95-865: Unstructured Data Analytics

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": ☀️🌧️💧☔️傘☔️傘☔️傘

```
Term           Frequency
            ☀️  0.1
            ☁️  0.2
            ☔️  0.3
            ☃️  0.4
```

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

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?
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.
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?
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

Probability of drawing “Elon Musk, Apple”?

Probability of drawing a card that says “Apple” on it?
## Co-occurrence table

<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>

Total: 11565
# Joint probability table

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

Sum to get $P(\text{Elon Musk})$.

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>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>

Recall: if events A and B are independent, $P(A, B) = P(A)P(B)$
Joint probability table 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>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>

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

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>0.00086</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>

<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)

Probability of $A$ and $B$ co-occurring

\[
\frac{P(A, B)}{P(A) \cdot P(B)}
\]

if equal to 1
$\Rightarrow A, B$ are indep.

Probability of $A$ and $B$ co-occurring \textit{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!
Recap: Use PMI to Rank Specific Person/Company Pairs

\[ \text{PMI}(A, B) = \log \frac{P(A, B)}{P(A) \cdot P(B)} \]

- 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:

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

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

Chi-squared = $N \times \Phi^2$

$N = \text{sum of all co-occurrence counts}$

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

0 $\Rightarrow$ pairs are all indep.

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

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

Cramér’s $V = \sqrt{\Phi^2 / [\min(#\text{rows}, #\text{cols})-1]}$
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/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
• 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)