Presentations

• Submit slides by 10am the morning of your presentation.

• Plan for 5 minutes of presentation + 3 mins of questions.

• Every group will be assigned another group from the opposite session to ask a question of.
Clustering

- Document clustering
- Token clustering (topic modeling)
Clustering

- Clustering is designed to learn structure in the data:
  - Hierarchical structure between data points
  - Natural partitions between data points
Hierarchical Clustering

- *Hierarchical* order among the elements being clustered

London. Michaelmas term lately over, and the Lord Chancellor sitting in Lincoln’s Inn Hall. Implacable November weather.

Among other public buildings in a certain town, which for many reasons it will be prudent to refrain from mentioning, and to which I will assign no fictitious name,

One January day, thirty years ago, the little town of Hanover, anchored on a windy Nebraska tableland, was trying not to be blown away.

It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife.

When men began to build cities vertically instead of horizontally there passed from our highways a picturesque figure, and from our language an expressive
Hierarchical clustering

Allison et al. 2009
Bottom-up clustering

Algorithm 1 Hierarchical agglomerative clustering

1: Data: \( N \) training data points \( x \in \mathbb{R}^F \)
2: Let \( X \) denote a set of objects \( x \)
3: Given some linkage function \( d(X, X') \rightarrow \mathbb{R} \)
4: Initialize clusters \( \mathcal{C} = \{C_1, \ldots, C_N\} \) to singleton data points
5: while data points not in one cluster do
6: \hspace{1cm} Identify \( X, Y \) as clusters with smallest linkage function among clusters in \( \mathcal{C} \)
7: \hspace{1cm} Create new cluster \( Z = X \cup Y \)
8: \hspace{1cm} remove \( X, Y \) from \( \mathcal{C} \)
9: \hspace{1cm} add \( Z \) to \( \mathcal{C} \)
10: end while
Similarity

\[ \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \rightarrow \mathbb{R} \]

- What are you comparing?

- How do you quantify the similarity/difference of those things?
Unigram probability

the a of love sword poison hamlet romeo king capulet be woe him most
Similarity

\[ \text{Euclidean} = \sqrt{\sum_{i}^{\text{vocab}} (P_{i}^{\text{Hamlet}} - P_{i}^{\text{Romeo}})^2} \]

Cosine similarity, Jensen-Shannon divergence...
Flat Clustering

- Partitions the data into a set of $K$ clusters

It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife.

When men began to build cities vertically instead of horizontally there passed from our highways a picturesque figure, and from our language an expressive London.


Among other public buildings in a certain town, which for many reasons it will be prudent to refrain from mentioning, and to which I will assign no fictitious name,

One January day, thirty years ago, the little town of Hanover, anchored on a windy Nebraska tableland, was trying not to be blown away.
**Algorithm 1** K-means

1: Data: training data $x \in \mathbb{R}^F$
2: Given some distance function $d(x, x') \rightarrow \mathbb{R}$
3: Select $k$ initial centers $\{\mu_1, \ldots, \mu_k\}$
4: **while** not converged **do**
5: \hspace{1em} **for** $i = 1$ to $N$ **do**
6: \hspace{2em} Assign $x_i$ to $\arg\min_c d(x_i, \mu_c)$
7: \hspace{1em} **end for**
8: \hspace{1em} **for** $i = 1$ to $K$ **do**
9: \hspace{2em} $\mu_i = \frac{1}{D_i} \sum_{j=1}^{D_i} x_i$
10: \hspace{1em} **end for**
11: **end while**
K-means

- Assignment
- Recomputation of means
Representation

$x \in \mathbb{R}^F$

[x is a data point characterized by F real numbers, one for each feature]

• This is a huge decision that impacts what you can learn
Representation

- Books (e.g., to learn genres)
- News articles (e.g., to learn articles about the same event)
sklearn.cluster.KMeans

sklearn.cluster.AgglomerativeClustering
Topic Models

• A probabilistic model for discovering hidden “topics” or “themes” (groups of terms that tend to occur together) in documents.

• Unsupervised (find *interesting structure* in the data)

• Clustering algorithm:

  How to tokens cluster into topics?
Topic Models

- **Input**: set of documents, number of clusters to learn.

- **Output**:
  - topics
  - topic ratio in each document
  - topic distribution for each word in doc
… The messenger, however, does not reach Romeo and, instead, Romeo learns of Juliet's apparent death from his servant Balthasar. Heartbroken, Romeo buys poison from an apothecary and goes to the Capulet crypt. He encounters Paris who has come to mourn Juliet privately. Believing Romeo to be a vandal, Paris confronts him and, in the ensuing battle, Romeo kills Paris. Still believing Juliet to be dead, he drinks the poison. Juliet then awakens and, finding Romeo dead, stabs herself with his dagger. The feuding families and the Prince meet at the tomb to find all three dead. Friar Laurence recounts the story of the two "star-cross'd lovers". The families are reconciled by their children's deaths and agree to end their violent feud. The play ends with the Prince's elegy for the lovers: "For never was a story of more woe / Than this of Juliet and her Romeo."
... The messenger, however, does not reach Romeo and, instead, Romeo learns of Juliet's apparent death from his servant Balthasar. Heartbroken, Romeo buys poison from an apothecary and goes to the Capulet crypt. He encounters Paris who has come to mourn Juliet privately. Believing Romeo to be a vandal, Paris confronts him and, in the ensuing battle, Romeo kills Paris. Still believing Juliet to be dead, he drinks the poison. Juliet then awakens and, finding Romeo dead, stabs herself with his dagger. The feuding families and the Prince meet at the tomb to find all three dead. Friar Laurence recounts the story of the two "star-cross'd lovers". The families are reconciled by their children's deaths and agree to end their violent feud. The play ends with the Prince's elegy for the lovers: "For never was a story of more woe / Than this of Juliet and her Romeo."
... The messenger, however, does not reach Romeo and, instead, Romeo learns of Juliet's apparent death from his servant Balthasar. Heartbroken, Romeo buys poison from an apothecary and goes to the Capulet crypt. He encounters Paris who has come to mourn Juliet privately. Believing Romeo to be a vandal, Paris confronts him and, in the ensuing battle, Romeo kills Paris. Still believing Juliet to be dead, he drinks the poison. Juliet then awakens and, finding Romeo dead, stabs herself with his dagger. The feuding families and the Prince meet at the tomb to find all three dead. Friar Laurence recounts the story of the two "star-cross'd lovers". The families are reconciled by their children's deaths and agree to end their violent feud. The play ends with the Prince's elegy for the lovers: "For never was a story of more woe / Than this of Juliet and her Romeo."
... The messenger, however, does not reach Romeo and, instead, Romeo learns of Juliet's apparent death from his servant Balthasar. Heartbroken, Romeo buys poison from an apothecary and goes to the Capulet crypt. He encounters Paris who has come to mourn Juliet privately. Believing Romeo to be a vandal, Paris confronts him and, in the ensuing battle, Romeo kills Paris. Still believing Juliet to be dead, he drinks the poison. Juliet then awakens and, finding Romeo dead, stabs herself with his dagger. The feuding families and the Prince meet at the tomb to find all three dead. Friar Laurence recounts the story of the two "star-cross'd lovers". The families are reconciled by their children's deaths and agree to end their violent feud. The play ends with the Prince's elegy for the lovers: "For never was a story of more woe / Than this of Juliet and her Romeo."
... The messenger, however, does not reach Romeo and, instead, Romeo learns of Juliet's apparent death from his servant Balthasar. Heartbroken, Romeo buys poison from an apothecary and goes to the Capulet crypt. He encounters Paris who has come to mourn Juliet privately. Believing Romeo to be a vandal, Paris confronts him and, in the ensuing battle, Romeo kills Paris. Still believing Juliet to be dead, he drinks the poison. Juliet then awakens and, finding Romeo dead, stabs herself with his dagger. The feuding families and the Prince meet at the tomb to find all three dead. Friar Laurence recounts the story of the two "star-cross'd lovers". The families are reconciled by their children's deaths and agree to end their violent feud. The play ends with the Prince's elegy for the lovers: "For never was a story of more woe / Than this of Juliet and her Romeo."

“Etc.”
The messenger, however, does not reach Romeo and, instead, Romeo learns of Juliet's apparent death from his servant Balthasar. Heartbroken, Romeo buys poison from an apothecary and goes to the Capulet crypt. He encounters Paris who has come to mourn Juliet privately. Believing Romeo to be a vandal, Paris confronts him and, in the ensuing battle, Romeo kills Paris. Still believing Juliet to be dead, he drinks the poison. Juliet then awakens and, finding Romeo dead, stabs herself with his dagger. The feuding families and the Prince meet at the tomb to find all three dead. Friar Laurence recounts the story of the two "star-cross'd lovers". The families are reconciled by their children's deaths and agree to end their violent feud. The play ends with the Prince's elegy for the lovers: "For never was a story of more woe / Than this of Juliet and her Romeo."
# Applications

## A Topic Model of Literary Studies Journals

### Overview

<table>
<thead>
<tr>
<th>List</th>
<th>Grid</th>
<th>Years</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 1889—2013

<table>
<thead>
<tr>
<th>topic ↓↑</th>
<th>top words</th>
<th>proportion of corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>see both own view role university further account critical particular</td>
<td>2.5%</td>
</tr>
<tr>
<td>2</td>
<td>other both two form same even each part experience process</td>
<td>2.6%</td>
</tr>
<tr>
<td>3</td>
<td>old beowulf english ic mid swa pe poet ond grendel</td>
<td>0.3%</td>
</tr>
</tbody>
</table>

\textbf{\( \mathbf{x} = \text{feature vector} \)}

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>follow clinton</td>
<td>0</td>
</tr>
<tr>
<td>follow trump</td>
<td>0</td>
</tr>
<tr>
<td>“republican” in profile</td>
<td>0</td>
</tr>
<tr>
<td>“democrat” in profile</td>
<td>0</td>
</tr>
<tr>
<td>“benghazi”</td>
<td>1</td>
</tr>
<tr>
<td>topic 1</td>
<td>0.55</td>
</tr>
<tr>
<td>topic 2</td>
<td>0.32</td>
</tr>
<tr>
<td>topic 3</td>
<td>0.13</td>
</tr>
</tbody>
</table>

\textbf{\( \beta = \text{coefficients} \)}

<table>
<thead>
<tr>
<th>Feature</th>
<th>( \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>follow clinton</td>
<td>-3.1</td>
</tr>
<tr>
<td>follow trump</td>
<td>6.8</td>
</tr>
<tr>
<td>“republican” in profile</td>
<td>7.9</td>
</tr>
<tr>
<td>“democrat” in profile</td>
<td>-3.0</td>
</tr>
<tr>
<td>“benghazi”</td>
<td>-1.7</td>
</tr>
<tr>
<td>topic 1</td>
<td>0.3</td>
</tr>
<tr>
<td>topic 2</td>
<td>-1.2</td>
</tr>
<tr>
<td>topic 3</td>
<td>5.7</td>
</tr>
</tbody>
</table>
Software

• Mallet
  http://mallet.cs.umass.edu/

• Gensim (python)
  https://radimrehurek.com/gensim/

• Visualization
  https://github.com/uwdata/termite-visualizations
Latent variables

• A latent variable is one that’s unobserved, either because:
  • we are predicting it (but have observed that variable for other data points)
  • it is unobservable
Probabilistic graphical models

- Nodes represent variables (shaded = observed, clear = latent)
- Arrows indicate conditional relationships
- The probability of $x$ here is dependent on $y$
- Simply a visual way of writing the joint probability:

$$P(x, y) = P(y) \cdot P(x \mid y)$$
document distribution over topics

topic indicators for words

words

topic distribution over words
Topic Models

- A document has distribution over topics
Topic Models

• A topic is a distribution over words

• e.g., $P(\text{“adore”} \mid \text{topic} = \text{love}) = .18$
P(topic | topic distribution)
P(topic | topic distribution)
P(topic | topic distribution)
P(topic | topic distribution)
P(topic | topic distribution)
war

love

chases

boats

aliens

family

"fights"

"alien"

"kills"

"marries"
P(topic | topic distribution)
P(topic | topic distribution)
P(topic | topic distribution)
P(topic | topic distribution)
P(topic | topic distribution)
# Inferred Topics

<table>
<thead>
<tr>
<th>{album, band, music}</th>
<th>{government, party, election}</th>
<th>{game, team, player}</th>
</tr>
</thead>
<tbody>
<tr>
<td>album</td>
<td>government</td>
<td>game</td>
</tr>
<tr>
<td>band</td>
<td>party</td>
<td>team</td>
</tr>
<tr>
<td>music</td>
<td>election</td>
<td>player</td>
</tr>
<tr>
<td>song</td>
<td>state</td>
<td>win</td>
</tr>
<tr>
<td>release</td>
<td>political</td>
<td>play</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>{god, call, give}</th>
<th>{company, market, business}</th>
<th>{math, number, function}</th>
</tr>
</thead>
<tbody>
<tr>
<td>god</td>
<td>company</td>
<td>math</td>
</tr>
<tr>
<td>call</td>
<td>market</td>
<td>number</td>
</tr>
<tr>
<td>give</td>
<td>business</td>
<td>function</td>
</tr>
<tr>
<td>man</td>
<td>year</td>
<td>code</td>
</tr>
<tr>
<td>time</td>
<td>product</td>
<td>set</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>{city, large, area}</th>
<th>{math, energy, light}</th>
<th>{law, state, case}</th>
</tr>
</thead>
<tbody>
<tr>
<td>city</td>
<td>math</td>
<td>law</td>
</tr>
<tr>
<td>large</td>
<td>energy</td>
<td>state</td>
</tr>
<tr>
<td>area</td>
<td>light</td>
<td>case</td>
</tr>
<tr>
<td>station</td>
<td>field</td>
<td>court</td>
</tr>
<tr>
<td>include</td>
<td>star</td>
<td>legal</td>
</tr>
</tbody>
</table>
Inference

• What are the topic distributions for each document?
• What are the topic assignments for each word in a document?
• What are the word distributions for each topic?

Find the parameters that maximize the likelihood of the data!
Inference

• Markov chain Monte Carlo (Gibbs sampling, Metropolis Hastings, etc.)

• Variational methods

• Spectral methods (Anandkumar et al. 2012, Arora et al. 2013)
Gibbs Sampling

- Markov chain Monte Carlo method for approximating the joint distribution of a set of variables (Geman and Geman 1984; Metropolis et al. 1953; Hastings et al. 1970)
Gibbs Sampling

1. Start with some initial value for all the variables

2. Sample a value for a variable conditioned on all of the other variables around it (using Bayes’ theorem)

\[ P(\theta|X) = \frac{P(\theta)P(X|\theta)}{\sum_\theta P(\theta)P(X|\theta)} \]
Inference
Inference

\[ P(\theta_d | \alpha, z_d) \]
\[ \propto P(\theta_d | \alpha) \prod_i P(z_i | \theta_d) \]
\[ \propto \text{Dir}(\theta | \alpha) \prod_i \text{Cat}(z_i | \theta) \]
\begin{align*}
P(z | \theta_d, w, \phi) \\
\propto P(z | \theta_d)P(w | z, \phi) \\
\propto \text{Cat}(z | \theta_d)\text{Cat}(w | z, \phi) \\
\propto \theta_d^z \times \phi_z^w
\end{align*}
Sampling

\[
\begin{align*}
\alpha & \quad \theta \\
\gamma & \quad \varphi \\
\varphi & \quad W
\end{align*}
\]

| \( z \) | \( P(z|\theta) \) | \( P(w|z) \) | \( \frac{P(z|\theta)}{P(w|z)} \) | norm |
|---|---|---|---|---|
| 1 | 0.100 | 0.010 | 0.001 | 0.019 |
| 2 | 0.200 | 0.030 | 0.006 | 0.112 |
| 3 | 0.070 | 0.020 | 0.001 | 0.026 |
| 4 | 0.130 | 0.080 | 0.010 | 0.193 |
| 5 | 0.500 | 0.070 | 0.035 | 0.651 |
Assumptions

- Every word has one topic
- Every document has one topic distribution
- No sequential information (topics for words are independent of each other given the set of topics for a document)
- Topics don’t have arbitrary correlations (Dirichlet prior)
- Words don’t have arbitrary correlations (Dirichlet prior)
- The only information you learn from are the identities of **words** and how they are divided into **documents**.
What if you want to encode other assumptions or reason over other observations?
Time is drawn from a Beta distribution $[0,1]$.

(Wang and McCallum 2006)
Time is drawn from a Normal distribution $[-\infty, \infty]$. 

\[ \psi_{\mu, \sigma^2}(x) \]
Time is drawn from a Multinomial distribution

\[ [1, \ldots, K] \]
Activity

• 17.clustering/TopicModeling