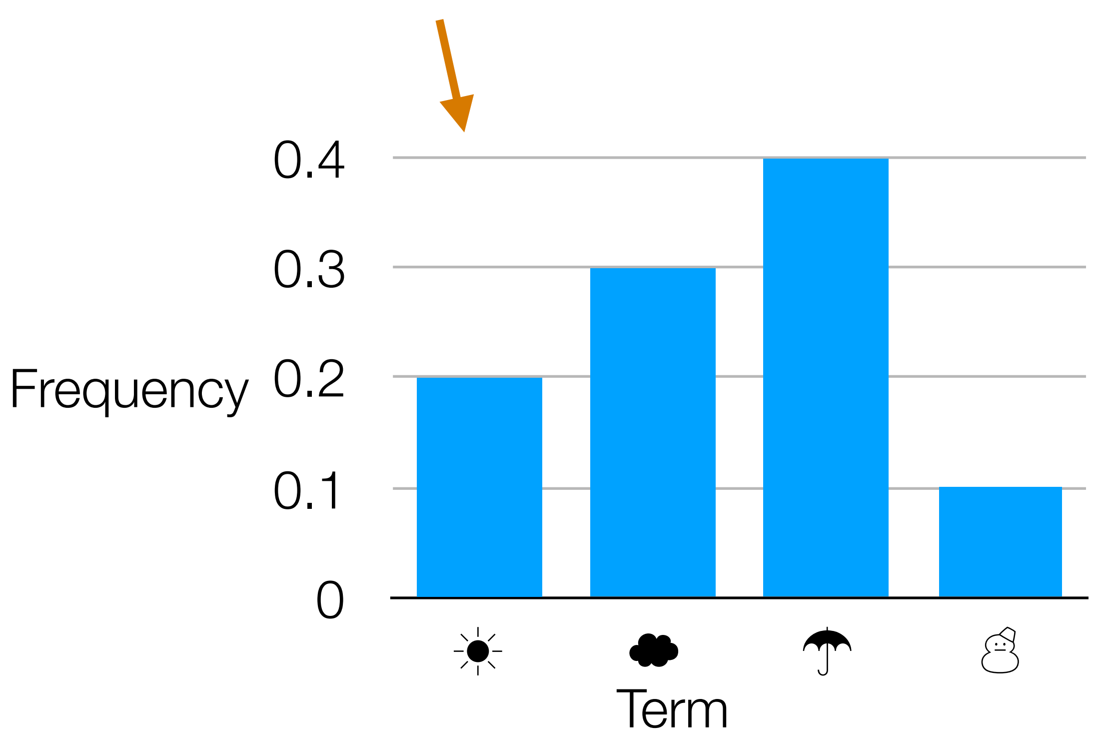
Recap: Basic Text Analysis

- Represent text in terms of “features”
  (e.g., how often each word/phrase appears, whether it’s a named entity, etc)
  - Can repeat this for different documents:
    represent each document as a “feature vector”

"Sentence": ☀☂☁☁☁☂☃☂☂☀



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

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

Example: Representing an Image

Go row by row and look at pixel values

0: black
1: white

$$\begin{bmatrix} 0 \\ \\ \\ \\ \\ \end{bmatrix}$$

Image source: starwars.com

Example: Representing an Image

0: black
1: white

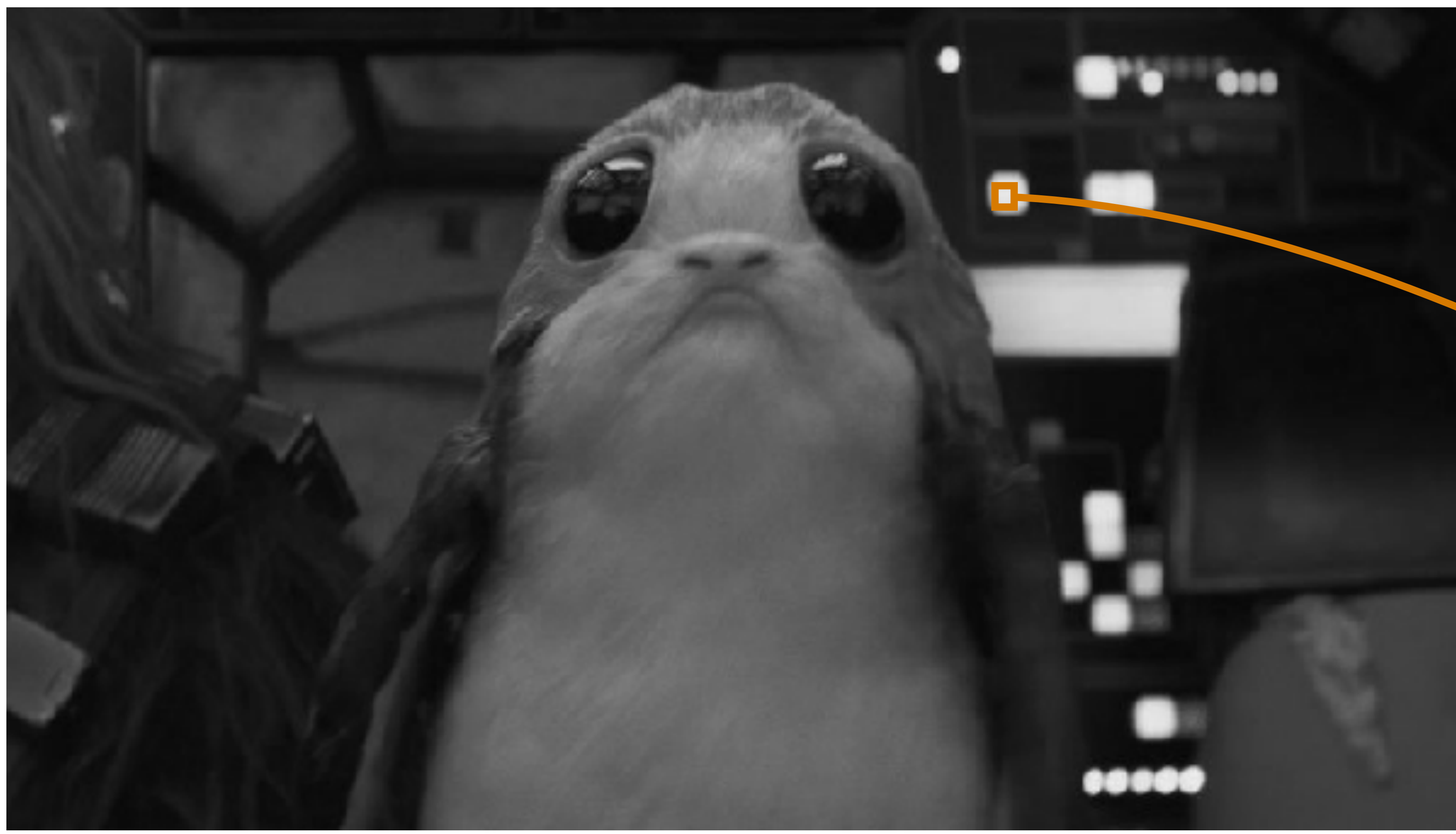
$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ \\ \\ \\ \end{bmatrix}$$

Go row by row and look at pixel values

Image source: starwars.com

Example: Representing an Image

0: black
1: white



$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0.9 \\ \vdots \\ \\ \end{bmatrix}$$

Go row by row and look at pixel values

Image source: starwars.com

Example: Representing an Image

0: black
1: white

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0.9 \\ \vdots \\ \\ 0.3 \end{bmatrix}$$

Go row by row and look at pixel values

dimensions = image width × image height

Very high dimensional!

Image source: starwars.com

Back to Text

Unigram bag of words model is already quite powerful:

- Enough to learn topics
  (each text doc: raw word counts without stopwords)
- Enough to learn a simple detector for email spam

These are HW2 problems

Finding Possibly Related Entities

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

	Elon Musk	Tesla	Apple	Tim Cook
Elon Musk		300	1	4
Tesla	300		5	1
Apple	1	5		195
Tim Cook	4	1	195	

Large values => possible related items

What does it mean for a named entity to co-occur with itself?

Example: could count # articles in which word appears ≥ 2 times

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

We aim to find *interesting* relationships by looking at co-occurrences

Black and white frequently co-occur, but is this relationship interesting?

	Green	White	Black
Green	1000	200	200
White	200	2000	350
Black	200	350	2000

How I'm counting: For each pixel, look at neighboring 4 pixels and compare their values
(1 of "green green", "green white", "green black", "white white", "white black", "black black")

Image source: awf.org/sites/default/files/media/gallery/wildlife/Plains%20Zebra/Z-Billy_Dodson_3.jpg?itok=rzMdZ7LM

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

Probability of drawing “White, Black”?

350/5750

Probability of drawing a card that has “White” on it?

(200+2000+350)/5750

1000 of these cards:	Green, Green
200 of these cards:	Green, White
200 of these cards:	Green, Black
2000 of these cards:	White, White
350 of these cards:	White, Black
2000 of these cards:	Black, Black

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

Probability of drawing "White, Black"?

350/5750

Probability of drawing a card that has "White" on it?

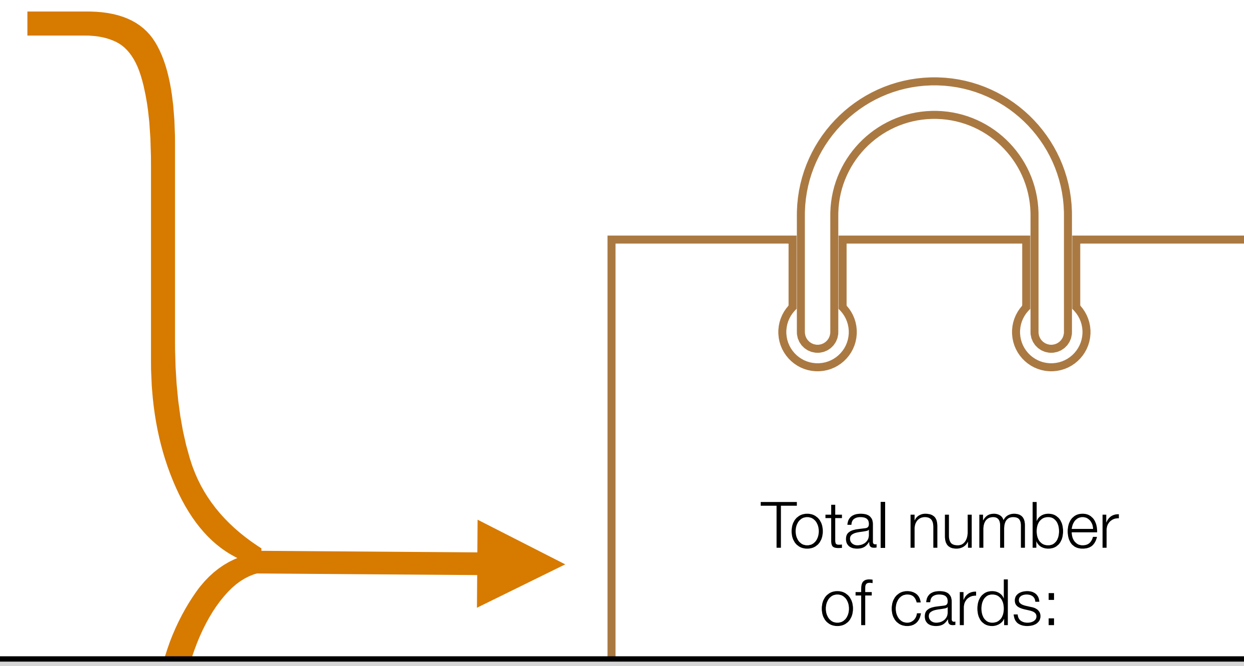
(200+2000+350)/5750

1000 of these cards: Green, Green

200 of these cards: Green, White

200 of these cards: Green, Black

2000 of these cards: White, White



$$P(\text{Green, White}) = \frac{200}{5750} \qquad P(\text{Green, Black}) = \frac{200}{5750} \qquad P(\text{White, Black}) = \frac{350}{5750}$$

$$P(\text{Green}) = \frac{1400}{5750} \qquad P(\text{White}) = \frac{2550}{5750} \qquad P(\text{Black}) = \frac{2550}{5750}$$

Measuring Association: Pointwise Mutual Information (PMI)

$$\text{PMI}(A, B) = \log_2 \frac{P(A, B)}{P(A)\ P(B)}$$

Base of log doesn't really matter (we'll use base 2)

PMI can be positive or negative

Higher PMI ➔ more "interesting"

$$\text{PMI}(\text{Green, White}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = \boxed{-1.63\ldots \text{ bits}}$$

$$\text{PMI}(\text{Green, Black}) = \log_2 \frac{200/5750}{(1400/5750)(2550/5750)} = \boxed{-1.63\ldots \text{ bits}}$$

-1.63 > -1.69

$$\text{PMI}(\text{White, Black}) = \log_2 \frac{350/5750}{(2550/5750)(2550/5750)} = -1.69\ldots \text{ bits}$$

$$P(\text{Green, White}) = \frac{200}{5750} \qquad P(\text{Green, Black}) = \frac{200}{5750} \qquad P(\text{White, Black}) = \frac{350}{5750}$$

$$P(\text{Green}) = \frac{1400}{5750} \qquad P(\text{White}) = \frac{2550}{5750} \qquad P(\text{Black}) = \frac{2550}{5750}$$

What is PMI Measuring?

Probability of A and B co-occurring

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

Probability of just A occurring

Probability of just B occurring

If A and B were "independent"

➔ probability of A and B co-occurring would be P(A)P(B)

What is PMI Measuring?

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

There are *lots* of connections of information theory to prediction

Rough intuition:
Something surprising ↔ less predictable ↔ more bits to store

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-square is between 0 and 1

0 ➔ pairs are all indep.

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

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

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

N = sum of all co-occurrence counts (in upper right of triangle earlier)

Example: Phi-Square Calculation

$$P(\text{Green, White}) = \frac{200}{5750} \qquad P(\text{Green, Black}) = \frac{200}{5750} \qquad P(\text{White, Black}) = \frac{350}{5750}$$

$$P(\text{Green}) = \frac{1400}{5750} \qquad P(\text{White}) = \frac{2550}{5750} \qquad P(\text{Black}) = \frac{2550}{5750}$$

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

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

N = 5750

Chi-square = N × Phi-square

N = sum of all co-occurrence counts (in upper right of triangle earlier)

Example: Phi-Square Calculation

$$P(\text{Green, White}) = \frac{200}{5750} \qquad P(\text{Green, Black}) = \frac{200}{5750} \qquad P(\text{White, Black}) = \frac{350}{5750}$$

$$P(\text{Green}) = \frac{1400}{5750} \qquad P(\text{White}) = \frac{2550}{5750} \qquad P(\text{Black}) = \frac{2550}{5750}$$

	Green	White	Black
Green	1000	200	200
White		2000	350
Black			2000

N = 5750

Sum comprises of 6 terms

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

Green, Green: $\dfrac{[\frac{1000}{5750} - (\frac{1400}{5750})(\frac{1400}{5750})]^2}{(\frac{1400}{5750})(\frac{1400}{5750})}$ $= 0.2216\ldots$

Green, White: $\dfrac{[\frac{200}{5750} - (\frac{1400}{5750})(\frac{2550}{5750})]^2}{(\frac{1400}{5750})(\frac{2550}{5750})}$ $= 0.0496\ldots$

Green, Black: $= 0.0496\ldots$

White, White: $= 0.1161\ldots$

$\vdots$

White, Black: $= 0.0937\ldots$

Black, Black: $= 0.1161\ldots$

Add these up to get:
Phi-square = 0.6470…

Interpretation: neighboring pixels not close to being indep.