Announcements

- **Final Projects:**
  - Presentation templates and instructions are on the final-project page on the website
  - Report/paper templates will go up later this weekend

- **Course evaluations**
  - Due Monday June 14th

- **Office hours**
  - 5-6 pm today
Announcements – HW04

▪ Due tonight

▪ Likelihoods
  • Words that don’t appear in training
  • Classifying document 1

▪ Twitter API
  • Lecture 13 slide 2
1. Create a Twitter developer account [https://developer.twitter.com/](https://developer.twitter.com/)
3. Click “Create an app”
4. Fill out the form, and click “Create”
5. A pop up window will appear for reviewing Developer Terms. Click the “Create” button again.

Instructions from [http://socialmedia-class.org/twittertutorial.html](http://socialmedia-class.org/twittertutorial.html)
Today’s agenda

Phrases

- n-grams

- Language models

- collocation
n-grams
N-grams

- **Unigram**
  - a single word

- **Bigram**
  - Two word phrase

- **Trigram**
  - Three word phrase

- **100-gram**
  - One hundred word phrase

- **n-gram**
  - *n-word phrase*
We can add even more columns to our DTM

\[
\begin{array}{cccccc}
    w_1 & w_2 & w_3 & w_4 & \ldots & \ldots & \ldots & \ldots & w_v \\
\end{array}
\]

\[
\begin{array}{cccccc}
    d_1 \\
    d_1 \\
    \ldots \\
    d_n \\
\end{array}
\]
We can add even more columns to our DTM

<table>
<thead>
<tr>
<th></th>
<th>( w_1 )</th>
<th>( w_2 )</th>
<th>( \ldots )</th>
<th>( \ldots )</th>
<th>( w_v )</th>
<th>( w_1, w_2 )</th>
<th>( w_1, w_3 )</th>
<th>( \ldots )</th>
<th>( w_{v-1}, w_v )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d_1 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( d_1 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \ldots )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( d_n )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Language Models
Given a corpus $C$, what is the probability of a word $w_i$?

$$P(w_i) = \frac{\text{count}(w_i)}{\sum_j \text{count}(w_j)}$$

Maximum Likelihood Estimation

Given a corpus $C$, what is the probability of a word “New”?

$$P(\text{New}) = \frac{\text{count}(\text{New})}{\sum_j \text{count}(w_j)}$$

Marginalizing
Given a corpus $C$, what is the probability of a word “New”?

$$P(New) = \frac{\text{count}(New)+1}{\sum_j \text{count}(w_j)+1}$$
Probability of a bigram

Given a corpus $C$, what is the probability of the phrase “New York”?

$$P(New) = \frac{\text{count}(New)}{\sum_j \text{count}(w_j)} \quad P(York) = \frac{\text{count}(York)}{\sum_j \text{count}(w_j)}$$

We can’t just combine these probabilities

$$P(New, York)$$

We also care about the order of the words

$$P(New) \quad \text{and the probability of } P(York \mid New)$$
Given a corpus $C$, what is the probability of the phrase “New York”?

$$P(\text{New})$$ and the probability of $$P(\text{York} \mid \text{New})$$

$$P(\text{New York}) = P(\text{New})P(\text{York} \mid \text{New})$$

$$P(\text{New}) = \frac{\text{count}(\text{New})}{\sum_j \text{count}(w_j)}$$

$$P(\text{York} \mid \text{New}) = \frac{\text{count}(\text{New York})}{\sum_j \text{count}(\text{New } w_j)}$$

$$= \frac{\text{count}(\text{New York})}{\text{count}(\text{New})}$$
Probability of a bigram

Given a corpus $C$, what is the probability of the phrase “New York”? 

$P(\text{New})$ and the probability of $P(\text{York} | \text{New})$

$$P(\text{New York}) = \frac{\text{count}(\text{New})}{\sum_j \text{count}(w_j)} \times \frac{\text{count}(\text{New York})}{\text{count}(\text{New})}$$
Probability of a sentence based on bigrams

\[ P(w_1 \ldots w_n) = \prod_i^n P(x_i|x_{i-1}) \]

Probability of a sentence based on trigram

\[ P(w_1 \ldots w_n) = \prod_i^n P(x_i|x_{i-1}, x_{i-2}) \]
Point-wise Mutual Information

\[ PMI(x, y) = \log \frac{P(x, y)}{P(x)P(y)} \]

\[ PMI(w_1, w_2) = \log \frac{P(w_1, w_2)}{P(w_1)P(w_2)} \]

\[ P(w_1, w_2) = P(w_2 | w_1) P(w_1) \]

\[ PMI(w_1, w_2) = \log \frac{P(w_2 | w_1) P(w_1)}{P(w_1)P(w_2)} \]

\[ PMI(w_1, w_2) = \log \frac{P(w_2 | w_1) P(w_1)}{P(w_1)P(w_2)} \]
Point-wise Mutual Information

\[ PMI(x, y) = \log \frac{P(y|x)}{P(y)} \]

\[ PMI(w_1, w_2) = \log \frac{P(w_2 | w_1)}{P(w_2)} \]

How likely are we to see \( w_1 \) followed by \( w_2 \) normalized by how likely are we to see \( w_2 \) in general