Lecture 3: Finding possibly related entities
Co-Occurrences

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

<table>
<thead>
<tr>
<th></th>
<th>Apple</th>
<th>Facebook</th>
<th>Tesla</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elon Musk</td>
<td>10</td>
<td>15</td>
<td>300</td>
</tr>
<tr>
<td>Mark Zuckerberg</td>
<td>500</td>
<td>10000</td>
<td>500</td>
</tr>
<tr>
<td>Tim Cook</td>
<td>200</td>
<td>30</td>
<td>10</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Apple</th>
<th>Facebook</th>
<th>Tesla</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elon Musk</td>
<td>10</td>
<td>15</td>
<td>300</td>
</tr>
<tr>
<td>Mark Zuckerberg</td>
<td>500</td>
<td>10000</td>
<td>500</td>
</tr>
<tr>
<td>Tim Cook</td>
<td>200</td>
<td>30</td>
<td>10</td>
</tr>
</tbody>
</table>

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?
<table>
<thead>
<tr>
<th></th>
<th>Apple</th>
<th>Facebook</th>
<th>Tesla</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Elon Musk</strong></td>
<td>10</td>
<td>15</td>
<td>300</td>
</tr>
<tr>
<td><strong>Mark Zuckerberg</strong></td>
<td>500</td>
<td>10000</td>
<td>500</td>
</tr>
<tr>
<td><strong>Tim Cook</strong></td>
<td>200</td>
<td>30</td>
<td>10</td>
</tr>
</tbody>
</table>

**Probability of drawing “Elon Musk, Apple”?**

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

- 10 of these cards: **Elon Musk, Apple**
- 15 of these cards: **Elon Musk, Facebook**
- 300 of these cards: **Elon Musk, Tesla**
- 10 of these cards: **Tim Cook, Tesla**

Place into bag
Co-occurrence table

<table>
<thead>
<tr>
<th></th>
<th>Apple</th>
<th>Facebook</th>
<th>Tesla</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Elon Musk</strong></td>
<td>10</td>
<td>15</td>
<td>300</td>
</tr>
<tr>
<td><strong>Mark Zuckerberg</strong></td>
<td>500</td>
<td>10000</td>
<td>500</td>
</tr>
<tr>
<td><strong>Tim Cook</strong></td>
<td>200</td>
<td>30</td>
<td>10</td>
</tr>
</tbody>
</table>

Total: 11565
### Joint probability table

<table>
<thead>
<tr>
<th></th>
<th>Apple</th>
<th>Facebook</th>
<th>Tesla</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Elon Musk</strong></td>
<td>10/11565</td>
<td>15/11565</td>
<td>300/11565</td>
</tr>
<tr>
<td><strong>Mark Zuckerberg</strong></td>
<td>500/11565</td>
<td>10000/11565</td>
<td>500/11565</td>
</tr>
<tr>
<td><strong>Tim Cook</strong></td>
<td>200/11565</td>
<td>30/11565</td>
<td>10/11565</td>
</tr>
</tbody>
</table>

Total: 11565

sum to get $P(\text{Elon Musk})$
### Joint probability table

<table>
<thead>
<tr>
<th></th>
<th>Apple</th>
<th>Facebook</th>
<th>Tesla</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Elon Musk</strong></td>
<td>0.00086</td>
<td>0.00130</td>
<td>0.02594</td>
</tr>
<tr>
<td><strong>Mark Zuckerberg</strong></td>
<td>0.04323</td>
<td>0.86468</td>
<td>0.04323</td>
</tr>
<tr>
<td><strong>Tim Cook</strong></td>
<td>0.01729</td>
<td>0.00259</td>
<td>0.00086</td>
</tr>
</tbody>
</table>

Bottom row totals: 0.06139, 0.86857, 0.07004

Recall: if events A and B are independent, $P(A, B) = P(A)P(B)$
Joint probability table \textit{if people and companies were independent}

<table>
<thead>
<tr>
<th></th>
<th>Apple</th>
<th>Facebook</th>
<th>Tesla</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Elon Musk</strong></td>
<td>0.00173</td>
<td>0.02441</td>
<td>0.00197</td>
</tr>
<tr>
<td><strong>Mark Zuckerberg</strong></td>
<td>0.05839</td>
<td>0.82614</td>
<td>0.06662</td>
</tr>
<tr>
<td><strong>Tim Cook</strong></td>
<td>0.00127</td>
<td>0.01802</td>
<td>0.00145</td>
</tr>
</tbody>
</table>

Recall: if events $A$ and $B$ are independent, $P(A, B) = P(A)P(B)$
What we actually observe

<table>
<thead>
<tr>
<th></th>
<th>Apple</th>
<th>Facebook</th>
<th>Tesla</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elon Musk</td>
<td>0.00086</td>
<td>0.00130</td>
<td>0.02594</td>
</tr>
<tr>
<td>Mark Zuckerberg</td>
<td>0.04323</td>
<td>0.86468</td>
<td>0.04323</td>
</tr>
<tr>
<td>Tim Cook</td>
<td>0.01729</td>
<td>0.00259</td>
<td>0.00086</td>
</tr>
</tbody>
</table>

What should be the case if people are companies were independent

<table>
<thead>
<tr>
<th></th>
<th>Apple</th>
<th>Facebook</th>
<th>Tesla</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elon Musk</td>
<td>0.00173</td>
<td>0.02441</td>
<td>0.00197</td>
</tr>
<tr>
<td>Mark Zuckerberg</td>
<td>0.05839</td>
<td>0.82614</td>
<td>0.06662</td>
</tr>
<tr>
<td>Tim Cook</td>
<td>0.00127</td>
<td>0.01802</td>
<td>0.00145</td>
</tr>
</tbody>
</table>
Pointwise Mutual Information (PMI)

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
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:
  
  $$\Phi^2 = \sum_{A, B} \left[ \frac{[P(A, B) - P(A)P(B)]^2}{P(A)P(B)} \right]$$

  Phi-square is between 0 and $\min(\#\text{rows, #cols})-1$.

  0 $\Rightarrow$ pairs are all indep.

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

  $\chi^2 = N \times \Phi^2$

  $N = \text{sum of all co-occurrence counts}$
PMI/Phi-Square/Chi-Square 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/rented/consumed more

- Products & locations
- Products & dates

- If you have a bunch of physical stores:
  anticipate where certain products are likely to be purchased/rented/consumed more

- If you're the police department:
  create "heat map" of where different criminal activity occurs

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)

- Crime reports & locations
Continuous Measurements

- So far, looked at relationships between *discrete* outcomes
- For pair of *continuous* outcomes, use a **scatter plot**

![Graph showing computing improvements: transistors per circuit](https://plot.ly/~MattSundquist/5405.png)

Of course, not all trends look like a line

(\textit{so don’t just do linear regression!})

Image source: https://plot.ly/~MattSundquist/5405.png
The Importance of Staring at Data

In general: not obvious what curve to fit (if any)

Not enough data => might think there's a pattern when it's just noise

In general: not obvious if some points are outliers and should be excluded
Correlation

Negatively correlated  Not really correlated  Positively correlated

Beware: Just because two variables appear correlated doesn't mean that one can predict the other
Correlation ≠ Causation

Moreover, just because we find correlation in data doesn't mean it has predictive value!

Important: At this point in the course, we are finding possible relationships between two entities.

We are *not* yet making statements about prediction (we'll see prediction later in the course).

We are *not* making statements about causality (beyond the scope of this course).
Causality

Studies in 1960's: Coffee drinkers have higher rates of lung cancer

*Can we claim that coffee is a cause of lung cancer?*

Back then: coffee drinkers also tended to smoke more than non-coffee drinkers (smoking is a **confounding variable**)

To establish causality, groups getting different treatments need to appear similar so that the only difference is the treatment

Image source: George Chen
Establishing Causality

If you control data collection

Users → Treatment Group
→ Control Group
Randomly assign

Compare outcomes of two groups
Randomized controlled trial (RCT)
also called A/B testing

Example: figure out webpage layout to maximize revenue (Amazon)
Example: figure out how to present educational material to improve learning (Khan Academy)

If you do not control data collection

In general: not obvious establishing what caused what
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 structure found in Part I*

- Classical classification methods
- Neural nets and deep learning for analyzing images and text